AN INFORMATION THEORETIC PERSPECTIVE ON DISENTANGLED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing works on disentangled representation learning usually lie on a common assumption: all factors in a disentangled representation should be independent. We argue that this assumption is not sufficient and another assumption is vital for disentangled representation learning: information contained in each factor of a disentangled representation is irrelevant to others, i.e. the containing information about data of factors is isolated. We formulate this assumption into two equivalent equations via mutual information, and theoretically show its relation with independence and conditional independence of factors in a representation. Meanwhile, we prove that conditional independence is satisfied in encoders of VAEs due to "no-sharing-parameter block" and reparameterization trick. To highlight the importance of the proposed assumption, we show in experiments that violating the assumption leads to decline of disentanglement. Based on this assumption, we further propose to split the deeper layers in encoder to ensure parameters in these layers are not shared for different factors. The proposed encoder, called *Split Encoder*, can be applied into other models and shows significant improvement in unsupervised learning of disentangled representations and reconstructions.

1 INTRODUCTION

Learning disentangled representations has been considered as an important step towards interpretable and more effective machine learning (Bengio et al., 2013; Bengio, 2007; Lake et al., 2017; Shanmugam, 2018; Tschannen et al., 2018; Schmidhuber, 1992). The disentangled representations are proved to be interpretable or semantically meaningful (Chen et al., 2016; Kumar et al., 2017), robust to adversarial attacks (Alemi et al., 2017), more generalizable (Steenbrugge et al., 2018) and correlated to fairness (Locatello et al., 2019). They are also useful to many downstream tasks, including sequential data generating (Yingzhen & Mandt, 2018), reinforcement learning (Higgins et al., 2017b; Nair et al., 2018), robot learning (Laversanne-Finot et al., 2018), transfer (Liu et al., 2018) and few shot learning (Janzing et al., 2012; Bengio et al., 2013), etc. Although there is no formal definition for a disentangled representation except some attempts (Higgins et al., 2018), we adopt one from (Bengio et al., 2013): a representation in which each factor corresponds to a single factor of variation in data, and meanwhile is invariant to other factors of variation.

There exist unsupervised approaches to learn disentangled representations, based on generative models including generative adversarial nets (GANs) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Kingma & Welling, 2014). And the VAE-based models have become mainstream due to the stability of VAEs. Although different VAE-based models for learning disentangled representations are proposed from different motivations, such as limiting the bottleneck capacity (Higgins et al., 2017a; Burgess et al., 2018), penalizing the total correlation (Kim & Mnih, 2018; Chen et al., 2018) and matching factorized priors (Kumar et al., 2017), they can be attributed to factorizing the distribution of representations (Locatello et al., 2018a;b; Kim & Mnih, 2018; Chen et al., 2018). Therefore, these models lie on a common assumption: the distribution of a disentangled representation is factorized, i.e. factors are independent. However, Locatello et al. (2018a) pointed out that this assumption cannot ensures disentanglement, and inductive biases are necessary.

In this work, we argue that the independence of factors in a representation is not sufficient for the disentanglement since it does not model the relation between representations and data, which should be the foundation of disentanglement. Thus we consider another assumption: the information about

data contained in a single factor of a disentangled representation is irrelevant to other factors, or factors in a disentangled representation contain isolated information. This assumption is formulated into two equivalent equations in terms of mutual information. Furthermore, we theoretically build up its relation with independence and conditional independence of factors in the representation, showing that conditional independence is also an important property for disentangled representations. Then we divide the encoders in VAEs into "sharing-parameter block" and "no-sharing-parameter block", and prove that conditional independence originates from no-sharing-parameter block and reparameterization trick (Kingma & Welling, 2014), which can be viewed as inductive biases for disentangled representation learning.

To verify our assumption, we perform experiments by violating the conditional independence of factors in representations and find it lead to decline of disentanglement, which demonstrates the importance of conditional independence and supports our assumption. Motivated by this, we propose the split encoder to improve model capacity of no-sharing-parameter block, which can help different factors learn isolated information, and thus facilitates disentanglement. Our experiments show that the split encoder can significantly improve disentanglement of representations learned by VAE and FactorVAE (Kim & Mnih, 2018) on dSprites (Higgins et al., 2017a) and Cars3D (Reed et al., 2015) datasets when the penalty for independence is strong, and also improves reconstructions.

The main contributions of this paper can be summarized as follows:

- We propose a novel and fundamental assumption for disentangled representation learning, and connect it with independence and conditional independence. We show the importance of both conditional independence and our assumption for disentanglement in experiments.
- We mathematically prove the sources of conditional independence are no-sharing-parameter block and reparameterization trick, which can be viewed as inductive biases on encoders.
- Based on the above analysis, we then develop a simple and effective architecture called split encoder to improve disentanglement and reconstructions, which can be applied into other models to improve their performance.
- Experimental results on dSprites and Cars3D show our approach combined with vanilla VAE and FactorVAE can significantly improve performance on disentanglement when the penalty for independence is strong enough, and it also improves reconstructions.

2 Theoretical Analysis

In this section, we first discuss the intuition and common assumption of disentangled representation learning. Then we point out that the common assumption is not sufficient for disentanglement, and motivated by this we propose another assumption from an information theoretic perspective. Finally, we obtain two equivalent equations to formulate the proposed assumption, and connect it with independence and conditional independence.

Although there is no formal definition for disentangled representations, the key intuition is that a disentangled representation should isolate the distinct and semantic factors of variations in data. In other words, a single factor in a disentangled representation is only sensitive to the changes of a single underlying factor of variation in data, while being relatively invariant when other underlying factors change. This statement lies on the relation between factors in representations and data, and also indicates the independence of factors in a disentangled representation. Unfortunately, underlying factors of variation in data commonly have no explicit expression, and thus the statement is hard to be formulated into mathematical expression.

Most variational auto-encoders (VAEs) (Kingma & Welling, 2014) based models for unsupervised disentanglement learning lie on a common assumption: factors in a disentangled representation are independent. VAEs assume a factorized prior $p(\mathbf{z})$ like standard normal distribution and utilize a generative network to generate data and parameterize $p(\mathbf{x}|\mathbf{z})$. Meanwhile, an inference network is used to learn representations from data distribution $q(\mathbf{x})$ and parameterize $p(\mathbf{z}|\mathbf{x})$. The objective of vanilla VAE is equivalent to the KL divergence $D_{\mathrm{KL}}(q(\mathbf{z},\mathbf{x})\|p(\mathbf{z},\mathbf{x}))$, thus the aggregate posterior $q(\mathbf{z})$ is pushed to match $p(\mathbf{z})$ and then becomes factorized – detailed discussion about this can be found in Appendix. Due to this property, most works focus on factorizing the aggregated posterior

$q(\mathbf{z})$ to improve disentanglement. Therefore, those proposed models based on VAEs (Higgins et al., 2017a; Burgess et al., 2018; Kumar et al., 2018; Kim & Mnih, 2018; Chen et al., 2018) are attributed to penalizing the divergence between $q(\mathbf{z})$ and $\prod_j q(\mathrm{z}_j)$, and thus enhancing independence of factors in representations, which coincides with the common assumption.

However, the common assumption is not sufficient for disentanglement of a representation. Because it is only related to the inner property of a representation. While as pointed out above, the relation between factors of a representation and data is also the foundation of disentanglement. Moreover Locatello et al. (2018a) prove the impossibility to learn disentangled representations without inductive biases in an unsupervised manner by simply ensuring the independence. To model the relation between factors of a disentangled representation with data, more assumptions should be considered.

2.1 THE PROPOSED ASSUMPTION

Our key insight is that the relation between factors in a representation with data can be described from the perspective of containing information. Since each factor in a disentangled representation corresponds to a single factor of variation in data, each factor only contains the information of the corresponding factor of variation in data. In other words, different factors in a disentangled representation contain absolutely different information about data, and the containing information about data in a single factor is irrelevant to other factors. Motivated by this insight, we aim at proposing an information-based assumption to describe the relation between factors in a disentangled representation with factors of variation in data. For mathematical analysis of the information terms, we adopt mutual information to measure the information of a variable contained by another:

$$I(\mathbf{x};\mathbf{z}) = H(\mathbf{z}) - H(\mathbf{z}|\mathbf{x}) \tag{1}$$

where entropy $H(\mathbf{z}) = -\mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})]$ and conditional entropy $H(\mathbf{z}|\mathbf{x}) = -\mathbb{E}_{q(\mathbf{z},\mathbf{x})}[\log q(\mathbf{z}|\mathbf{x})]$ are measures of uncertainty. Hence mutual information measures the uncertainty reduction of one variable when another is given. The mutual information satisfies the symmetry $I(\mathbf{x};\mathbf{z}) = I(\mathbf{z};\mathbf{x})$.

Therefore, for a single factor z_j in a disentangled representation $\mathbf{z}$, its containing information about the corresponding underlying factor of variation in data $\mathbf{x}$ can be expressed as $I(\mathbf{z};\mathbf{x})$. To represent the irrelevance of z_j to other information in data, we involve conditional mutual information:

$$I(\mathrm{z}_j;\mathbf{x}|\mathbf{z}_{i\in S_{-j}}) = H(\mathrm{z}_j|\mathbf{z}_{i\in S_{-j}}) - H(\mathrm{z}_j|\mathbf{x},\mathbf{z}_{i\in S_{-j}}) \tag{2}$$

where $j \in \{1,2,\ldots,J\}$, S_{-j} is any subset of $\{1,\ldots,j-1,j+1,\ldots,J\}$ and $\mathbf{z}_{i\in S_{-j}}$ denotes all factors with subscript index in S_{-j}. The conditional mutual information is the difference of two conditional entropy, representing the containing information of z_j on the corresponding underlying factor of variation in data $\mathbf{x}$ when other factors $\mathbf{z}_{i\in S_{-j}}$ are given. Considering the corresponding factors of variation in data of $\mathbf{z}_{i\in S_{-j}}$, they are fixed when $\mathbf{z}_{i\in S_{-j}}$ are given. Hence the conditional mutual information term above can describe the their relation with z_j.

Now we give more strict formulation. Since z_j only contains information of the corresponding factor of variation in data, denoted by $I(\mathrm{z}_j;\mathbf{x})$, its containing information should be irrelevant to other factors of variation. Therefore, the conditional mutual information term $I(\mathrm{z}_j;\mathbf{x}|\mathbf{z}_{i\in S_{-j}})$ should be equal to $I(\mathrm{z}_j;\mathbf{x})$. To conclude, we propose an assumption as follows:

Assumption. *Suppose $\mathbf{z} \in \mathbb{R}^J$ is a disentangled representation of data x and $I(\mathbf{z};\mathbf{x}) > 0$, then for any single factor z_j, its containing information in data is irrelevant to other factors $\mathrm{z}_{i\in S_{-j}}$, i.e.*

$$I(\mathrm{z}_j;\mathbf{x}|\mathbf{z}_{i\in S_{-j}}) = I(\mathrm{z}_j;\mathbf{x}) \tag{3}$$

where $j \in \{1,2,\ldots,J\}$ and S_{-j} is any subset of $\{1,\ldots,j-1,j+1,\ldots,J\}$.

For further investigation on our proposed assumption, we consider turning Eq. 3 into a more intuitive one. Since the assumption is motivated by our insight that different factors in a disentangled representation contain different information in data, the Eq. 3 should intuitively reflect the separation of those information terms. Using the chain rule of mutual information, we prove that Eq. 3 is equivalent to the decomposition of mutual information as follows:

Lemma 1. *$I(\mathrm{z}_j;\mathbf{x}|\mathbf{z}_{i\in S_{-j}}) = I(\mathrm{z}_j;\mathbf{x})$ for any j is equivalent to the following equation:*

$$I(\mathbf{z}_{i\in S};\mathbf{x}) = I(\mathrm{z}_j;\mathbf{x}) + I(\mathbf{z}_{i\in S_{-j}};\mathbf{x}) \tag{4}$$

where $S = \{j\} \cup S_{-j}$ is any subset of $\{1,2,\ldots,J\}$, j is any single element in S.

Based on our assumption and the lemma, we can draw an intuitive conclusion that in a disentangled representation $\mathbf{z}$, the containing information of z_j is different from those of $\mathbf{z}_{i\in S_{-j}}$, and thus their corresponding factors of variation in data are distinct. Furthermore, by iteratively using Eq. 1, we can decompose the mutual information $I(\mathbf{z}_{i\in S};\mathbf{x})$ into the sum of $I(\mathrm{z}_i;\mathbf{x})$, which shows that the containing information of different factors in a disentangled representation are different and coincide with our insight. Hence our assumption can be summarized into an equivalent statement:

Assumption. *Suppose $\mathbf{z}\in\mathbb{R}^J$ is a disentangled representation of data x and $I(\mathbf{z};\mathbf{x})>0$, then for any subset S of $\{1,2,\ldots,J\}$, the containing information of $\mathrm{z}_i\in S$ in data is isolated, i.e.*

$$I(\mathbf{z}_{i\in S};\mathbf{x})=\sum_{i\in S}I(\mathrm{z}_i;\mathbf{x}) \tag{5}$$

2.2 Relation with Independence and Conditional Independence

From the containing information perspective, an intuitive assumption of disentangled representations has been proposed to describe the relation between factors in a disentangled representation and data. Two equivalent expressions Eq. 3 and Eq. 5 are also derived. Although the two equations have clear physical meanings, the structured relation between a disentangled representation and data in the proposed assumption is still implicit. Hence in this subsection, we aim at investigating the structure of the proposed assumption from Eq. 5.

To analyze Eq. 5, we investigate the conditions of establishing Eq. 5 by considering $I(\mathbf{z}_{i\in S};\mathbf{x})-\sum_{i\in S}I(\mathrm{z}_i;\mathbf{x})$. We find that this term is tightly related to independence and conditional independence of factors in a representation. Specifically, we build up the following lemma:

Lemma 2. *For any subset S of $\{1,2,\ldots,J\}$, we have:*

$$I(\mathbf{z}_{i\in S};\mathbf{x})-\sum_{i\in S}I(\mathrm{z}_i;\mathbf{x})=\mathbb{E}_{q(\mathbf{x})}[D_{\mathrm{KL}}(q(\mathbf{z}_{i\in S}|\mathbf{x})\|\prod_{i\in S}q(\mathrm{z}_i|\mathbf{x}))]-D_{\mathrm{KL}}(q(\mathbf{z}_{i\in S})\|\prod_{i\in S}q(\mathrm{z}_i)) \tag{6}$$

In the lemma above, the left term is equal to the difference of two KL divergences. The first divergence measures the level of conditional independence for factors $\mathbf{z}_{i\in S}$ conditioned on data $\mathbf{x}$, and the second one measures the level of independence for $\mathbf{z}_{i\in S}$. Specifically, when the conditional distribution $q(\mathbf{z}_{i\in S}|\mathbf{x})$ is more factorized, the first divergence is closer to zero, which indicates higher level of conditional independence. Similarly, when the distribution $q(\mathbf{z}_{i\in S})$ is more factorized, the second divergence is more closer to zero, and hence the level of independence is higher.

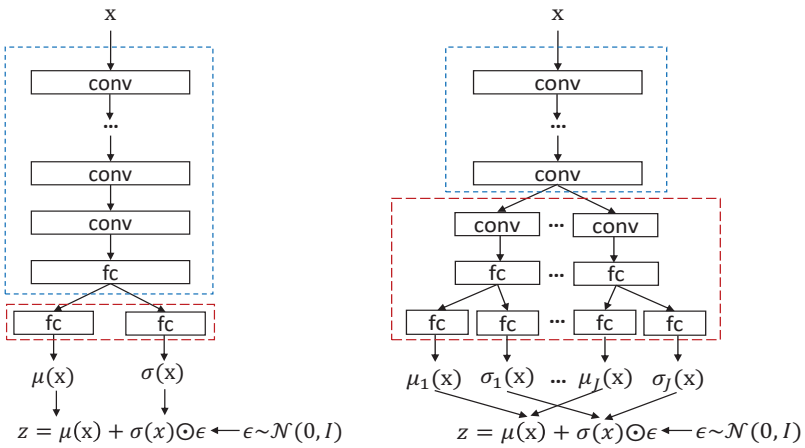
Eq. 6 shows the importance of conditional independence and independence in our assumption: when the factors in a representation are independent and conditionally independent, our assumption is satisfied and thus the representation is oriented towards disentanglement. Note that conditional independence and independence are sufficient but not necessary to our assumption. This shows that our assumption is different and independent to the common assumption.

It is important to note that our assumption is necessary but not sufficient for learning meaningful disentangled representations. For example, if $q(\mathbf{z}|\mathbf{x})=q(\mathbf{z})$, then the two divergences in Eq. 6 are equal, and thus Eq. 5 holds and our assumption is satisfied, but in this case z does not contain any information about data $\mathbf{x}$. To prevent this meaningless case, we should pay attention to the reconstruction error, which reflects the containing information of a representation about data. Lower reconstruction error indicates that more information about data is contained in representation.

As mentioned above, the common assumption highlights the independence of factors in a disentangled representation, which is enhanced via the objective in VAEs. Given the establishment of the common assumption, our proposed assumption is equivalent to conditional independence, which is exactly the structured relation between a disentangled representation and data. Motivated by this consideration, we investigate the inductive biases on model about conditional independence, and successfully find the sources of conditional independence in encoders of VAEs.

3 Inductive Biases on Encoders

On condition of independence, the proposed assumption is equivalent to conditional independence. In this section we analyze the inductive biases about conditional independence, and then propose a simple architecture to improve disentanglement based on our assumption and the inductive biases.



(a) Structure of vanilla encoder (b) Structure of split encoder

Figure 1: **Structure of vanilla encoder and split encoder.** The blue and short-dash box denotes sharing-parameter block, and the red and long-dash box represents no-sharing-parameter block.

3.1 Sources of Conditional Independence

Motivated by our proposed assumption, we find that there are some inductive biases on encoders of VAEs to ensure conditional independence. Considering $q(\mathbf{z})$ is pushed to match a factorized prior $p(\mathbf{z})$ in VAE-based models and thus encourages independence and the common assumption is satisfied, this discovery shows that our assumption is also satisfied in these models.

In VAEs, the encoders encode data into a Gaussian distribution of representation $q(\mathbf{z}|\mathbf{x})$ via reparameterization trick. Specifically, the data is encoded into a mean $\mu(\mathbf{x})$ and a variance $\sigma(\mathbf{x})$, then the representation is sampled from $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}(\mathbf{x}), \sigma^2(\mathbf{x})\boldsymbol{I})$ with reparameterization trick:

$$\boldsymbol{z} \sim q(\mathbf{z}|\mathbf{x}) \Leftrightarrow \boldsymbol{z} = \mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \epsilon, \epsilon \sim \mathcal{N}(0, \boldsymbol{I}) \tag{7}$$

For analyzing the outputs of those encoders, we divide them into two blocks: if the parameters of several layers are shared for different factors in a representation, the block consists of these layers is called "sharing-parameter block". Conversely, those layers without shared parameters for different factors form "no-sharing-parameter block". In vanilla VAE, the last layers in encoders are fully-connected nets, which are no-sharing-parameter block (see Fig. 1). Obviously, factors in the output of no-sharing-parameter block are independent conditioned on the input, thus factors in $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are conditionally independent. Considering the independence of factors in ϵ, we have:

$$q(\mu|\mathbf{x}) = \prod_{j=1}^{J} q(\mu_j|\mathbf{x}) \quad q(\sigma|\mathbf{x}) = \prod_{j=1}^{J} q(\sigma_j|\mathbf{x}) \quad q(\epsilon) = \prod_{j=1}^{J} q(\epsilon_j) \tag{8}$$

From these equations, we can obtain the conditional independence of factors in a representation:

$$q(\mathbf{z}|\mathbf{x}) = \prod_{j=1}^{J} q(\mathbf{z}_j|\mathbf{x}) \tag{9}$$

To conclude, conditional independence originates from no-sharing-parameter block and the independence of noise in reparameterization trick, which can be regarded as inductive biases on model. These two inductive biases are very important for disentangled representation learning. We will see in experiments that without no-sharing-parameter block or independent noise, the learned representations are more entangled, which indicates that our proposed assumption is vital for disentanglement.

3.2 Split Encoder

In this subsection, we aim at devising a simple and effective architecture to enhance disentanglement in encoders based on our assumption and the inductive biases about conditional independence.

According to the proposed assumption, factors in a disentangled representation contain isolated information about data. Intuitively no-sharing-parameter block is beneficial to learn isolated information from input, but in previous works, the no-sharing-parameter block is a single fully-connected layer, which limits the model ability to learn isolated information. A limited no-sharing-parameter block may encourage the conditional distribution $q(\mathbf{z}|\mathbf{x})$ to match $q(\mathbf{z})$ without capturing any information about data, which violates the essence of our assumption.

To tackle this problem, we propose split encoder, which improves model capacity of the no-sharing-parameter block without increasing the depth of encoder. As Fig. 1 shows, in split encoder more convolution layers and fully-connected layers are created for each single factor in representation without parameter sharing, then the outputs are concatenated to calculate a representation with reparameterization trick. In experiments, we only create one convolution layer and two fully-connected layers for each factor due to the increase of parameters.

From the perspective of containing information, split encoder coincides with the essence of the proposed assumption: factors in a disentangled representation contain isolated information. But in the case of $q(\mathbf{z}|\mathbf{x}) = q(\mathbf{z})$, the containing information of factors are zero. This trivial case is excluded to ensure our assumption meaningful. To conclude, the split encoder is to prevent the trivial case, rather than improve conditional independence.

The split encoder can be viewed as an architecture to improve the flexibility for $\mathbf{z}$. Improving flexibility of representation in split encoder is beneficial to independence of factors and thus improves disentanglement when the penalty for independence is strong enough. But when the penalty is too weak, flexibility for factors might lead to smaller probability to be independent, and thus causes less disentanglement. As will be shown in the experiment part, these phenomena described from the two perspective above are observed in experiments of learning disentangled representations. In addition, the relation of split encoder KL term in objective of VAEs is discussed in Appendix A.2.

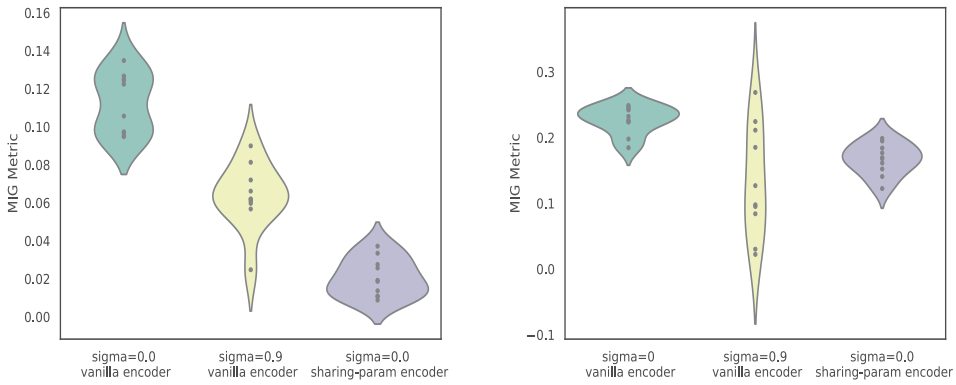
4 Related Works

There are many works related to disentangled representation with different objectives. A line of works focus on learning disentangled representations from typical types of data (Yingzhen & Mandt, 2018), some pay attention to supervised learning (Mathieu et al., 2016) or semi-supervised learning of disentangled representations (Spurr et al., 2017; Siddharth et al., 2017), others explore other objectives (Rubenstein et al., 2018; Zhao et al., 2019). There are also works on the evaluation of disentanglement (Eastwood & Williams, 2018). Since our assumption and analysis lie on the foundation of disentangled representation learning, in this section we focus on those related to the understanding of disentanglement or basic unsupervised models for disentangled representation learning.

Some works on unsupervised learning of disentangled representations are proposed based on early generative models. Schmidhuber (1992) proposes a variant of auto-encoder to learn disentangled representations by minimizing the predictablity of one factor in representation when other factors are fixed, this model obviously is motivated by the independence of factors, i.e. the common assumption. Desjardins et al. (2012) and Reed et al. (2014) propose variants of (Restricted) Boltzmann Machine in which interactions act to entangle the factors.

Recent works on unsupervised learning of disentangled representations are mainly based on GANs and VAEs. In line with GANs, InfoGAN (Chen et al., 2016) penalizes the mutual information of representations, and qualitatively shows that different factors in a representation correspond to different visual concepts. Brakel & Bengio (2018) propose to penalize the Jensen-Shannon divergence between the distribution of representations and its factorized distribution with a discriminator, based on Independent Component Analysis.

The mainstream models of disentangled representation learning are variants of VAE due to its stability. β-VAE (Higgins et al., 2017a) introduces to encourage the encoder to learn a disentangled representation by penalizing the KL term in the objective of vanilla VAE. AnneledVAE (Burgess et al., 2018) proposes to progressively increase the bottleneck capacity of VAE to encourage the encoder to learn different factors of variation when capacity grows. FactorVAE (Kim & Mnih, 2018) uses a discriminator to penalize the KL divergance between the distribution of representations and its factorized distribution (total correlation) via ratio trick to enhance independence of factors in representations. DIP-VAE (Kumar et al., 2017) matches the distribution of representations with dis-



(a) Vanilla VAE (b) FactorVAE

Figure 2: **MIG scores of verifying experiments in vanilla VAE and FactorVAE on dSprite.** In subfigures, the first violin graph is the baseline, the second one is a vanilla encoder with correlated noise ($\sigma = 0.9$), and the third one is a sharing-parameter encoder with uncorrelated noise.

entangled priors. In TC-VAE (Chen et al., 2018), the authors decompose the objective of VAE and argue that the total correlation term is the source of disentanglement, then they derive a mini-batch estimator for the total correlation term and penalize it to enhance disentanglement. Most VAE-based models can be attributed to enhance independence, which coincides with the common assumption.

5 Experiments

The objective in this section is not only to show our trick can improve unsupervised learning of disentangled representations, but also to verify the importance of the proposed assumption by investigating the change of disentanglement when conditional independence is violated.

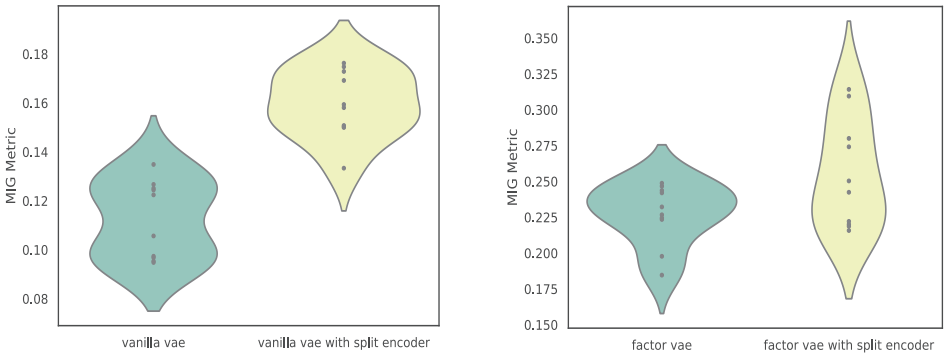
Datasets: We use dSprites (Higgins et al., 2017a) and a more complex dataset Cars3D (Reed et al., 2015). DSprites is a set of 737280 64*64 images in black and white, generated from fives independent latent factors. Cars3D consists of 199 colorful 3D car models in shape of 128*128*3*24*4.

Models: We select two models as baselines: vanilla VAE and FactorVAE. Since most VAE-based models are attributed to enhance independence of factors in representations, it might be redundant to consider more VAE-based models. Vanilla VAE and FactorVAE are weakly and strongly penalize the total correlation respectively, hence we believe that this choice is sufficient for showing the importance of our assumption and the power of our proposed split encoder. Additional experiment results of β-VAE, which does not explicitly penalize total correlation, are reported in Appendix.

Metrics: Following (Watters et al., 2019), we use Mutual Information Gap (MIG) (Chen et al., 2018) to evaluate disentanglement for its usefulness and rationality. MIG is defined by first computing the normalized mutual information of each factor in representations and each ground truth factor, then computing the gap between the highest two normalized mutual information values along factors in representations, and finally returning the gap averaged along ground truth factors. We also use reconstruction error to reflect the quality of the containing information for representations. Lower reconstruction error means more meaningful information are contained in representations. Additional results of FactorVAE Score (Kim & Mnih, 2018) are reported in Appendix.

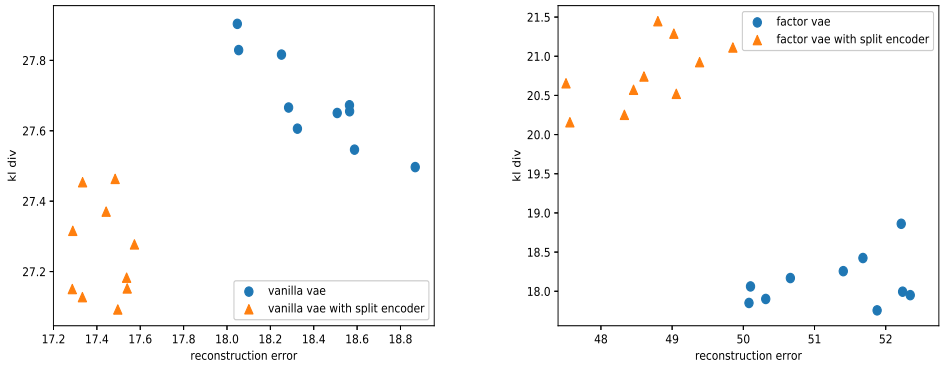
5.1 Violating Conditional Independence

When independence holds, our assumption is equivalent to conditional independence. According to our analysis, conditional independence originates from no-sharing-parameter block and the independence of noise in reparameterization trick. Therefore, for verifying the importance and necessity of our assumption, in this section we adopt two ways to violate conditional independence and compare the disentanglement of learned representations in terms of MIG: involving a encoder without no-sharing-parameter block, and a correlated noise for reparameterization.



(a) Vanilla VAE (+ split encoder) (b) FactorVAE (+ split encoder)

Figure 3: **MIG scores of vanilla VAE and FactorVAE (+ split encoder) on dSprites.**



(a) Vanilla VAE (+ split encoder) (b) FactorVAE (+ split encoder)

Figure 4: **KL vs. reconstruction error of vanilla VAE, FactorVAE (+ split encoder) on dSprites.**

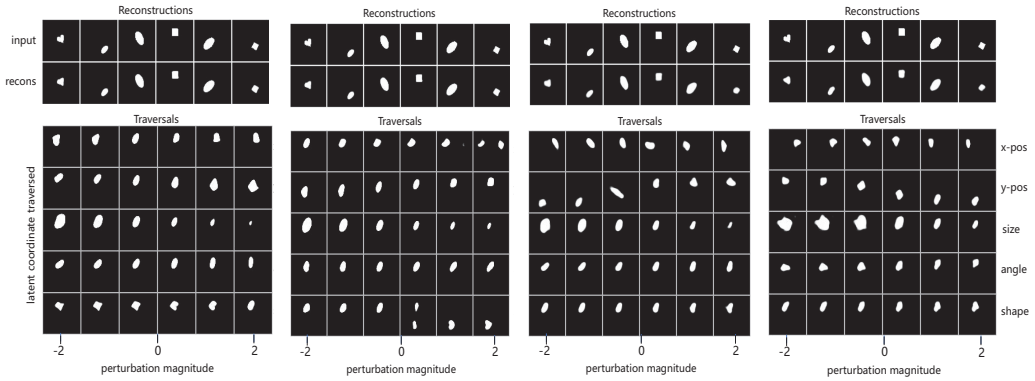
To show the necessity of no-sharing-parameter block, we remove it and develop a sharing-parameter encoder: the inputs are extracted features by a sequence of convolution neural networks, then the features are straightly flattened into representations. We compare the disentangling performance of sharing-parameter encoder and the vanilla encoder on dSprites.

To show the importance of the independence of noise in reparameterization trick, we introduce a correlated noise for comparison: $\epsilon \sim \mathcal{N}(0, \Sigma), \Sigma = (1-\sigma)\boldsymbol{I} + \sigma \mathbf{1}\mathbf{1}^T$ where $\mathbf{1}$ is an identity column vector, and $\sigma \in [0, 1)$ is correlation weight. Larger σ corresponds to higher correlation. We set $\sigma = 0.9$.

The two designs are applied into vanilla VAE and FactorVAE, and compare with the original models in terms of MIG metric. The results are shown in Fig.2, from which we can see that both using a correlated noise and removing the no-sharing-parameter block lead to the significant decrease of MIG metric in both vanilla VAE and FactorVAE. These results show that without uncorrelated noise and no-sharing-parameter block the learned representations will become more entangled, i.e. conditional independence and our assumption are vital for disentanglement. And rarely ensuring independence like FactorVAE is not sufficient for learning disentangled representations.

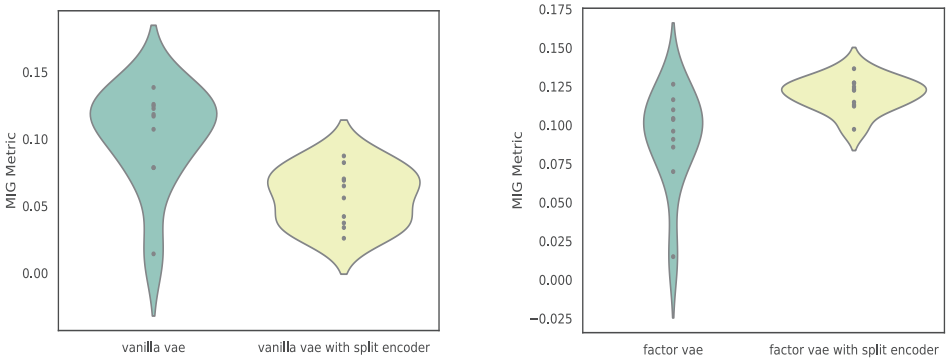
5.2 Unsupervised Learning of Disentangled Representations

In this subsection, we compare the performance of vanilla VAE and FactorVAE with ones combined with split encoder on dSprites and Cars3D. We not only focus on the MIG metric to reflect the performance of disentanglement, but also concern the reconstruction loss which indicates the quality of containing information in representations, and meanwhile observe the quality of reconstructed images and traversals for intuitive understanding. Traversals are generated from a data point, which



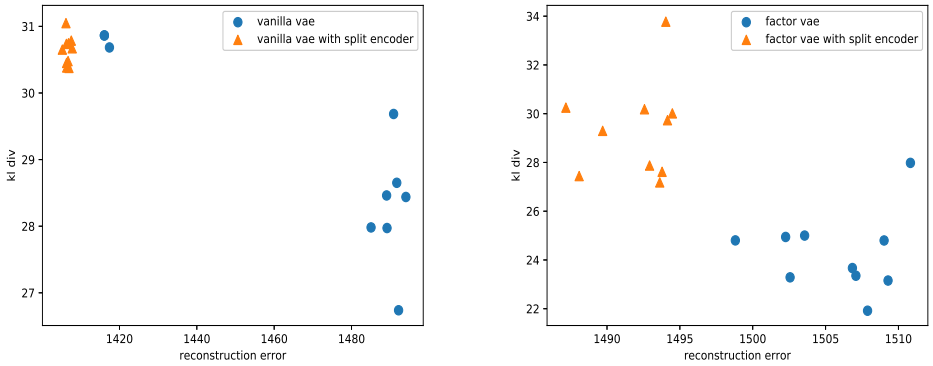
(a) Vanilla VAE baseline (MIG=0.135) (b) Vanilla VAE + split encoder (MIG=0.176) (c) FactorVAE baseline (MIG=0.249) (d) FactorVAE + split encoder (MIG=0.314)

Figure 5: **Reconstructions and traversals of VAE and FactorVAE (+ split encoder) on dSprites.**



(a) Vanilla VAE (+ split encoder) (b) FactorVAE (+ split encoder)

Figure 6: **MIG scores of vanilla VAE and FactorVAE (+ split encoder) on Cars3D.**



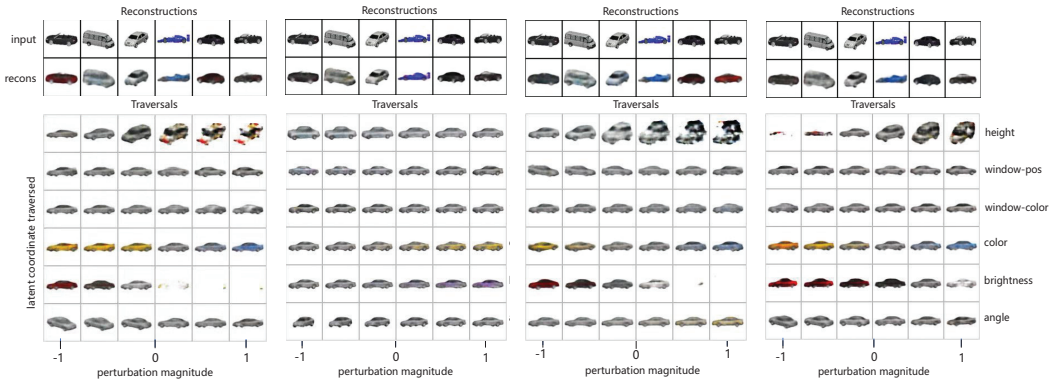
(a) Vanilla VAE (+ split encoder) (b) FactorVAE (+ split encoder)

Figure 7: **KL vs. reconstruction error of vanilla VAE, FactorVAE (+ split encoder) on Cars3D.**

is first encoded into a representation, then for each row only a single factor in the representation is changed in $[-2, 2]$ and then reconstruct the data point. For the fairness of comparison, we show the reconstruction images and traversals from output model with the highest MIG score for each case.

5.2.1 Performance on dSprite

As shown in Fig.3, combing split encoder with both vanilla VAE and FactorVAE significantly improve disentanglement in terms of MIG on dSprites. This result indicates that dSprites is simple and



(a) Vanilla VAE baseline (MIG=0.138) (b) Vanilla VAE + split encoder (MIG=0.087) (c) FactorVAE baseline (MIG=0.126) (d) FactorVAE + split encoder (MIG=0.136)

Figure 8: **Reconstructions and traversals of VAE and FactorVAE (+ split encoder) on Cars3D.**

the penalty for independence of representation in vanilla VAE is strong enough, so that improving flexibility of representations by involving split encoder can significantly improve disentanglement.

And Fig.4 shows that involving split encoder leads to smaller reconstruction error. This result means that split encoder is beneficial to learn meaningful information from data, which coincides with our assumption. Note that the KL term in vanilla VAE with split encoder is smaller, while in FactorVAE with split encoder it is larger. This phenomenon prove that the improvement of disentanglement in split encoder cannot be explained as the results of implicitly penalizing the KL term, which does not coincide with the analysis for β-VAE (Higgins et al., 2017a; Burgess et al., 2018).

The reconstructions and traversals are shown in Fig. 5. We can see the four models work well in dSprites, and the learned representations are well disentangled except some small flaws.

5.2.2 Performance on Cars3D

The results on Car3D are different from those on dSprites. As shown in Fig.6, vanilla VAE combined with split encoder leads to lower MIG, while FactorVE combined with split encoder outperforms FactorVAE. This result is not surprising, which means that Car3D is too complex and the penalty of independence in vanilla VAE is not strong enough, and involving split encoder leads to flexibility for representation and smaller probability of being disentangled. While for FactorVAE, the penalty of independence is strong enough, and thus involving split encoder leads to better performance.

As shown in Fig. 7, involving split encoder can lead to smaller reconstruction error, which indicates the containing information in representations is more meaningful. While the KL terms are larger in both models with split encoder. This suggests the effect of split encoder is not penalizing the KL terms, but increasing the quality of information contained in representations and flexibility.

The reconstructions and traversals results on the four models are different, as shown in Fig. 8. Models with split encoder obviously have better reconstructions. And traversals in FactorVAE with split encoder are more disentangled than FactorVAE.

6 Conclusion

We argue that an assumption from the perspective of information is vital for disentanglement: factors in a disentangled representation should contain isolated information about data. We formulate this assumption into mathematical equation and connect it with independence and conditional independence of factors. Then we find two inductive biases are the key to conditional independence: no-sharing-parameter block and uncorrelated noise in reparameterization trick. Inspired by this, we propose split encoder to improve model capacity of no-sharing-parameter block and thus improve disentanglement. Experimental results demonstrate the importance of our assumption and the power of split encoder for improving disentanglement and reconstructions.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *international conference on learning representations*, 2017.

Yann Lecun Bengio. Scaling learning algorithms towards ai. *Large-scale Kernel Machines*, 34(5): 1–41, 2007.

Yoshua Bengio, Aaron C Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv: Machine Learning*, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv: Machine Learning*, 2018.

Tian Qi Chen, Xuechen Li, Roger B Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *neural information processing systems*, pp. 2610–2620, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: interpretable representation learning by information maximizing generative adversarial nets. *neural information processing systems*, pp. 2180–2188, 2016.

Guillaume Desjardins, Aaron C Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv: Machine Learning*, 2012.

Cian Eastwood and Christopher K I Williams. A framework for the quantitative evaluation of disentangled representations. *international conference on learning representations*, 2018.

Ian J Goodfellow, Jean Pougetabadie, Mehdi Mirza, Bing Xu, David Wardefarley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. *neural information processing systems*, pp. 2672–2680, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *international conference on machine learning*, 2017a.

Irina Higgins, Arka Pal, Andrei A Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: improving zero-shot transfer in reinforcement learning. *international conference on machine learning*, pp. 1480–1490, 2017b.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Jimenez Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv: Learning*, 2018.

Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, Joris M Mooij, and Bernhard Sch Lkopf. On causal and anticausal learning. pp. 459–466, 2012.

Deepmind Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *international conference on machine learning*, pp. 2649–2658, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *international conference on learning representations*, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *international conference on learning representations*, 2017.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *international conference on learning representations*, 2018.

Brenden M Lake, Tomer Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:1–101, 2017.

Adrien Laversanne-Finot, Alexandre Péré, and Pierre-Yves Oudeyer. Curiosity driven exploration of learned disentangled goal spaces. *Conference on Robot Learning*, 2018.

Alexander H Liu, Yencheng Liu, Yuying Yeh, and Yuchiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *neural information processing systems*, pp. 2590–2599, 2018.

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *international conference on machine learning*, 2018a.

Francesco Locatello, Damien Vincent, Ilya Tolstikhin, Gunnar Ratsch, Sylvain Gelly, and Bernhard Scholkopf. Competitive training of mixtures of independent deep generative models. *arXiv: Learning*, 2018b.

Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Scholkopf, and Olivier Bachem. On the fairness of disentangled representations. *arXiv: Learning*, 2019.

Michael Mathieu, Junbo Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann Lecun. Disentangling factors of variation in deep representations using adversarial training. pp. 5047–5055, 2016.

Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *neural information processing systems*, pp. 9191–9200, 2018.

Scott E Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. *international conference on learning representations*, pp. 1431–1439, 2014.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *neural information processing systems*, pp. 1252–1260, 2015.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. *neural information processing systems*, pp. 185–194, 2018.

Paul K Rubenstein, Bernhard Scholkopf, and Ilya Tolstikhin. Learning disentangled representations with wasserstein auto-encoders. *international conference on learning representations*, 2018.

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

Ramalingam Shanmugam. Elements of causal inference: foundations and learning algorithms. *Journal of Statistical Computation and Simulation*, 88(16):3248–3248, 2018.

N Siddharth, Brooks Paige, Janwillem Van De Meent, Alban Desmaison, Noah D Goodman, Pushmeet Kohli, Frank Wood, and Philip H S Torr. Learning disentangled representations with semi-supervised deep generative models. *neural information processing systems*, pp. 5927–5937, 2017.

Adrian Spurr, Emre Aksan, and Otmar Hilliges. Guiding infogan with semi-supervision. *European conference on Machine Learning*, pp. 119–134, 2017.

Xander Steenbrugge, Sam Leroux, Tim Verbelen, and Bart Dhoedt. Improving generalization for abstract reasoning tasks using disentangled feature representations. *arXiv: Learning*, 2018.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv: Learning*, 2018.

Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv: Learning*, 2019.

Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. *international conference on machine learning*, pp. 5656–5665, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. pp. 5885–5892, 2019.

A DISCUSSION ABOUT OBJECTIVE OF VAE

The objective of VAE is usually derived as a lower bound of log-likelihood:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{x})q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \mathbb{E}_{q(\mathbf{x})}[D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))] \leq \mathbb{E}_{p(x)}[\log q(x)]$$

where the first term (negative reconstruction error) encourages reconstructions when the objective is maximized and the second term (KL term) is a regularizer.And obviously the objective is equivalent to the negative KL divergence between two joint distributions:

$$\mathcal{L} = -D_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{x})\|p(\mathbf{z}, \mathbf{x})) + \mathbb{E}_{q(\mathbf{x})}[\log q(\mathbf{x})]$$

Hence maximizing the objective $mathcalL$ is equivalent to minimizing the KL divergence between joint distributions of encoder and decoder. From these two equivalent objectives, we can further understand VAEs from different perspectives.

A.1 SOURCE OF INDEPENDENCE IN VAES

Factors in representations of VAEs are encouraged to be in dependent, which is attributed to total correlation penalty in KL term (Chen et al., 2018). Here we demonstrate this property from the decomposition of $D_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{x})\|p(\mathbf{z}, \mathbf{x}))$:

$$D_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{x})\|p(\mathbf{z}, \mathbf{x})) = D_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z})) + \mathbb{E}_{p(\mathbf{x})}[D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x}))]$$

When minimizing the $D_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{x})\|p(\mathbf{z}, \mathbf{x}))$, the two KL divergences on the right are minimized. Using the factorizing of $p(z)$, the first one can be further decomposed into two KL divergences:

$$D_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z})) = D_{\mathrm{KL}}(q(\mathbf{z})\|\prod_j q(\mathrm{z}_j)) + \sum_j D_{\mathrm{KL}}(q(\mathbf{z}_j)\|p(\mathbf{z}_j)$$

where the first term is exactly the total correlation, which is the source of independence.

To conclude, the objective $D_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{x})\|p(\mathbf{z}, \mathbf{x}))$ can be decomposed into three KL divergences, and thus they are minimized when minimizing $D_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{x})\|p(\mathbf{z}, \mathbf{x}))$. One term is the total correlation, which is exactly the source of independence.

A.2 KL TERM AND SPLIT ENCODER

Chen et al. (2018) decompose the KL term into three KL divergences:

$$D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) = D_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{x})\|q(\mathbf{z})q(\mathbf{x})) + D_{\mathrm{KL}}(q(\mathbf{z})\|\prod_j q(\mathrm{z}_j)) + \sum_j D_{\mathrm{KL}}(q(\mathrm{z}_j)\|p(\mathrm{z}_j))$$

The fist term is mutual information between data with representation, the second term is total correlation, and the third one is divergence between distribution of representations with prior. Hence minimizing the KL term leads to dropping information, independence and matching prior for representations due to the three KL divergences, respectively.

Our split encoder encourages learning isolated information for factors in representations, prevents dropping information and improves the flexibility of representations. Therefore, split encoder can be regarded as an inductive bias, which limits minimizing of the first term, and meanwhile enhances minimizing the other terms. Experiments on dSprites and Cars3D shows that the KL term in models with split encoder can be larger or smaller than the original models, but reconstructions are improved significantly. These results support our opinion.

B PROOFS OF LEMMAS

Lemma 1. *$I(\mathrm{z}_j; \mathbf{x}|\mathbf{z}_{i\in S_{-j}}) = I(\mathrm{z}_j; \mathbf{x})$ for any j is equivalent to the following equation:*

$$I(\mathbf{z}_{i\in S}; \mathbf{x}) = I(\mathrm{z}_j; \mathbf{x}) + I(\mathbf{z}_{i\in S_{-j}}; \mathbf{x})$$

where $S = \{j\} \cup S_{-j}$ is any subset of $\{1, 2, \ldots, J\}$, j is any single element in S.

Proof: This conclusion is a straight corollary of chain rule of mutual information. For further understanding, here we derive it from scratch using Bayes rule:

$$\begin{aligned}
I(\mathbf{z}_j; \mathbf{x}|\mathbf{z}_{i\in S_{-j}}) &= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{z}_j, \mathbf{x}|\mathbf{z}_{i\in S_{-j}})}{q(\mathbf{z}_j|\mathbf{z}_{i\in S_{-j}})q(\mathbf{x}|\mathbf{z}_{i\in S_{-j}})}] \\
&= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{x}|\mathbf{z}_{i\in S})}{q(\mathbf{x}|\mathbf{z}_{i\in S_{-j}})}] \\
&= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{x}, \mathbf{z}_{i\in S})q(\mathbf{z}_{i\in S_{-j}})}{q(z_{i\in S})q(\mathbf{x}, \mathbf{z}_{i\in S_{-j}})}] \\
&= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{x}, \mathbf{z}_{i\in S})}{q(z_{i\in S})q(\mathbf{x})} - \log \frac{q(\mathbf{x}, \mathbf{z}_{i\in S_{-j}})}{q(\mathbf{z}_{i\in S_{-j}})q(\mathbf{x})}] \\
&= I(\mathbf{z}_{i\in S}; \mathbf{x}) - I(\mathbf{z}_{i\in S_{-j}}; \mathbf{x})
\end{aligned}$$

Hence we have:

$$I(\mathbf{z}_j; \mathbf{x}|\mathbf{z}_{i\in S_{-j}}) = I(\mathbf{z}_j; \mathbf{x}) \Leftrightarrow I(\mathbf{z}_{i\in S}; \mathbf{x}) = I(\mathbf{z}_j; \mathbf{x}) + I(\mathbf{z}_{i\in S_{-j}}; \mathbf{x})$$

Lemma 2. *For any subset S of $\{1, 2, \ldots, J\}$, we have:*

$$I(\mathbf{z}_{i\in S}; \mathbf{x}) - \sum_{i\in S} I(\mathbf{z}_i; \mathbf{x}) = \mathbb{E}_{q(\mathbf{x})}[D_{\mathrm{KL}}(q(\mathbf{z}_{i\in S}|\mathbf{x}) \| \prod_{i\in S} q(\mathbf{z}_i|\mathbf{x}))] - D_{\mathrm{KL}}(q(\mathbf{z}_{i\in S}) \| \prod_{i\in S} q(\mathbf{z}_i))$$

Proof:

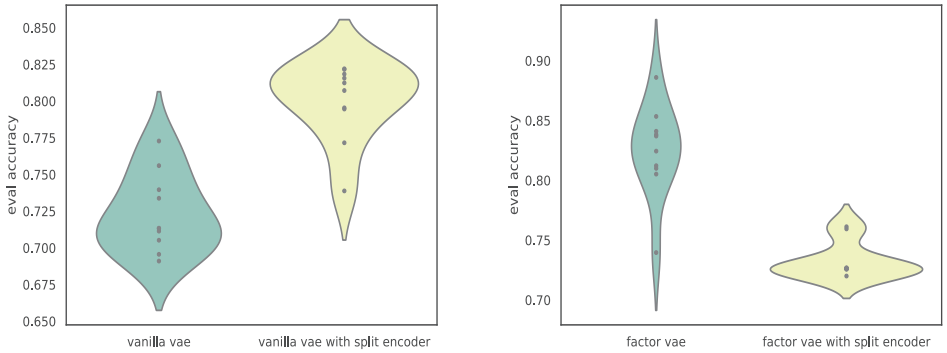
$$\begin{aligned}
I(\mathbf{z}_{i\in S}; \mathbf{x}) - \sum_{i\in S} I(\mathbf{z}_i; \mathbf{x}) &= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{z}_{i\in S}, \mathbf{x})}{q(\mathbf{z}_{i\in S})q(\mathbf{x})}] - \sum_{i\in S} \mathbb{E}_{q(\mathbf{z}_i, \mathbf{x})}[\log \frac{q(\mathbf{z}_i, \mathbf{x})}{q(\mathbf{z}_i)q(\mathbf{x})}] \\
&= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{z}_{i\in S}, \mathbf{x})}{q(\mathbf{z}_{i\in S})q(\mathbf{x})} - \sum_{i\in S} \log \frac{q(\mathbf{z}_i, \mathbf{x})}{q(\mathbf{z}_i)q(\mathbf{x})}] \\
&= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{z}_{i\in S}|\mathbf{x})}{q(\mathbf{z}_{i\in S})} - \sum_{i\in S} \log \frac{q(\mathbf{z}_i|\mathbf{x})}{q(\mathbf{z}_i)}] \\
&= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{z}_{i\in S}|\mathbf{x}) \prod_{i\in S} q(\mathbf{z}_i)}{q(\mathbf{z}_{i\in S}) \prod_{i\in S} q(\mathbf{z}_i|\mathbf{x})}] \\
&= \mathbb{E}_{q(\mathbf{z}_{i\in S}, \mathbf{x})}[\log \frac{q(\mathbf{z}_{i\in S}|\mathbf{x})}{\prod_{i\in S} q(\mathbf{z}_i|\mathbf{x})} - \log \frac{q(\mathbf{z}_{i\in S})}{\prod_{i\in S} q(\mathbf{z}_i)}] \\
&= \mathbb{E}_{q(\mathbf{x})}[D_{\mathrm{KL}}(q(\mathbf{z}_{i\in S}|\mathbf{x}) \| \prod_{i\in S} q(\mathbf{z}_i|\mathbf{x}))] - D_{\mathrm{KL}}(q(\mathbf{z}_{i\in S}) \| \prod_{i\in S} q(\mathbf{z}_i))
\end{aligned}$$

C MODELS PERFORMANCE IN TERMS OF FACTORVAE SCORE

There are several metrics for evaluating disentanglement, including Mutual Information Gap (MIG) (Chen et al., 2018), DCI Disentanglement (Ridgeway & Mozer, 2018), BetaVAE metric (Higgins et al., 2017a), FactorVAE Score (Kim & Mnih, 2018), SAP Score (Kumar et al., 2018) and Modularity (Ridgeway & Mozer, 2018). Locatello et al. (2018a) report that MIG and DCI Disentanglement strongly correlated with each other, and so do BetaVAE Score and FactorVAE Score, but SAP and Modularity are not correlated with other metrics. This result indicates that not all metrics are suitable for evaluating disentanglement. We believe that MIG is a suitable metric due to its usefulness and rationality, so we mainly report experimental results in terms of MIG. However, for a more complete comparison, here we report models performance in terms of evaluation accuracy in FactorVAE score as shown in Fig. 9 and Fig. 10, which is not well correlated with MIG.

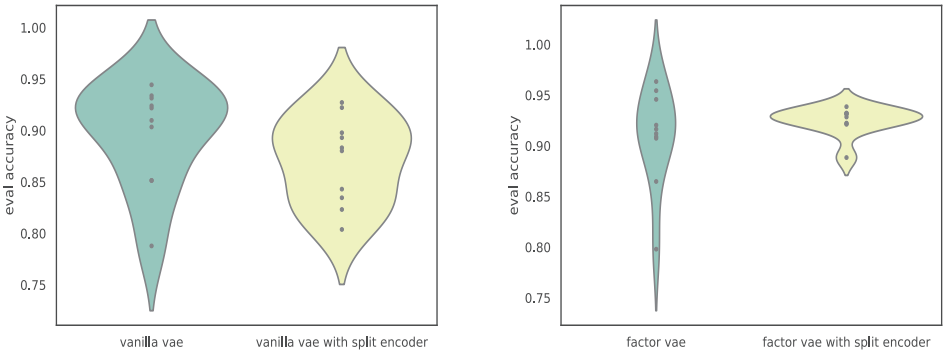
D EXPERIMENTAL RESULTS OF BETA-VAE ON DSPRITES

In Fig. 11, we can see that split encoder significantly improves the disentanglement performance of β-VAE in terms of MIG. And the KL term in β-VAE is larger while the reconstruction error is smaller. These results supports our analysis on the effect of split encoder in Section A.



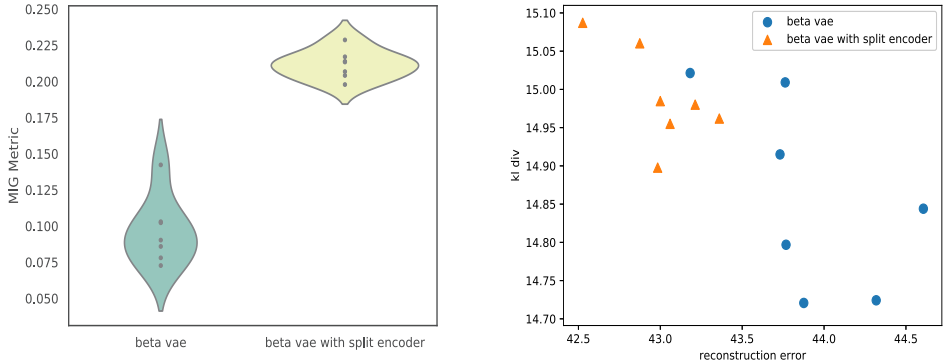
(a) Vanilla VAE (+ split encoder) (b) FactorVAE (+ split encoder)

Figure 9: **FactorVAE Score of vanilla VAE and FactorVAE (+ split encoder) on dSprites.**



(a) Vanilla VAE (+ split encoder) (b) FactorVAE (+ split encoder)

Figure 10: **FactorVAE Score of vanilla VAE and FactorVAE (+ split encoder) on Cars3D.**



(a) MIG score (b) KL term vs. reconstruction error

Figure 11: **Experimental results of β-VAE (+ split encoder) on dSprites.**