The Visual Computer (2024) 40:1229–1243
https://doi.org/10.1007/s00371-023-02843-9

ORIGINAL ARTICLE

mcVAE: disentangling by mean constraint

Ming-fei Hu[1] · **Ze-yu Liu**[1] · **Jian-wei Liu**[1]

Accepted: 11 March 2023 / Published online: 6 April 2023

Abstract
Disentanglement tends to automatically learn and separate the interpretable factors of variation hidden in the data. Disentangled representations are more transferable and robust for the chosen model, and they are commonly used in image attack detection and anti-fraud, as well as classification and recommendation systems in special situations. As a popular method for learning unsupervised disentanglement, β-VAE re-weights the KL divergence by an adjustable hyperparameter. However, good disentangled representations always lead to blurry reconstructions and mode collapse on complex datasets. We find that the variance vector of the variational posterior is related to the nature of the dataset and representation space, limiting its value to 1 is not reasonable enough. More importantly, constraining mean variable alone can achieve better disentanglement and reconstruction performance. Therefore, we introduce mean constraint VAE, a simple and effective replacement of the β-VAE for improving the poor reconstruction and learning a higher degree of disentanglement. In addition, a classifier-free measure of disentanglement called variance proportion metric is proposed. Experiments show that our framework outperforms β-VAE on several benchmark datasets.

Keywords Variational autoencoder · Disentanglement · Representation learning

1 Introduction

Disentangled representations can be specified as ones where single latent units are sensitive to changes in single generative factors while being relatively invariant to changes in other factors [1]. This definition is somewhat controversial, [2] provide a formal definition of disentanglement, and they argue that disentangled representations should capture the symmetry transformations of the world state. Learning such representations for training data are an important precursor for a variety of standard downstream tasks, which are more transferable [3] and robust [4] for the chosen model, and it may help artificial intelligence to understand the world in the same way that humans do. For example, a disentangled factor of face objects could control one distinct attribute like smile or age, and we can use it to classify facial attributes [5], video synthesis [6] and detect face presentation attacks [7, 8]. However, learning disentangled representations are still a difficult problem, and good disentangled effect accompanies with risks in the degradation of reconstruction quality and mode collapse for the existing framework such as β-VAE [9].β-VAE is a framework for learning disentangled representations based on the Variational Autoencoder (VAE) [10, 11], it uses a modified version of the VAE by re-weighting the KL divergence. β-VAE claims that a large weight ($\beta > 1$) of the KL divergence between the variational posterior and the prior is necessary to achieve good disentanglement performance. Analogously, InfoGAN [12] argues that encouraging the latent variables more interpretable help to learn disentangled representations by rewarding the mutual information between the observations and a subset of latent variables.

The main drawback of β-VAE is that the reconstruction quality (compared to VAE) must be sacrificed for good disentangled representations due to the restricted capacity of the latent information. Our goal is to identify the sources of disentanglement and reconstruction quality in KL divergence, and find a better constraint term; then, we propose mean constraint Variational Auto-Encoding(mcVAE), which learns better disentangled representations and lower reconstruction losses, simultaneously. Assuming that the prior is

✉ Jian-wei Liu
liujw@cup.edu.cn; 2236677012@qq.com

Ming-fei Hu
hmfzsy@gmail.com

Ze-yu Liu
2275045480@qq.com

[1] Department of Automation, College of Information Science and Engineering, China University of Petroleum Beijing, 260 Mailbox, Changping District, Beijing 102249, China

an isotropic standard normal distribution $p(z) = \mathcal{N}(0, \mathbf{I})$, the KL divergence tries to match the variational posterior $q(z|x) = \mathcal{N}(\mu, \sigma^2)$ (μ and σ^2 denote mean and variance, respectively) to the prior $p(z)$ that can control the capacity of the representation space and embody the disentangled factors. However, the variance vector σ^2 is greatly affected by the properties of the dataset and representation space, and it is usually much larger than 1 to guarantee enough information in representation space. More importantly, we find that limiting the mean or variance individually can reduce the reconstruction loss significantly, limiting the mean alone can help the representation explore more disentangled factors. Therefore, we propose mean constraint VAE (mcVAE) that uses an additional mean constraint term instead of the re-weighting KL divergence, and our new framework is a feasible and simple way to improve the poor reconstruction of β-VAE while exploring a higher degree of disentanglement. In summary, we make the following contributions:

1. Through analyzing the sources of disentanglement and reconstruction in β-VAE, we find that increasing the weight of mean constraint or variance constraint can reduce reconstruction loss, and the mean constraint helps to explore more disentangled factors.
2. We introduce mcVAE, a simple method for disentangling that gives higher disentanglement scores than β-VAE with better reconstruction quality.
3. We identify the weaknesses of classifier-based disentanglement metrics and propose a simple and more comprehensible alternative: Variance Proportion Metric.
4. We give quantitative comparisons of mcVAE, vcVAE and β-VAE for disentanglement and reconstruction indexes.

2 Related work

In this paper, we focus on discussing the disentangled representations without supervision. Early works have shown successful disentanglement in limited settings with few factors, such as penalizing predictability of one latent dimension given the others in autoencoder [13], disentangling two factors of variation in Boltzmann Machine [14] and a multilinear generalization of factor analyzers named Tensor Analyzers [15].

Many recent works explore frameworks to learn disentanglement in the latent variables. The first framework combines the concepts of variational autoencoder and generative adversarial network by using a discriminator to optimize the divergence and encourage independence factors. Adversarial autoencoder (AAE) [16] uses the GAN framework to optimize the reconstruction error and KL divergence, and it shows that AAE can disentangle the style and content of images on the semi-supervised classification and unsupervised clustering task. PixelGAN Autoencoders [17] show the same objective of AAE and different decompositions of information between the latent coding and the decoder, and it combines a generative PixelCNN with a GAN inference network to impose arbitrary priors on the latent coding of VAE. Adversarial objectives also can be used to penalize the Jensen-Shannon Divergence between the distribution of codes and the product of its marginals and learn disentangled features in the context of nonlinear ICA source separation only [18]. FactorVAE [19] encourages the distribution of representations to be disentangled by a Multi-Layer Perceptron (MLP) classifier as discriminator that distinguishes whether the input was drawn from the marginal code distribution or the product of its marginals. ID-GAN [20] maximizes the mutual information between the latent variable of VAE and the output by a generator, the additional generator is effective for high-fidelity synthesis.

The second framework learns disentangled factors through mutual information or information bottleneck. InfoGAN [12] rewards the mutual information between a small subset of the latent variables and the observation to learn disentangled representations. However, due to the training stability issues of GAN, there has been few empirical comparisons between VAE-based methods. Information bottleneck GAN [21] constrains the mutual information between the input and the generated output of the additional encoder; the samples generated by IB-GAN have better reconstruction quality than InfoGAN. InfoVAE [22] is argued that an additional mutual information loss between latent variables and the observation can improve the quality of the variational posterior, which is a conflicting conclusion with InfoGAN. Achille and Soatto [23] attempt to promote the creation of optimal disentangled representations simply by using information bottleneck; they claim that the information dropout algorithm can be extended to the VAE setting, but there are any experiments on disentangling to support the theory. Hu and Liu [24] prove that penalizing the mutual information between the capsule, and the observation can help to learn disentangled representations from information bottleneck of view. However, the experiment was only limited to the capsule network [25] without further promotion.

The third framework for learning disentanglement depends on the punishment for KL divergence and its decomposition term. Mathieu et al. [26] attempt to generalize β-VAE to a general framework by decomposed latent variable. Based on the β-VAE and evidence lower bound decomposition [27], β-TCVAE carries out a decomposition of the variational lower bound [28] and uses the total correlation term to explain the success of β-VAE in learning disentanglement. In a concurrent work, FactorVAE encourages an equivalent total correlation penalty to the β-TCVAE with different training methods, and it requires an auxiliary

discriminator network which can be seemed as a combination of β-VAE and GAN for learning disentanglement. Gyawali et al. [29] investigate the effect of the posterior density on the disentangling ability, utilizing a nonparametric latent factor model named Indian Buffet Process (IBP) method to balance poor reconstruction and disentanglement. Joint-VAE learns disentangled jointly continuous and discrete representations for disentangling the factors of different categories on supervised data [30].

3 A novel metric for disentanglement

Most prior works have resorted to measure disentanglement by latent traversals: visualizing the change in reconstructions while traversing one factor of the latent variable and fixing others at a time, then judging whether the factor is disentangled according to the changes of the reconstructions. Although visualizing latent traversals are an essential and observable indicator, comparing different disentangling algorithms without proper metrics is difficult, having a human in the loop to assess disentanglement is also too time-consuming and subjective [19]. The absence of a proper quantitative metric is a major obstacle for learning disentangled representations.

When the true underlying generative factors are known, a classifier-based metric is proposed by [9] to quantify disentanglement. It dates back to [13] learning a predictor to quantify predictability between the different dimensions of representations and [31] using a linear map from representations to factors in the context of linear ICA. The classifier-based metric score is the accuracy of a linear classifier that can achieve by identifying a fixed ground truth factor as follows: Assuming that the ground truth factors $v \in \mathrm{R}^K$ and conditionally dependent factors $w \in \mathrm{R}^H$ are known, the images x are generated by the factors $x = f(v, w)$. To train the classifier, choose a factor v_k and its label y_k, generate two images $\{x_1, x_2\}$ as a pair with v_k fixed but all other factors varying randomly. Then, we can obtain their latent representations $\{z_1, z_2\}$ and take the absolute value of the pairwise differences of these representations $z_{\text{diff}} = |z_1 - z_2|$, if we generate L pair data, each training data point is an aggregation over L samples $\sum_{l=1}^{L} z_{\text{diff}}$, and the fixed factor label y_k is the corresponding training output; the accuracy of the classifier is disentanglement metric score.

The score is 100% if the representation is perfectly disentangled. The classifier-based metric is relatively simple in design and generalizable that has been used many [19, 32, 33]. To ensure that the classifier does not overfit, the right parameters of the low VC-dimension linear classifier must be carefully selected, because the classifier is sensitive to the hyperparameters such as training iterations, initialization and optimizer. To make the classifier stable, β-VAE uses ten replicas of the model with the same hyperparameters and each of the replicas was evaluated three times with different random seeds to initialize the linear classifier, and the top half of the thirty resulting scores remain, but discarding low scores makes it not objective and reasonable enough. FactorVAE modifies z_{diff} by taking the empirical variance in each dimension of these normalized representations and taking the dimension with the lowest variance as training input for the classifier. However, the classifier-based metrics are loosely interpreted as measuring the reduction in entropy of z, choosing different classifiers still has an impact on the score of disentanglement. In addition, when a factor is fixed and all other factors are varied, the pairwise differences of the representations $|z_1 - z_2|$ will bring a lot of randomness; the uncertainty is difficult to compensate through sampling more data, and it is the main factors that causes the inaccuracy of the classifier measuring.

Based on the shortcomings of the above metrics, we try to propose a stable and easy-to-understand method. Disentanglement is defined as that the change in one dimension of the latent variable corresponds a change in a truth factor, and it means that a set of data with one factor varied and other factors fixed should generate latent representations with only one dimension changed. For instance, in 2D shapes dataset, generating a set of samples with scale varied and other factors fixed, then, this group of images is different in scale only and others are the same, and its corresponding latent representations should only have one dimension changed if the latent variable is disentangled.

According to the above analysis, we propose a new disentanglement metric without classifier as shown in Fig. 1, and the specific derived process is as follows:

1. Choose a factor v_k and generate a group of data $\left\{x^l_{v_k}, \cdots, x^l_{v_k}\right\}$ containing l samples with only v_k varying.
2. Obtain the corresponding representations $\left\{z^l_{v_k}, \cdots, z^l_{v_k}\right\}$ of $\left\{x^l_{v_k}, \cdots, x^l_{v_k}\right\}$ and normalize each dimension by empirical mean and standard deviation over the group for rescaling the representations.
3. Calculate the variance of each dimension of these normalized representations, the proportion of the highest variance to the sum of variances is the disentanglement score of our algorithm.

We can call this metric of measuring disentanglement Variance Proportion Metric (VP Metric). For a ground truth factor, the selection of the data in a group is random, sampling multiple groups ensure the adequacy of the data and reduces randomness. We demonstrate the effect of the number of groups on the VP scores of each truth factor in Fig. 2: when the number of the groups is less than 20 groups, there

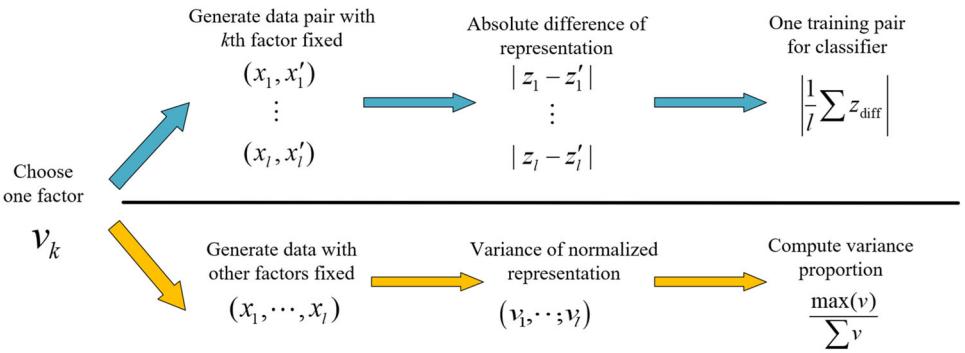


Fig. 1 Top: classifier-based metric in β-VAE. Bottom: our metric

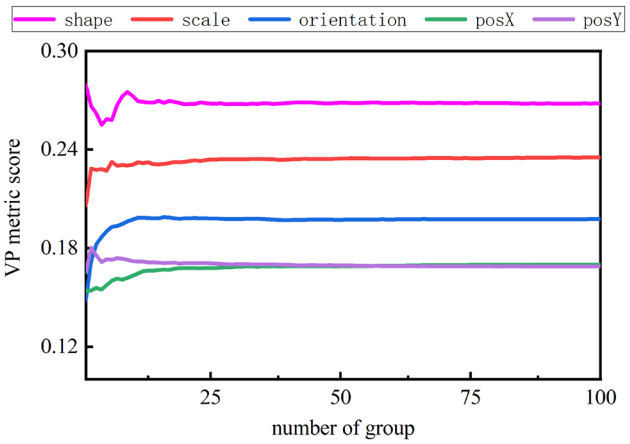


Fig. 2 As the number of groups grows, the disentanglement score of our VP metric gradually stabilizes on all truth factors

are obvious errors in the disentanglement score. If the number of the groups is more than 50, the score is sufficiently reliable and stable.

Compared to the classifier-based metrics and MIG [19], our VP metric uses a simpler concept to directly judge the change of the representations by perturbing the truth factors, and it can give a more reliable measure score without hyperparameter. More importantly, the VP metric can achieve the scores of each truth factor separately, it is more reasonable due to each factor with a different degree of disentanglement. For instance, 'shape' and 'orientation' are ground truth factors of dSprites dataset, it is always easier to learn 'orientation' factor than 'shape' in most disentanglement frameworks, and they should have different disentanglement scores.

We think for future research, developing an unsupervised metric to measure disentanglement without truth factors is an important direction, and it could help us deal with more complex datasets, rather than 2d shapes or 3d face. We believe that a reliable metric has the same value with an effective disentanglement algorithm.

4 Learning disentangled representations by mean constraint

We motivate our approach by analyzing the sources of disentanglement and reconstruction quality in the β-VAE objective function. First, by analyzing the source of poor reconstruction and some experimental results, we find that the variance vector of the latent variable is related to its dimension and properties of input datasets. Second, we analyze the source of the disentanglement and design the learning process of the disentangled representation, and we propose a novel framework mean constraint Variational Auto-Encoding. Last, we sum up the relationship between information bottleneck, mutual information and KL divergence based on VAE framework by generalizing and unifying these constraint algorithms.

4.1 Source of poor reconstruction in β-VAE

Variational autoencoder (VAE) [10, 11] is a generative model that pairs an inference network as encoder and a generator as decoder. The purpose of VAE is to learn the distribution of the input data x, by learning latent representations z where such that z can reconstruct x as much as possible. Instead of directly performing the intractable marginal likelihoods $\log p(x)$, VAE optimizes the evidence lower bound:

$$L = \mathrm{E}_{q(z|x)}[\log p(x|z)] - D_{\mathrm{KL}}(q(z|x)||p(z)) \tag{1}$$

where D_{KL} denotes Kullback–Leibler divergence, $q(z|x)$ denotes an approximation to the intractable true posterior $p(z|x)$. According to β-VAE, the disentanglement comes from the KL divergence term, the KL divergence with larger weight β leads to a higher quality of disentanglement and poorer reconstruction, and it means that better disentanglement is always at the expense of the reconstruction fidelity in β-VAE, which may cause mode collapse. For example, in celebA dataset, the output faces reconstructed from different input faces have high homogeneity, and most of the detailed features from input images are lost. Although we can observe disentangled attributes, the poor reconstruction and the lost

information of the latent variable will have a negative impact on downstream tasks.

Therefore, we attempt to explore the sources of the poor reconstruction and disentanglement; then, we propose a new framework that learns better disentanglement with sharper reconstruction. The KL divergence constraint D_{KL} of evidence lower bound is integrated analytically without sampling, and we can assume the prior $p(z) = \mathcal{N}(0, \mathbf{I})$ and the approximate posterior $q(z|x) = \mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 denote mean and variance, respectively, and $q(z|x)$ is closer to $\mathcal{N}(0, \mathbf{I})$ by increasing the value of β. In the best setting, the values of mean variable and variance variable are 0 and 1, and its disentanglement has the highest metric score.

However, the strict constraint of β-VAE makes the representation space so small that the model loses most detailed features during training. To circumvent it, the first thing is to investigate whether it is reasonable to force $q(z|x)$ to $\mathcal{N}(0, \mathbf{I})$. Several VAE models are trained on 3D chairs and celebA datasets to validate the value of the variance vector, and we use different dimensions of the latent variable and observe the changes of variance variable as shown in Fig. 3.

It can be seen that the variance variable is related to the nature of data and the dimension: The face samples on celebA are more complex than chair samples, its representation needs to contain more features for reconstructing, and the representation on the celebA requires a larger variance if the dimensions are the same. Similarly, when the input data are the same, and the dimension of the latent variable is reduced, the encoder has to learn larger variance to increase the representation space and reduce reconstruction error. Therefore, we think that the excessive variance constraint in β-VAE is the main reason of poor reconstruction.

4.2 mcVAE: improve the quality of disentanglement and reconstruction

In β-VAE, mean constraint and variance constraint compress the representation space at the same time, squeezing the information in the latent variable and losing a lot of features. Fortunately, we find that moving the parameter β of the KL term to mean constraint term is a feasible way to improve β-VAE's poor reconstruction while learning a higher degree of disentanglement. We propose mean constraint Auto-Encoding Variational (mcVAE), and mean constraint is an additional regularization term constraining the values of the mean alone. According to the above assumptions, the KL divergence can be integrated analytically as follows:

$$\begin{aligned} D_{KL}(q(z|x)||p(z)) &= -\int N(z;\, \mu,\, \sigma^2) \log N(z;\, 0,\, \mathbf{I}) \mathrm{d}z \\ &\quad + \int N(z;\, \mu,\, \sigma^2) \log N(z;\, \mu,\, \sigma^2) \mathrm{d}z \\ &= -\frac{1}{2} \sum \left(1 + \log(\sigma^2) - \mu^2 - \sigma^2\right) \end{aligned} \tag{2}$$

We can assume that the variance variable is an identity matrix, the mean constraint term is:

$$\begin{aligned} L_{mc} &= \int N(z;\, \mu,\, \mathbf{I}) \log N(z;\, 0,\, \mathbf{I}) \mathrm{d}z \\ &\quad - \int N(z;\, \mu,\, \mathbf{I}) \log N(z;\, \mu,\, \mathbf{I}) \mathrm{d}z \\ &= \frac{1}{2} \sum \left(1 + \log \mathbf{I} - \mu^2 - \mathbf{I}\right) \end{aligned} \tag{3}$$

Then the objective function of our mcVAE is:

$$L = L_r - D_{KL} + \beta L_{mc} \tag{4}$$

where $L_r = E_{q(z|x)}[\log p(x|z)]$ denotes reconstruction error. Similarly, if the variance constraint is used as a regularization term, we can assume that the mean is 0; the objective function of variance constraint Variational Auto-Encoding is:

$$L = L_r - D_{KL} + \beta L_{vc} \tag{5}$$

where $L_{vc} = \frac{1}{2} \sum \left(1 + \log(\sigma^2) - \sigma^2\right)$ is the variance constraint.

We are trying to explain why the mean constraint makes better disentanglement than variance constraint and KL term. Assuming the traversal of all dimensions is the space of representations, the space learned by VAE is shown in Fig. 4a; now we use orange circles and yellow circles to denote attribute variations expressed in different dimensions. A direct manifestation of disentangled representation is that the change of an attribute can be observed when traversing a dimension, and we guess that the attribute variations of a dimension are randomly distributed in the representation space without disentangled constraint; we cannot observe the corresponding feature changes when traversing a dimension. However, effective disentanglement constraints help to explore semantically attributes and limit their variations into meaningful alignments during training.

As shown in Fig. 4b, mean constraint forces the variations of each attribute to appear in the traversal interval of the corresponding dimension, which has little impact on representation space and reconstruction quality. Even though the KL divergence and variance constraint can rearrange the attribute variations, over-squeezed representation space would loss many features which leads to a decrease in generating quality and disentangling effect. It explains why our mcVAE can learn more disentangled attributes with higher reconstruction quality.

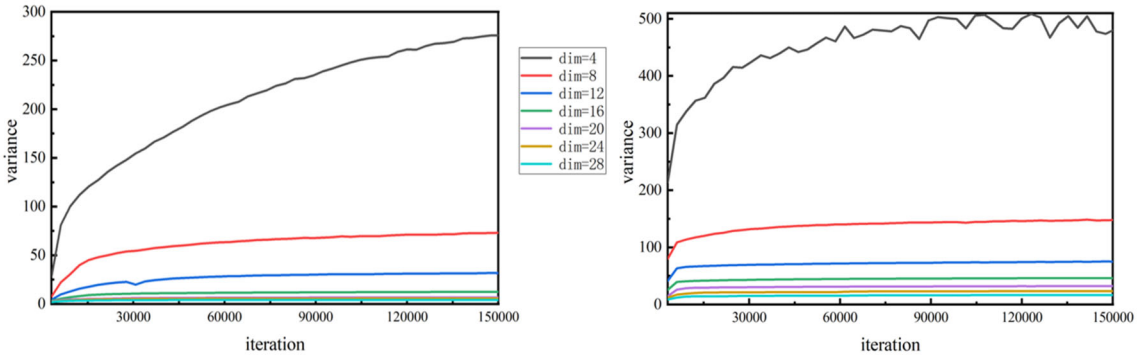
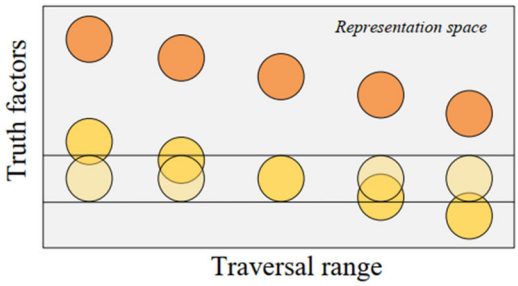
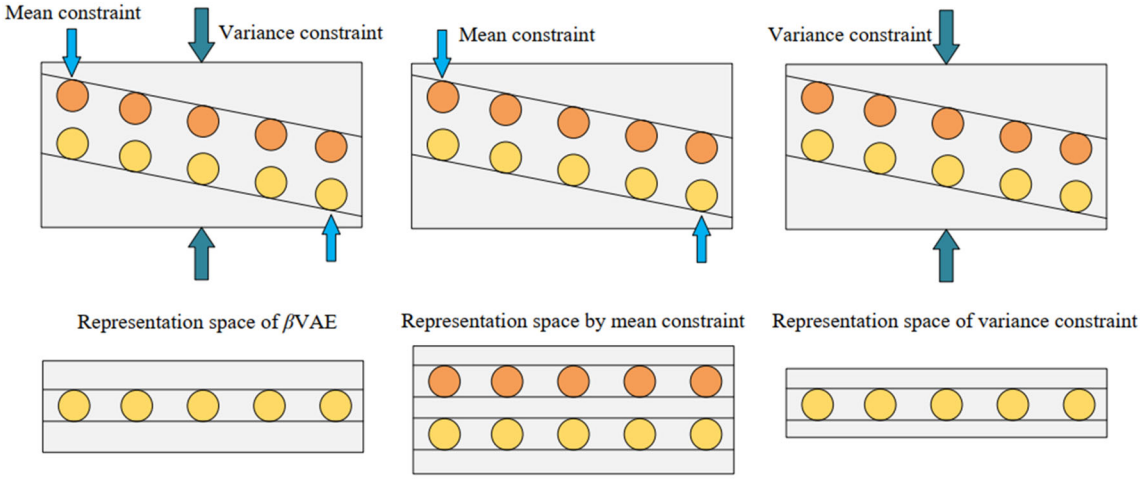


Fig. 3 Variance variable with different dimensions on celebA (left) and 3D chairs (right)



(a) A schematic diagram of representation space



(b) Limiting attribute variations by KL term, mean constraint and variance constraint

Fig. 4 Learn disentangled representation by different framework

4.3 Analyze disentanglement through information bottleneck and mutual information

Optimal disentangled representations [23] are created by enforcing a factorized prior in perspective of information bottleneck, and it states that there is a correlation between disentanglement and information bottleneck. InfoGAN maximizes the mutual information between a small subset of the latent variables and the input data, and it achieves a higher degree of disentanglement than GAN. InfoVAE improves the quality of the variational posterior by adding mutual information term between the representation and the input data. In this subsection, we attempt to explain the relationship between disentanglement, information bottlenecks and mutual information in the context of VAE.

Given the input data x, we want to compute a representation z that has some desirable properties given task y.

Table 1 Disentanglement scores of five truth factors by VP metric on dSprites dataset

Model	Shape	Scale	Orientation	Position X	Position Y
β-VAE ($\beta = 4$)	23.45	17.34	15.83	14.32	14.16
vcVAE ($\beta = 6$)	18.76	**25.18**	15.98	17.33	17.36
mcVAE ($\beta = 6$)	**28.17**	23.87	**20.84**	**18.12**	**17.93**

The best values are highlighted in bold

Information bottleneck [34, 35] suggests that the representation z should be a sufficient feature, and its relevant information in the input variable about the target is minimum. Formally, z is sufficient for the task y means that the mutual information $I(z, y)$ should be maximized, and z is minimum for the input data x means that $I(z, x)$ should be minimized. The objective function of information bottleneck is equivalent to solve the optimization problem:

$$L_{\text{IB}} = I(z, y) - \beta I(x, z) \tag{6}$$

where β is a positive constant managing the trade-off between sufficiency and minimality. In the context of VAE, a sufficient representation is reflected in inference accuracy and reconstruction fidelity [36]. Therefore, $I(z, y)$ need to be replaced by a reconstruction error, and $I(z, x)$ is used to constrain the reconstruction as an additional regularization term. Now the objective function is:

$$L = E_{q(z|x)}[\log p(x|z)] - \beta I(x; z) \tag{7}$$

We can find that the constraint from the information bottleneck is a mutual information term. However, $I(z, x)$ is difficult to compute due to the joint probability distribution $p(z, x)$. Variational inference is an elegant method to estimate it while constructing a variational bound, and $I(z, x)$ can be seen as the uncertainty in x given z in information theory:

$$\begin{aligned} I(z, x) &= H(z) - H(z|x) \\ &= -\int p(z) \log p(z) \mathrm{d}z \\ &\quad + \int \int p(x, z) \log p(z|x) \mathrm{d}z \mathrm{d}x \end{aligned} \tag{8}$$

where $H(\cdot)$ denotes the Shannon entropy, and $H(x|z)$ is conditional entropy. Let $q(z)$ be a variational approximation to $p(z)$:

$$\begin{aligned} &\int p(z) \log p(z) \mathrm{d}z \approx \mathrm{KL}(p(z)||q(z)) \geq 0 \\ \Rightarrow &\int p(z) \log p(z) \mathrm{d}z \geq \int p(z) \log q(z) \mathrm{d}z \end{aligned} \tag{9}$$

Then, we can get the variational bound:

$$\begin{aligned} &\int p(z) \log p(z) \mathrm{d}z \approx \mathrm{KL}(p(z)||q(z)) \geq 0 \\ \Rightarrow &\int p(z) \log p(z) \mathrm{d}z \geq \int p(z) \log q(z) \mathrm{d}z \end{aligned} \tag{10}$$

Now the objective function of the marginal likelihood with mutual information or information bottleneck constraint is:

$$L = E_{q(z|x)}[\log p(x|z)] - \beta D_{\mathrm{KL}}(p(z|x)||q(z)) \tag{11}$$

We can find Eq. (11) is the same as β-VAE, and it proves that there is an approximate equivalence between the information bottleneck (mutual information constraint) and the KL divergence term. Increasing the weight β of mutual information $I(z, x)$ can improve the degree of disentanglement; it is an explanation of the objective function of β-VAE and provides a theoretical source. Stronger mutual information constraint limits model capacity, it is a restriction on the representation that forces the model to infer in a narrow space, and it is similar to our analysis of poor reconstruction.

5 Experimental results

We perform a series of quantitative and qualitative experiments, showing that mcVAE can achieve higher disentanglement scores compared to some baselines, while leading to lower reconstruction error and sharper reconstructions.

5.1 Disentanglement scores

We analyze the disentangling performance of β-VAE, mcVAE and vcVAE on dSprites [37]. The dSprites dataset is designed for disentanglement testing, and it contains 737,280 binary 64 × 64 images of 2D shapes with 5 ground truth factors: shape [4], scale [11], orientation [40], position X [5] and position Y [5]. We perform quantitative evaluations with 5 factors as shown in Table 1.

We see that our mcVAE gives much better disentanglement scores than β-VAE on all factors, 4 out of 5 factors have higher scores than vcVAE. This experiment demonstrates that our mean constraint creates more disentangled representations than KL divergence and variance constraint.

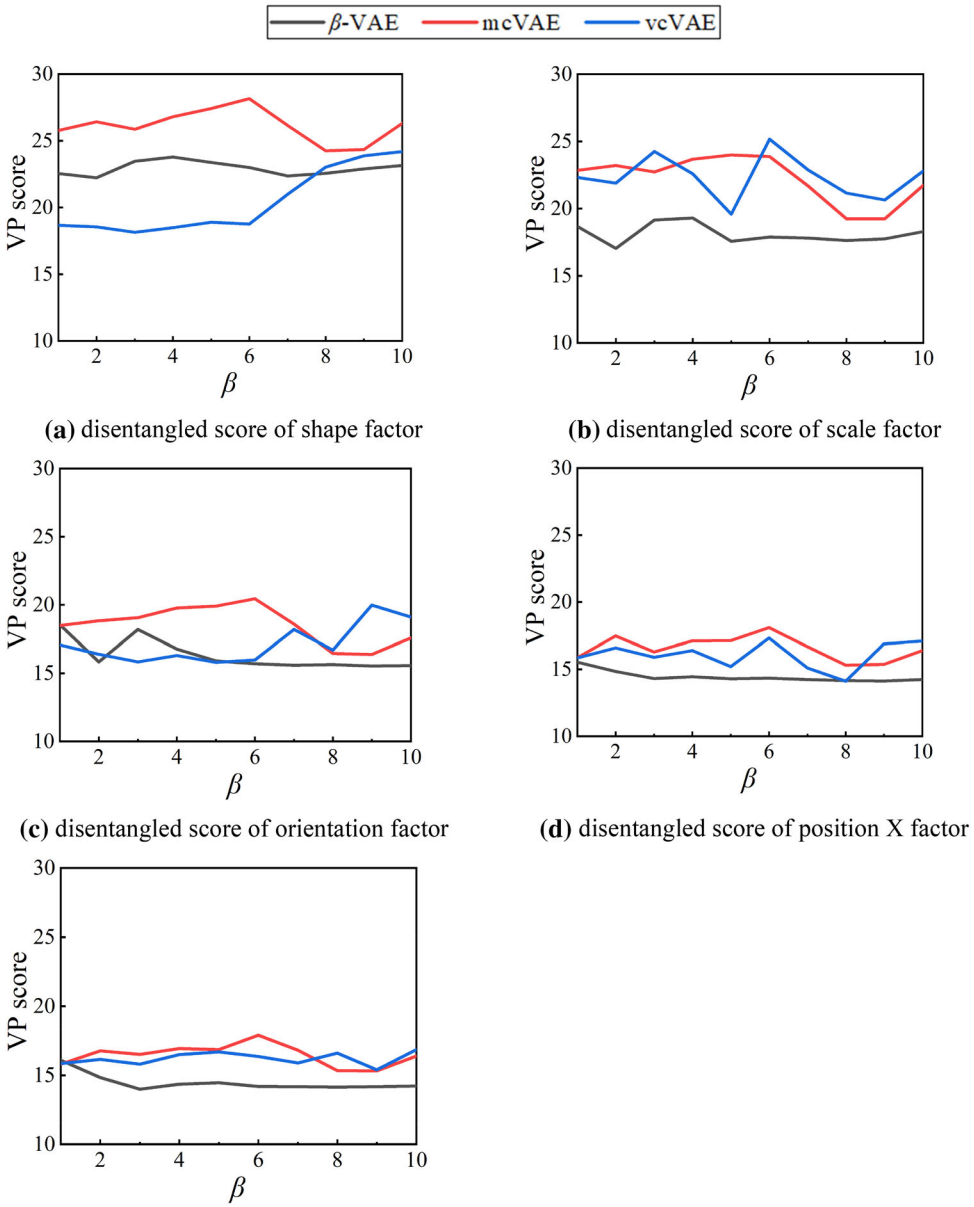


Fig. 5 Disentangled scores of five truth factors measured by VP metric

(a) disentangled score of shape factor

(b) disentangled score of scale factor

(c) disentangled score of orientation factor

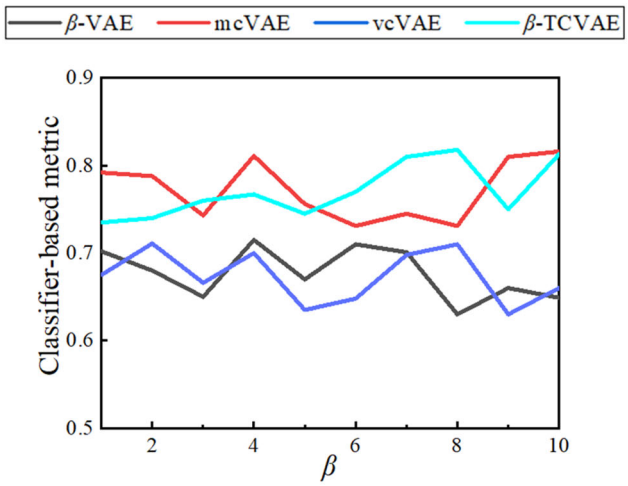
(d) disentangled score of position X factor

(e) disentangled score of position Y factor

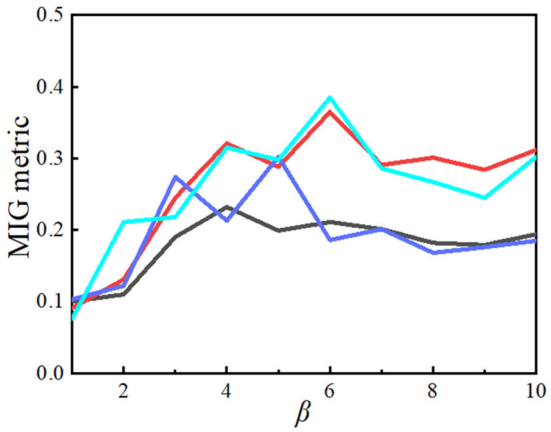
Then, we would like to see how the β affects different modeling algorithms, and we train three models using a range of values [1, 15]. Each truth factor measured by VP metric is shown in Fig. 5.

The best disentanglement scores for mcVAE are noticeably better than β-VAE and vcVAE on most factors. While increasing β often leads to the scores of each factor to increase and then decrease, the trend of the curve is the same as the classifier-based metric [9] and MIG metric [28]. When the value of β in mcVAE is 6, the scores of all truth factors can achieve maximum values, and β-VAE achieves the maximum when the β is 3 or 4. However, the scores of most factors in vcVAE are low and inconsistent for different truth factors, confirming that the variance constraint has little effect on disentanglement learning. For example, when we use vcVAE framework to learn scale factor, the corresponding disentanglement scores vary greatly with increasing β, but the score remains constant on position Y factor.

In Fig. 6, we compare the above three methods and β-TCVAE by a classifier-based metric and MIG metric. Due to the poor stability of the original classifier-based metric [9], we choose the modified metric proposed by [19]. We can see that mcVAE and β-TCVAE give much better disentangled scores than β-VAE and vcVAE on both metrics, and the trend of scores is consistent with our VP metric which illustrates the consistency of these metrics.



(a) Disentangled scores of classifier-based metric



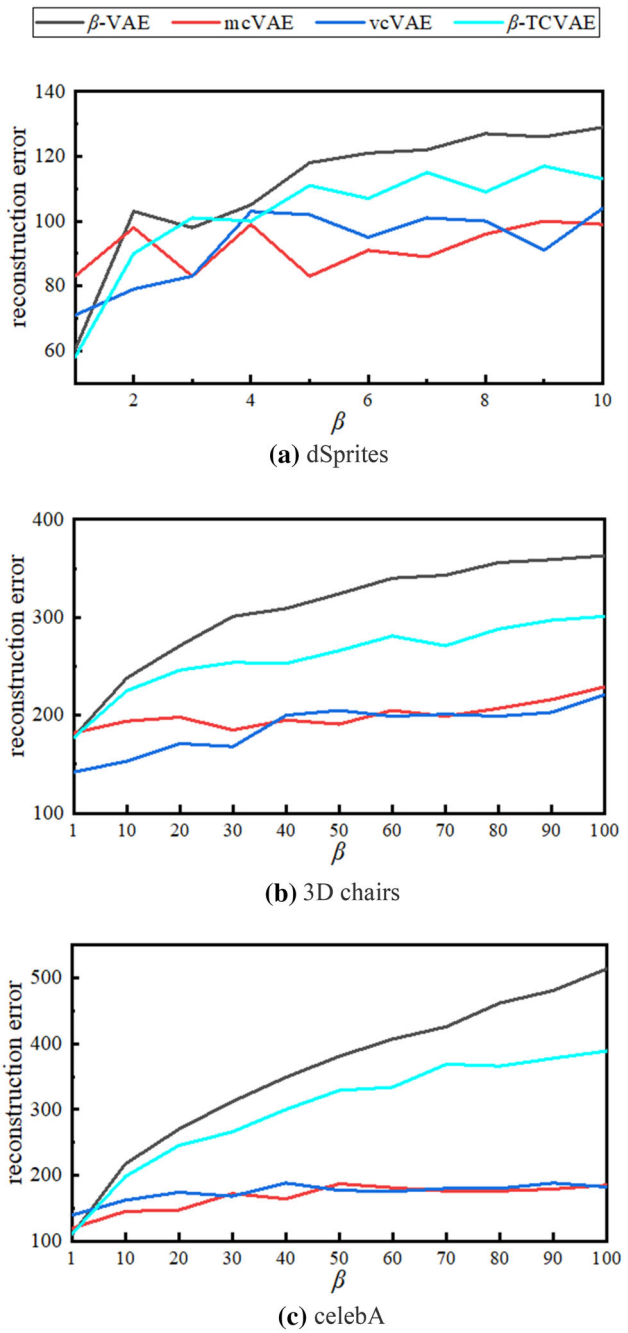
(b) Disentangled score of MIG metric

Fig. 6 Disentangled scores of five truth factors measured by VP metric

However, our disentanglement metric is advantageous than classifier-based metric and MIG metric. VP metric computes disentanglement scores for each factor individually, and we can see that mcVAE is capable of finding shape, orientation, position X and position Y, but struggle to disentangle scale compared to vcVAE. Each model has different disentanglement abilities for different truth factors, and measuring the scores of each factor separately are beneficial to compare disentangling algorithms.

5.2 Disentanglement and reconstruction trade-off

The objective function of mcVAE is a lower bound on the evidence lower bound, and the hyperparameter β is used to trade-off the reconstruction and disentanglement, and we would like to see how the choice of β affects these learning algorithms. The training results using a range of values on dSprites dataset are illustrated in Fig. 7.



(a) dSprites

(b) 3D chairs

(c) celebA

Fig. 7 Reconstruction losses of four methods

We find that both mcVAE and vcVAE can provide a better trade-off between reconstruction and disentanglement than β-VAE and β-TCVAE. With higher values of β, the KL constraint imposing on the mean and variance variables in β-VAE is too strong to learn enough usefulness of the representation, the reconstruction errors have risen dramatically which always leads to blurry and homogeneous reconstructions. In contrast, the reconstruction errors of our methods have a gentle upward trend with increasing β, constraining

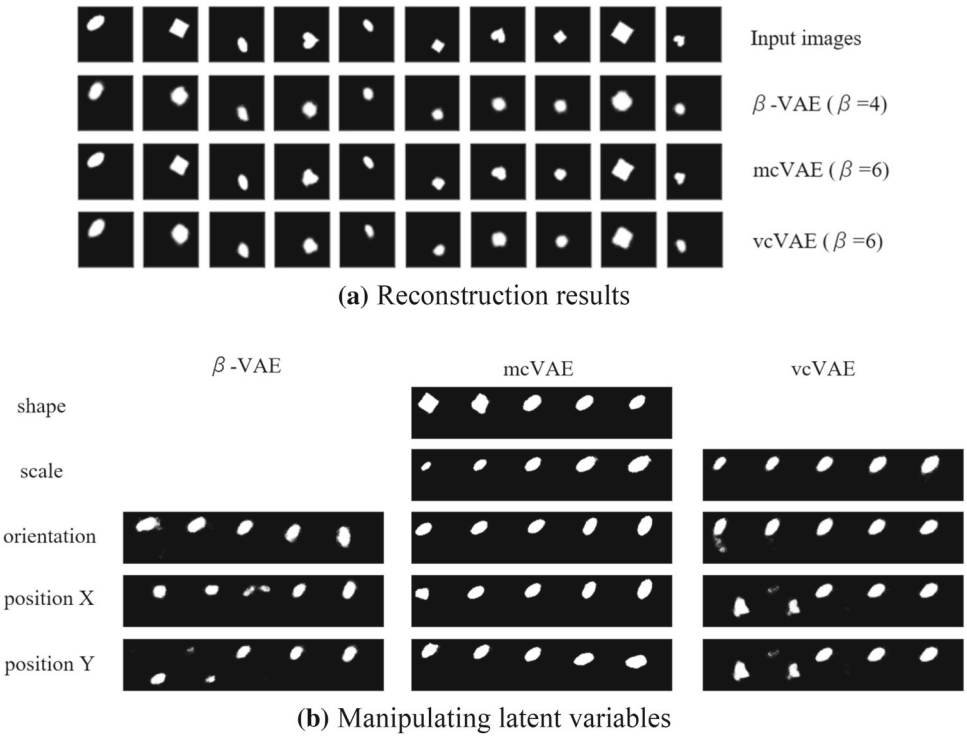


Fig. 8 Qualitative results comparing reconstruction and disentangling performance on dSprites

the mean vector or variance vector individually can adjust the other vector automatically during training. For instance, the values of the mean vector would decrease if the weight of the mean constraint increases, and our algorithm tends to increase the variance slightly to ensure representation space; it guarantees that the representations obtain more sufficient features to complete the reconstruction task.

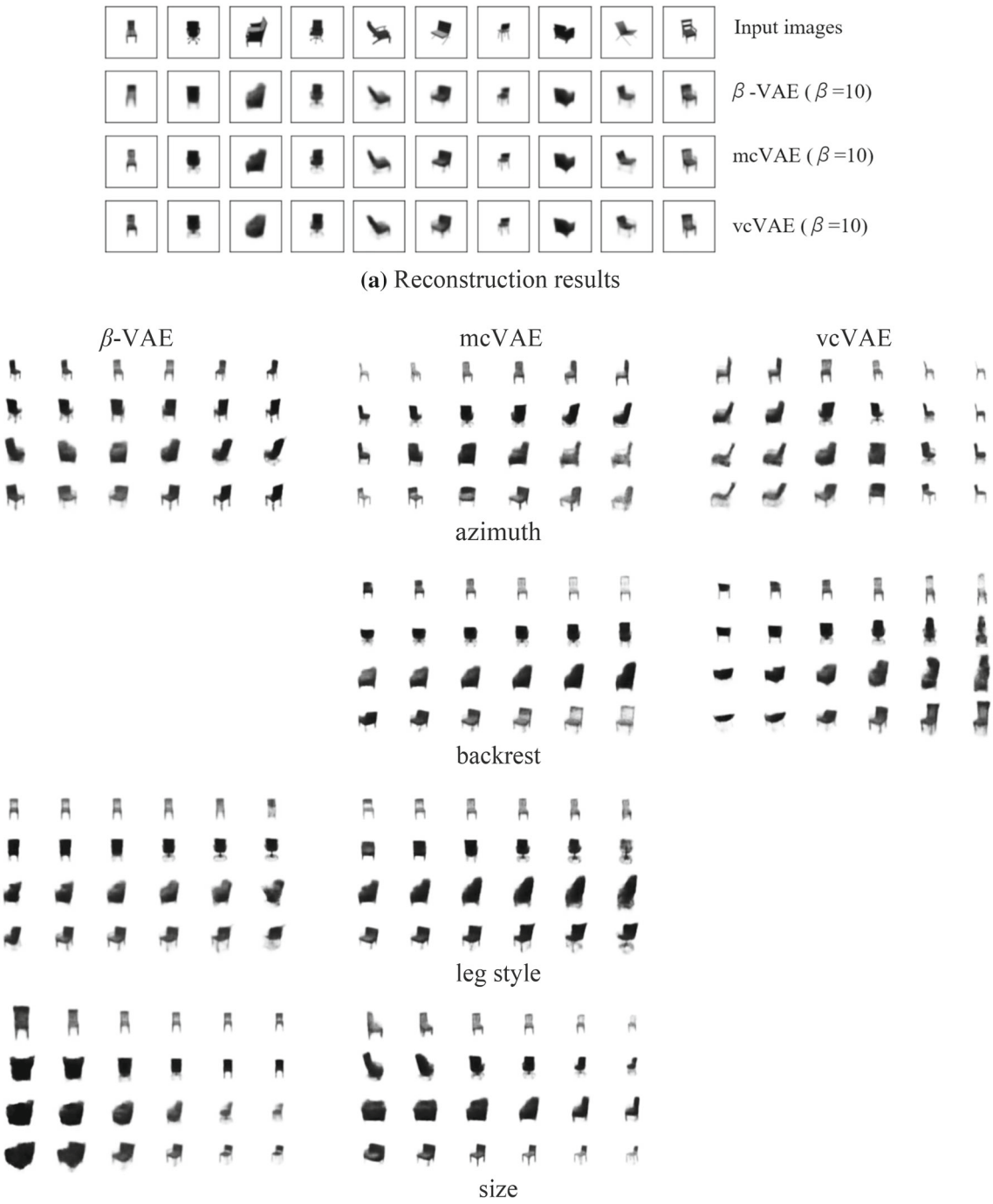
Comprehensive comparison of Figs. 5, 6, and 7, we see that our mcVAE gives much better disentanglement scores than β-VAE while barely sacrificing reconstruction error, and it shows that the disentangling and reconstructing effect of adding the mean constraint term to the VAE objective. Compared with TCVAE, our method demonstrates comparable disentanglement performance and better reconstruction capability.

5.3 Qualitative comparisons

We examined qualitatively the representations and disentangling performance learned by our mcVAE, β-VAE and vcVAE on datasets of dSprites, 3D chairs and celebA.dSprites: Fig. 8 shows reconstruction results and traversals in latent variables learned by β-VAE, mcVAE and vcVAE, the traversal ranges are $[-3, 3]$, $[-2,2]$ and $[-5,5]$, respectively. From Fig. 8a, we see that β-VAE has blurry reconstructions which are unable to recurrence shapes other than a circle, it is because that stronger constraint from KL divergence severely compresses the representation space, leading to the loss of the ability to reconstruct other shapes. In contrast, mcVAE is sensitive to subtle details, such as orientation of the first column and shape of the second column.

In Fig. 8b, β-VAE has shown to be capable of learning three truth factors: orientation, position X and position Y, these factors are entangled with each other. (e.g., orientation is entangled with position Y), vcVAE learns four factors that are entangled and damaged, mcVAE can learn all the truth factors with clear outlines. Furthermore, we find that learning position X and position Y seem to be more difficult than other factors in the three models, it conforms that position X and position Y have lower VP scores in Fig. 5.

3D Chairs: Fig. 9a shows the reconstruction results of three models. The reconstructions of β-VAE are blurry, many of which are wrong, such as the leg style in the second column. In contrast, the reconstructions of mcVAE recovered more details than β-VAE and vcVAE. For instance, in the first column, images from VAE and vcVAE have the wrong azimuth, but the reconstruction of mcVAE is almost the same as that of the input samples. Disentangling performance of three models by manipulating latent variables is shown in Fig. 9b. β-VAE traversal is over the $[-2, 2]$ range which can learn three attributes, but some factors do not seem to be consistent for all inputs such as the leg style. Our mcVAE (the traversal range is $[-1, 1]$) can learn more attributes than β-VAE and vcVAE (the traversal range is $[-5, 5]$) with sharper images, and it can be proved that mean constraint is capable of learning sensible factors of variation.



(a) Reconstruction results

(b) Manipulating latent variables

Fig. 9 Qualitative results comparing reconstruction and disentangling performance on 3D chairs

CelebA: Comparing the reconstruction results of the three datasets in Fig. 10a, we see that the reconstruction of β-VAE has stronger homogeneity if the dataset is complex. The extremely compressed representation space loses most features, resulting in all reconstructed samples being very similar. For example, the third input image and its reconstruction have completely different characteristics, whether it is angle, gender or expression. Although it learns some disentangled attributes, these attributes do no correlation with the input sample; we can regard it as invalid disentanglement.

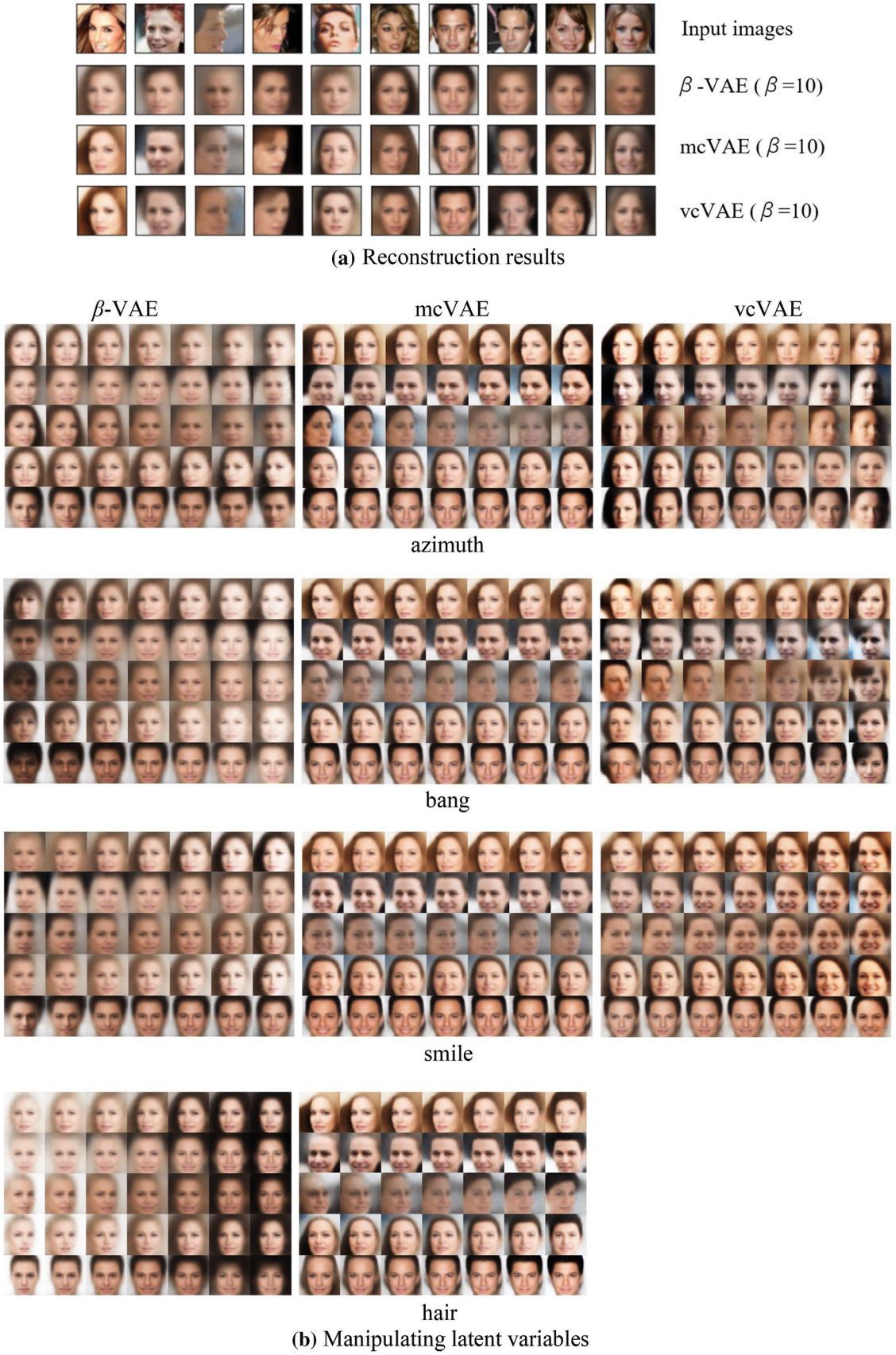
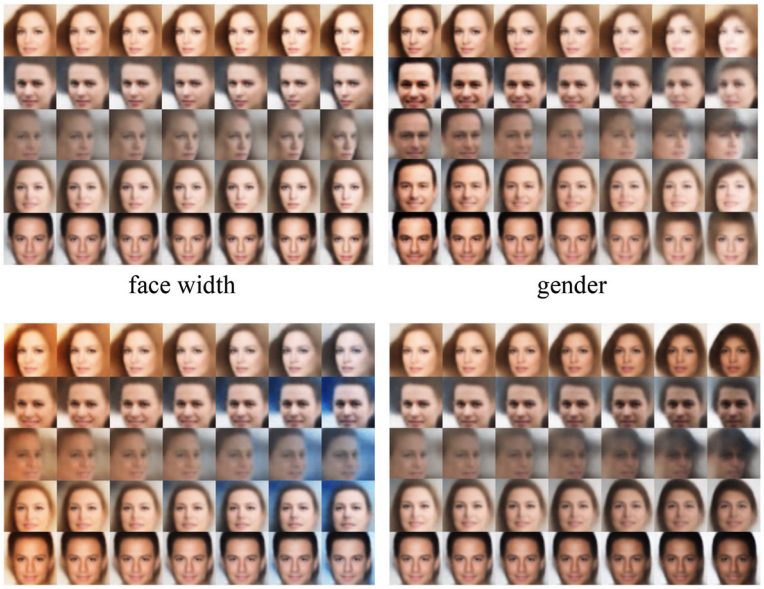


Fig. 10 Qualitative results comparing reconstruction and disentangling performance on celebA dataset

Fig. 10 continued



(c) New disentangled factors by mcVAE

Figure 10b shows that 4 attributes are discovered by three VAE frameworks without supervision; the same original faces in Fig. 10a are tested for three kinds of VAE models. The disadvantage of β-VAE is that the traversal is ambiguous, and all factors are entangled with nuances, vcVAE learns only 3 attributes, and the changes of the traversal are exaggerated and unreal. In contrast, our mcVAE shows better disentangled and reconstruction performance than β-VAE and vcVAE. For instance, in azimuth, the right column images generated by β-VAE are fuzzy and disturbed by noise seriously; the images generated by vcVAE are blurry, and most reconstructions have minor changes about azimuth; in contrast, all the images from mcVAE keep a higher degree of disentanglement and reconstruction quality, the azimuth in the reconstructed sample is disentangled with other factor.

In β-VAE, the bang is always entangled with skin color and noise, other features like the outline and size of the face also change significantly, and the lower degree of disentanglement is seen in smile and hair attributes. The changes of bang generated by vcVAE are inauthentic and entangled with many factors such as gender and azimuth. However, the reconstructions of mcVAE are clearer than β-VAE and vcVAE, and only the shape of bang changes when manipulating latent variables. Among other attributes, mcVAE also has better disentanglement performance than other methods.

In addition, mcVAE can generate rare samples including face width, gender, hue and skin color as depicted in Fig. 10c, showing the ability to meaningful generalization and extrapolation. Noting that each attribute has a different degree of influence on the image. For example, when we manipulate latent variables of skin color, the overall structure of the image changes such as the outline of the face and the shape of the hair, some details such as the mustache when transforming from female to male. In contrast, skin has little effect on the image, and other attributes remain fixed.

6 Conclusion and future work

We introduce mcVAE, a novel method for a higher degree of disentanglement than β-VAE with better reconstruction quality. By analyzing the source of poor reconstruction in KL divergence, we find that an additional mean constraint term can help to achieve better disentanglement scores and learn more properties than β-VAE. To quantitatively evaluate our approach, we identified the weaknesses of the classifier-based metrics; then, we proposed an alternative metric, named Variance Proportion Metric that is comprehensible and classifier-free without hyperparameters.

Learning disentanglement representations in a completely unsupervised manner are still a difficult problem; some key problems need to be solved urgently: The essence of disentanglement is unknown, and all the existing metrics are not suitable or applicable for the real-world applications due to uncertain truth factors. We hope that the mean constraint can bring new insights into the nature of disentanglement. The low dimensional latent variable is a basic shortcoming of VAE, which is hard to reconstruct clear samples like GAN-based methods. Although we have improved the reconstruction ability of the disentanglement representation as much as possible, it is difficult for our method to learn

the interpretable factors of complex data due to the limited capacity of VAE model. Therefore, improving disentangled performance with real world is also our expectation.

Declarations

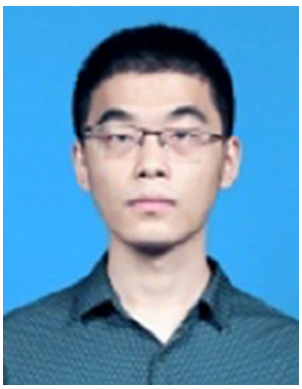
Conflict of interest We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, and there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled "mcVAE: Disentangling by Mean Constraint".

Data availability No new data were created during the study.

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
2. Higgins, I., Amos, D., Pfau, D., et al.: Towards a Definition of Disentangled Representations. arXiv:1812.02230 (2018)
3. Liu, X., Huang, H., Wang, W., et al.: Multi-view 3D shape style transformation. Vis. Comput. **38**(6), 669–684 (2021)
4. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: International Conference on Learning Representations (2017)
5. Park, S., Hwang, S., Kim, D., et al.: Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. Proc. AAAI Conf. Artif. Intell. **35**(3), 2403–2411 (2021)
6. Huang, X., Wang, M., Gong, M.: Fine-grained talking face generation with video reinterpretation. Vis. Comput. **37**, 1–11 (2020)
7. Zhang, K.Y., Yao, T., Zhang, J., et al.: Face anti-spoofing via disentangled representation learning. In: European Conference on Computer Vision, pp. 641–657. Springer, Cham (2020)
8. Wang, G., Han, H., Shan, S., et al.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6678–6687 (2020)
9. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: Beta-vae: learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations (2017)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (2014)
11. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: ICML (2014)
12. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: NIPS (2016)
13. Schmidhuber, J.: Learning factorial codes by predictability minimization. Neural Comput. **4**(6), 863–879 (1992)
14. Desjardins, G., Courville, A., Bengio, Y.: Disentangling factors of variation via generative entangling. arXiv:1210.5474 (2012)
15. Tang, Y., Salakhutdinov, R., Hinton, G.: Tensor analyzers. In: International Conference on Machine Learning. PMLR, pp. 163–171 (2013)
16. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv:1511.05644 (2015)
17. Makhzani, A., Frey, B.J.: Pixelgan autoencoders. In: Advances in Neural Information Processing Systems (2017)
18. Brakel, P., Bengio, Y.: Learning independent features with adversarial nets for non-linear ica. arXiv:1710.05050 (2017)
19. Kim, H., Mnih, A.: Disentangling by factorizing. In: International Conference on Machine Learning. PMLR, pp. 2649–2658 (2018)
20. Lee, W., Kim, D., Hong, S., et al.: High-fidelity synthesis with disentangled representation. In: European Conference on Computer Vision, pp. 157–174. Springer, Cham (2020)
21. Jeon, I., et al.: Ib-gan: disengangled representation learning with information bottleneck generative adversarial networks. In: 35th AAAI Conference on Artificial Intelligence (2021)
22. Zhao, S., Song, J., Ermon, S.: Infovae: information maximizing variational autoencoders. arXiv:1706.02262 (2017)
23. Achille, A., Soatto, S.: Information dropout: learning optimal representations through noisy computation. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
24. Hu, M.F., Liu, J.W.: Optimal representations of CapsNet by information bottleneck. In: ICANN (2021)
25. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, pp. 3856–3866 (2017)
26. Mathieu, E., Rainforth, T., Siddharth, N., et al.: Disentangling disentanglement in variational autoencoders. In: International Conference on Machine Learning. PMLR, pp. 4402–4412 (2019)
27. Hoffman, M.D., Johnson, M.J.: Elbo surgery: yet another way to carve up the variational evidence lower bound. In: Workshop in Advances in Approximate Bayesian Inference, NIPS, vol. 1, No. 2 (2016)
28. Chen, R.T.Q., Li, X., Grosse, R.B., et al.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
29. Gyawali, P., Li, Z., Knight, C., et al.: Improving disentangled representation learning with the beta Bernoulli process. In: IEEE International Conference on Data Mining, pp. 1078–1083 (2019)
30. Dupont, E.: Learning disentangled joint continuous and discrete representations. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
31. Yang, H.H., Amari, S.-i: Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information. Neural Comput. **9**(7), 1457–1482 (1997)
32. Eastwood, C., Williams, C.K.: A framework for the quantitative evaluation of disentangled representations. In: International Conference on Learning Representations (2018)
33. Karaletsos, T., Belongie, S., Rätsch, G.: Bayesian representation learning with oracle constraints. arXiv:1506.05011 (2015)
34. Tishby, N., Pereira, F.C.N., Bialek, W.: The information bottleneck method. CoRR, vol. physics/0004057 (2000)
35. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. In: 2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26–May 1, 2015.1em plus 0.5em minus 0.4em, pp. 1–5 (2015)
36. Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information. CoRR, vol. abs/1703.00810 (2017)
37. Matthey, L., Higgins, I., Hassabis, D., Lerchner, A.: dsprites: disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/ (2017)

Ming-Fei Hu born in 1993. He received the B.S. degree in 2015. He is currently working toward the Ph.D. degree in control theory and control engineering with the department of automation, College of Information Science and Engineering, China University of Petroleum, Beijing. His research interests include deep learning and pattern recognition and intelligent Systems.



Ze-yu Liu born in 1993. He received the B.S. degree from Jilin University and M.S. degree from the Institute of Software, Chinese Academy of Sciences. He is currently working toward the Ph.D. degree in control theory and control engineering with the department of automation, College of Information Science and Engineering, China University of Petroleum Beijing. His research interests include deep learning, pattern recognition and intelligent Systems.

Jian-Wei Liu born in 1966. He received the Ph.D. degree in control theory and control engineering from DongHua University in 2006. He is now an associate professor with the department of automation, College of Information Science and Engineering, China University of petroleum, Beijing. His research interests include pattern recognition and intelligent Systems, machine learning, analysis, prediction and control of complex nonlinear system. In these areas, he has published over 210 papers in international journals or conference proceedings.