Machine Learning (2022) 111:2275–2295
https://doi.org/10.1007/s10994-022-06134-7

DEFT: distilling entangled factors by preventing information diffusion

Jiantao[1,2] · **Lin Wang**[1,2] · **Bo Yang**[1,2] · **Fanqi Li**[1] · **Chunxiuzi Liu**[1] · **Jin Zhou**[1]

Received: 3 May 2021 / Revised: 19 January 2022 / Accepted: 6 February 2022 /
Published online: 29 March 2022

Abstract
Disentanglement is a highly desirable property of representation owing to its similarity to human understanding and reasoning. Many works achieve disentanglement upon information bottlenecks. Despite their elegant mathematical foundations, the IB branch usually exhibits lower performance. In order to provide an insight into the problem, we develop an annealing test to calculate the information freezing point (IFP), which is a transition state to freeze information into the latent variables. We also explore this clue or inductive bias for separating the entangled factors according to the differences in the IFP distributions. We found the existing approaches suffer from the information diffusion problem, according to which the increased information diffuses in all latent variables. Based on this insight, we propose a novel disentanglement framework, termed the distilling entangled factor (DEFT), to address the information diffusion problem by scaling backward information. DEFT applies a multistage training strategy, including multigroup encoders with different learning rates and piecewise pressure, to disentangle the factors stage by stage. We evaluate DEFT on three variants of dSprites and SmallNORB, which shows low-variance and high-level disentanglement scores. Furthermore, the experiment under the correlative factors demonstrates incapable of TC-based approaches. DEFT also exhibits a competitive performance in the unsupervised setting.

Keywords Disentanglement · Information Bottleneck · VAE · Representation learning · Information diffusion

1 Introduction

An understanding and reasoning about the world based on a limited set of observations is important in the field of artificial intelligence. For instance, we can infer the movement of a ball in motion at a single glance, as the human brain is capable of disentangling positions

Communicated by Editors: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautmann.

✉ Lin Wang
wangplanet@gmail.com

[1] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan 250022, China

[2] Quancheng Shandong Laboratory, Jinan 250100, China

from a set of images without supervision. Therefore, disentanglement learning is highly desirable to build intelligent applications. A disentangled representation has been proposed to be beneficial for a large variety of downstream tasks (Schölkopf et al., 2012). According to Kim and Mnih (2018), a disentangled representation promotes interpretable semantic information, resulting in substantial advancement, which includes but is not limited to reducing the performance gap between humans and AI approaches (Higgins et al., 2018b; Tenenbaum, 2018). Other instances of disentangled representations include semantic image understanding and generation (Lample et al. 2017), zero-shot learning (Zhu et al., 2019), and reinforcement learning (Higgins et al., 2017b).

As depicted in the seminal paper by Bengio et al. (2013), humans can understand and reason from a complex observation, after which they can induce the explanatory factors. The observations are generated by explanatory ground-truth factors c, which are invisible from the observations. The task of disentanglement learning aims to obtain a disentangled representation that separates these factors from the observations. The notion of disentanglement remains an open topic (Do and Tran, 2020; Higgins et al., 2018a), and we follow a strict version of discourse that one and only one latent variable z_i represents one corresponding factor, c_j (Burgess et al., 2017).

Locatello et al. (2019) proved the impossibility of disentanglement learning without inductive biases on the model and data. One popular inductive bias on the model assumes that the latent variables are independent. These approaches, penalizing total correlation (TC), dominate visual disentanglement learning (Chen et al., 2018; Kumar et al., 2018). This assumption is correct when the factors are sampled uniformly; however, the independent factors show statistical relevance in reality (Träuble et al., 2021). For instance, we observe that men are more likely to have short hair, and based on the observations, there is a correlation between gender and hair length. However, a man who is not bald may grow long hair if desired. In other words, sex does not determine hair length, and they are two independent factors. Therefore, the exploration of disentanglement approaches beyond the independence assumption is vital to reality applications.

Another popular research approaches are based on information theory (Jeon et al., 2021; Chen et al., 2016). They hypothesize that the gradually increased information bottleneck (IB) leads to a better disentanglement (Burgess et al., 2017; Dupont, 2018). Unfortunately, in practice, the approaches based on IB usually exhibit lower performance than those penalizing the TC (Locatello et al., 2019). However, it is important to understand whether this means that the total correlation beats the IB. It is believed that the answer is negative. In this research, we investigate the reason for which IBs fall behind TC in practice. We found that the information diffusion (ID) problem is an invisible hurdle that should be addressed in the IB community.

Information diffusion indicates that one factor's information diffuses into two or more latent variables; thus, the disentanglement scores fluctuate during training. Figure 1 shows the disentanglement scores of three approaches with the best hyperparameter settings, and it is observed that numerous trials have a high variance[1]. We bridge the ID problem with the instability of the current approaches in Sect. 3.

In this paper, we trace the ID problem by measuring the NMI1 and the NMI2, see Equation 8. The learned information may diffuse into other latent variables when IB-based approaches, such as AnnealedVAE (Burgess et al., 2017) and CascadeVAEC (Jeong and

[1] We use the pretrained models in disentanglement lib by Locatello et al. .

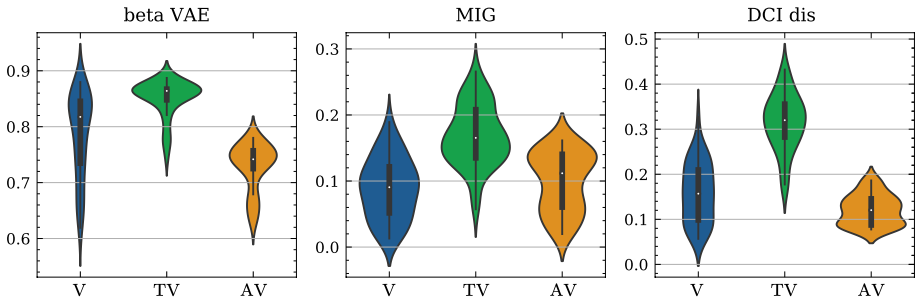


Fig. 1 The distribution of beta VAE metric, MIG, and DCI disentanglement on dSprites. Models are abbreviated (V=β-VAE, TV=β-TCVAE, AV=AnnealedVAE), and 50 trials are run with different random seeds

Song, 2019), learn new information. It is crucial to detect the components with different contributions to the objective for increasing the IB gradually. To do that, we have developed the annealing test to measure information freezing point (IFP) that the critical value for learning information from inputs. We also find that one factor is easy to be disentangled if the IFP distribution is distinguished from others.

Inspired by distillation[2] in chemistry, we can divide the training process into several stages and disentangle one component at each stage. In particular, we propose a framework, called the distilling entangled factor (DEFT), to disentangle factors stage-by-stage. DEFT chooses selective pressure to enable some information to pass through the IB according to the IFP distribution at each stage. In addition, DEFT reduces the backward information of the first $m-1$ sub-encoders by scaling the learning rate to relieve the ID problem at the m -th stage. We evaluate DEFT on four datasets, which shows robust performances. We also examine DEFT on the dataset with correlative factors. Our codes and all experimental settings are published in dlib for PyTorch forked from disentanglement lib. Our contributions are summarized in the following:

- We hypothesize that the ID problem is one reason for the low performances of IB-based approaches.
- We propose DEFT, a multistage disentangling framework, to address the ID problem by blocking partial information and scaling the backward information.

2 Preliminary

2.1 Disentanglement approaches

Variational autoencoder In variational inference, posterior $p(z|x)$ is intractable. The variational autoencoder (VAE) (Kingma and Welling, 2014) uses a neural network $q_\phi(z|x)$ (encoder) to approximate the posterior $p(z|x)$. The other neural network $p_\theta(x|z)$

[2] Distillation is the process of separating a mixture into its components by heating at an appropriate temperature, such that components boil and freeze into the target containers.

(decoder) rebuilds the observations. The objective of the VAE is to optimize the evidence lower bound (ELBO):

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

β-VAE Higgins et al. discovered the relationship between the disentanglement and the Kullback-Liebler (KL) divergence penalty strength. They proposed the β-VAE to introduce an additional pressure on the KL term:

$$\mathcal{L}^1(\theta, \phi; \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(x|z)] - \beta D_{\mathrm{KL}}(q_\phi(z|x)||p(z)), \tag{2}$$

where β controls the pressure for the posterior $q_\phi(z|x)$ to match the factorized unit Gaussian prior $p(z)$. However, there is a trade-off between the quality of the reconstructed images and the performance of disentanglement.

AnnealedVAE Burgess et al. (2017) proposed the AnnealedVAE, which progressively increases the information capacity of the latent variables while training:

$$\mathcal{L}^2(\theta, \phi; C) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(x|z)] - \gamma \left| D_{\mathrm{KL}}(q_\phi(z|x)||p(z)) - C \right|, \tag{3}$$

where γ is a sufficiently large constant (usually 1, 000) to constrain the latent information, and C controls the capacity that gradually increases from zero to a large number.

β-TCVAE The TC (Watanabe, 1960) quantifies the dependency among variables. β-TCVAE (Chen et al., 2018) decomposed the KL term into three parts: mutual information (MI), total correlation (TC), and dimensional-wise KL (DWKL). The TC can be penalized to achieve both high reconstruction quality and disentanglement:

$$\begin{aligned} \mathcal{L}^3(\theta, \phi; \beta) =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(x|z)] - \mathbb{E}_{q(z,n)}\left[\log \frac{q_\phi(z \mid n)p(n)}{q_\phi(z)p(n)}\right] \\ & - \beta \mathbb{E}_{q_\phi(z)}\left[\log \frac{q_\phi(z)}{\prod_j q_\phi(z_j)}\right] - \sum_j \mathbb{E}_{q_\phi(z_j)}\left[\log \frac{q_\phi(z_j)}{p(z_j)}\right]. \end{aligned} \tag{4}$$

CascadeVAEC Jeong and Song provided another total correlation penalization through information cascading. They proved that $TC(z) = \sum_{i=2}^{d} I(z_{1:i-1}; z_i)$. CascadeVAEC, the continuous version, sequentially relieves one latent variable at one stage, encouraging the model to disentangle one factor during the i-th stage:

$$\begin{aligned} \mathcal{L}^4(\theta, \phi; \beta_l, \beta_h) =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(x|z)] \\ & - \beta_l D_{\mathrm{KL}}(q_\phi(z_{1:i}|x)||p(z_{1:i})) - \beta_h D_{\mathrm{KL}}(q_\phi(z_{i+1:d}|x)||p(z_{i+1:d})), \end{aligned} \tag{5}$$

where β_l is a small value for opening the information flow, β_h is a large value for blocking information, and d is the number of dimensions.

Relevant but not compared approaches ICA (Comon, 1994) and PCA (Wold et al., 1987) guarantee the independence mathematically, and the nonlinear versions are helpful to disentanglement (Sorrenson et al., 2020). However, they require the factors satisfying a factorized prior distribution. Learning factorial codes (Schmidhuber, 1992) is limited in the cases with binary codes. Merely encouraging independence is insufficient

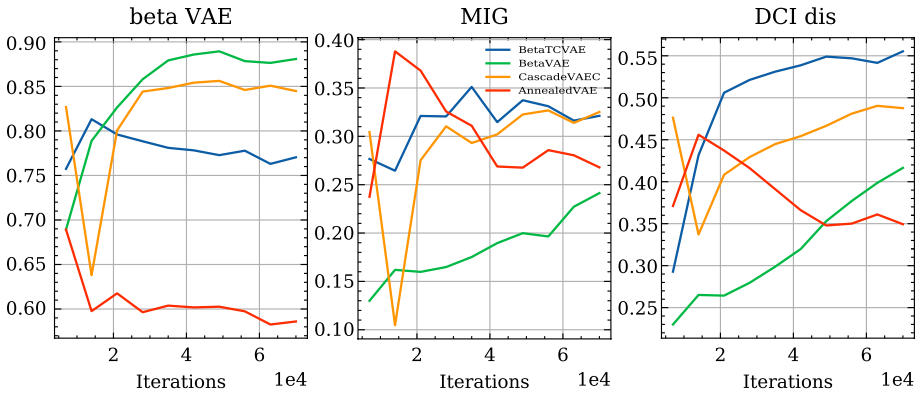


Fig. 2 Disentanglement fluctuation of the IB-based approaches. AnnealedVAE and CascadeVAEC could degenerate into lower disentanglement scores

to disentangle factors theoretically, and the inductive biases on the data and the model should be explored explicitly (Locatello et al., 2019).

2.2 Disentanglement evaluation

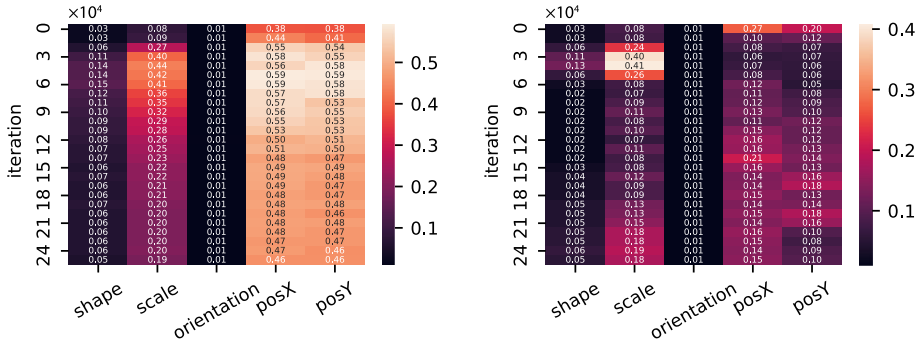
Several metrics have been proposed to evaluate the disentanglement, including the Beta-VAE metric (Higgins et al., 2017a), FactorVAE metric (Kim and Mnih, 2018), MI gap (Chen et al., 2018), modularity (Ridgeway and Mozer, 2018), DCI (Eastwood and Williams, 2018), and SAP score (Kumar et al., 2018). Shannon MI is an information-theoretic quantity that measures the amount of information shared between two variables. Based on that, the MIG (Chen et al., 2018) measures the gap between the top two latent variables with the highest MI to evaluate the performance of disentanglement:

$$\text{MIG} = \frac{1}{\|c\|} \sum_{i=1}^{\|c\|} \text{NMI}(c_i, 1) - \text{NMI}(c_i, 2), \tag{6}$$

where $\text{NMI}(c_k, m)$ is the m-th largest normalized MI (NMI) between z_j and c_k. The calculation can be:

$$\text{NMI}(c_k, m) = \frac{1}{H(c_k)} I(z_{j^m}; c_k), \tag{7}$$

where z is the vector of latent variables, c is the vector of ground-truth factors, and j^m denotes the index of the m-th largest element ($j^1 = \arg\max_i I(z_i; c_k)$). $\text{NMI}(c_k, 1)$ measures how best one variable can learn for the factor c_k, and $\text{NMI}(c_k, 2)$ indicates the diffused information into other variables. Therefore, the gap of $\text{NMI}(c_k, 1)$ and $\text{NMI}(c_k, 2)$ should be large for the disentanglement.



(a) NMI$(c, 1)$ **(b)** NMI$(c, 2)$

Fig. 3 The change of NMI over training process on dSprites. The NMIs on many factors decrease gradually after the largest values have been captured at the early stage, especially on the factor scale

3 Motivation

Locatello et al. conducted a survey of current disentanglement approaches, and the results show that these approaches have a high variance of disentanglement scores. They concluded that "tuning hyperparameters matters more than the choice of the objective function" (See Fig. 7 in their paper). A reliable and robust approach should therefore have a consistently high performance and low variance. We investigated the performance of β-VAE ($\beta = 4$), β-TCVAE ($\beta = 6$), and AnnealedVAE ($C = 25$) on dSprites, and traced the disentanglement scores through the training processes. Fig. 2 shows the curves of three metrics (beta VAE metric, MIG, and DCI disentanglement) for four models (β-VAE, β-TCVAE, CascadeVAEC, and AnnealedVAE). AnnealedVAE, CascadeVAEC, and β-TCVAE show significant improvements in the very first iteration. However, CascadeVAEC has a sharp decrement in the 10, 000 iteration, and AnnealedVAE shows a downward trend after 10, 000 iteration. The training process did not consistently enhance the model being disentangled, resulting in poor performance.

One solution to address fluctuation is to block some information by using a narrow information bottleneck and then assign the increased information to a new latent variable by increasing the bottleneck. AnnealedVAE and CascadeVAEC follow this concept; however, they differ in terms of expanding the IB. AnnealedVAE directly controls the capacity of the latent variables by an annealed increasing parameter, C. CascadeVAEC increases the capacity by relieving the pressure on the i-th latent variable at the i-th stage, opening the information flow. Ideally, these approaches that are based on IB should have a steady growth of disentanglement; however, they also show fluctuation.

A perfect disentangled representation should project one factor into one latent variable. In other words, the largest NMI (NMI$(c, 1)$) reaches the maximum 1, and the second largest NMI (NMI$(c, 2)$) is close to 0. Therefore, the decrement of NMI$(c, 1)$ implies that the information of one factor diffuses into another latent variable, which we define as information diffusion (ID). The representation can be said to re-entangle in the case of an ID.

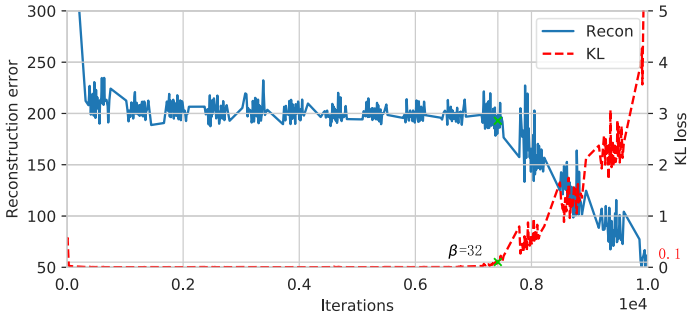


Fig. 4 Information freezing. The model starts to learn information at iteration 7500 ($\beta = 32$), where KL increases and the reconstruction error decreases

Though the final disentanglement score is desirable, it is insufficient to indicate the problems in the learning process. Monitoring of metrics probably clears the way to reveal the hidden problems on disentanglement during training the model. To do that, we monitored NMI($c, 1$) and NMI($c, 2$) during training with AnnealedVAE on dSprites (training details in Sect. 5.1), as shown in Fig. 3. We computed the NMIs for the five factors every 10, 000 iterations and presented them in one row. Ideally, the expanded capacity would promote the model to learn new information. Oppositely, NMI($c, 1$) (scale) decreased after 5e4 iterations. AnnealedVAE suffered the ID problem, which caused the low performance.

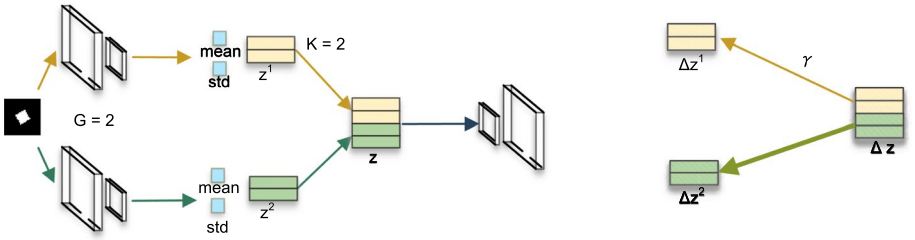
4 Method

4.1 Information freezing

Burgess et al. proposed that the value of beta in β-VAE controls the IB between inputs and latent variables, similar to the role of temperature in distillation; a low value of beta encourages the MI $I(x;z)$, and more information condenses on the latent space. The IFP is a critical point at which the model starts to learn information from observations. It is an intrinsic property of a dataset and almost invariant. Thus, different factors can be identified by IFPs.

Definition 1 The IFP is the maximum value of β, such that $I(x;z) > 0$ for the β-VAE objective.

We introduce *the annealing test to determine the IFP for a given dataset*. The objective of the annealing test is the same as that of β-VAE, except that it uses an annealing β from a high value to 1 (i.e., it starts with value 200 and ends with value 1). While the pressure of the KL term decays, there exists a critical point where $I(x;z)$ increases and the reconstruction error decreases. For example, we trained the model with an annealing β from 200 to 0 in 100,000 iterations in Fig. 4. One can see that the IFP is approximately 32 at iteration



(a) Forward **(b)** Backward

Fig. 5 Illustration of DEFT with two sub-encoders (G = 2), and each sub-encoder has K = 2 latent variables. DEFT has isolated sub-encoders and scales partial backward information

7400. Roughly, we regard the IFP as the value of beta where the model learns information ($I(x;z)$ is over 0.1).

4.2 DEFT

Distillation is the process of separating a mixture into its components by heating to an appropriate temperature, such that components boil and freeze into the target containers. Inspired by distillation in chemistry, this paper proposes a novel disentanglement approach based on β-VAE, which distills independent components into several isolated sub-encoders. There exists a suitable pressure on β that makes one component with high IFP being separated from the components with lower IFP. Therefore, an iterative algorithm is concluded to **D**istill (disentangle) the **E**ntangled **F**ac**T**or, named **DEFT**. Specifically, it splits the latent variables into G groups which per group has K latent variables, and there are G $\times$ K latent variables in total. The decoder takes the concatenation of latent variables of all groups as inputs, which is the same as the conventional decoder. DEFT also divides the training process into G stages, so that the model extracts components according to IFPs per stage by assigning a different β, β^i for the i-th stage. Apart from that, DEFT scales the gradients of the old sub-encoders to prevent the ID problem; that is, the backward gradients of the first $i-1$ sub-encoders are scaled by multiplying a scaling coefficient γ. Overall, the architecture of DEFT (G = 2, K = 2) at stage 2 is shown in Fig. 5. The forward part of DEFT takes one image as inputs to two isolated sub-encoders and then concatenates the outputs of two groups of sub-encoders into a four-dimensional vector which is fed to the decoder. The backward part of the DEFT scales the backward gradients for the old variables. For example, $\nabla z^1 = \gamma \nabla z_{1:2}, \nabla z^2 = \nabla z_{3:4}$ at stage 2. In addition, the algorithm of DEFT is shown in Algorithm 1, where $q_{\phi_i}(z^i|x)$ denotes one sub-encoder, $p_\theta(x|z)$ denotes the decoder, and $\mathcal{L}$ denotes the β-VAE objective.

DEFT chooses a suitable value of beta to separate factors, that act as like the temperature, such that the desired factor's information passes the bottleneck and freezes into the latent variables. Furthermore, backward information scaling is performed for these old variables to prevent the information from diffusing into others.

Table 1 Lite encoder, standard encoder, and decoder architecture for all experiments. For dSprites and SmallNORB, $c = 1$. For Color and Scream, $c = 3$

Lite Encoder	Standard Encoder	Decoder
4×4 conv. Eight Stride 2	4×4 conv. 32 stride 2	FC.256
4×4 conv. Eight Stride 2	4×4 conv. 32 stride 2	FC. $4 \times 4 \times 64$
4×4 conv. Sixteen stride 2	4×4 conv. 64 stride 2	4×4 upconv. 64 stride 2
4×4 conv. Sixteen stride 2	4×4 conv. 64 stride 2	4×4 upconv. 32 stride 2
FC. 64	FC. 256	4×4 upconv. 32 stride 2
FC. K	FC. $K \times G$	4×4 upconv. c stride 2

Algorithm 1: The algorithm of DEFT. We use default values of $\alpha = 0.0005,\ \beta_1 = 0,\ \beta_2 = 0.99$.

Input : The number of stages G, pressures on KL divergence $B = \{\beta^i\}_{i=1}^{G}$, the scaling coefficient of gradients for old sub-encoders γ, the number of training iterations per stage N, Adam hyperparameters α, β_1, β_2

Initialize $\theta, \{\phi_i\}_{i=1}^{G}$ in $p_\theta(x|z),\ \{q_{\phi_i}(z^i|x)\}_{i=1}^{G}$;
for $i \leftarrow 1$ **to** G **do**
  repeat
    Sample a batch of observations x;
    `/* the forward part */`
    for $j = 1$ **to** G **do** $\mu^j, \sigma^j \leftarrow q_{\phi_j}(x)$;
    $\mu, \sigma \leftarrow$ `Concatenate(`$\{\mu^j\}_{j=1}^{G}$`)`, `Concatenate(`$\{\sigma^j\}_{j=1}^{G}$`)`;
    Sample $z \sim \mathcal{N}[\mu, \sigma]$;
    $\tilde{x} \leftarrow p_\theta(z)$;
    $\mathcal{L} \leftarrow \log(p(\tilde{x})) - \sum_{j=1}^{i} \beta^i D_{\mathrm{KL}}(q_{\phi_j}(z^j|x)||p(z^j))$;
    `/* the backward part */`
    $\nabla\theta, \{\nabla\phi_j\}_{j=1}^{i} \leftarrow \nabla_{\theta, \{\nabla\phi_j\}_{j=1}^{i}} \mathcal{L}$;
    $\theta \leftarrow$ `Adam(`$\nabla\theta, \theta, \alpha, \beta_1, \beta_2$`)`;
    for $j \leftarrow 1$ **to** $i - 1$ **do** Scale the gradients
      $\phi_j \leftarrow$ `Adam(`$\gamma\nabla\phi_j, \phi_j, \alpha, \beta_1, \beta_2$`)`;
  until N *times*;

5 Experiment

5.1 Settings

In this work, there are two types (standard and lite) of sub-encoders and one type of decoder architecture, as shown in Table 1. For the encoder part, DEFT uses the lite architecture—the dimension of z is $K \times G$ in total; the other approaches use the standard architecture. All models use the same decoder architecture. All layers are activated by ReLU. The optimizer is Adam with a learning rate of 5e-4, $\beta_1 = 0$, $\beta_2 = 0.99$. The batch size is 256, which accelerates the training process.

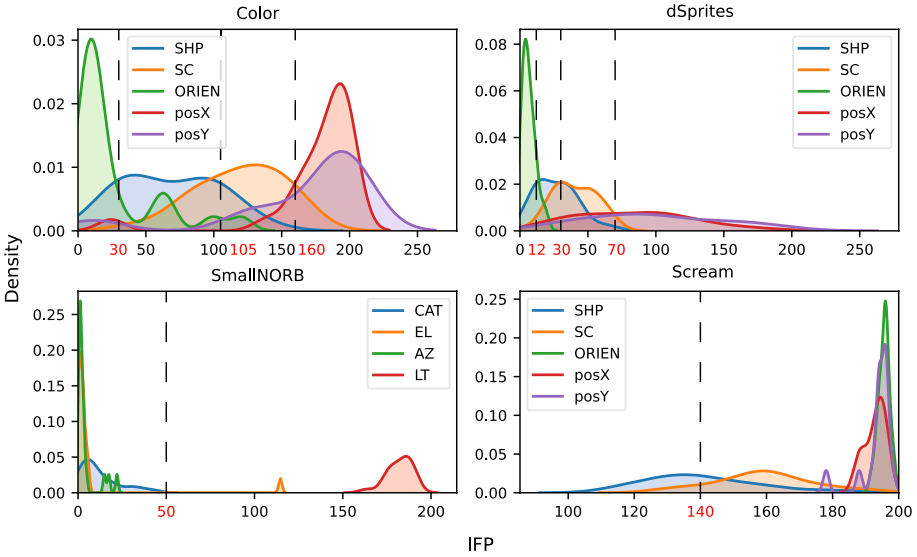


Fig. 6 The distribution of IFP on four datasets. The red number denotes the pressure required to separate these factors. There are four factors on SmallNORB—category: CAT, elevation: EL, azimuth: AZ, and lighting condition: LT. There are five factors on three variants of dSprites—shape: SHP, scale: SC, orientation: ORIEN, position X: posX, position Y: posY

5.2 Supervised problem

Dataset detail We compared DEFT with others on dSprites (Matthey et al., 2017), color dSprites (color for short), scream dSprites (scream for short) (Locatello et al., 2019), and SmallNORB (LeCun et al., 2004). The images of dSprites are strictly generated by the five factors. It has three shapes: square, ellipse, and heart; six scale values: 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0, 40 orientation values in [0, 2 pi], 32 position X values, and 32 position Y values. Two variants of dSprites (Color and Scream), which introduce random noise, were closer to the true situation. SmallNORB is generated from 3D objects and is much more complex than 2D shapes. It contains five generic categories, namely, four-legged animals, human figures, airplanes, trucks, and cars; nine elevation values, i.e., 30, 35, 40, 45, 50, 55, 60, 65, and 70; eighteen azimuth values, 0, 20, 40, ..., 340; and six lighting conditions.

Information freezing point The ideal situation is to find *a set of β to isolate IFPs into several parts without overlaps*. To obtain the distribution of IFPs with respect to a factor c_i, we enumerate all possible values of factor c_i for a random sample, and calculate its IFP using the algorithm introduced in Sect. 4.1. Then, we repeated the above procedure 50 times to estimate the IFP distribution of c_i. We measured the IFPs of the factors on the four datasets, as shown in Fig. 6. dSprites and Color had more separable IFPs than Scream and SmallNORB. Although the three variants of dSprites have the same factors, their IFPs are different. The difference in IFP distributions explains why current approaches fail to transfer hyperparameters across different problems in Locatello et al. (2019). Note that the IFP distributions of factors are almost separable for dSprites and Color; the ground-truth factors are independent of the four datasets. In summary, four datasets are all independent; Scream and SmallNORB are inseparable. Based on the distribution of IFPs, we summarize

Table 2 Experimental settings for DEFT. γ is always 0.1, see in Sect. 5.6. The number of iterations per stage (N) is sufficiently large such that the objective converges. The number of latents per sub-encoder (K) is not less than the size of the newly learned factors. The number of sub-encoders (G) is determined by the number of separable areas in Fig. 6

	G	K	N	β^i
Color	4	3	20, 000	160,105,30,4
dSprites	4	3	20, 000	70,30,12,4
SmallNORB	2	5	40, 000	50,1
Scream	2	5	40, 000	140,1

Table 3 Experimental settings for the compared approaches

	Color	dSprites	Scream	SmallNORB
C (AnnealedVAE)	10	5	25	5
β (β-TCVAE)	10	10	6	1
β (β-VAE)	16	16	6	1
β_h (CascadeVAEC)	10	10	10	10

the optimal training settings for the DEFT in Table 2. We tune the hyperparameters of compared approaches with the highest MIG and show these settings in Table 3.

5.3 Performance

We trained each model 50 times and compared our model with the other four disentanglement approaches on dSprites, Color, Scream, and SmallNORB.

Disentanglement metric We show the performances of disentanglement metrics in Fig. 7. All approaches have a lower performance on Scream and SmallNORB, where the distributions of IFP are inseparable. β-VAE and β-TCVAE have similar performances on four datasets. CascadeVAEC shows high performances on three variants of dSprites but has high variances for most cases. DEFT outperforms others for most cases and has lower variances. A downside of DEFT is the reduction of the searching space for more possible solutions, better or worse. As a result, DEFT has lower performances for the best models. The distributions of MIG at different stages are shown in Fig. 8. All experimental results on the four datasets reveal that DEFT obtains low scores at the first stage and gradually improves disentanglement in the following stages.

Reconstruction quality We also show the distributions of the reconstruction error in Fig. 9. CascadeVAEC and DEFT generally have higher qualities on rebuild images. Note that, though CascadeVAEC beats DEFT in some cases (dSprites and SmallNORB), the improved values are negligible compared with the overall errors (10% for dSprites, 2% for SmallNORB), and the differences are merely indistinguishable to human eyes. In general, DEFT reduces the variance by blocking partial information and achieves both a high image quality and disentanglement.

Failure rate We define the failure rate as the percentage of models that fail to learn a disentangled representation if the MIG score is lower than 0.1. Table 4 shows the failure rate. It can be seen that DEFT has the lowest average failure rates. Although AnnealedVAE

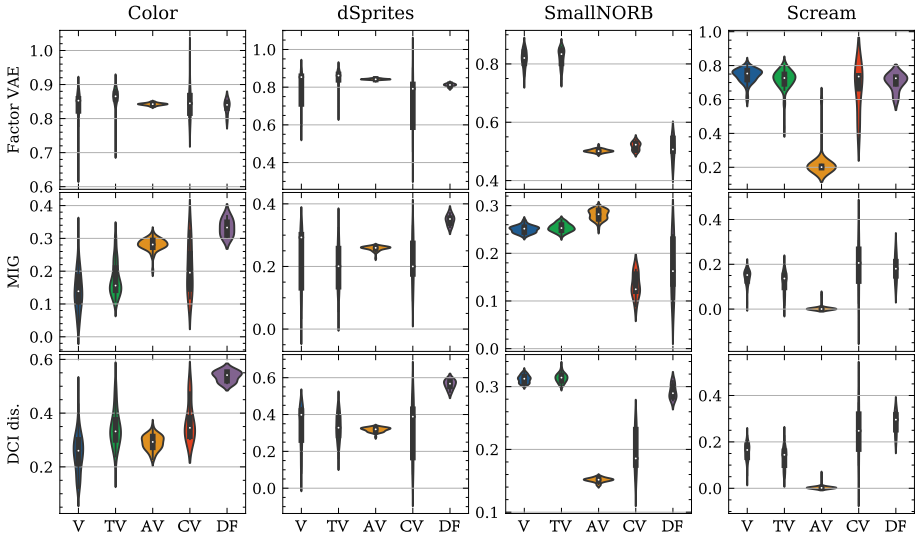


Fig. 7 The performances of three disentanglement metrics (Factor VAE (Kim and Mnih, 2018), MIG (Chen et al., 2018), DCI dis. (Eastwood and Williams, 2018)) for five approaches (V=β-VAE, TV=β-TCVAE, AV=AnnealedVAE, CV=CascadeVAEC, DF=DEFT) on four datasets (Color, dSprites, SmallNORB, Scream)

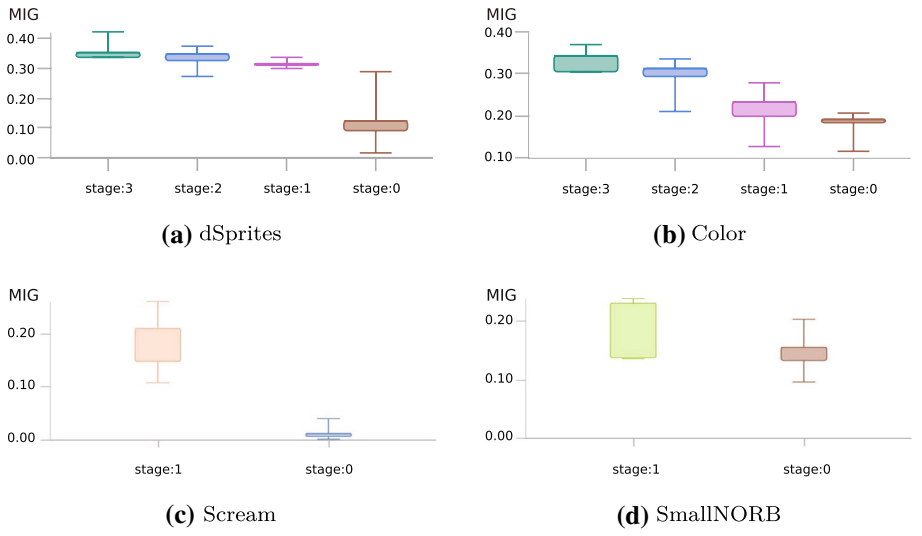


Fig. 8 MIG distribution of DEFT on four datasets for different stages

success to disentangle factors on three datasets, it fails to disentangle factors on Scream for most cases. Note that it is possible to reduce the failure rate for AnnealedVAE on Scream, but we have tried six settings, and none of them outperforms on all datasets. Overall, the failure rates of DEFT are lower than others. From the IFP distributions in Fig. 6, we can see that SmallNORB has a separable factor that is easy to be disentangled. That causes

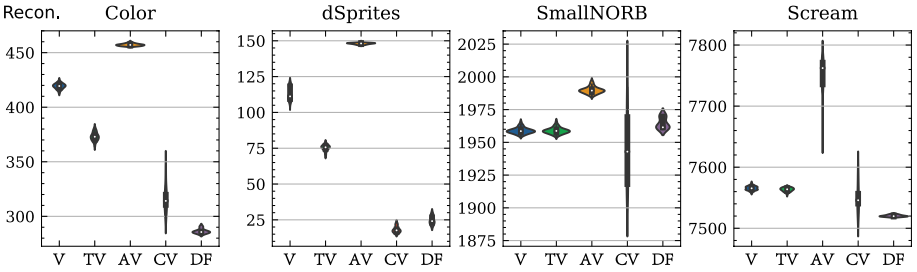
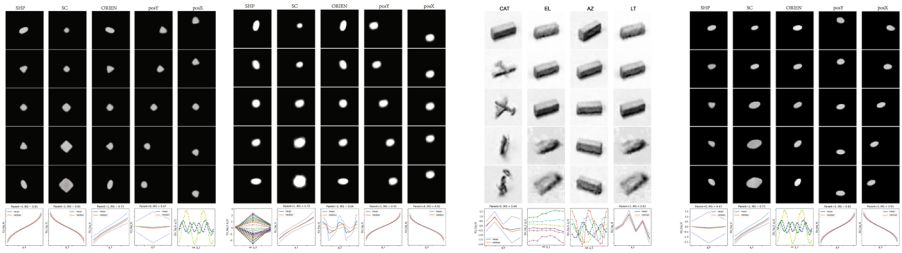


Fig. 9 Reconstruction error for different approaches and datasets. Five approaches respectively denote V=β-VAE, TV=β-TCVAE, AV=AnnealedVAE, CV=CascadeVAEC, DF=DEFT

Table 4 Failure rate (%) for each approach (column) and dataset (row)

	DEFT	β-VAE	β-TCVAE	AnnealedVAE	CascadeVAEC
Color	0	24	0	0	8
dSprites	8	16	2	0	0
SmallNORB	0	0	0	0	10
Scream	12	12	26	80	25

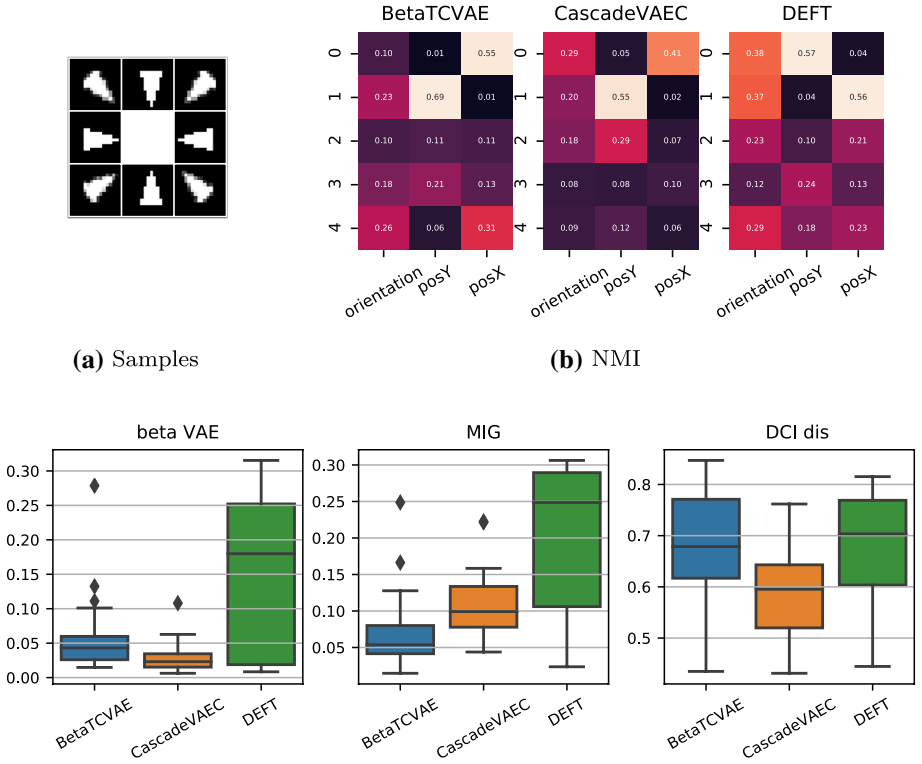


(a) Color (0.39) **(b)** dSprites (0.41) **(c)** SmallNORB (0.31) **(d)** Scream (0.29)

Fig. 10 Latent traversal of DEFT on four datasets (MIG score). Each column shows the images of traversing a latent variablez_i representing a factor and its VIR (last row) (Suter et al. 2019). We choose the variable having the highest MI for the factor. The same variable has the highest MI for both elevation and light condition

SmallNORB to have a high lower bound of disentanglement but get a low overall score. Generally, DEFT significantly decreased the failure rate compared to the other approaches.

Visualization Higgins et al. (2017a) introduced the latent traversal to visualize the generated images through the traversal of a single latent z_i. Fig. 10 shows the latent traversal of the best model with the highest MIG score. One can see the intrinsic relationship between IFP and disentanglement. Orientation has the lowest IFP among all factors; meanwhile, it is the hardest one to be disentangled for all approaches. For SmallNORB, the lighting condition is separable with others, which is easy to be disentangled. For Scream, three factors have similar IFP distributions, and it is also a hard problem for the disentanglement approaches.



(a) Samples **(b)** NMI

(c) Disentanglement score distribution for different approaches.

Fig. 11 **a** Dataset visualization. **b** NMI matrix $I(z_i;c_j)$ for three approaches. **c** Disentanglement scores for different approaches

5.4 Correlative but separable

To demonstrate the superiority of the IB approaches, we built a dataset of a triangle with three factors (posX, posY, and orientation), where posX and posY are independent, and the triangle always points to the center of the canvas $\theta = \arctan(\text{posY} - 16, \text{posX} - 16)$. Figure 11a shows the samples from this toy dataset. We trained CascadeVAEC, β-TCVAE ($\beta = 6$), and DEFT ($\text{K} = 2, \text{G} = 2$) within 10,000 steps and repeated 10 times. From Fig. 11 (b), all three approaches disentangle posX and posY successfully. However, only DEFT extracts orientation information ($I(z_4;\text{orientation})$ is high, $I(z_4;\text{posX})$ and $I(z_4;\text{posY})$ are low). DEFT has higher disentanglement scores for all three metrics, as shown in Fig. 11c. The latent traversal in Fig. 12 shows that DEFT has a high image quality and separated orientation information. The correlation makes it difficult for β-VAE to disentangle orientation.

5.5 Unsupervised problem

3D Chairs (Aubry et al., 2014) is an unlabeled dataset containing 1394 3D models from the Internet.

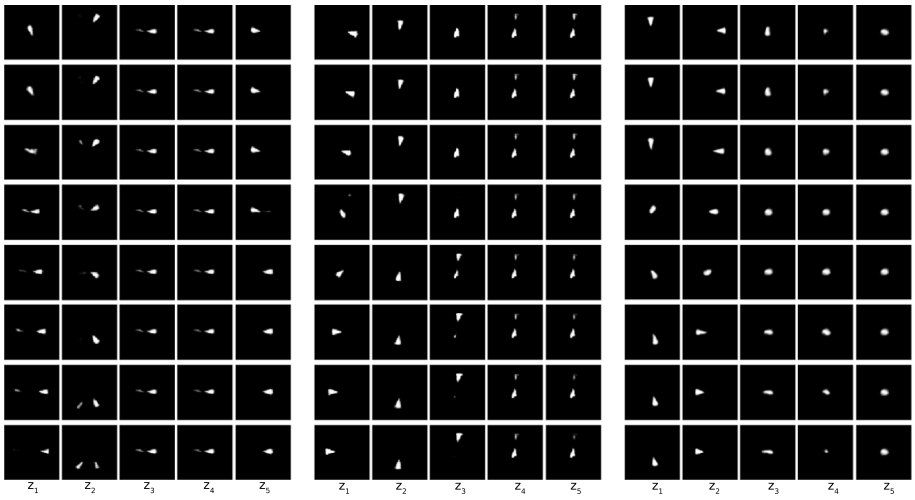


Fig. 12 Latent traversal of CascadeVAEC, β-TCVAE and DEFT on the separable but correlative dataset. Each column denotes the rebuild images by traversing the variable from -2 to 2

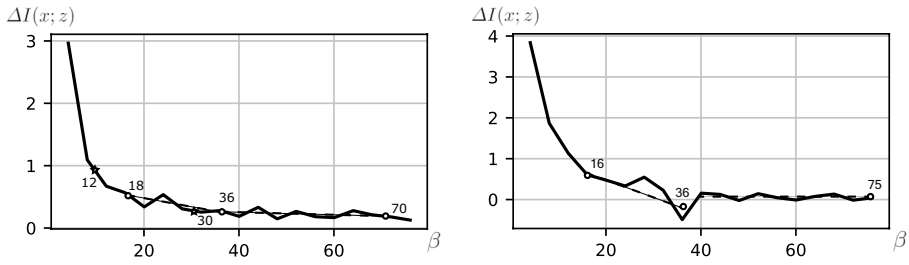
Annealing test without supervision The label information is unavailable for common situations. Therefore, the factor's IFP distribution is hard to be obtained. Alternatively, we calculate the upper bound of IFP distribution for the unsupervised setting. Intuitively, the rate of information increment changes if there is a new factor starting to freeze. We conducted an annealing test on dSprites and 3D chairs without labels and plotted the curves of beta vs. $\Delta I(x;z)$ in Fig. 13. This method is in agreement with the upper bound of the IFP distribution for position and scaling, as shown in Fig. 13a. One can recognize four points where the latent information suddenly increases: 36 and 16 from Fig. 13b. Though this method needs human participation, we only show the potency to develop a fully unsupervised procedure for the separations. Therefore, we set $\mathrm{G}=3, \mathrm{K}=3, \beta_j=\{36,16,1\}$ for 3D Chairs and trained the DEFT 20 epochs per stage. We compared the performance with β-TCVAE and CascadeVAEC on 3D Chairs, as shown in Fig. 14. We notice that DEFT can learn one additional interpretable property compared with CascadeVAEC— leg orientation.

5.6 Analysis

We introduce the following metrics to evaluate the problems on disentanglement during training in detail:

$$\mathrm{NMI1}=\frac{1}{\|c\|}\sum_{i=1}^{\|c\|}\mathrm{NMI}(c_i,1),\ \mathrm{NMI2}=\frac{1}{\|c\|}\sum_{i=1}^{\|c\|}\mathrm{NMI}(c_i,2). \tag{8}$$

NMI1 denotes the major information representing the factors, which should be as large as possible (1 at maximal). In contrast, NMI2 indicates the diffused information from the major latent variables, which should be as small as possible (0 at minimal).



(a) dSprites **(b)** 3D Chairs

Fig. 13 Information increment variability. The broken line denotes the tendency of the growth increment of mutual information. The dot denotes the mutation point of the mutual information increment. The star point denotes the selective separation of the IFP distributions

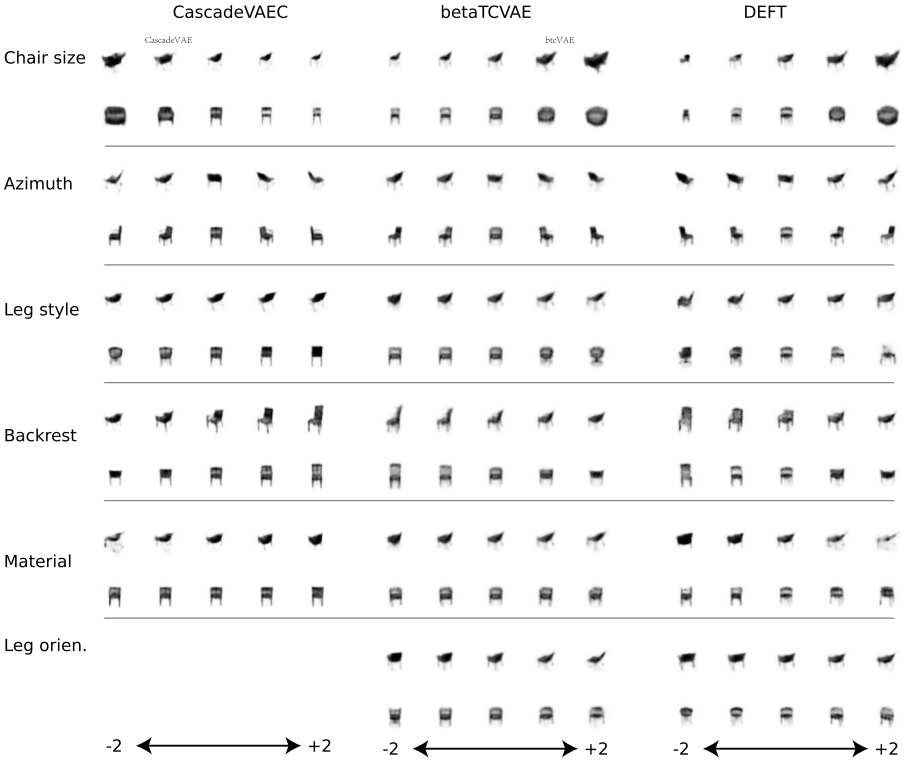
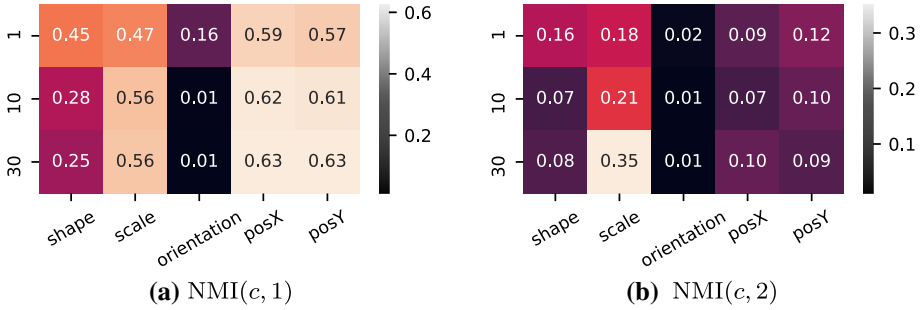


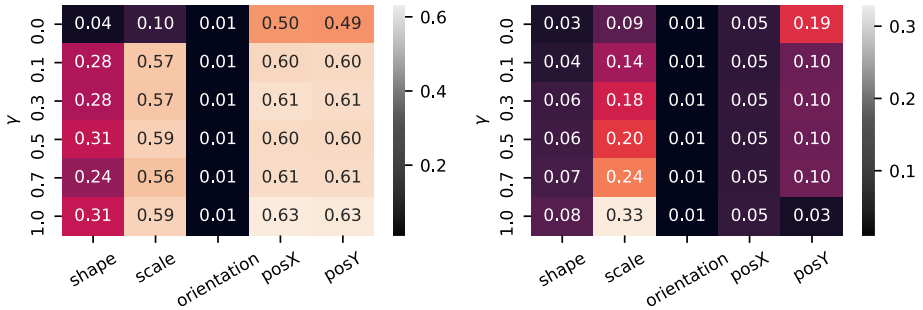
Fig. 14 Latent traversal on 3D Chairs. Each row shows the rebuild images by traversing the corresponding variable from -2 to 2. We show two samples for each factor separated by a line

To analyze the effects of techniques applied in DEFT, a simple model with only two stages is examined on dSprites. Experiments use the same setting at the first stage and apply specific settings for different purposes. At the first stage, the model with $\beta^1 = 70$ was trained 15, 000 iterations so that the model could only learn a disentangled representation of posX and posY according to the IFP distribution in Fig. 6.



(a) NMI$(c, 1)$ **(b)** NMI$(c, 2)$

Fig. 15 Each row shows the NMI$(c, 1)$ or NMI$(c, 2)$ in an independent trail with different values of β



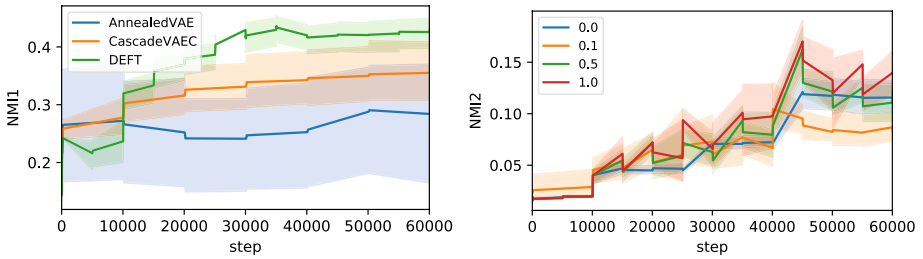
(a) NMI$(c, 1)$ **(b)** NMI$(c, 2)$

Fig. 16 Each row shows the NMI$(c, 1)$ or NMI$(c, 2)$ in an independent trail with different values of γ

Piecewise pressure At the second stage, the model was trained 15, 000 iterations with different values of β^2. The experimental results in Fig. 15 show that a lower β^2 helps the model to learn the insignificant factors with small IFPs (shape and orientation) but violates the model to improve the factors with high IFPs (scale, posX, and posY). β^2 plays the role of a valve for passing information, and impure information is harmful to the disentanglement. Note that, increasing the β^2 brings the problem of the larger NMI$(c, 2)$, which is incapable of improving the disentanglement solely.

Backward information scaling At the second stage, we train the model with $\beta^2 = 30$ for 15, 000 iterations across different values of γ. As shown in Fig. 16, the diffused information (scale) is descending as reducing γ, relieving the ID problem. However, NMI$(c, 1)$ reaches the lowest value when all backward information is clipped $\gamma = 0$, violating the model to extract new factors. NMI$(c, 2)$ and γ are simultaneously increased; A small value of γ is sufficient to learn the majority information, and it also prevents information from diffusing into another variable. In conclusion, β controls the passing information, and a large one is used to generate pure information; γ retards the increment of NMI2; the disentanglement can benefit from both two techniques by relieving the ID problem.

Comparison To see the overall effects of DEFT, we compare DEFT with AnnealedVAE and CascadeVAEC on dSprites 10 times, see details in Table 2. Note that we use a standard DEFT in this part. From Fig. 17a, one can see that there is a declination of NMI1 during iteration 1, 000 to 3, 000 for AnnealedVAE and an overall low level of NMI1 for CascadeVAEC. DEFT shows a steady improvement and a high level on NMI1. We also show the



(a) Model **(b)** γ

Fig. 17 Comparison of three models and four values of γ

Table 5 The computational cost (second) for the normal encoder and the fractional encoder

G	10	100	500	1, 000
Normal	0.044 ± 0.002	0.058 ± 0.003	0.121 ± 0.034	0.225 ± 0.069
Fractional	0.046 ± 0.004	0.062 ± 0.004	0.314 ± 0.026	0.622 ± 0.057

NMI2 with four values of γ: 0, 0.1, 0.5, 1, in Fig. 17b. The curves with error regions from 10 trails demonstrate that $\gamma = 0.1$ achieves a lower NMI2.

5.7 Complexity

The difference between DEFT and other approaches is mainly in the encoder part: DEFT uses a fractional encoder that has several sub-encoders. We assume that $\boldsymbol{W}_{(\mathrm{G}\times\mathrm{K})\times\mathrm{M}}$ represents the parameters of a normal encoder, and $\boldsymbol{W}^i_{\mathrm{K}\times\mathrm{M}}$ represents the parameters of a sub-encoder in a fractional encoder, where M is the dimension of inputs. The computational costs for both should be the same ideally, $J(\boldsymbol{W}x) = \sum_{i=1}^{\mathrm{G}} J(\boldsymbol{W}^i x)$. However, there are some extra operations, such as the iterative loop and the concatenation of the latent variables. To make a fair comparison, we set the dimension of latent variables to 1, 000 and change the number of channels in convolutional layers so that the total parameters of the fractional encoder and the normal one are equal. Each trail generates a batch of samples (256) randomly and then runs a forward and a backward process. Table 5 shows the mean and the standard deviation of runtime (second) for 100 trails. Overall, the increased cost of the fractional encoder is only about 6.9% for G = 100, which is acceptable in practice. The extra computational cost is acceptable for the common disentanglement tasks, which usually have less than ten ground-truth factors.

6 Conclusion

Based on existing studies involving IBs, we have developed new insights into the reason for which these approaches have lower performances than the TC-based ones. In particular, we identified the IFP distribution for each factor by performing an annealing test, and a dataset was easily disentangled if the IFP distributions were separable. Furthermore, we

found that the ID problem is an invisible hurdle that prevents steady improvements in disentanglement. We proposed DEFT to retain the learned information by blocking partial information. In addition, scaling the backward information is also helpful to relieve the ID problem. Our results show that approaches that are based on IBs are competitive and have the potential to solve problems with correlative factors.

We verified the ID problem that causes the low performance of IB-based approaches. However, as a plain solution, the DEFT method still needs to be further improved. In the future, an automatic way to adjust the best separation of IFP distribution is highly required.

Author Contributions Jiantao conceived of the presented idea, carried out the experiments, and wrote the manuscript with support from Lin, Bo, Chunxiuzi, and Jin. Lin encouraged Jiantao to investigate the ID problem and supervised the findings of this work. Lin, Bo, and Jin provided funding supports for this project. Fanqi processed the experimental data and designed the figures. Chunxiuzi also provided constructive advice for improving the manuscript and the experimental design. All authors provided critical feedback and helped shape the research, analysis, and manuscript. We confirm that all authors agree with the results and contributed to the final manuscript.

Funding This work was supported by National Natural Science Foundation of China under Grant No. 61872419, No. 62072213, No. 61873324, No. 61903156. Shandong Provincial Natural Science Foundation No. ZR2020KF006, No. ZR2019MF040, No. ZR2018LF005. Taishan Scholars Program of Shandong Province, China, under Grant No. tsqn201812077. "New 20 Rules for University" Program of Jinan City under Grant No. 2021GXRC077

Data availability We confirm that all data are openly available in public repositories. Specifically, dSprites is included in Matthey et al. (2017); Color and Scream are included in Locatello et al. (2019); SmallNORB is included in LeCun et al. (2004).

Declarations

Conflicts of interest We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Ethics approval Not Applicable.

Consent to participate We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

Consent for publication We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Code availability Our codes and all experimental settings are published in DistillationVAE, depending on dlib for PyTorch.

References

Aubry, M., Maturana, D., Efros, A.A., Russell, B.C., & Sivic, J. (2014). Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of CAD models. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014* (pp 3762–3769). IEEE Computer Society.

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

Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2017). Understanding disentangling in β-vae. In *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems 2017, NeurIPS 2017, December 4–9, 2017, Long Beach, CA, USA*.

Chen, T.Q., Li, X., Grosse, R.B., & Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp 2615–2625).

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain* (pp 2172–2180).

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*(3), 287–314.

Do, K., & Tran, T. (2020). Theory and evaluation metrics for learning disentangled representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/

Dupont, E. (2018). Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp 708–718).

Eastwood, C., & Williams, C.K.I. (2018). A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings*. https://openreview.net/

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017a). beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. https://openreview.net/

Higgins, I., Pal, A., Rusu, AA., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., & Lerchner, A. (2017b). DARLA: Improving zero-shot transfer in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, PMLR* (Vol. 70, pp 1480–1490).

Higgins, I., Amos, D., Pfau, D., Racanière, S., Matthey, L., Rezende, D.J., & Lerchner, A. (2018a). Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230

Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C.P., Bosnjak, M., Shanahan, M., Botvinick, M., Hassabis, D., & Lerchner, A. (2018b). SCAN: Learning hierarchical compositional visual concepts. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings*. https://openreview.net/

Jeon, I., Lee, W., Pyeon, M., & Kim, G. (2021). Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks. In *Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence(AAAI), ASSOC Advancement Artificial Intelligence* (pp 7926–7934).

Jeong, Y., & Song, HO. (2019). Learning discrete and continuous factors of data via alternating disentanglement. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, PMLR* (Vol. 97, pp. 3091–3099).

Kim, H., & Mnih, A. (2018). Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, PMLR* (Vol. 80, pp. 2654–2663).

Kingma, D.P., & Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014*. Conference Track Proceedings.

Kumar, A., Sattigeri, P., & Balakrishnan, A. (2018). Variational inference of disentangled latent concepts from unlabeled observations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings*. https://openreview.net/

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., & Ranzato, M. (2017). Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA* (pp 5967–5976).

LeCun, Y., Huang, F.J., & Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June–2 July 2004, Washington, DC, USA*. IEEE Computer Society.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, PMLR* (Vol. 97, pp 4114–4124).

Matthey, L., Higgins, I., Hassabis, D., & Lerchner, A. (2017). dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/

Ridgeway, K., & Mozer, M.C. (2018). Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp 185–194).

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

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J.M. (2012). On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26–July 1, 2012*. https://icml.cc/

Sorrenson, P., Rother, C., & Köthe, U. (2020). Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. https://openreview.net/

Suter, R., Miladinovic, Đ., Schölkopf, B., & Bauer, S. (2019). Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, PMLR* (Vol. 97, pp. 6056–6065).

Tenenbaum, J. (2018). Building machines that learn and think like people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10–15* (p. 5).

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., Bauer, S. (2021). On disentangled representations learned from correlated data. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event, PMLR* (Vol. 139, pp. 10401–10412).

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development, 4,* 66–82.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 2*(1–3), 37–52.

Zhu, Y., Xie, J., Liu, B., Elgammal, A. (2019). Learning feature-to-feature translator by alternating backpropagation for generative zero-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019* (pp. 9843–9853). IEEE.