Variational Autoencoders with Decremental Information Bottleneck for Disentanglement

Jiantao Wu[12]
jiantao.wu@surrey.ac.uk
Shentong Mo[4]
shentonm@andrew.cmu.edu
Xingshen Zhang[5]
zhxshd@outlook.com
Muhammad Awais[12]
muhammad.awais@surrey.ac.uk
Sara Atito[12]
sara.atito@surrey.ac.uk
Zhenghua Feng[123]
z.feng@surrey.ac.uk
Lin Wang[5]
wangplanet@gmail.com
Xiang Yang[6]
anson@mintyea.com

[1] Surrey Institute for People Centred AI, University of Surrey, Guildford, UK
[2] Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK
[3] School of Computer Science and Electronic Engineering, University of Surrey, Guildford, UK
[4] Carnegie Mellon University, Pittsburgh, USA
[5] University of Jinan, Jinan, China
[6] Mintyea Tech, Inc. Hangzhou, China

Abstract

One major challenge of disentanglement learning with variational autoencoders is the trade-off between disentanglement and reconstruction fidelity. Previous studies, which increase the information bottleneck during training, tend to lose the constraint of disentanglement, leading to the information diffusion problem. In this paper, we present a novel framework for disentangled representation learning, DeVAE, which utilizes hierarchical latent spaces with decreasing information bottlenecks across these spaces. The key innovation of our approach lies in connecting the hierarchical latent spaces through disentanglement-invariant transformations, allowing the sharing of disentanglement properties among spaces while maintaining an acceptable level of reconstruction performance. We demonstrate the effectiveness of DeVAE in achieving a balance between disentanglement and reconstruction through a series of experiments and ablation studies on dSprites and Shapes3D datasets. Code is available.

1 Introduction

Unsupervised learning [25] is essential for bridging the gap between human and machine intelligence. Disentanglement learning is a promising approach for obtaining explanatory

representations from observations without supervision, mimicking human intelligence [1]. Variational autoencoders (VAEs) [13] are widely used for disentanglement learning, with methods like beta-VAE [8] introducing a penalty (weighted by β) on the Kullback–Leibler (KL) divergence to promote disentanglement. However, there is a trade-off between disentanglement and reconstruction fidelity in beta-VAE.

To address this trade-off, some methods utilize a dynamic controlling strategy for β [3, 20, 24]. Generally, a high initial β value is set to enforce VAEs disentangle at the beginning. Then, the value of β is gradually reduced to facilitate reconstruction. Since β controls the Information Bottleneck (IB) [3, 21], these methods are called *incremental VAEs*, where the IB increases during training. As a result, incremental VAEs achieve a good balance by optimizing disentanglement and reconstruction in separate time spans.

In this work, we propose an alternative approach to address the conflict between optimizing disentanglement and reconstruction. Our primary motivation is to optimize disentanglement and reconstruction simultaneously by creating multiple latent spaces. Each latent space focuses on different tasks, either optimizing disentanglement or reconstruction, while our framework ensures these spaces share disentanglement properties. This approach enables simultaneous optimization of both disentanglement and reconstruction.

Specifically, we introduce DeVAE, a VAE framework with hierarchical latent spaces (HiS) that applies a novel IB-decremental strategy and a disentanglement-invariant transform (DiT) operator. DeVAE gradually decreases the information bottleneck across latent spaces, constrains the first space for reconstruction, and learns factors in subsequent spaces using narrow IBs. The disentanglement-invariant transform operator guarantees that the representations in these latent spaces disentangle the same factors.

Our contributions can be summarized as follows:

- We introduce a novel framework, DeVAE, which employs hierarchical latent spaces with decreasing information bottlenecks across the spaces, offering a new approach to balance disentanglement and reconstruction fidelity.
- We develop the disentanglement-invariant transformation, a key innovation that connects hierarchical latent spaces and enabling the sharing of disentanglement properties among them while maintaining a high level of reconstruction performance.
- We conduct comprehensive experiments and ablation studies on benchmark datasets, i.e. dSprites and Shapes3D, demonstrating the effectiveness of DeVAE in achieving a balance between disentanglement and reconstruction.

2 Related Work

Disentanglement Learning. Disentanglement learning aims to learn generative factors existing in the dataset [1]. Although the formal definition of disentanglement is still an open topic, it is widely accepted that the redundancy between latent variables diminishes disentanglement [6]. Penalizing the Total Correlation (TC) [22] is an important direction in disentanglement learning, and many state-of-the-art (SOTA) methods are based on it [4]. Predictability Minimization (PM) algorithm [18] promotes factorial codes but only works for binary codes; Though ICA [5] and PCA [23] ensure independence theoretically, they extract linear representations. Recently, deep learning has made this more feasible. FactorVAE [11] applies an adversarial training method to approximate and penalize the TC term.

β-TCVAE [4] decomposed the KL term into three parts: mutual information (MI), total correlation (TC), and dimensional-wise KL (DWKL). They achieve good performance by optimizing the TC term and avoiding penalizing the MI term. However, the TC-based methods introduce a strong assumption that generative factors are independent, which is impractical for real-world problems.

Information Bottleneck. Information bottleneck theory [19, 21] plays a vital role in interpreting neural networks. Some methods encourage disentanglement by increasing the IB during training [3]. These methods differ in the way they expand the IB. CascadeVAE [10] sequentially relieves one latent variable at each stage to increase the IB. DynamicVAE [20] designs a non-linear PI controller for manipulating β to control the steadily increasing IB. DEFT [24] applies a multi-stage training strategy with separated encoders to extract factors separately at different stages. However, the above incremental models, which increase the IB during training, suffer from the information diffusion (ID) problem [24], as the disentangled representation may diffuse the learned information into other variables when expanding the IB.

Hierarchical Latent spaces. Normalizing Flow [14, 17] uses hierarchical latent spaces to generate an arbitrary distribution. Unlike Normalizing Flow, each space in our model aims to encourage disentanglement or reconstruction. Additionally, Normalizing Flow gradually increases the complexity of the output distribution after entering a new space. In contrast, our model reduces the complexity space by space.

3 Methodology

3.1 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 that 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 [13] computes the representation function by introducing $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and optimizing the variational lower bound (ELBO). β-VAE [8] 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 [3]. We can set a large β to learn a disentangled representation and a small β to learn an informative representation. However, β-VAE suffers a trade-off between disentanglement and reconstruction, which means that β can only optimize one of these two goals.

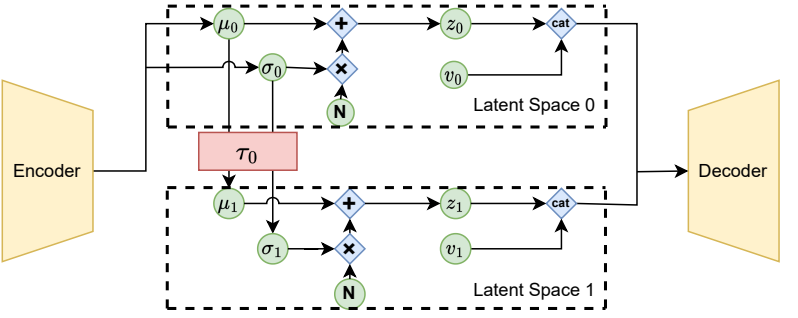

Figure 1: Illustration of our Decremental Variational Autoencoder (DeVAE). Each space has a pressure β_i to control the capacity of IB. τ_i connects two latent spaces. The first space is our main space to represent inputs. The subsequent spaces are minor spaces to improve disentanglement.

3.2 Hierarchical Latent Spaces with Decremental Information Bottleneck

To maintain the disentanglement constraint while optimizing reconstruction fidelity, we introduce a Hierarchical Latent Space (HiS) with K spaces and assign a pressure β_i to the i-th space $\mathcal{Z}_i$. Each space promotes disentanglement or reconstruction through a suitable value of β. The objective of the i-th space is given by:

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

where the first space $q_\phi(\boldsymbol{z}_0|\boldsymbol{x})$ is a conditional Bayesian network, v_i denotes a K-D vector to indicate the index of space, and the subsequent spaces 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 during processing. Therefore, we gradually decrease the IB in the sequential spaces, i.e., $\beta_{i+1} > \beta_i$. Typically, we set $\beta_0 = 1$ to encourage the first space to focus on reconstructing the original inputs. In this way, sequential spaces aim to disentangle factors of variation by setting narrow bottlenecks.

3.3 Disentanglement-invariant Transformation

In this part, we discuss the transformation τ_i which is vital to optimizing disentanglement and reconstruction simultaneously. If the transformation is arbitrary, the spaces will optimize their goal independently. Therefore, we need a mechanism to connect these goals to balance disentanglement and reconstruction in one space. To share disentanglement across all latent spaces, we propose a disentanglement-invariant transformation (DiT) denoted as τ:

$$\boldsymbol{\mu}_{i+1} = h(\boldsymbol{w}_i^1)\boldsymbol{\mu}_i, \quad \boldsymbol{\sigma}_{i+1} = h(\boldsymbol{w}_i^2)\boldsymbol{\sigma}_i, \tag{4}$$

where $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$, $\boldsymbol{w}_i^1, \boldsymbol{w}_i^2$ are learnable diagonal matrices of the i-th space, $h(\boldsymbol{w}) = e^{\boldsymbol{w}} > 0$ is an exponential function to make sure the scale values greater than 0.

We prove that scaling the latent space will not change disentanglement in Theorem 1, see proof in Appendix A.2.

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

3.4 Optimization Algorithm

According to Equation 4, we derive the parameters of latent variables for i-th space:

$$\boldsymbol{\mu}_i = h(\sum_{j=0}^{i-1} w_j^1)\boldsymbol{\mu}_0, \quad \boldsymbol{\sigma}_i = h(\sum_{j=0}^{i-1} w_j^2)\boldsymbol{\sigma}_0, \quad i > 0. \tag{5}$$

Applying the chain law, we get the i-th KL divergence:

$$D_{\mathrm{KL}_i} = \frac{1}{2}(1 + 2\sum_{j=0}^{i-1} w_j^2 + 2\log(\boldsymbol{\sigma}_0) - h(2\sum_{j=0}^{i-1} w_j^2)\boldsymbol{\sigma}_0^2 - h(2\sum_{j=0}^{i-1} w_j^1)\boldsymbol{\mu}_0^2) \tag{6}$$

The final objective of DeVAE is:

$$\mathcal{L}(\theta, \phi) = \sum_{i=0}^{K-1} \mathbb{E}_{q_\phi(\boldsymbol{z}_i|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_i, \boldsymbol{v}_i)] - \sum_{i=0}^{K-1} \beta_i D_{\mathrm{KL}_i}. \tag{7}$$

In this work, we aim to prove the validity of the proposed HiS with DiT for optimizing disentanglement and reconstruction simultaneously in different latent spaces. The algorithm of our method is shown in Algorithm 1. Figure 1 illustrates the architecture of DeVAE with two spaces. We set $K = 2$ for simplicity, and we find it is effective in practice. The main space applies $\beta_0 = 1$ to work as a vanilla VAE. We set a high value of β_1, adjusting according to problems, to encourage disentanglement.

Algorithm 1 DeVAE: Hierarchical Latent Spaces with Decremental Information Bottleneck

Require: Data $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N$, epochs T, learning rate η, pressure parameters $\beta_0 = 1, \beta_1$

```
Initialize the encoder and decoder networks φ and θ
for t = 1 to T do
    for each x in D do
        Compute μ_0, σ_0 using the encoder network q_φ(z_0|x)
        Sample z_0 ~ N(μ_0, σ_0)
        Compute μ_{i+1} = h(w_i^1)μ_i, σ_{i+1} = h(w_i^2)σ_i, using DiT
        Sample z_1 ~ N(μ_1, σ_1)
        Compute reconstruction loss L_rec = Σ_{i=0}^{1} E_{q_φ(z_i|x)}[log p_θ(x|z_i, v_i)]
        Compute KL divergence losses D_KL0 and D_KL1
        Compute total loss L(θ, φ) = L_rec − β_0 D_KL0 − β_1 D_KL1
        Update φ and θ using gradient descent with learning rate η
    end for
end for
```

4 Experiments

4.1 Experimental Setup

Datasets. The experiment section assesses the proposed DeVAE method on two widely-used datasets, dSprites [16] and Shapes3D [2]. dSprites has 737,280 binary 64 × 64 x 1

dataset	model	MIG	DCI dis.	FactorVAE	Recon.
dSprites	DeVAE	0.34± 0.02	0.53± 0.02	0.80± 0.03	48.31± 27.98
	DynamicVAE	0.35± 0.01	0.53± 0.01	0.82± 0.05	19.25± 1.85
	β-TCVAE(12.0)	0.29± 0.09	0.47± 0.08	0.73± 0.08	73.04± 3.41
	β-VAE(6.0)	0.17± 0.05	0.30± 0.07	0.74± 0.05	48.75± 2.84
shapes3D	DeVAE	0.53± 0.11	0.71± 0.02	0.79± 0.02	46.81± 13.97
	DynamicVAE	0.54± 0.04	0.68± 0.03	0.87± 0.10	31.02± 3.56
	β-TCVAE(12.0)	0.49± 0.11	0.73± 0.07	0.78± 0.01	44.53± 5.69
	β-VAE(6.0)	0.42± 0.18	0.68± 0.06	0.82± 0.06	34.95± 2.34

Table 1: Quantitative benchmarks on dSprites and shapes3D.

images generated from five factors: shape (3), orientation (40), scale (6), position X (32), and position Y (32). Shapes3D has 480,000 RGB $64 \times 64 \times 3$ 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. To evaluate the performance of disentanglement, three disentanglement metrics are applied. **MIG** [4]: the mutual information gap between two variables with the highest and the second-highest mutual information. **FactorVAE metric** [11]: the error rate of the classifier, which predicts the latent variable with the lowest variance. **DCI Dis.**: abbreviation for DCI Disentanglement [7], a matrix of relative importance by regression. **Recon.**: abbreviation for Reconstruction Error. We use Squared Error for RGB images (Shapes3D) and Binary Cross Entropy for binary images (dSprites).

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 [12] with a learning rate of $1e^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We employed a large batch size of 256 to accelerate the training process. All experiments train 300, 000 iterations by default. For the hyper-parameters, we set $\beta = 12$ for β-TCVAE, $\beta = 6$ for β-VAE, and $K_i = 0.001, K_p = 0.01$ for DynamicVAE, and $\{\beta_i\} = [1, 40]$ for DeVAE. We set $\beta_0 = 1$ to reconstruct image details and set $\beta_1 = 40$ to filter hard factors (shape, orientation) according to DEFT [24].

4.2 Comparison to Prior Work

To demonstrate the effectiveness of the proposed DeVAE, we compare it to three typical disentanglement methods: 1) β-VAE [8]: the baseline model for disentanglement and also the special case of DeVAE when $\beta_0 = \beta_1$; 2) β-TCVAE [4]: the SOTA method for penalizing TC; 3) Dynamic-VAE [20]: the SOTA method for incremental VAEs.

Disentanglement & Reconstruction. In comparison to prior work, DeVAE demonstrates effectiveness in achieving a balance between disentanglement and reconstruction. We conducted experiments on dSprites and Shapes3D where each trail was repeated 10 times with different random seeds and evaluated by MIG, FactorVAE, DCI disentanglement, and reconstruction error. We expect higher values for these metrics except recon. On the dSprites

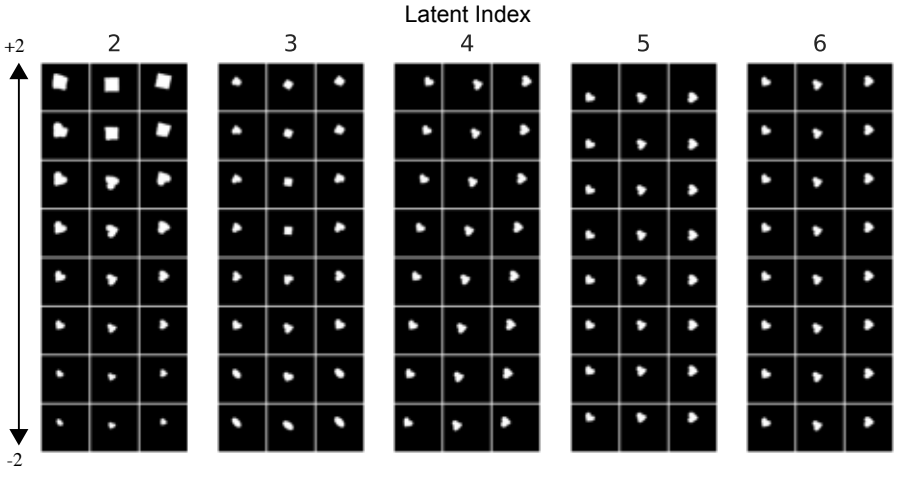

Figure 2: Latent traversal on dSprites. Each block shows the generated images of traversing the latent variable (title) from -2 to 2 with three different random sampling.

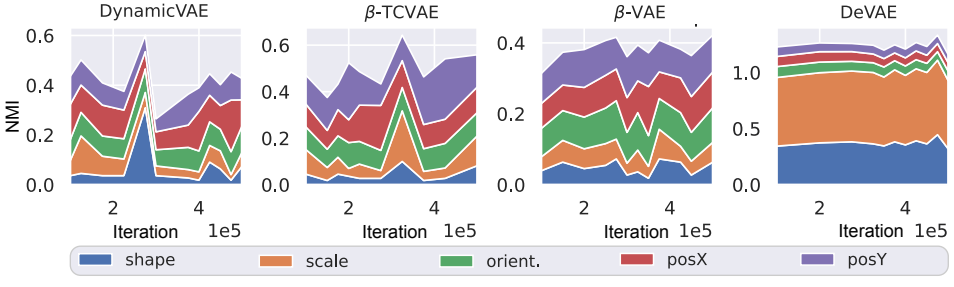

Figure 3: Comparison results of information diffusion. Each colored curve denotes the learned information that belongs to one factor over training iterations.

dataset, DeVAE achieves an average improvement of 12% in disentanglement compared to β-TCVAE and 38% compared to β-VAE. Furthermore, the reconstruction error is only half of that in β-TCVAE. The reconstruction drop on shapes3D is not that kind of large, because we use l2 loss instead of Bernoulli loss. DeVAE gains remarkable disentanglement with accepable reconstruction drop. Overall, DeVAE is competitive with Dynamic-VAE and surpasses both β-TCVAE and β-VAE.

Qualitative Visualization. Qualitative analysis is conducted to assess disentanglement by visualizing latent traversals [9] as shown in Figure 2. Specifically, each row reveals the reconstruction images from one dimension of the latent space systematically varied from -2 to 2 while keeping the others fixed. For each variation, the decoder of the VAE generates new images with three random seeds. We choose the top5 dimensions with the highest KL divergence to visualize their latent traversals. DeVAE successfully disentangles position X and position Y by latent 4 and 5. The hard factors, shape, scale, and orientation, are still a challenge in this domain. More examples can be found in the Appendix A.4.

MS	HiS	DiT	MIG			Recon.		
			space0	space1	space2	space0	space1	space2
✗	✗	✗	0.19	-	-	23.49	-	-
✓	✗	✗	0.24	0.32	**0.35**	**22.21**	**40.79**	**62.40**
✓	✓	✗	0.24	0.29	0.30	38.82	45.48	63.78
✓	✓	✓	**0.35**	**0.35**	**0.35**	43.29	75.11	175.99

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

Preventing Information Diffusion. Information diffusion is a phenomenon where one factor's information diffuses into other latent variables during training, leading to fluctuations in disentanglement scores [24]. We argue that our framework can solve the problem effectively due to removing the dynamic controlling strategy. Figure 3 demonstrates the changes in mutual information for the latent variable with the highest KL during training. NMI refers to the normalized mutual information, calculating the mutual information between one latent variable and one factor divided by the maximum information. The results show that Dynamic-VAE loses information significantly at iteration 3e5, indicating that the learned structure of representation is destroyed when expanding the information bottleneck (IB). On the other hand, DeVAE demonstrates a relatively steady trend of increasing information, thanks to consistent regularization. DeVAE overcomes the drawbacks of traditional IB-based methods by maintaining the constraint of disentanglement.

4.3 Experimental Analysis

In this section, we conduct ablation studies to evaluate the benefits of the proposed Hierarchical Latent Spaces (HiS) and Disentanglement-invariant Transformation (DiT). We also explore the effect of these spaces on the balance between disentanglement and reconstruction.

HiS & DiT. To demonstrate the effectiveness of the proposed Hierarchy Latent Spaces (HiS) and Disentanglement-invariant Transformation (DiT), we performed ablation experiments on the following scenarios: 1) HiS and DiT are removed, which equals to β-VAE; 2) HiS is replaced with multiple symmetric encoders instead of the hierarchy encoder, where latent spaces are independent; 3) DiT is replaced with Linear Transformation ($\tau_i(\boldsymbol{z}_{i+1}|\boldsymbol{z}_i) = \boldsymbol{w}\boldsymbol{z}_i$), where $\boldsymbol{w}$ is an arbitrary matrix. 4) The proposed model DeVAE. Unlike previous experiments, we compared these models on the dSprites dataset using three spaces ($\{\beta_i\} = [1, 10, 40]$) to show how DiT affects the connection between spaces. Table 2 shows the MIG and reconstruction for each space. From the results, we can see that MS and HiS without DiT improve disentanglement slightly. Adding DiT can make sure all latent spaces have same disentanglement. DeVAE achieves the best balance through sharing disentanglement at the third space and learning reconstruction at the first space. Thus, the key to DeVAE lies in connecting HiS through DiT.

Pressure on Space. We argue that the primary role of the first space is to optimize reconstruction and the second space is to optimize disentanglement. We investigated DeVAE with

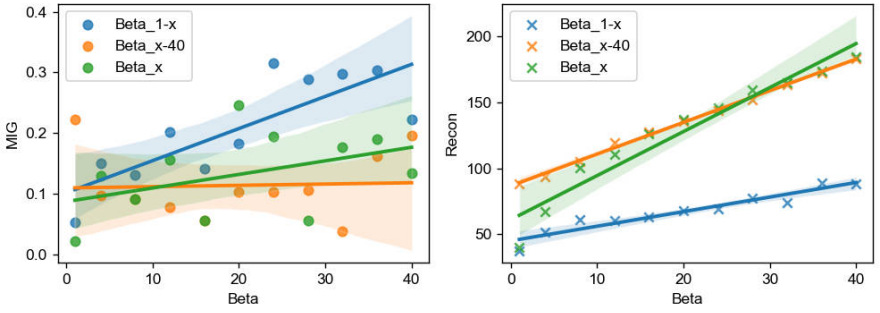

Figure 4: The effects of increasing β_i on latent spaces.

Dataset	betas	MIG	Recon.	Runtime (min)
dSprites	[1, 10, 20, 40, 80]	0.30±0.03	79.65±16.06	134
	[1, 10, 40]	**0.35±0.02**	51.99±26.99	109
	[1, 10]	0.16±0.11	**38.19±02.35**	101
Shapes3D	[1, 10, 20, 40, 80]	0.53±0.07	70.93±24.98	144
	[1, 10, 40]	**0.56±0.01**	56.09±4.39	119
	[1, 10]	0.55±0.04	**41.43±5.89**	103

Table 3: The effect of redundant spaces.

two latent spaces and applied the following rules to increase beta: 1) Beta_x: two spaces apply the same β, which equals to β-VAE. 2) Beta_1-x: only change the pressure of the second space. 3) Beta_x-40: only change the pressure of the first space. Figure 4 demonstrates the MIG and reconstruction by increasing beta. Each point denotes one experiment with corresponding beta. One can see that the DeVAE has few reconstruction drop to get a high MIG score. β_0 and β_1 have strong positive correlations with reconstruction error and MIG score respectively, meanwhile, the relationships to MIG and reconstruction are weaker. Therefore, β_0 controls reconstruction and β_1 promotes disentanglement.

Increasing Spaces. The number of spaces is a crucial hyperparameter in our framework. Although the setting $K = 2$ achieves remarkable performance, increasing the number of spaces may provide more opportunities to find an optimal solution. However, more spaces require additional computational resources and make it more challenging to optimize the neural network. In Table 3, we compared tree settings: $\{\beta_i\} = [1, 10, 20, 40, 80]$, $\{\beta_i\} = [1, 10, 40]$, $\{\beta_i\} = [1, 10]$. Fortunately, redundant betas slightly reduce the performance, which means we can create redundant latent spaces spanning a wide range of β values to obtain a good model without tuning the hyperparameter extensively.

5 Conclusion

In this paper, we propose a novel framework featuring hierarchical latent spaces, where the information bottleneck decreases across spaces. These latent spaces are connected through disentanglement-invariant transformations which are the key components to sharing disen-

tanglement among the spaces. Unlike incremental methods that optimize disentanglement and reconstruction in separate time spans, our work offers insights into optimizing these objectives simultaneously in hierarchical latent spaces. As an original contribution, we have demonstrated how to decouple the two goals, disentanglement and reconstruction, into different latent spaces.

Limitation. One limitation of the hierarchical latent spaces is the degradation of reconstruction, which occurs because these spaces are connected and share certain properties, such as disentanglement. Therefore, it is highly desirable to develop a better transformation between latent spaces that results in lower degradation. Future research could focus on improving this aspect of the model to further enhance the balance between disentanglement and reconstruction performance.

References

[1] 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.

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

[3] 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.

[4] 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.

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

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

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

[8] 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.

[9] Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

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

[11] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018.

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

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

[14] 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.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE Computer Society, 2015.

[16] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset, 2017.

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

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

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

[20] 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*, pages 19250–19259, 2022.

[21] 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.

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

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

[24] 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.

[25] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.