VARIATIONAL AUTOENCODERS WITH DECREMENTAL INFORMATION BOTTLENECK FOR DISENTANGLEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

One major challenge of disentanglement learning with variational autoencoders is the trade-off between disentanglement and reconstruction fidelity. Previous methods, spreading the conflict of disentanglement and reconstruction in time, will lose the constraint of disentanglement when expanding the information bottleneck, which causes the information diffusion problem. To tackle this issue, we present a novel decremental variational autoencoder with disentanglement-invariant transformations to spread the conflict on multiple latent spaces, termed DeVAE, for balancing disentanglement and reconstruction fidelity by decreasing the information bottleneck of diverse latent spaces gradually. Benefiting from the multiple latent spaces, DeVAE allows simultaneous optimization of multiple objectives to optimize reconstruction while keeping the constraint of disentanglement, avoiding information diffusion. DeVAE is also compatible with large models with high-dimension latent space. Experimental results on dSprites and Shapes3D that DeVAE achieves the best performance on both disentanglement and reconstruction.

1 INTRODUCTION

Unsupervised learning for sensing the properties of objects is crucial to reduce the gap between humans and machines intelligence. Inline with human intelligence disentanglement learning (Bengio et al., 2013) is considered to be a promising direction to obtain explanatory representations from observations to understand and reason objects without any supervision.

In the recent years, various approaches (Higgins et al., 2017; Chen et al., 2018; Kim & Mnih, 2018; Burgess et al., 2018; Chen et al., 2016) have been proposed to successfully extract basic properties of objects, such as position, color, orientation, and scale (Burgess & Kim, 2018; Matthey et al., 2017). The commonly-used methods are based on variational autoencoder (VAE) (Kingma & Welling, 2014). In particular, β-VAE (Higgins et al., 2017) introduced an extra parameter β on the Kullback-Leibler (KL) divergence to promote disentanglement. However, there is a trade-off between disentanglement and reconstruction fidelity on β-VAE, which is a problem to be solved in the following works.

One common direction for dealing with the trade-off is to penalize the Total Correlation (TC) between latent variables, avoiding reducing the mutual information, such as FactorVAE (Kim & Mnih, 2018), β-TCVAE (Chen et al., 2018), and DIPVAE (Kumar et al., 2018). As pointed out in (Träuble et al., 2020; Dittadi et al., 2020), TC-based VAEs have a strong prior assumption that the factors are statistically independent. Beyond that, when it comes to high-dimension latent space, the estimation of TC becomes inaccurate due to curse of dimensionality, as our experiments observed in Section 3.2. The realistic problems usually have numerous factors, therefore it would need a large model with high latent space to extract representations. For example, the popular deep model ResNet50 (He et al., 2016) has 2048 dimensional feature space. However, in this work, we get rid of calculating TC by leveraging the narrow information bottleneck (Tishby et al., 1999; Burgess et al., 2018) to find efficient codes for representing the data, which promotes disentanlement.

In the meanwhile, previous information bottleneck (IB)-based methods (Burgess et al., 2018; Shao et al., 2022; Wu et al., 2022) have tried to solve the obstacle of trade-off between disentanglement and reconstruction fidelity. In general, they first set a high pressure with a narrow IB to encourage disentanglement and then expands the IB gradually to promote reconstruction fidelity under a latent

space, termed *incremental methods*. For example, DynamicVAE (Shao et al., 2022) initiated β with a large value at the beginning of training for disentanglement and stably increase the KL divergence for reconstruction by a non-linear PI controller. However, they lost the constraint of disentanglement when expanding the IB, which causes the information diffusion problem (Wu et al., 2022). In this work, to avoid information diffusion, we aim to optimize reconstruction while keeping the constraint of disentanglement.

Different from IB-Incremental based approaches listed above, our key motivation is to optimize disentanglement and reconstruction simultaneously.Previous methods spread the targets of disentanglement and reconstruction over time and optimize only one target at a time. To optimize both two targets, we spread targets over spaces by creating multiple latent spaces. In this way, each latent layer has its own objective to optimize disentanglement or reconstruction. Furthermore, our framework constrain these latent spaces to share disentanglement so that the first latent space achieves simultaneous optimization of disentanglement and reconstruction.

To achieve this, we propose a simple yet effective VAE framework composed of multiple continuous latent sub-spaces with a novel IB-Decremental strategy and a disentanglement-invariant transform operator, which we call DeVAE. Specifically, we decrease the information bottleneck of each latent space layer by layer, where we constrain the first space for informativeness to recover the input image, and other disentangled spaces for learning factors of the image by narrow IBs. Furthermore, we introduce the disentanglement-invariant transform operator to ensure simultaneous optimization of disentanglement across continuous latent sub-spaces, which avoids the information diffusion. Our decremental model avoids ID by keeping the constraints of disentanglement and reconstruction simultaneously. Furthermore, DeVAE is capable of large models with high dimensional space. We also conducted comprehensive comparisons with popular methods quantitatively and qualitatively.

Our contributions can be summarized as follows:

- We introduce several latent spaces sharing disentanglement by disentanglement-invariant transformations.
- We propose a novel diagram for disentanglement learning by decreasing IB, termed decremental VAE (DeVAE). Our decremental model can handle large-scale problems and show robustness on several datasets.

2 METHODOLOGY

2.1 PRELIMINARIES

Problem Setup & Notations. Disentanglement learning aims to learn the factors of variation which raises the change of observations. Given a set of samples $\boldsymbol{x} \in \mathcal{X}$, they can be uniquely described by a set of ground-truth factors $\boldsymbol{c} \in \mathcal{C}$. Generally, the generation process $g(\cdot)$ is invisible $\boldsymbol{x} = g(\boldsymbol{c})$. We say a representation for factor $\boldsymbol{c}_i$ is disentangled if it is invariant for the samples with $\boldsymbol{c}_j$. We use variational inference to learn the disentangled representation for a given problem. $p(\boldsymbol{z}|\boldsymbol{x})$ denotes the probability of $\boldsymbol{z} = f(\boldsymbol{x})$, $p(\boldsymbol{x}|\boldsymbol{z})$ denotes the probability of $\boldsymbol{x} = g(\boldsymbol{z})$. The representation function is a conditional Bayesian network of the form $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to estimate $p(\boldsymbol{z}|\boldsymbol{x})$. The generative model is another network of the form $p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. ϕ, θ are trainable parameters.

Revisit VAE & β-VAE. The VAE framework (Kingma & Welling, 2014) computes the representation function by introducing $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and optimizing the variational lower bound (ELBO). β-VAE (Higgins et al., 2017) introduces the hyperparameter β to control the IB:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta D_{\mathrm{KL}}(q_\phi(\boldsymbol{x}|\boldsymbol{z})||p(\boldsymbol{z})). \tag{1}$$

Consider using β-VAE to learn a representation of the data; the representation will be disentangled but lose information when β is large (Burgess et al., 2018). We can set a large β to learn a disentangled representation and a small β to learn an informative representation.

However, previous disentanglement methods (Higgins et al., 2017; Chen et al., 2018; Burgess et al., 2018) are limited in low-dimension latent space and poorly deal with the trade-off between disentanglement and reconstruction. Current state-of-the-art approach (Shao et al., 2022) with an annealing manner from high pressure to low pressure will loosen the constraint of disentanglement

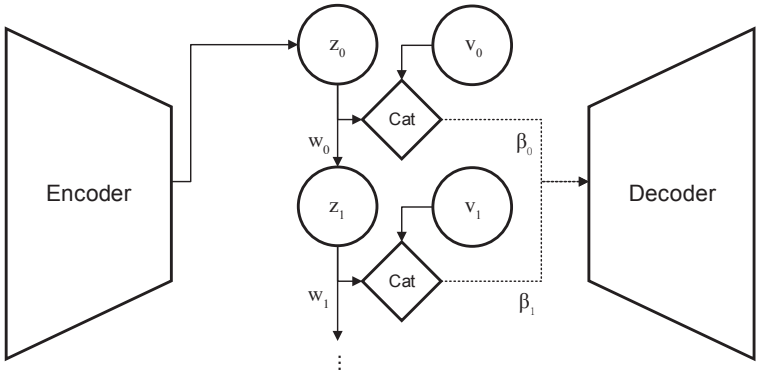

Figure 1: Illustration of our Decremental Variational Autoencoder (DeVAE). The solid lines denote the information flow of the encoding process. The dash lines denote the decoding process which randomly selects one layer's representation and concatenates the corresponding embedding vector. Each layer has a pressure β_i to control the capacity of IB.

when reducing the pressure. To address this issue, we propose a novel decremental variational autoencoder with hierarchical latent spaces, namely DeVAE, to optimize disentanglement and reconstruction fidelity simultaneously, which can handle high-dimensional latent spaces, as shown in Figure 1. Our DeVAE applies a hierarchical structure with a decremental information bottleneck and disentanglement-invariant transformation to produce latent variables layer by layer. The decoder part randomly selects one layer's latents concatenating an embedding vector to generate images.

2.2 Hierarchical Latent Spaces with Decremental Information Bottleneck

In order to retain the disentanglement constraint while optimizing the reconstruction fidelity, we introduce a Hierarchical Latent Space (HiS) with K layers and assign a pressure β_i to the $i-\mathrm{th}$ layer $\mathcal{Z}_i$. Each layer will promote disentanglement or reconstruction by a suitable pressure. The objective of the i-th layer is

$$\mathcal{L}_i(\theta,\phi) = \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_i)] - \beta_i D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}_i|\boldsymbol{x})||p(\boldsymbol{z})), \tag{2}$$

where the first layer $q_\phi(\boldsymbol{z}_0|\boldsymbol{x})$ is a conditional Bayesian network, and the following layers can be calculated by

$$q(\boldsymbol{z}_{i+1}|\boldsymbol{x}) = \tau_i(\boldsymbol{z}_{i+1}|\boldsymbol{z}_i)q(\boldsymbol{z}_i|\boldsymbol{x}), i \neq 0, \tag{3}$$

where τ_i denotes a transformation from $\mathcal{Z}_i$ to $\mathcal{Z}_{i+1}$.

According to information theory, information can only decrease while processing, therefore *We gradually decrease the IB in the sequential layers, i.e.,* $\beta_{i+1} > \beta_i$. Usually, we set $\beta_0 = 1$ to encourage the first layer to focus on reconstructing the original inputs. In this way, the sequential layers aim to disentangle factors of variation by setting narrow bottlenecks.

2.3 Disentanglement-invariant Transformation

Though we create multiple latent spaces, these objectives only encourage the local representations to be disentangled or informative. We need a mechanism to connect these objectives for balancing disentanglement and reconstruction in one layer. In order to make sure disentanglement across all latent layers, we propose a disentanglement-invariant transformation (DiT) denoted as τ.

Theorem 1 *$w \cdot \boldsymbol{z}$ is disentangled if $\boldsymbol{z}$ is disentangled, w is a diagonal matrix.*

Proof in Appendix A.2.

According to Theorem 1, we can scale the latent space to keep disentanglement. However, scaling the posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ violates the generation process which wants the marginal distribution $q(\boldsymbol{z}) = \sum q_\phi(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})$ to be close to a standard normal distribution. Besides, most downstream tasks

```
def loss_fn(x,encoder,decoder,W,embeddings,betas,K):
    idx = np.random.randint(K)
    mu, logvar = encoder(x)
    w1, w2 = W
    for i in range(idx):
        mu = torch.exp(w1[i]) * mu
        logvar = logvar + w2[i]
    z = sample(mu, logvar) # re-parameter trick
    recon = decoder(torch.cat([z,embeddings(idx)],1))
    loss = F.mse(recon,x) + betas[idx] * kld(mu,logvar)
    return loss
```

Algorithm 1: PyTorch-like implementation of DeVAE loss.

use the mean representation instead of sampled representation. Therefore, we only need the mean representation disentanglement-invariant. Furthermore, we add an extra learnable diagonal matrix w^2 to adjust the marginal distribution. The disentanglement-invariant transformation of i-th layer can be

$$\boldsymbol{z_{i+1}} \sim \mathcal{N}(h(w_i^1)\boldsymbol{\mu_i}, h(w_i^2)\boldsymbol{\sigma_i}) = \tau_i(\boldsymbol{z_i} \sim \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\sigma_i})), \tag{4}$$

where w_i^1, w_i^2 are learnable diagonal matrices of the i-th layer, $h(w) = e^w > 0$ is an exponential function to make sure the scale values greater than 0. Therefore, it's easy to get the parameters of i-th latent variables

$$\boldsymbol{\mu_i} = h(\sum_{j=0}^{j-1} w_j^1)\boldsymbol{\mu_0}, \quad \boldsymbol{\sigma_i} = h(\sum_{j=0}^{j-1} w_j^2)\boldsymbol{\sigma_0}, \quad i > 0. \tag{5}$$

And the i-th KL divergence is

$$D_{\mathrm{KL}_i} = \frac{1}{2}(1 + 2\sum_{j=0}^{j-1} w_j^2 + 2\log(\boldsymbol{\sigma_0}) - h(2\sum_{j=0}^{j-1} w_j^2)\boldsymbol{\sigma_0^2} - h(2\sum_{j=0}^{j-1} w_j^1)\boldsymbol{\mu_0^2}) \tag{6}$$

2.4 Optimization Algorithm

In this section, we combine the above components and introduce the optimization algorithm for the multiple objectives. We use a random process to optimize one layer's objective from K latent spaces:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{p(\boldsymbol{z}_i)}[\mathcal{L}_i(\theta, \phi)], \tag{7}$$

where $p(\boldsymbol{z}_i) = (1-s)s^i/(1-s^{i+1}), s \neq 1; \frac{1}{K}, s = 1$, s is a hyperparameter to weight the objective of each layer. In practice, $s = 1$ has good performance as shown in Section 3.3. Note that we do not aggregate the objectives into a loss, instead, we optimize one layer's objective in one mini-batch.

In our model, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ are modelled by two neural networks, a K-size sequence 'betas' denotes the penalties on the KL divergences of corresponding layers, w^1, w^2 stores the learnable parameters for transforming latent spaces. First, we randomly sample a mini-batch and choose a target layer to optimize. Then use the algorithm introduced in Section 2.3 to obtain the representation of the target layer and reconstruct the corresponding images. Instead of using K separated decoders to rebuild images, we apply a shared decoder with layer embeddings. In particular, it concatenates the latent variables and a layer embedding vector as the inputs of the decoder. The PyTorch-like algorithm is shown in Algorithm 1.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate our method on two widely-used datasets (dSprites, Shapes3D). dSprites (Matthey et al., 2017) has 737,280 binary 64 × 64 x 1 images generated from five factors: shape (5), orientation (40), scale (6), position X (32), and position Y (32). Shapes3D (Burgess

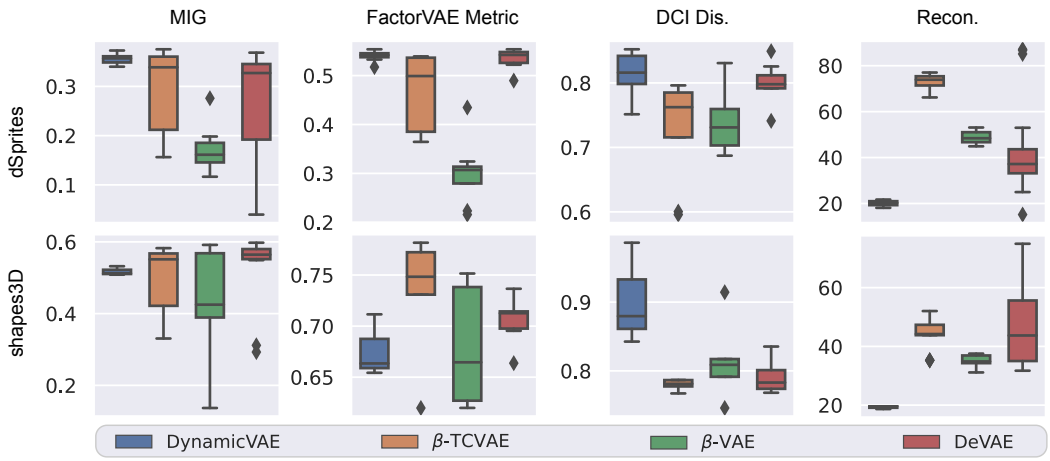

Figure 2: Box plots of quantitative benchmarks MIG, FactorVAE, Disentanglement, and reconstruction error on dSprites and Shapes3D.

& Kim, 2018) has 480,000 RGB 64 × 64 × 3 images of 3D shapes generated from six factors: floor color (10), wall color (10), object color (10), object size (8), object shape (4), and azimuth (15).

Evaluation Metrics. We apply the following metrics to evaluate the performance of disentanglement and reconstruction. **MIG** (Chen et al., 2018): the mutual information gap between two variables with the highest and the second-highest mutual information. **FactorVAE metric** (Kim & Mnih, 2018): the error rate of the classifier, which predicts the latent variable with the lowest variance. **DCI Dis.**: abbreviation for DCI Disentanglement (Eastwood & Williams, 2018), a matrix of relative importance by regression. **Recon.**: abbreviation for Reconstruction Error, a measure of the distance between images; we use Mean Squared Error for RGB images and Binary Cross Entropy for binary images.

Implementation. We use a convolutional neural network as the encoder and a deconvolutional neural network as the decoder. Detailed architecture can be found in Appendix A.1. The activation function is ReLU. The optimizer is Adam (Kingma & Ba, 2015) with a learning rate of 1e-4, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size is 256, which accelerates the training process. All experiments train 300, 000 iterations by default. For the hyperparameters, we set $\beta = 12$ for β-TCVAE, $\beta = 6$ for β-VAE, and $K_i = 0.001$, $K_p = 0.01$ for DynamicVAE, and $\{\beta_i\} = [1, 10, 40]$, $s = 1$ for DeVAE.

3.2 Comparison to Prior work

To demonstrate the effectiveness of the proposed DeVAE, we compare it to previous all types of baselines: 1) β-VAE (Higgins et al., 2017): the popular method for disentanglement and also the baseline model for DeVAE when there is only one latent layer; 2) β-TCVAE (Chen et al., 2018): the TC-based method with a good balance of simplicity and effectiveness; 3) Dynamic-VAE (Shao et al., 2022): the latest method with incremental information bottleneck.

Disentanglement & Reconstruction. We conducted experiments on dSprites and Shapes3D to compare the above methods. Each trail was repeated 10 times with different random seeds. We draw the distributions of three disentanglement scores and reconstruction errors in Figure 2. Experimental results reveal that DeVAE achieves an average improvement of 8% comparing to β-TCVAE and 47% to β-VAE on dSprites for disentanglement. DeVAE surpasses 2% for β-TCVAE and 9% for β-VAE. Usually, DeVAE has better reconstruction fidelity than β-TCVAE. Though DynamicVAE achieves the best overall results, it still suffers from ID problems and is incapable of dealing with high-dimensional space.

Qualitative Visualization. We also conducted a qualitative analysis to assess disentanglement. We show the selected latent traversals whose KL divergence is larger than 0.5 in Figure 3. We can

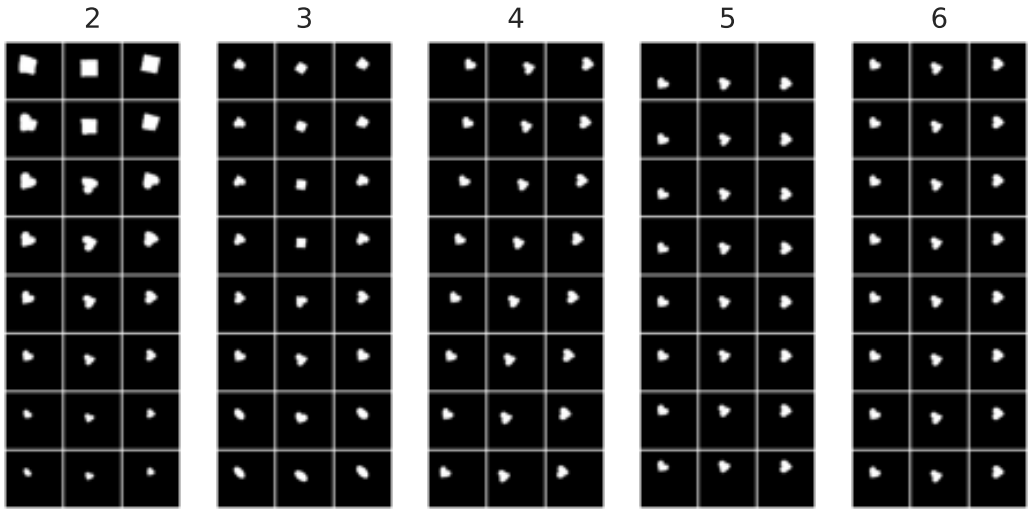

Figure 3: Latent traversal on dSprites. Each block shows the generated images of traversing the latent variable (title) from -2 to 2 with three different random sampling.

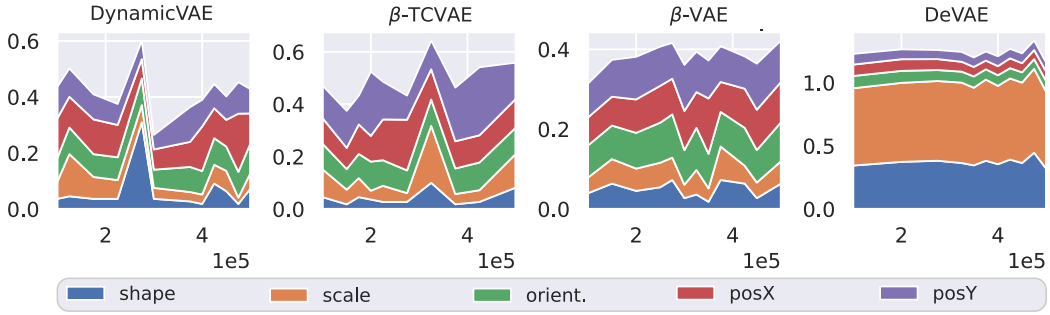

Figure 4: Comparison results of information diffusion. Each colored curve denotes the learned information that belongs to one factor over training iterations.

see that DeVAE disentangles position X and position Y perfectly. Shape, scale, and orientation are hard to be disentangled. We show the latent traversals with the highest MIG in Appendix A.4.

Preventing Information Diffusion. Information diffusion is a phenomenon of disentangling that one factor's information diffuses into other latent variables while training, causing the disentanglement scores to fluctuate during training (Wu et al., 2022). We hypothesize that losing the constraint of disentanglement is the reason for ID.

To prove it, we monitored the changes in mutual information during training. From Figure 4, we see that DynamicVAE has a significant trend of losing information on iteration 3e5. It means that the learned structure of representation was destroyed when expanding the IB. In contrast, DeVAE shows a relatively steady trend of increasing information for consistent regularizing. DeVAE overcomes the drawback of traditional information bottleneck-based methods by keeping the constraint of disentanglement.

Scaling to High-dimensional Latent Space. Most disentanglement methods evaluate their performance on simple scenes with only one object and few factors. It is a challenge to extend these methods to complex scenes. However, whether these methods adapt to a large latent space to fit more factors is questionable. In particular, the dimension of latent space affects the estimation accuracy of MI for the TC-based methods.

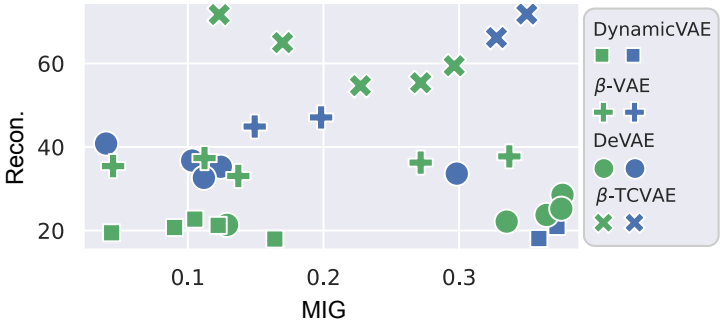

Figure 5: Distributions of MIG scores and reconstruction errors for low-dimensional space (blue) and high-dimensional space (green). The points in the bottom right have a better balance of disentanglement and reconstruction.

Table 1: The estimated MI of FactorVAE and the real MI on high dimensional spaces. The cases having large error are bold. ρ denotes the correlation of two random variables.

Dim	ρ	Estimated TC	TC	Error
10	0.3	0.23	0.24	0.03
	0.6	1.08	1.12	0.03
	0.9	4.14	4.15	0.00
100	0.3	2.17	2.36	0.08
	0.6	10.27	11.16	0.08
	0.9	23.39	41.52	0.44
1000	0.3	8.62	23.58	0.63
	0.6	17.40	111.57	0.84
	0.9	**22.47**	**415.18**	**0.95**

To study the effect of high-dimensional latent space on estimating TC, we first generate samples from a D-dimensional multi-variable normal distribution $\boldsymbol{x}$ which is divided into two groups $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ with D/2 dimensions. The variables in a group are independent $\mathrm{Cov}(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m) = 0, i \neq j$; the variables between groups are correlative $\mathrm{Cov}(\boldsymbol{x}_i^m, \boldsymbol{x}_i^n) = \rho, m \neq n$; each variable is a standard normal distribution. So, the TC of $\boldsymbol{x}$ is

$$\mathrm{TC}(\boldsymbol{x}) = -\frac{\mathrm{D}}{4}\log(1-\rho^2). \tag{8}$$

We trained a discriminator for 2000 iterations to estimate the TC introduced in FactorVAE (Kim & Mnih, 2018). We compared the estimated TC and the real TC over dimensions and ρ. Each trail was repeated 10 times, and we report the average results as shown in Table 1. One can see that increasing ρ or dimension diminishes the accuracy of estimation, and the estimators always have low errors when the dimension is 10. However, the estimation becomes extremely inaccurate when the dimension raises to 1000, which means such estimation will fail to penalize the TC for large models.

We further conduct experiments on dSprites to validate the above conclusion. The experimental settings are the same except for increasing the dimension of latent space to 1024. From Figure 5, we can see that β-TCVAE and DynamicVAE have significant performance decay. β-VAE and DeVAE are resistant to increasing the dimension of latent space. Higher dimensional space increases the complexity of calculating the TC and leads to significant estimation errors. Our model show robustness in high-dimensional latent spaces.

Table 2: Ablation Study on Multiple Space (MS), Hierarchical Structure (HiS) and Disentanglement-invariant Transformation (DiT).

MS	HiS	DiT	MIG Layer1	MIG Layer2	MIG Layer3	Recon. Layer1	Recon. Layer2	Recon. Layer3
✗	✗	✗	0.19	-	-	47.49	-	-
✓	✓	✗	0.24	0.29	0.30	38.82	**45.48**	63.78
✓	✗	✗	0.24	0.32	**0.35**	**22.21**	40.79	**62.40**
✓	✓	✓	**0.35**	**0.35**	**0.35**	43.29	75.11	175.99

Table 3: Exploration study of betas on disentanglement (MIG) and reconstruction (Recon.). l1, l2, l3 denote [1,10,20,40,80], [1,10,40], [1,10] respectively. s is fixed to 1.

Dataset	No. betas	MIG	Recon.
dSprites	l1	0.30±0.03	79.65±16.06
	l2	**0.35±0.02**	51.99±26.99
	l3	0.16±0.11	**38.19±02.35**
Shapes3D	l1	0.54±0.06	65.01±25.37
	l2	**0.57±0.01**	43.24±11.41
	l3	0.55±0.04	**39.31±06.96**

Table 4: Comparison of scale. We report the mean±std of MIG and reconstruction for 5 trails on dSprites and Shapes3D. The sequence of betas is fixed to [1,10,40].

scale	MIG	Recon.
0.3	0.21±0.14	**16.01±01.14**
0.5	0.29±0.09	22.40±01.78
1.0	**0.35±0.02**	51.99±26.99
0.3	0.55±0.02	**24.43±01.37**
0.5	**0.57±0.02**	28.48±03.31
1.0	0.57±0.01	43.24±11.41

3.3 Experimental Analysis

In this section, we performed ablation studies on the benefit of the proposed Hierarchical Latent Spaces (HiS) and Disentanglement-invariant Transformation (DiT). We also conducted extensive experiments to explore the effect of β and s on disentanglement and reconstruction performance.

Hierarchical Latent Spaces & Disentanglement-invariant Transformation. To demonstrate the effectiveness of the introduced Hierarchy Latent Spaces (HiS) and Disentanglement-invariant Transformation (DiT), we performed ablation experiments on the following situations: 1) The model has one single latent space; 2) We replace DiT with Linear Transformation ($\tau_i(\boldsymbol{z_i}) = w\boldsymbol{z_i}$), w is an arbitrary matrix; 3) The model applies a parallel structure instead of the hierarchy that latent spaces are independent; 4) The proposed model DeVAE.

We report the MIG and Recon. for each layer in Table 2. We can see that multiple latent spaces with different objectives promote disentanglement. Adding a hierarchical structure without DiT is incapable of promoting disentanglement further. Therefore, the key of DeVAE is to connect the multiple latent spaces by DiT to form a hierarchical structure with a decremental IB.

Effect of β. More latent layers mean more chance to explore disentanglement solutions but need more time to converge. Though Wu. etc. (Wu et al., 2022) proposes the Annealing Test to determine the value of β, it requires labels to learn the information freezing point (IFP). Choosing a suitable β for each layer is difficult without knowing the information of factors. Fortunately, DeVAE is insensitive to the choice of β, which means we can create redundant latent layers to cover all suitable βs. In Table 3, we compared tree cases: redundant betas ([1,10,20,40,80]), just betas ([1,10,40]), insufficient betas ([1,10]). Redundant betas slightly diminish the performance of disentanglement and reconstruction.

Effect of Scale s. Increasing s will add the weights of higher beta, encouraging disentanglement more than reconstruction fidelity. It is a crucial hyperparameter to balance the objectives of latent layers. Note that our model equals the vanilla VAE when $s = 0$. In Table 4, we compared the effects

of choosing s and reported the mean$\pm$std scores of MIG (Chen et al., 2018) and reconstruction. For most cases, $s = 1$ is a good choice.

3.4 LIMITATION

Since our model creates several diverse latent spaces, it is a challenge to optimize multiple objectives. Though there are numerous combinations for setting pressures and weighting these objectives, we only search a limited range of hyper-parameters. Even so, DeVAE shows compatible performance on the benchmarks. Though we validated that our model is adequate for high-dimensional space, we did not test it on real problems. It is challenging to train a disentanglement model on large-scale problems, such as ImageNet.

4 RELATED WORK

Disentanglement Learning. Disentanglement learning aims at learning generative factors existing in the dataset, that is, disentangled representation learning. Though the definition of disentanglement is still an open topic (Kumar et al., 2018; Do & Tran, 2020; Abdi et al., 2019; Duan et al., 2020), it is widely accepted that the redundancy between latent variables diminishes disentanglement. Penalizing the Total Correlation (TC) (Watanabe, 1960) is an important direction in disentanglement learning, and many SOTA methods are based on it (Chen et al., 2018; Kim & Mnih, 2018; Esmaeili et al., 2019;?; Kumar et al., 2018; Wei et al., 2021). PM algorithm promotes factorial codes but only works for binary codes (Schmidhuber, 1992); Though ICA (Comon, 1994) and PCA (Wold et al., 1987) ensure independence theoretically, they extract linear representations. Until recently, deep learning has made it workable. FactorVAE (Kim & Mnih, 2018) applies an adversarial training method to approximate and penalize the TC term. β-TCVAE (Chen et al., 2018) decomposed the KL term into three parts: mutual information (MI), total correlation (TC), and dimensional-wise KL (DWKL). However, these methods rely on the estimation of TC, which is extremely hard for high-dimensional spaces.

Information Bottleneck. Information bottleneck theory (Tishby et al., 1999; Shannon, 1948) plays a vital role in interpreting neural networks. Some methods encourage disentanglement by increasing the information bottleneck while training (Jeong & Song, 2019; Burgess et al., 2018; Shao et al., 2022; Dupont, 2018; Wu et al., 2022). These methods vary in the way of expanding the IB. Cascade-VAE (Jeong & Song, 2019) sequentially relieves one latent variable at one stage to increase the IB. DynamicVAE (Shao et al., 2022) designs a non-linear PI controller for manipulating β to control IB steadily increasing. DEFT (Wu et al., 2022) applies a multi-stage training strategy with separated encoders to extract one factor at one stage according to its information freezing point (IFP). However, the above incremental models, increasing the IB while training, suffer from the information diffusion (ID) problem (Wu et al., 2022) that the disentangled representation may diffuse the learned information into other variables. This work presents a novel framework with a decremental information bottleneck to solve the ID problem.

Hierarchical Latent Variables. Normalizing Flow (Rezende & Mohamed, 2015; Kingma et al., 2016) also uses hierarchical latent layers to generate an arbitrary distribution. Unlike Normalizing Flow, each layer aims to encourage disentanglement or reconstruction. Besides, Normalizing Flow gradually increases the complexity of the output distribution after entering a new layer. In contrast, our model reduces the complexity layer by layer.

5 CONCLUSION

We propose a novel framework with a decremental information bottleneck for disentanglement. Hierarchical latent spaces with disentanglement-invariant transformation are the key to overcoming the problem of losing disentanglement constraint while expanding the information bottleneck. The decremental method is compatible with high-dimensional problems and reduces the information diffusion problem.

Broader Impact Unlike previous works that spread the conflict of the trade-off over time, our work demonstrates a novel direction to solve the trade-off by spreading the conflict in spaces. Our work provides insights on balancing disentanglement and reconstruction.

REFERENCES

Amir H Abdi, Purang Abolmaesumi, and Sidney Fels. A preliminary study of disentanglement with insights on the inadequacy of metrics. *arXiv preprint arXiv:1911.11791*, 2019. 9

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018. 1, 4

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. In *International Conference on Machine Learning (ICML)*, 2018. 1, 2, 9

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 5, 9

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Neural Information Processing Systems (NeurIPS)*, pp. 2172–2180, 2016. 1

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. 9

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. *arXiv preprint arXiv:2010.14407*, 2020. 1

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2020. 9

Sunny Duan, Loic Matthey, Andre Saraiva, Nick Watters, Christopher Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 9

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *Neural Information Processing Systems (NeurIPS)*, pp. 708–718, 2018. 9

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018. 5

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2525–2534. PMLR, 2019. 9

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2, 5

Yeonwoo Jeong and Hyun Oh Song. Learning discrete and continuous factors of data via alternating disentanglement. In *International Conference on Machine Learning (ICML)*, 2019. 9

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018. 1, 5, 7, 9

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. 5

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 1, 2

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016. 9

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 9

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. 1, 4

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015. 9

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992. ISSN 08997667. 9

Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948. 9

Huajie Shao, Yifei Yang, Haohong Lin, Longzhong Lin, Yizhuo Chen, Qinmin Yang, and Han Zhao. Rethinking controllable variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19250–19259, 2022. 1, 2, 5, 9

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *In Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999. 1, 9

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. *arXiv preprint arXiv:2006.07886*, 2020. 1

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4:66–82, 1960. 9

Yuxiang Wei, Yupeng Shi, Xiao Liu, Zhilong Ji, Yuan Gao, Zhongqin Wu, and Wangmeng Zuo. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6721–6730, October 2021. 9

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 9

Jiantao Wu, Lin Wang, Bo Yang, Fanqi Li, Chunxiuzi Liu, and Jin Zhou. DEFT: distilling entangled factors by preventing information diffusion. *Mach. Learn.*, 111(6):2275–2295, 2022. doi: 10.1007/s10994-022-06134-7. URL https://doi.org/10.1007/s10994-022-06134-7. 1, 2, 6, 8, 9