Rethinking Controllable Variational Autoencoders

Huajie Shao[1*†], Yifei Yang[2†], Haohong Lin[3†], Longzhong Lin[2],
Yizhuo Chen[2], Qinmin Yang[2], Han Zhao[4]
[1] College of Willam and Mary, [2] Zhejiang University
[3] Carnegie Mellon University, [4] University of Illinois at Urbana-Champaign

Abstract

The Controllable Variational Autoencoder (ControlVAE) combines automatic control theory with the basic VAE model to manipulate the KL-divergence for overcoming posterior collapse and learning disentangled representations. It has shown success in a variety of applications, such as image generation, disentangled representation learning, and language modeling. However, when it comes to disentangled representation learning, ControlVAE does not delve into the rationale behind it. The goal of this paper is to develop a deeper understanding of ControlVAE in learning disentangled representations, including the choice of a desired KL-divergence (i.e, set point), and its stability during training. We first fundamentally explain its ability to disentangle latent variables from an information bottleneck perspective. We show that KL-divergence is an upper bound of the variational information bottleneck. By controlling the KL-divergence gradually from a small value to a target value, ControlVAE can disentangle the latent factors one by one. Based on this finding, we propose a new DynamicVAE that leverages a modified incremental PI (proportional-integral) controller, a variant of the proportional-integral-derivative (PID) algorithm, and employs a moving average as well as a hybrid annealing method to evolve the value of KL-divergence smoothly in a tightly controlled fashion. In addition, we analytically derive a lower bound of the set point for disentangling. We then theoretically prove the stability of the proposed approach. Evaluation results on multiple benchmark datasets demonstrate that DynamicVAE achieves a good trade-off between the disentanglement and reconstruction quality. We also discover that it can separate disentangled representation learning and reconstruction via manipulating the desired KL-divergence.

*Corresponding author: Huajie Shao
†Authors contribute equally

1. Introduction

Variational Autoencoders (VAEs) have been widely used in various applications, such as language modeling, image generation, and representation learning. In particular, many variants of VAEs, such as β-VAE [11], FactorVAE [17] and β-TCVAE [6], have been recently proposed to learn the disentangled representations from the observations. Disentangled representation learning aims to encode input data into a low-dimensional space that preserves information about the salient factors of variation, so that each dimension of the representation corresponds to a distinct and explanatory factor in the data [3, 26, 45, 46]. Learning disentangled representations benefits a variety of downstream tasks [8, 11, 21, 24, 27, 28], including abstract visual reasoning [45], zero-shot transfer learning [5] and image generation [30, 48].

One major challenge of disentanglement learning is that there exists a trade-off between reconstruction quality of the input signal and the degree of disentanglement in the latent representations. To address this issue, researchers have developed a controllable variational antoencoder (ControlVAE) [39, 40] that combines control theory and the basic VAE to control the output KL via dynamically tuning the weight on the KL term in the VAE objective. While ControlVAE shows its good ability to disentangle latent variables, it still remains to explain the rationale behind it. The question is, why does it perform well on disentanglement learning via controlling the KL-divergence?

In addition, the core component of ControlVAE is the designed non-linear PI controller, a variant of the proportional-integral-derivative (PID) algorithm [1, 43]. The PI controller is able to stabilize the output KL-divergence to a specified value via dynamically adjusting the weight on the KL term. It thus can achieve a good trade-off between disentanglement and reconstruction. Since deep VAE model is complicated, its training may become unstable after incorporating the PI controller. In automatic control, tuning the hyperparameters of a PI/PID controller is regarded as the most challenging task [31]. Hence, obtaining a feasible region of the hyperparameters to ensure

the stability of ControlVAE remains a challenging question.

Moreover, ControlVAE does not explicitly provide the target KL-divergence C (i.e., set point) that can disentangle all the latent factors from observations. Namely, it is hard for us to tune the target KL-divergence for different datasets. Another question is, how to choose the set point of the target KL-divergence for different datasets? Is it possible to tune them in an unsupervised manner?

Our Contributions: This paper seeks to rethink ControlVAE in order to address the unanswered questions above for better understanding it. We attempt to offer an explanation of its good performance on disentangled representation learning via controlling the KL-divergence, analyze the set point of the target KL-divergence C, and discuss parameter tuning of the PI controller to ensure its stability. The main contributions of this paper are summarized as follows.

- We fundamentally explain why controlling the value of KL-divergence can learn disentangled representations from an information bottleneck perspective. We show that KL-divergence is an upper bound of the mutual information between inputs and their encodings.
- We propose a new model, DynamicVAE, that leverages an incremental PI controller, hybrid annealing and moving average to smoothly evolve the desired KL-divergence along a trajectory that can achieve high-quality disentanglement and low reconstruction error.
- We analytically derive a lower bound of the set point for the target KL-divergence of DynamicVAE and ControlVAE.
- We provide theoretical conditions on parameters of the PI controller to guarantee stability of DynamicVAE.
- Extensive experiments on benchmark datasets demonstrate that DynamicVAE achieves higher reconstruction quality and disentanglement than ControlVAE and the other baselines. Importantly, we discover that the proposed method can separate disentangled representation learning and reconstruction optimization.

2. Preliminaries

2.1. Variational Autoencoders (VAEs)

A Variational Autoencoder (VAE) [19, 36] is comprised of an encoder and a decoder. The encoder maps the observed data $\mathbf{x}$ into a low-dimensional latent space $\mathbf{z}$ while the decoder attempts to reconstruct the observations by sampling the data from the latent space. However, due to the intractable posterior inference, the basic VAE model is trained by optimizing the following variational lower bound (ELBO):

$$\mathcal{L}_{vae} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \quad (1)$$

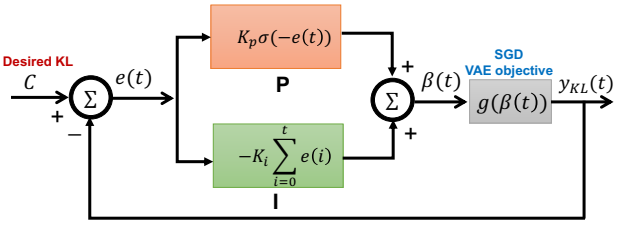


Figure 1. Diagram of the PI controller of ControlVAE.

where $p_\theta(\mathbf{x}|\mathbf{z})$ is the reconstruction of the observed data $\mathbf{x}$ given the latent variable $\mathbf{z}$; $q_\phi(\mathbf{z}|\mathbf{x})$ is a posterior distribution of latent variable $\mathbf{z}$ given $\mathbf{x}$; and $p(\mathbf{z})$ is a prior, such as the standard Gaussian.

2.2. ControlVAE

ControlVAE [40] is a new framework of VAE that combines control theory with the basic VAE to stabilize the KL-divergence to a desired value (i.e., set point), as illustrated in Fig. 1. It designs a PI controller to dynamically tune the weight β in the following β-VAE [12] objective to balance the disentanglement learning and reconstruction.

$$\mathcal{L}_c = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})). \quad (2)$$

Different from β-VAE that assigns a fixed weight to the KL term, ControlVAE adopts a positional PI controller to compute the weight β using the actual KL-divergence as feedback during training as follows:

$$\beta(t) = K_p\sigma(-e(t)) - K_i \sum_{j=0}^{t} e(j) + \beta_{min}, \quad (3)$$

where $e(t) = C - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))(t)$, which is the difference between desired KL-divergence C and the actual one at training step t; $\sigma(.)$ is a sigmoid function; β_{min} is an application-specific constant, such as 0; K_p and K_i are positive coefficients for the P term and I term respectively. Next, we will describe the basic idea of a PI controller.

2.3. PID Control Algorithm

The PID is a simple yet effective control algorithm that can stabilize system output to a desired value via feedback control [1, 43]. The PID algorithm calculates an error, $e(t)$, between a set point (in this case, the desired KL-divergence) and the current value of the controlled variable (in this case, the actual KL-divergence), then applies a correction in a direction that reduces that error. The correction is the weighted sum of three terms, one *proportional* to the error (called P), one that is the *integral* of error (called I), and one that is the *derivative* of error (called D); thus, the term PID. The derivative term is not recommended for noisy systems, such as ours, reducing the algorithm to PI control. The canonical form of a PI controller (applied to control

$\beta(t)$) is the following:

$$\beta(t) = K_p e(t) + K_i \sum_{j=0}^{t} e(j), \tag{4}$$

where $\beta(t)$ is the output of a controller, which (in our case) is the used β during training at time t; $e(t)$ is the error between the output value and the desired value at time t; K_p, K_i denote the coefficients for the P term and I term, respectively. Eq. (4) may be rewritten in incremental form, as follows:

$$\beta(t) = \Delta\beta(t) + \beta(t-1), \tag{5}$$

where $\beta(0)$ can be set as needed (as we show later), and:

$$\Delta\beta(t) = K_p[e(t) - e(t-1)] + K_i e(t). \tag{6}$$

Different from ControlVAE with *positional PI controller*, this paper adopts a nonlinear *incremental* form of the PI controller, as described later in Section 4.

3. Explanation of ControlVAE's Ability to Disentanglement Learning

In this section, we offer an explanation about the good ability of ControlVAE to disentangle the latent variables from the observations. ControlVAE leverages annealing method with step function to gradually change the KL-divergence from a small value to a large target value C. While it shows excellent performance for the learning of disentangled representations, the main reason remains unclear. The following proposition can help us better understand its good performance through mutual information.

Proposition 3.1. *The KL-divergence in the objective of ControlVAE, $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$, is an upper bound of the mutual information (MI) between the observed data $\mathbf{x}$ and the latent variables $\mathbf{z}$, denoted by $\mathcal{I}(\mathbf{x}, \mathbf{z})$. Namely, $\mathcal{I}(\mathbf{x}, \mathbf{z}) \leq \mathbb{E}_{p(\mathbf{x})}\left[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))\right]$.*

Please refer to the proof in Appendix A

According to Proposition 3.1, we can find that controlling the desired KL-divergence is equivalent to controlling the variational information bottleneck (VIB) [33] for information transmission. As the capacity of VIB is increased gradually, the simple and common latent factors in the observed data will transmit the latent channels for reconstruction. After all the latent factors are disentangled, ControlVAE tends to optimize the reconstruction as the target KL-divergence is gradually increased to some extend. That is why ControlVAE can balance disentanglement learning and reconstruction optimization via dynamically controlling the KL-divergence. Inspired by this, below, we try to improve the ControlVAE to smoothly increase the KL-divergence along a trajectory for disentanglement learning.

4. The DynamicVAE Model

Motivated by Section 3, we propose a new DynamicVAE model that controls the output KL-divergence from a small value to a target value smoothly.

We first review the objective of ControlVAE for disentangled representation learning. The basic idea is to maximize the log likelihood and simultaneously stabilize the KL-divergence to a target value C. It can be formulated as the following constrained optimization problem:

$$\begin{aligned} \max_{\phi,\theta} \quad & \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \\ \text{s.t.} \quad & D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) = C \end{aligned} \tag{7}$$

The prior work [43] has illustrated PI controller outperforms the Lagrange Multiplier (LM) for solving the constrained optimization problem, since LM may suffer from oscillation and constraint violations [34]. Hence, ControlVAE designs a PI controller to dynamically adjust $\beta(t)$ in the VAE objective in Eq. (2) to stabilize the KL-divergence to the desired value C:

While ControlVAE can disentangle latent representations via controlling KL-divergence, sometimes the output of KL divergence is not very stable during model training, as shown in Fig. 7 (a). In order to evolve the KL-divergence smoothly along a good trajectory in a tightly controlled fashion, we propose a novel DynamicVAE model based on control theory that can achieve a good trade-off between disentanglement and reconstruction. To reach this goal, we need to address the following two challenges:

1. KL-divergence should be increased smoothly from a small value to a large one. To this end, $\beta(t)$ should dynamically change from a large value to small one during model optimization. Specifically, at the beginning of training, $\beta(t)$ should be large enough to control the information bottleneck for disentangling the latent factors. After that, $\beta(t)$ is required to gradually drop to a small value to optimize the reconstruction.

2. $\beta(t)$ should not change too fast or oscillates too frequently. There is an interplay between the output KL-divergence and $\beta(t)$. $\beta(t)$ is computed from the feedback of the output KL-divergence while $\beta(t)$ influences the output KL-divergence during optimization. When $\beta(t)$ drops too fast or oscillates, it may cause KL-divergence to grow with a large value. Consequently, some latent factors may emerge earlier so that they can potentially be entangled with each other.

In this paper, we propose methods to deal with these two challenges, summarized below.

A non-linear incremental PI controller: As mentioned earlier, we need a large $\beta(t)$ in the beginning to change the KL-divergence smoothly from a small value to a large target value so that the information can be transmitted through

the latent channels per data sample. Accordingly, we adopt an *incremental form of the PI controller* in Eq. (3), and initialize it to a large value:

$$\beta(t) = \Delta\beta(t) + \beta(t-1), \tag{8}$$

where

$$\Delta\beta(t) = K_p[\sigma(-e(t)) - \sigma(-e(t-1))] - K_i e(t). \tag{9}$$

and $\beta(0)$ is a large initial value. When the PI controller is initialized to a large value $\beta(0)$, it can quickly produce a (small) KL-divergence during initial model training, preventing emergence of entangled factors.

Moving average: Since our model is trained with mini-batch data, it often contains noise that causes $\beta(t)$ to oscillate. In particular, when $\beta(t)$ plunges during training, it would cause KL-divergence to rise too quickly. This may lead to multiple latent factors coming out together to be entangled. To mitigate this issue, we adopt moving average method to smooth the output KL-divergence as the feedback of PI controller below.

$$y(t) = \sum_{i=t-T}^{t} \alpha_i y_{KL}(i), \tag{10}$$

where α_i denotes weight and T denotes the window size of past training steps.

Hybrid annealing: Control systems with step (input) function (i.e., those where the set point can change abruptly) often suffer from an overshoot problem [38]. An overshoot is temporary overcompensation, where the controlled variable oscillates around the set point. In our case, it means that the actual KL-divergence may significantly (albeit temporarily) exceed the desired value, when set point is abruptly changed. This effect would cause some latent factors to come out earlier than expected, and be entangled, thereby producing poor-quality disentanglement. To address this problem, we develop a hybrid annealing method that changes the set point more gradually, as illustrated in Fig. 7(b) in Appendix. It combines step function with ramp function to smoothly increase the target KL-divergence in order to prevent overshoot and thus better disentangle latent factors one by one.

The combination of the above three methods allows DynamicVAE to evolve $\beta(t)$ along a favorable trajectory to separate disentanglement learning and reconstruction optimization. We summarize the proposed incremental PI algorithm in Algorithm 1.

Algorithm 1 Incremental PI Control algorithm.

Input: desired KL C, coefficients K_p, K_i, β_{min}, iterations N, window T
Output: weight $\beta(t)$ at training step t
Initialization: $\beta(0) = 150\ (100), y_{KL}(0) = 0$
for $t = 1$ **to** N **do**
  Sample KL-divergence, $y_{KL}(t)$
  $y(t) = \sum_{t-T}^{t} \alpha_i y_{KL}(i)$
  $e(t) \leftarrow C - y(t)$
  $dP(t) \leftarrow K_p[\sigma(e(t)) - \sigma(e(t-1))]$
  $dI(t) \leftarrow K_i e(t)$
  if $\beta(t-1) < \beta_{min}$ **then**
    $dI(t) \leftarrow 0$ // wind up
  end if
  $d\beta(t) \leftarrow dP(t) + dI(t)$
  $\beta(t) \leftarrow d\beta(t) + \beta(t-1)$
  if $\beta(t) < \beta_{min}$ **then**
    $\beta(t) \leftarrow \beta_{min}$
  end if
  Return $\beta(t)$
end for

4.1. Set Point Guidelines

In this section, we fundamentally analyze how to choose a set point for the target KL-divergence C of DynamicVAE (same to ControlVAE). In DynamicVAE, latent factors transmit through the information bottleneck for the reconstruction of input data. In information theory, one bit, also called the information entropy of a binary random variable, is often used to encode data or transmit information. Hence, we adopt one bit theory to analyze the mutual information $\mathcal{I}(\mathbf{x}, \mathbf{z})$ between the input data $\mathbf{x}$ and the latent variable $\mathbf{z}$, which can be used as a lower bound of the desired KL-divergence for DynamicVAE.

Let M denote the capacity of variational information bottleneck for the complete reconstruction of N_x data samples. Then we have

$$2^M \geq N_x \Longrightarrow M \geq \log_2 N_x \tag{11}$$

Therefore, we can derive a lower bound of a set point for the target KL-divergence C, satisfying

$$C \geq M \geq \log_2 N_x \tag{12}$$

We will further verify this result empirically by conducting a set of experiments in Section 6.3.

5. Stability Analysis of DynamicVAE

We further theoretically analyze the stability of the proposed DynamicVAE with PI controller. This work is the *first* to offer the necessary and sufficient conditions that control hyperparameters should satisfy in order to guarantee the stability of KL-divergence, when β is manipulated dynamically during the training process of a (variant of) β-VAE.

To this end, our first step is to build the state space model for our control system. Throughout the paper, the state variable at training step t is defined as $x(t) = \beta(t)$.

Accordingly, the model of incremental PI controller can be written as:

$$x(t+1) - x(t) = K_p[\sigma(-e(t)) - \sigma(-e(t-1))] - K_i e(t), \tag{13}$$

where error $e(t)$, as shown in Fig. 1(a), is given by $e(t) = C - y(t-1)$. Here $y(t)$ is a dynamic model about the time response of the output KL divergence, $y_{KL}(t)$. According to [23], stochastic gradient descent (SGD) for optimizing an objective function can be described by a first-order dynamic model. Our experiment, as illustrated in Fig. 1(b), also shows that the response $y(t)$ in the open loop system approximately meets a negative exponential function, which further verifies that our system is a first-order dynamic system. We hence use the first-order dynamic model to describe it below.

$$\frac{dy}{dt} + ay = ag(x), \tag{14}$$

where a is a positive hyperparameter to describe the dynamic property, and $g(x)$ is a mapping function between the actual KL-divergence and $\beta(t)$. Since DynamicVAE is a discrete control system with sampling period $T_s = 1$, the above first-order dynamic model can be reformulated as

$$\begin{aligned} &y(t) - y(t-1) + ay(t) = ag(x(t)) \Longrightarrow \\ &y(t) = \frac{1}{1+a}y(t-1) + \frac{a}{1+a}g(x(t)). \end{aligned} \tag{15}$$

Now let $x_1(t) = x(t), x_2(t) = y(t-1), x_3(t) = y(t-2)$, then Eqs. (13) and (15) can be rewritten as the following state space equations.

$$\begin{aligned} x_1(t+1) &= x_1(t) - K_i[C - x_2(t)] + K_p[\sigma(x_2(t) - C) \\ &\quad - \sigma(x_3(t) - C)] \triangleq f_1(x_1(t), x_2(t), x_3(t)) \\ x_2(t+1) &= \frac{a}{1+a}g(x_1(t)) + \frac{1}{1+a}x_2(t) \\ &\triangleq f_2(x_1(t), x_2(t), x_3(t)) \\ x_3(t+1) &= x_2(t) \triangleq f_3(x_1(t), x_2(t), x_3(t)) \end{aligned} \tag{16}$$

In order to analyze the stability of the above non-linear state space model, one commonly used method is to linearize it at an equilibrium point [14]. In this paper, we use the following equilibrium point:

$$x^* = (x_1^*, x_2^*, x_3^*) = (g^{-1}(C), C, C), \tag{17}$$

where $g^{-1}(\cdot)$ denotes the inverse function and $x_2^* = x_3^*$. Next, we apply the first-order Taylor expansion to the above Eq. (16), yielding

$$X(t+1) = AX(t), \tag{18}$$

where

$$X(t) = [x_1(t) - x_1^*, x_2(t) - x_2^*, x_3(t) - x_3^*]^T, \tag{19}$$

and A is the Jacobian matrix at equilibrium point x^*, as defined in Eq. (22) in Appendix C. After this linearization, we can prove the stability of the proposed method as the modulus of eigenvalue λ of A is smaller than 1, as described in the following theorem.

Theorem 5.1. *Let $a > 0$ and assume $g'(x) < 0, \forall x > 0$. Then DynamicVAE is stable at the equilibrium point C if and only if the parameters of the PI controller, K_i and K_p, satisfy the following conditions*

$$\begin{cases} K_p + K_i < -\dfrac{4(1+a)}{ag'(x_1^*)} \\ 0.5K_p^2 ag'(x_1^*)^2 + 2[K_p - 8K_i(1+a)]g'(x_1^*) - 8(1+a) < 0 \\ K_i > 0, K_p > 0 \end{cases} \tag{20}$$

We provide the detailed proof in Appendix C.

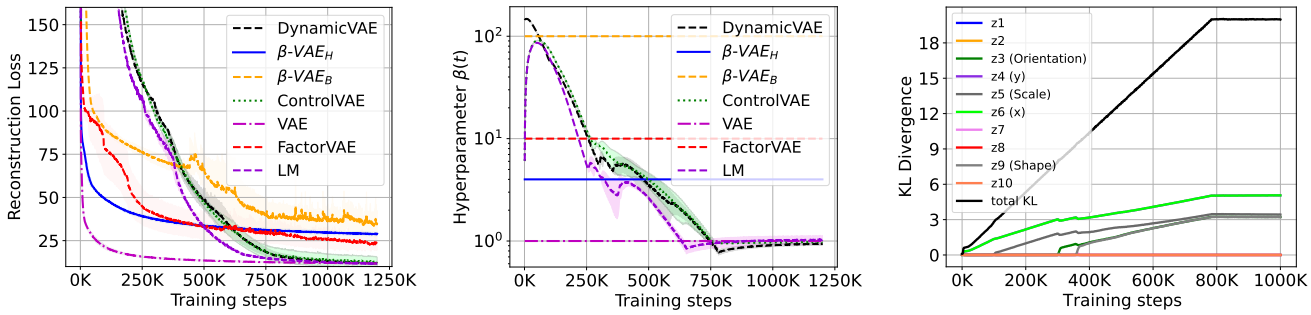
Remark 5.1. *The assumption of $g'(x) < 0, \forall x$ basically asks that the KL term in the objective to be a monotonously decreasing function of the coefficient $\beta(t)$, and we also further empirically corroborate its validity on two benchmark datasets as shown in Appendix C.1. In addition, we choose K_p and K_i that meet the above conditions (20) to verify the stability of DynamicVAE in Appendix C.1.*

6. Experiments

We evaluate the performance of DynamicVAE and compare it against existing baselines, including ControlVAE [40], β-VAE$_H$ [12], β-VAE$_B$ [5], FactorVAE [17], Lagrange Multiplier (LM) [37], and VAE [19]. We conduct experiments on multiple benchmark datasets: dSprites [5], MNIST [7], smallNORB [20], and 3D Chairs [2]. The detailed model configurations and hyperparameter settings are presented in Appendix D.

6.1. Results and Analysis

Dsprites Dataset: We first evaluate the performance of DynamicVAE on learning disentangled representations using *dSprites*. Fig. 2 (a) and (b) illustrate the comparison of reconstruction error and the hyperparameter $\beta(t)$ (using 5 random seeds) for different approaches. We can observe from Fig. 2 (a) that DynamicVAE (KL=20) has much lower reconstruction error (about 11.8) than β-VAE and FactorVAE, and is comparable to the basic VAE, LM, and ControlVAE. This is because DynamicVAE dynamically adjusts the weight, $\beta(t)$, to balance the disentanglement and reconstruction. Specifically, DynamicVAE automatically assigns a large $\beta(t)$ at the beginning of training in order to obtain good disentanglement, and then its weight gradually drops to about 1 at the end of optimization, as shown in Fig. 2 (b). In contrast, β-VAE and FactorVAE have a large and fixed weight in the objective so that their optimization algorithms tend to optimize the KL-divergence term (total correlation term for FactorVAE), leading to higher reconstruction error. For ControlVAE, it can also dynamically tune $\beta(t)$ to control the value of KL-divergence, but its disentanglement performance is worse than DynamicVAE as illustrated in



(a) Reconstruction loss. (b) $\beta(t)$. (c) KL-divergence.

Figure 2. (a) shows the comparison of reconstruction error on dSprites using 5 random seeds. DynamicVAE (KL=20) has comparable reconstruction errors to the basic VAE. (b) shows that DynamicVAE turns the weight of β-VAE into a small value less than 1. (c) shows an example of DynamicVAE on disentangling factors as the total KL-divergence increases.

Table 1. RMIG for different methods averaged over 5 random seeds. The higher the better. Note that we dismiss that FactorVAE suffers from **total correlation (TC) collapse** [17].

Models/Metric	pos. x	pos. y	Shape	Scale	Orientation	RMIG	MIG	BetaVAE Score
VAE	0.0359	0.0243	0.0116	0.1507	0.0039	0.0452	0.0539	0.8636
ControlVAE (KL=20)	0.6802	0.6597	0.0956	0.6040	0.1081	0.4295	0.5233	0.9608
FactorVAE ($\gamma = 10$)	0.7482	0.7276	0.1383	0.6262	0.1412	0.4763	0.5479	0.9801
β-VAE$_B$ ($\gamma = 100$)	0.5666	0.5763	0.4353	0.3814	0.0631	0.4045	0.4001	0.9940
β-VAE$_H$ ($\beta = 4$)	0.1635	0.1047	0.1391	0.3958	0.0127	0.1632	0.1687	0.8831
Lagrange Multiplier (LM)	0.6234	0.6177	0.0831	0.5850	0.0365	0.3891	0.4706	0.9977
DynamicVAE (KL=20)	0.7166	0.7179	0.2004	0.6530	0.1024	**0.4781**	**0.5578**	**0.9981**

Table 1. Fig. 2(c) illustrates an example of per-factor KL-divergence in the latent code as the total information capacity (KL-divergence) increases from 0.5 to 20. We can see that DynamicVAE disentangles all the five data generative factors, starting from position (x and y) to scale, followed by orientation and shape.

Next, we use three disentanglement metrics, robust mutual information gap (RMIG) [9], MIG and BetaVAE Score [25], to evaluate the disentanglement of different methods. We can observe from Table 1 that DynamicVAE has slightly better RMIG, MIG and BetaVAE Score than the FactorVAE, but it has much lower reconstruction error as illustrated in Fig. 2. However, FactorVAE may suffer from *total correlation collapse* [17] which is dismissed in our results. Moreover, DynamicVAE has higher RMIG, MIG and BetaVAE Score than β-VAE and the Lagrange Multiplier (LM). This is because LM does not inherently ensure convergence as it may suffer from oscillations and constraint violations [34, 43], as shown in training step around 370K in Fig. 2 (b). Our method also achieves much better disentanglement than ControlVAE for comparable reconstruction accuracy. Hence, DynamicVAE is able to improve the reconstruction quality yet obtain good disentanglement.

Qualitatively, we also visualize the disentanglement results of different models in Fig. 3. We can observe that DynamicVAE disentangles all the five generative factors on dSprites. However, ControlVAE is not very effective to disentangle all the factors when its KL-divergence is set to a large value, such as 20. Furthermore, β-VAE$_B$ ($\gamma = 100$) disentangles four generative factors and mistakenly combines the scale and shape factors together (in the third row). The other methods do not perform well for disentanglement.

3D Chairs and Other Datasets We also evaluate the proposed DynamicVAE on the other datasets: 3D Chairs, MNIST, and smallNORB. Fig. 4 illustrates the disentangled factors for DynamicVAE on 3D Chairs. We can observe from it that it disentangles six different latent factors, such as wheels, leg height, and azimuth. The proposed method can also disentangle diferent latent factors on smallNORB dataset as shown in Fig. 10 in Appendix E. Besides, we demonstrate that DynamicVAE can uncover many different latent factors on MNIST dataset, as illustrated in Fig. 11 in Appendix E. We can observe that our method achieves better disentanglement compared to the other approaches.

6.2. Separating Reconstruction and Disentanglement Learning

Additionally, we show that the proposed DynamicVAE is able to separate the reconstruction and disentanglement learning into two phases, mitigating the issue of balancing the trade-off between reconstruction and disentanglement. Fig. 5 illustrates the RMIG score and reconstruction loss with the increase of training steps after all the factors are

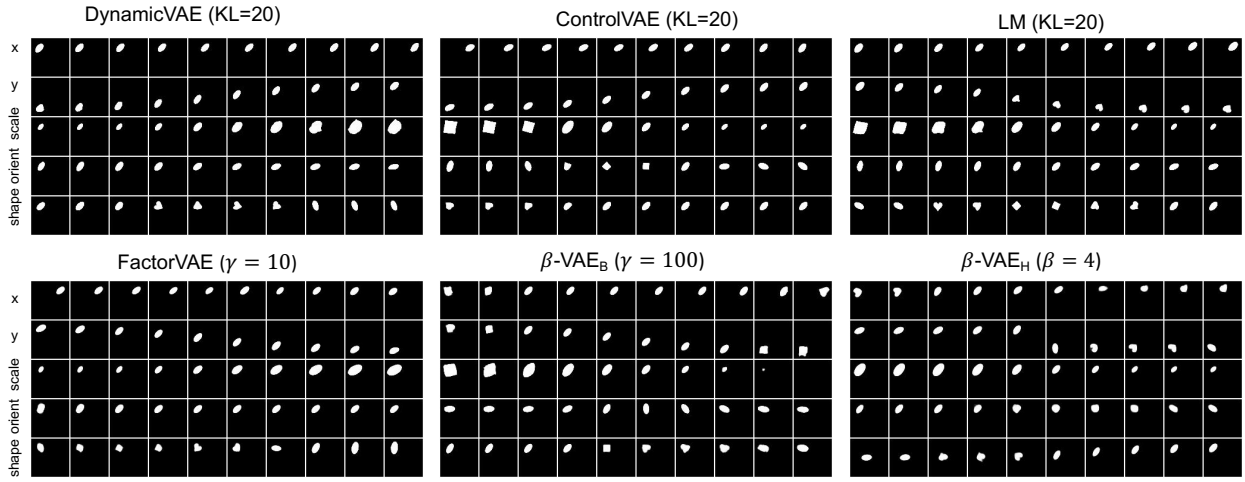


Figure 3. Rows: latent traversals ordered by the value of KL-divergence in a descending order. We initialize the latent representation from a seed image, and then traverse a single latent code in a range of $[-3, 3]$, while keeping the remaining latent code fixed.

disentangled (after $800,000$). It can be seen that RMIG score of our method remains stable as the reconstruction loss drops. Therefore, the proposed method barely introduces a conflict between reconstruction optimization and disentangled representation learning.

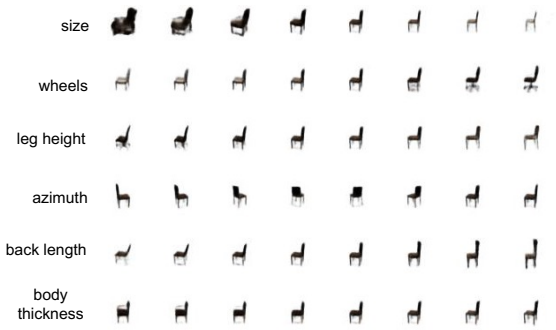


Figure 4. Sample traversals for the six latent factors on 3D Chairs.

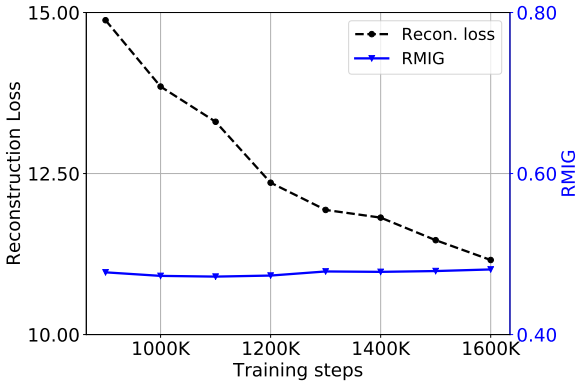


Figure 5. Averaged RMIG score and reconstruction loss vary with training steps.

6.3. Lower Bound of Set Point

Next, we conduct extensive experiments on benchmark datasets to verify the lower bound of the set point for the target value C in Section 4.1. In our experiments, we set the target KL-divergence C to 20 (using $\log_e$), and then measure the mutual information (MI) between input data $\mathbf{x}$ and latent variable $\mathbf{z}$, $\mathcal{I}(\mathbf{x}, \mathbf{z})$. Fig. 6 shows the mutual information (using $\log_2$) for different benchmark datasets with different number of samples. We can observe from it that when the proposed model is trained on dSprites with $737,280$ and 3000 data samples, the corresponding MI is about 19.33 and 14.82 respectively. They are very close to the corresponding ground truth: 19.49 and 14.87. We also conduct experiments on dSprites, 3DChairs and MNIST with different latent factors using same $30,000$ data samples. The results show their mutual information is about 14.74, 14.70, and 14.95, which are very close to the theoretical ground truth: 14.87. Thus, we can choose the target value C based on the lower bound in Eq. (12).

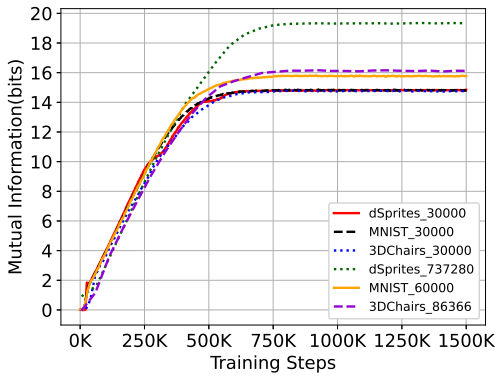


Figure 6. Mutual information $\mathcal{I}(\mathbf{x}, \mathbf{z})$ for benchmark datasets.

Table 2. RMIG for different methods averaged over 5 random seeds. The higher is better.

Models/Metric	pos. x	pos. y	Shape	Scale	Orientation	RMIG
DynamicVAE	0.7166	0.7179	0.2004	0.6530	0.1024	**0.4781** $\pm$ 0.0172
DynamicVAE-P	0.7376	0.7317	0.0992	0.6400	0.1120	0.4641 $\pm$ 0.0240
DynamicVAE-step	0.7209	0.7143	0.0664	0.6218	0.1543	0.4555 $\pm$ 0.0355
DynamicVAE-t	0.7152	0.7110	0.0997	0.6267	0.1322	0.4570 $\pm$ 0.0182

6.4. Ablation Studies

To compare the performance of DynamicVAE and its variants, we perform following ablation studies:

- DynamicVAE-P: it uses positional PI controller with no initialization to a large $\beta(0)$, instead of the incremental PI initialized to a large $\beta(0)$, to tune the weight on KL term in the VAE objective.
- DynamicVAE-step: it solely adopts step function without ramp function for our annealing method.
- DynamicVAE-t: this model directly uses the output KL-divergence at time t as a feedback of PI controller without using moving average to smooth it.

Table 2 shows the comparison of RMIG score for DynamicVAE and its variants. It can be observed that DynamicVAE outperforms the other methods in terms of overall RMIG score. We also find that DynamicVAE-step does not perform well because the ramp function is removed from our annealing method, leading to overshoot of PI controller. As a result, it makes the other factors come out earlier and entangled to each other. Thus, we can conclude the importance of adding ramp function for our annealing method. In addition, we can see that the proposed moving average and incremental PI control algorithm also play a critical role to improve the disentanglement.

7. Related Work

Supervised disentanglement learning [18, 29, 35, 42]. This method requires the prior knowledge of some data generative factors from human annotation to train the model. Some studies [24, 25, 27] figure out that it is hard to achieve reliable and good disentanglement without supervision. For supervised learning, the limited labeling information can help ensure a latent space of the VAE with desirable structure w.r.t to the ground-truth latent factors. In order to reduce human annotations, researchers tried to develop weakly supervised learning [4, 13, 26] to learn disentangled representations. However, these methods still require explicit human labeling or assume the change of the two observations is small. In practice, it is unrealistic for initial learners to discover the data generative factors in most real world scenarios.

Unsupervised disentanglement learning. The recent approaches mainly build on Variational Autoencoders (VAEs) [19, 49] and Generative Adversarial Networks (GANs) [10, 32, 41]. InfoGAN [22] is the first scalable unsupervised learning method for disentangling. It, however, suffers from training instabilities and does not perform well in disentanglement learning [11], so most recent works are largely based on VAEs models. The VAE models, such as β-VAE ($\beta > 1$), FactorVAE and β-TCVAE [6], often suffer from high reconstruction errors in order to obtain better disentanglement, since they add a large weight to terms in the objective. To address this problem, recent studies adopted the Lagrange Multiplier (LM) method to dynamically adjust the weight on the KL term [37, 44]. However, LM may suffer from oscillations on its way to the steady state, leading to constraint violations [43] and thus high errors. To mitigate this issue, researchers proposed ControlVAE that leverages PI control algorithm to ensure better/faster convergence of KL divergence to a desired value [40]. While ControlVAE shows the good ability to disentanlement learning, it does not offer a clear explanation. In this paper, we aim to better understand ControlVAE and then propose a new DynamicVAE that can achieve better disentanglement yet obtain lower reconstruction error.

8. Conclusion

This paper aimed to develop a deep understanding of ControlVAE for disentangled representation learning. From information bottleneck theory, we offered an explanation about why it performs well on disentanglement learning via stabilizing the output KL-divergence to different set points. Then we theoretically derived a lower bound of the set point for the target KL-divergence. It was further validated via conducting extensive experiments. In order to evolve the output KL-divergence smoothly along a good trajectory, we further proposed a novel model, DynamicVAE, for better disentanglement learning. Specifically, we leveraged an incremental PI controller, moving average and a hybrid annealing to stabilize the KL-divergence to separate the disentanglement learning and reconstruction optimization. We further theoretically prove the stability of the proposed method. The evaluation results demonstrate DynamicVAE can significantly improve the reconstruction accuracy meanwhile achieving better disentanglement than ControlVAE and the other baselines.

Acknowledgement This work was supported by 2022 Pre-Tenure Faculty Summer Grant at William and Mary.

References

[1] Karl Johan Åström, Tore Hägglund, and Karl J Astrom. *Advanced PID control*, volume 461. ISA-The Instrumentation, Systems, and Automation Society Research Triangle ..., 2006. 1, 2

[2] Mathieu Aubry, Daniel Maturana, Alexei A Efros, Bryan C Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014. 5

[3] 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

[4] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 8

[5] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018. 1, 5, 14

[6] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018. 1, 8

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016. 5

[8] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423, 2017. 1

[9] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *Proceedings of ICLR*, 2020. 6

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 8

[11] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 1, 8

[12] Irina Higgins, Loic 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. *ICLR*, 2(5):6, 2017. 2, 5

[13] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, pages 2506–2513, 2019. 8

[14] Jonathan L Hughes. Applications of stability analysis to nonlinear discrete dynamical systems modeling interactions. 2015. 5, 11

[15] IS Isa, BCC Meng, Z Saad, and NA Fauzi. Comparative study of pid controlled modes on automatic water level measurement system. In *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pages 237–242. IEEE, 2011. 13

[16] Eliahu Ibraham Jury. Theory and application of the z-transform method. 1964. 11

[17] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2654–2663, 2018. 1, 5, 6

[18] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 8

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 5, 8

[20] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004. 5

[21] Seunghun Lee, Sunghyun Cho, and Sunghoon Im. Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15252–15261, June 2021. 1

[22] Zinan Lin, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *international conference on machine learning*, 2020. 8

[23] Guan-Horng Liu and Evangelos A Theodorou. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint arXiv:1908.10920*, 2019. 5

[24] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pages 14611–14624, 2019. 1, 8

[25] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124, 2019. 6, 8

[26] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020. 1, 8

[27] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *In-*

ternational Conference on Learning Representations, 2019. 1, 8

[28] Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412, 2019. 1

[29] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in neural information processing systems*, pages 5040–5048, 2016. 8

[30] Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debhath, Anjul Patney, Ankit B Patel, and Anima Anandkumar. Semi-supervised stylegan for disentanglement learning. *Proceedings of ICML*, pages arXiv–2003, 2020. 1

[31] Yoshikazu Nishikawa, Nobuo Sannomiya, Tokuji Ohta, and Haruki Tanaka. A method for auto-tuning of pid control parameters. *Automatica*, 20(3):321–332, 1984. 1

[32] Lili Pan, Peijun Tang, Zhiyong Chen, and Zenglin Xu. Contrastive disentanglement in generative adversarial networks. *arXiv preprint arXiv:2103.03636*, 2021. 8

[33] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018. 3

[34] John C Platt and Alan H Barr. Constrained differential optimization for neural networks. 1988. 3, 6

[35] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439, 2014. 8

[36] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 2

[37] Danilo Jimenez Rezende and Fabio Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018. 5, 8

[38] Mohammad Shamsuzzoha and Sigurd Skogestad. The setpoint overshoot method: A simple and fast closed-loop approach for pid tuning. *Journal of Process control*, 20(10):1220–1234, 2010. 4

[39] Huajie Shao, Zhisheng Xiao, Shuochao Yao, Aston Zhang, Shengzhong Liu, and Tarek Abdelzaher. Controlvae: Tuning, analytical properties, and performance analysis. *arXiv preprint arXiv:2011.01754*, 2020. 1

[40] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. Controlvae: Controllable variational autoencoder. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. 1, 2, 5, 8

[41] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. *arXiv preprint arXiv:2101.02477*, 2021. 8

[42] Narayanaswamy Siddharth, Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pages 5925–5935, 2017. 8

[43] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. *arXiv preprint arXiv:2007.03964*, 2020. 1, 2, 3, 6, 8

[44] Ties van Rozendaal, Guillaume Sautiere, and Taco S Cohen. Lossy compression with distortion constrained optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 166–167, 2020. 8

[45] Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pages 14245–14258, 2019. 1

[46] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602, 2021. 1

[47] Jerzy Zabczyk. Mathematical control theory. *An Introduction*, 1992. 12

[48] Yutong Zheng, Yu-Kai Huang, Ran Tao, Zhiqiang Shen, and Marios Savvides. Unsupervised disentanglement of linear-encoded facial semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3917–3926, June 2021. 1

[49] Xinqi Zhu, Chang Xu, and Dacheng Tao. Where and what? examining interpretable disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5861–5870, June 2021. 8