Closed-Loop Unsupervised Representation Disentanglement with β-VAE Distillation and Diffusion Probabilistic Feedback

Xin Jin[1] *, Bohan Li[1,2], Baao Xie[1],
Wenyao Zhang[1,2], Jinming Liu[1,2], Ziqiang Li[1,2], Tao Yang[3], and Wenjun Zeng[1]

[1] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[2] Shanghai Jiao Tong University, Shanghai, China
[3] Xi'an Jiaotong University, Xi'an, China
{jinxin, wenjunzengvp, bxie}@eitech.edu.cn,
{bohan_li, wy_zhang, jmliu206, ziqiangli}@sjtu.edu.cn,
yt14212@stu.xjtu.edu.cn

Abstract. Representation disentanglement may help AI fundamentally understand the real world and thus benefit both discrimination and generation tasks. It currently has at least three unresolved core issues: (i) heavy reliance on label annotation and synthetic data — causing poor generalization on natural scenarios; (ii) heuristic/hand-craft disentangling constraints make it hard to adaptively achieve an optimal training trade-off; (iii) lacking reasonable evaluation metric, especially for the real label-free data. To address these challenges, we propose a **C**losed-**L**oop unsupervised representation **Dis**entanglement approach dubbed **CL-Dis**. Specifically, we use diffusion-based autoencoder (Diff-AE) as a backbone while resorting to β-VAE as a co-pilot to extract semantically disentangled representations. The strong generation ability of diffusion model and the good disentanglement ability of VAE model are complementary. To strengthen disentangling, VAE-latent distillation and diffusion-wise feedback are interconnected in a closed-loop system for a further mutual promotion. Then, a self-supervised **Navigation** strategy is introduced to identify interpretable semantic directions in the disentangled latent space. Finally, a new metric based on content tracking is designed to evaluate the disentanglement effect. Experiments demonstrate the superiority of CL-Dis on applications like real image manipulation and visual analysis.

1 Introduction

Disentangled representation learning (DRL) [29] learns the underlying explainable factors behind the observed data, where the different latent factors correspond to different properties, respectively. This matches the way humans perceive this world that we understand an object from its various properties (e.g.,

* Corresponding author

shape, size, color, etc.) [2, 20, 21, 58, 88]. Thus, DRL is thought to be one of the possible ways for AI to understand the world fundamentally, thus achieving Artificial General Intelligence (AGI) [3, 46, 79].

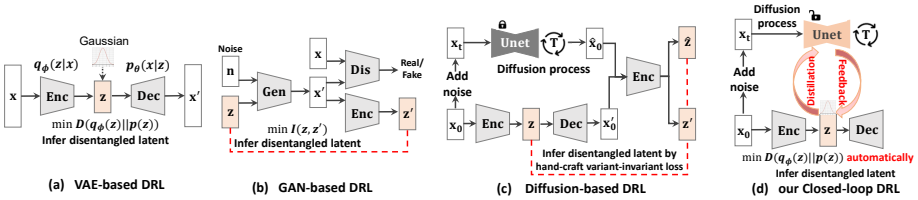


Fig. 1: The illustration of disentangled representation learning (DRL) frameworks: (a) VAE-based, (b) GAN-based, (c) Diffusion-based, and (d) our proposed Closed-loop approach. We can intuitively see that previous works all relied on heuristic hand-craft loss constraints to infer disentangled latent, only our method leverages two kinds of generative branches to build a cycle system for mutually-promoting automatically disentangled learning.

Most existing DRL methods learn the disentangled representation based on generative models, such as VAE [4, 8, 39, 42, 66], GAN [9, 24, 51, 74, 94], and even recently popular diffusion [45,82,83,86,90]. Among them, VAE-based DRL methods always encourage the disentanglement of the latent variables by forcing the variational posterior to be closer to the factorized Gaussian prior (Figure 1(a)), but they generally need to balance a hard inherent trade-off between the disentangling ability and generating quality [8, 30, 39]. As shown in Figure 1(b), the GAN-based DRL methods tend to use the mutual information as disentanglement constraints, but they typically suffer from the problem of reconstruction quality due to the gan-inversion difficulty [78,90]. Besides, VAE- and GAN-based generative models are prone to learn the factorized or conditioned statistics bias [32], which essentially conflicts with the ultimate purpose of a generalizable disentanglement that makes AI understand the world like humans.

Moreover, given the recently popular diffusion probabilistic models (DPMs) show superiority in image generation quality [31, 70] and unsupervised controlling (i.e., learning meaningful representations) of images [62], researchers have started to investigate the possibility of DRL in DPMs. Asyrp [45] proposes an asymmetric reverse process to discover the semantic latent space in frozen pretrained diffusion models, which reveals that DPMs already have a semantic latent space and thus more suitable for learning disentangled semantics. The co-current work of DisDiff [90] shown in Figure 1(c) uses the gradient field of frozen DPMs to achieve a vector-wise disentangled representation learning. However, this work applies an external heuristic hand-craft variant-invariant loss [54, 90] to enforce the disentanglement processes, resulting in reduced model flexibility and posing challenges for optimization.

In this paper, we propose to leverage two kinds of generative branches of the diffusion model and VAE model, to build a cycle system for mutually-promoting DRL. As shown in Figure 1(d)), we construct an unsupervised **C**losed-**L**oop **Dis**entanglement framework dubbed **CL-Dis**, which is built upon a diffusion-based autoencoder (Diff-AE) backbone while combining with a co-pilot of β-VAE, to together learn the meaningful disentangled representations. To col-

laborate two branches well and facilitate feature disentangling, we encourage Diff-AE and β-VAE to **complement each other, achieving a win-win effect:** we use the pre-trained latent of VAE to guide a semantic-aware reverse diffusion process with a distillation loss. Then, the gradually increasing information capability during the diffusion process is taken as feedback, in turn, to progressively enforce the inner latent of VAE more disentangling. Next, a self-supervised Navigation strategy is introduced to clearly identify each factor's semantic meanings, where shifting along a disentangled semantic direction will result in *continuously changed generations*. Finally, we further design a label-free metric based on **"changes tracking"** to quantitatively measure the disentanglement of learned features, wherein we use optical flow to reflect image variation degree objectively. Our main contributions are summarized:

- We build a new unsupervised representation disentanglement framework CL-Dis, which has a mutually-promoting closed-loop architecture driven by diffusion and VAE models. They complement each other with a distillation loss and a new feedback loss, and thus empower a more controllable and stronger DRL capability.
- We introduce a self-supervised method in CL-Dis to clearly identify each factor's semantic meanings by navigating directions in the learned disentangled latent space, leading to coherent generated variations. This makes the generation more explainable and fine-controllable.
- We design a new label-free metric to quantitatively measure the interpretability and disentanglement of learned features, and provide an example of immediate practical benefit from our work. Namely, we experimentally show how to exploit our CL-Dis for generative manipulation of real images and visual discrimination tasks.

2 Related Works

2.1 Disentangled Representation Learning (DRL)

The concept of DRL was introduced by Bengio et al. [3] in 2013 and was believed helpful in different tasks in practical applications like image generation [8, 9, 39, 40, 42, 44, 51, 84], image editing/translation [23, 47, 53, 80], NLP [11, 27, 81] and multimodal applications [33, 56, 75, 87, 93]. Different from image generation or editing, the core of this task is to fully understand the latent factors of the model to enhance its fine-grained controllability [8, 39, 40, 44, 84].

Most VAE-based DRL approaches like β-VAE [30], DIP-VAE [44], Factor-VAE [39], etc., achieve unsupervised disentanglement by the direct constraints on probabilistic distributions [8, 30, 39]. While straightforward, Locatello et al. [54] point out that is not enough and emphasize the need for extra inductive bias. Except for VAE, there are also some DRL works based on GAN, e.g., leveraging mutual information [9], self-supervised contrastive regularizer [51], spatial constriction [94], and combining with other pre-trained generative models [49, 63].

Fast-developing generative diffusion probabilistic models have accelerated the DRL exploration. LCG-DM [41] uses a vanilla VAE in DPMs to extract semantics codes for controllable generation, but ignores the interaction between these two structures. However, they uses gradient fields as an inductive bias at all time steps to achieve DRL, which relies on an external heuristic hand-craft disentangling loss, lacking a self-driven learning ability and is hard to optimize. In contrast, our CL-Dis builds a closed-loop DRL framework with a mutual-promoting mechanism driven by a two-branch complementary diffusion and β-VAE architecture, encouraging stronger disentangled feature auto-learning.

2.2 Knowledge Distillation and Feature Feedback

Our closed-loop CL-Dis is bridged by these two techniques. Knowledge distillation (KD) is originally developed for model compression and acceleration [12,25, 77], or used in transfer learning [1,35]. Feature feedback [15,16,38,55,61] enhances the robustness of learning systems by using explainable information, which is gaining more and more attention in RLHF [14, 43], and embodied AI [18, 64]. For instance, ControlVAE [66] achieves much better reconstruction quality with a variant of the proportional-integral-derivative (PID) control, using the output KL-divergence as feedback. Following this idea, we explore taking the strong generation capability of diffusion models as information-increasing feedback to automatically and adaptively enhance disentangled feature learning.

2.3 Latent Semantics Discovery (LSD)

This task emerges to move around the latent code (of GAN [13, 72, 76], Diffusion [45], etc), such that only one factor varies during the traversal generation. They mostly rely on human annotations (i.e., segmentation masks, attribute categories, 3D priors, and text descriptions) to define the semantic labels [10, 22, 50, 60, 67, 69]. Unsupervised LSD [68, 71, 76] often learn a set of directions or a classifier/matrix (e.g., Hessian, Jacobian regularization, etc.) to identify latent semantics. Wu et al. [82] find semantics in the stable diffusion model by partially changing the text embeddings. However, these methods are not strictly equivalent to DRL, most of them just rely on already disentangled (text) features to get coherent traversal generation results.

3 Background and Motivation

3.1 Understanding disentangling in β-VAE

Due to the poor disentanglement of vanilla VAE on complex data, β-VAE [30] adds explicit inductive bias to strengthen the independence constraint of the variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ with a β penalty coefficient in ELBO:

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

when β=1, β-VAE degenerates to the original VAE, and larger β encourages more disentangled representations but harms the performance of reconstruction. Unfortunately, it is practically intractable to obtain the optimal β to balance the trade-off between reconstruction quality and the disentangling capability [8, 44]. Thus, Burgess et al. [4] understand β-VAE from the perspective of information bottleneck theory, and propose improving disentangling in β-VAE with controlled information capacity increase:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - \beta \big| D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\big) - C \big|, \tag{2}$$

where C gradually increases from 0 to a value large enough during the training to guarantee the expressiveness of latent representations, thus improving disentangling. However, such a hand-craft increasing strategy is hard to control and might hurt disentangling performance, especially for various scenarios on different datasets. *Thus, in our work, we learn from this idea but upgrade it by exploiting the essence of the Diffusion Probabilistic Model (DPM) in automatically increasing information capacity [48, 59] for disentangling enhancement.*

3.2 DPM and Diffusion Autoencoder (Diff-AE)

Here, we first take DDPM [31] as a vanilla DPM example for illustration, which adopts a sequence of fixed variance distributions $q(x_t|x_{t-1})$ as the noise-adding forward process to collapse the image distribution $p(x_0)$ to $\mathcal{N}(0, I)$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{3}$$

The reverse process is fitting by other distributions of θ, and it is a progressively increasing process of information capacity (i.e., entropy decreasing) as we mentioned before:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_t; \mu_\theta(x_t, t), \sigma_t I), \tag{4}$$

where $\mu_\theta(x_t, t)$ is parameterize by a Unet $\epsilon_\theta(x_t, t)$. Its optimization is to minimize the variational upper bound of negative log-likelihood $\mathcal{L}_\theta = \mathbb{E}_{x_0,t,\epsilon} \|\epsilon - \epsilon_\theta(x_t, t)\|$.

Furthermore, Diff-AE [62] is proposed for representation learning based on DPM and Autoencoder. It uses a learnable **semantic encoder** for discovering the high-level meaningful semantics $\mathbf{z_{sem}} = E_{sem}(x_0)$, and a DPM-based **decoder** $p(x_{t-1}|x_t, \mathbf{z_{sem}})$ that is conditioned on $\mathbf{z_{sem}}$ for reconstruction. However, the latent representation $\mathbf{z_{sem}}$ learned by Diff-AE does not explicitly respond to the underlying factors of the data, or say, its disentangling potential has not been fully unleashed.

4 Methodology: Closed-Loop DRL (CL-Dis)

Based on the analysis above, we propose to leverage VAE and diffusion to unleash their strengths for a mutually promoting DRL. As shown in Figure 2, our CL-Dis

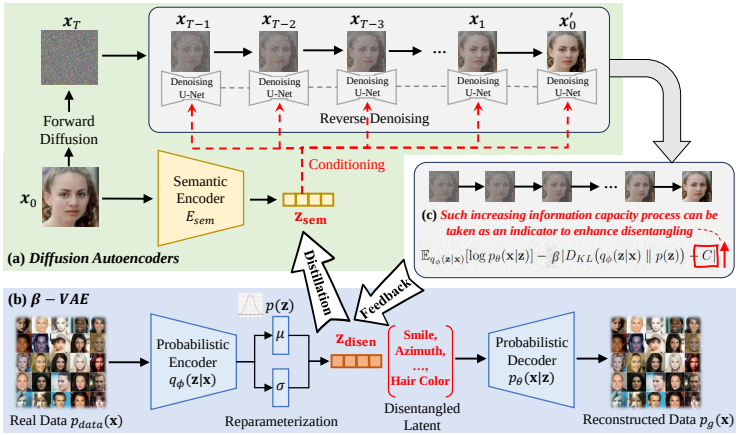


Fig. 2: Pipeline of the proposed CL-Dis. We explore using **(a) diffusion-based autoencoder (Diff-AE)** as a backbone while resorting to **(b)** β**-VAE** as a co-pilot via knowledge distillation to help E_{sem} extract a semantic disentangled representation $\mathbf{z}_{\mathbf{sem}}$ of an input image x_0 (which is taken as condition for diffusion-based generation). To further facilitate feature disentangling, we further take **(c) the reverse diffusion process as an increasing information capacity indicator feedback** to, in turn, enhance the disentanglement of feature $\mathbf{z}_{\mathbf{disen}}$ in β-VAE. As a result, Diff-AE and β-VAE are interconnected as a closed-loop cycle system with VAE-latent distillation and diffusion-wise feedback controlling, achieving a win-win effect and well-disentangled fine-controllable representation learning.

is a dual-branch cycle framework, involving multiple iterative phases with (a) diffusion autoencoder (Diff-AE), (b) β-VAE, and the core (c) knowledge distillation & feature feedback process. In Section 4.1, we present the formulation of closed-loop DRL and the overview of CL-Dis. Then, we introduce different phases of CL-Dis, including Diff-AE pre-training, β-VAE pre-training, and knowledge distillation & feature feedback in Section 4.2. Then, we propose a self-supervised Navigation strategy to identify disentangled semantic meanings and a new well-designed DRL metric (Section 4.3) for disentanglement measurement.

4.1 Problem Formulation and Overview of CL-Dis

Given a practical dataset $\mathcal{D}$ derived by N underlying factors $\mathcal{C} = \{1, 2, \ldots, N\}$, the target of disentanglement for our CL-Dis is to learn a semantic encoder E_{sem}, for each factor $c \in \mathcal{C}$, when varying its latent value, its reconstructed generation results will have the corresponding changes.

In Figure 2, our CL-Dis takes Diff-AE as the backbone, which has exactly a semantic encoder E_{sem} to encode the raw data x_0 into the semantic representations $\mathbf{z}_{\mathbf{sem}}$ for conditioning the diffusion reverse distributions $p(x_{t-1}|x_t, \mathbf{z}_{\mathbf{sem}})$. To empower E_{sem} disentangling capability, we additionally construct a β-VAE as co-pilot to help $\mathbf{z}_{\mathbf{sem}}$ become a semantic *disentangled* representation. Given a dataset $\mathcal{X}$ from a distribution $p_{data}(\mathbf{x})$, a standard β-VAE framework [30] is built and ϕ, θ parametrize the VAE encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and the decoder $p_\theta(\mathbf{x}|\mathbf{z})$, respectively. The prior $p(\mathbf{z})$ is typically set to the isotropic unit Gaussian $\mathcal{N}(0, 1)$, using a differentiable "reparametrization trick ($z = \mu + \sigma$)" [42]. With the objective of

Eq. 1, we get a preliminary disentangled representations $\mathbf{z_{disen}}$. Then, we use a distillation loss $\mathcal{L}_{dt}$ to transfer such disentangling capability to the semantic latent $\mathbf{z_{sem}}$ of Diff-AE,

$$\mathcal{L}_{dt} = D_{KL}(\ \mathbf{z_{sem}} \parallel \mathbf{z_{disen}}\ ) = \sum \mathbf{z_{sem}} \cdot \log(\frac{\mathbf{z_{sem}}}{\mathbf{z_{disen}}}), \tag{5}$$

As discussed before, gradually increasing the information capacity of the VAE latent with a controlled way could improve disentangling [4]. The reverse diffusion process of Diff-AE (Eq. 4) is exactly a procedure of information capability increasing. Thus, we take it as an indicator (see Figure 2(c)) in turn to enhance β-VAE disentangling by adaptively adjusting the objective of Eq. 2, where we change the hand-craft C to a variable that can automatically change in the optimization (Eq. 8). Such a mutually promoting mechanism forms a closed-loop optimization flow.

4.2 Training of the Whole CL-Dis Framework

Phase-1: Diff-AE and β-VAE Pre-training Following [62], the entire Diff-AE model is pre-trained on the various datasets [36,37,92] to cover more semantic factors in the latent space of $\mathbf{z_{sem}}$. But at this phase, $\mathbf{z_{sem}}$ is still far from well-disentangled, which is just taken as a kind of coarse condition for guiding diffusion generation [90].

Moreover, the co-pilot β-VAE is also pre-trained on the corresponding datasets (i.e., if Diff-AE was trained on the facial data, β-VAE did the same), following the objective of Eq. 1, its latent $\mathbf{z_{disen}}$ has already enabled a preliminary disentangling capability, i.e., can distinguish facial features of *smile, azimuth, hair color, etc* (Figure 2) to some degree.

Phase-2: Knowledge Distillation & Diffusion Feedback To transfer the preliminary disentangling capability of $\mathbf{z_{disen}}$ into $\mathbf{z_{sem}}$, we use a distillation loss shown in Eq. 5 to smoothly bridge Diff-AE and β-VAE. But this interaction is currently one-way, which is still limited by the intrinsic shortcomings of β-VAE discussed after Eq. 1 that it is hard to balance a trade-off and thus hurt the final performance.

Therefore, we propose a novel diffusion feedback-driven loss $\mathcal{L}_{fd}$ to further facilitate the disentanglement of the representations ($\mathbf{z_{disen}}$, $\mathbf{z_{sem}}$) and break the optimization bottleneck of the entire CL-Dis framework. Basically, inspired by the work of understanding β-VAE [4] that modified the vanilla objective by adding a hand-craft increasing controller C (as shown in Eq. 2), we propose to leverage the reverse diffusion of Diff-AE that adaptively increases information capacity to build constraints to replace the original format. Formally, we make the original C auto-changeable, and control its value dynamically according to the information capacity of the reverse diffusion process as follows:

$$C_{dyn} = f(E_{x_t}) = \begin{cases} C_{base} \cdot (\frac{E_{x_0}}{E_{x_t}}), & \text{if } 0 < f(E_{x_t}) < C_{max} \\ C_{max}, & \text{if } f(E_{x_t}) \geq C_{max} \end{cases} \tag{6}$$

$$E_{x_0} = -\sum_{p_{x_0}\in\mathbb{R}} p_{x_0} log p_{x_0}, \quad E_{x_t} = -\sum_{p_{x_t}\in\mathbb{R}} p_{x_t} log p_{x_t}, \tag{7}$$

where E_{x_0} and E_{x_t} denote the information entropy of the static target image (i.e., Ground Truth) x_0 and the intermediate predicted diffusion denoising image result x_t, respectively. The probability values of p_{x_0}, p_{x_t} are calculated by flattening and probabilizing the values of the image matrices across the dimension of $\mathbb{R}^{C\times H\times W}$. Besides, C_{base}, C_{max} are the original empirical start-point and the maximum upper-bound according to [4], and their influences are studied in supplementary. We can intuitively see that with the denoising diffusion processing, the entropy E_{x_t} of the predicted image result x_t will gradually drop (i.e., information capacity increasing), and thus the ratio of $\frac{E_{x_0}}{E_{x_t}}$ will increase to raise the influence of C_{dyn}. So, the objective of β-VAE changes from Eq. 2 to the following $\mathcal{L}_{fd}$,

$$\begin{aligned}\mathcal{L}_{fd}(\theta,\phi,C_{dyn};\mathbf{x},\mathbf{z}) =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}|\mathbf{z}) - \\ & \beta\big|D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})\big) - C_{dyn}\big|,\end{aligned} \tag{8}$$

the above physical meaning is that: when KL-divergence drops to a certain extent (i.e., disentangling ability encounters bottleneck in a naive VAE latent), the dynamic controller C_{dyn} will counteract this change by increasing the value of the second item of Eq. 8. Intuitively, C_{dyn} that is controlled by diffusion feedback gradually highlights the penalty for KL-divergence, thus improving the latent disentangling. Under this closed-loop optimization flow (see Figure 2(c)), the gradually increasing information capacity of diffusion adaptively enhances the disentanglement of latent $\mathbf{z_{disen}}$, so as to further benefit the semantic code of $\mathbf{z_{sem}}$. Please see supplementary for more theoretical analysis.

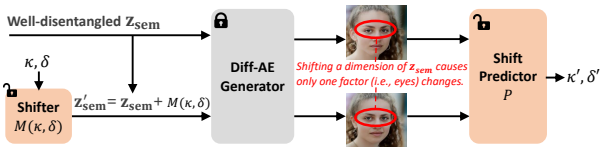


Fig. 3: A self-supervised Navigation strategy is to discover the interpretable shifts in the well-disentangled space of $\mathbf{z_{sem}}$. A training sample consists of two images, produced by the Diff-AE with original and shifted inputs. The images are given to a predictor P that predicts the direction index k and shift magnitude δ.

Phase-3: Semantics Directions Navigation After getting the well-disentangled representations of $\mathbf{z_{sem}}$, the next step is to identify its semantic meanings. Shifting a certain dimension of the well-disentangled condition $\mathbf{z_{sem}}$ will result in a generated result that only the corresponding property/factor changes. Thus, we adopt an outer navigation branch to identify the traversal directions in a global view. We employ a learnable matrix M (Figure 3) to manually append a shift on $\mathbf{z_{sem}}$. The shifted code $\mathbf{z'_{sem}}$, together with the original $\mathbf{z_{sem}}$, is fed into our Diff-AE decoder to generate a pair of images, following the reverse conditional diffusion process of $p(x_{t-1}|x_t, \mathbf{z_{sem}})$. A learnable predictor P is then

devised to predict the shift (i.e., semantic direction) according to such generated paired images, with a reconstruction loss. In this way, we clearly identify the inner disentangled semantics directions in our CL-Dis for further fine-controllable generations.

4.3 New Disentanglement Metric

The conditional diffusion-based generation in our CL-Dis defines a clear mapping between the disentangled representations and the generated image results, so we can quantitatively measure the performance of disentanglement with the generation difference. Following this idea, we quantify this difference by using an optical flow tracking technique [73]: Given a well-disentangled representation $\mathbf{z}^{\mathbf{a}}_{\mathbf{sem}}$ and its shift variant $\mathbf{z}^{\mathbf{a}'}_{\mathbf{sem}}$, we calculate the optical flow map F between their corresponding generated images. Then, we use a continuously changed hyper-parameter as a dynamic threshold to distinguish these "*changed characteristics*" caused by shifted factors from the rest "*frozen ones*". The average ratio of these two parts actually reflects the disentanglement performance, the smaller the better. Due to the space limitations, more operation details are in supplementary.

5 Experiment

5.1 Implementation Detail and Experiment Setting

For the backbone of Diff-AE, we take pre-trained weights in [62] as the initial parameters with Adam optimizer and the learning rate of 1e-4. For the co-pilot β-VAE, we adopt the same Adam optimizer and learning rate. For the navigation branch, we refer to the GAN-based editing methods [13, 76] for a fine-controllable generation. Differently, our CL-Dis has already in advance got the well-disentangled latent $\mathbf{z}_{\mathbf{sem}}$. The whole CL-Dis framework is trained on 2 NVIDIA V100 GPUs with a batch size of 32. In inference, the co-pilot β-VAE is discarded for efficiency. The dimension of $\mathbf{z}_{\mathbf{sem}}$ is set to 512 in our framework. To evaluate disentanglement, we conduct experiments on both generation & perception tasks, on synthetic & real data.

Datasets & Baselines. For the image generation task, we conduct experiments on FFHQ [37], CelebA [36], LSUN Horse [92], LSUN Cars [92], Shape3D [39]. We compare our proposed CL-Dis with VAE-based, GAN-based, and Diffusion-based baselines, including β-VAE [30], FactorVAE [39], Beta-TCVAE [8], GAN Edit v1 [76], GAN Edit v2 [13], InfoGAN-CR [51], Asyrp [45], GANspace (GS) [26], NaviNeRF [85], DisCo [63], Diff-AE [62] and vector-based DisDiff [90].

For the perception task, we just choose the classic face recognition for instantiation due to limited space and conduct experiments on CASIA-Webface [91] and VGGFace2 [5] in comparison with the baseline of LFW [34].

Evaluation Metrics. To alleviate errors, we have multiple runs for each method. We use 4 representative metrics of Factor-VAE score [39], DCI [19], FID [28] and classification accuracy in face recognition, and our additionally proposed

new metric for validation. Note that our CL-Dis belongs to dimension-/scalar-based DRL methods, so we just discuss but do not directly compare with many vector-based ones [7, 17, 52, 74, 89] in the next experiment sections.

5.2 Comparisons with SOTA DRL Methods

Table 1: Quantitative comparison results on CelebA dataset.

Methods	FID
VAE [42]	132.80
β-VAE [30]	136.23
β-TCVAE [8]	139.07
FactorVAE [39]	134.52
CL-Dis (Ours)	**6.54**

Table 2: Quantitative comparison results on FFHQ dataset.

Methods	FID
pi-GAN [6]	87.00
GIRAFFE [57]	38.00
NaviNeRF [85]	13.00
CL-Dis (Ours)	**12.15**

Quantitative Results. Table 1 and Table 2 report the results of Frechet Inception Distance (FID) [28] to measure the quality of the generated images. We see that our CL-Dis outperforms other typical disentangled methods in generation quality, especially on the real-world face images of CelebA.

Table 3: Comparisons of disentanglement on the FactorVAE score and DCI disentanglement metrics (mean ± std, higher is better). Our scalar-based CL-Dis even achieves comparable results with the co-current factor-based vector-wise DisDiff [90].

Method		Shapes3D	
		FactorVAE score	DCI
VAE-based:	FactorVAE [39]	0.840 ± 0.066	0.611 ± 0.082
	β-TCVAE [8]	0.873 ± 0.074	0.613 ± 0.114
GAN-based:	InfoGAN-CR [51]	0.587 ± 0.058	0.478 ± 0.055
Pre-trained GAN-based:	GAN Edit v1 [76]	0.805 ± 0.064	0.380 ± 0.062
	GANspace [26]	0.788 ± 0.091	0.284 ± 0.034
	DisCo [63]	0.877 ± 0.031	0.708 ± 0.048
Diffusion-based:	DisDiff [90]	0.902 ± 0.043	0.723 ± 0.013
	CL-Dis (Ours)	**0.952 ± 0.017**	**0.731 ± 0.045**

Besides, the quantitative comparison results of disentanglement under different metrics are shown in Table 3. As shown in the table, CL-Dis outperforms other competitors, demonstrating the method's superior disentanglement ability. Compared with the VAE-based methods, these methods suffer from the the trade-off between generation and disentanglement [4, 30], but CL-Dis does not have this problem due to its automatic optimization mechanism. As for the GAN-based methods are typically limited by the latent space [90]. The diffusion-based DisDiff is a vector-wise method, in which representations $\{z^c\}$ could express more information than our scalar-based ones $\mathbf{z_{sem}}$, but our CL-Dis still outperforms DisDiff on Shape3D. And, our scalar-based disentangled representations are more light, generic, robust, and easy to apply to different scenarios.

Qualitative Results. Unlike most previous papers that only test on the toy synthetic datasets [8,30,39], our CL-Dis can be naturally applied to real images. Figure 4 shows the effectiveness of our method on various datasets. To clearly

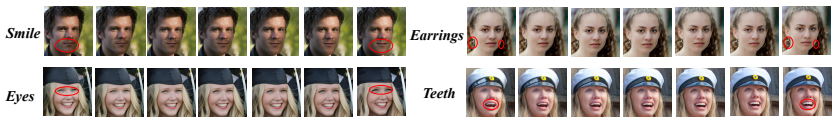


Fig. 4: Fine-grained Disentanglement Results. The qualitative results present the results of attribute manipulation, which demonstrate the semantic manipulation on the FFHQ dataset. To clearly show the isolated and "disentangled" effects, we use red circles to mark the corresponding changed regions.

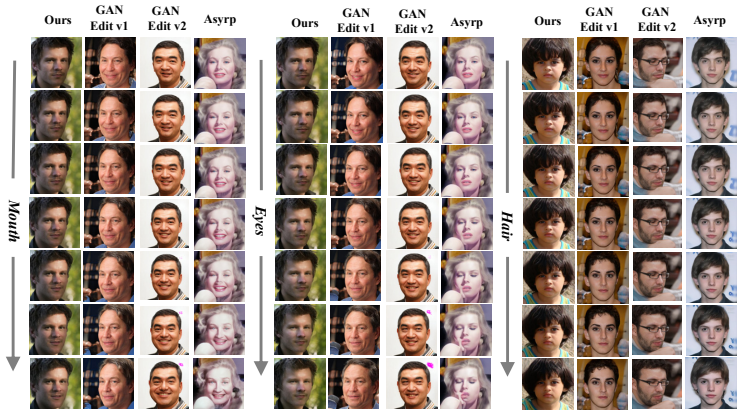


Fig. 5: Comparison of the "mouth", "eyes" and "hair" attributes with different methods. Since different methods pre-train on different datasets, we thus compare the disentanglement with different images according to the previous methods [13, 45, 76].

show the isolated and "disentangled" effects, we use red circles to mark the corresponding changed regions. We can observe that CL-Dis can disentangle and change the *isolated* attributes continuously in an unsupervised manner, such as *serious expression* → *smile, closing eyes* → *opening eyes, w/o earrings* → *w/ earrings, etc.*

Moreover, to validate the superiority of CL-Dis in the disentanglement capability, Figure 5 also provides results for changing human facial attributes to different characteristics along certain directions like "Mouth", "Eyes" and "hair", in comparison with three baselines of *GAN Edit v1* [76], *GAN Edit v2* [13], and *Asyrp* [45]. We can see that other methods like Asyrp are prone to generate artifacts (bottom row of Figure 5), and their editing results are not well-disentangled. For example, the whole face content has changed for the schemes of *GAN Edit v1* [76] and *GAN Edit v2* [13]). In contrast, our method consistently keeps the identity while just changing the disentangled properties, achieving an "isolated and disentangled" effect.

To demonstrate the generalization ability of our CL-Dis, we also use different datasets and different backbones (supplementary) for validation. As shown in Figure 6, even for cars, horses, and buildings, CL-Dis still obtains a group of fine-grained disentanglement results, where only the "color" attribute changes across the traversal direction, proving the disentanglement generalization capability.

Fig. 6: Generalization evaluation on extensive datasets. We visualize the results of shifting the latent $\mathbf{z}_{sem}$ with our proposed method along the "glasses", "hair" "mouth" directions on FFHQ (left three rows), and common "Color" direction on LSUN Horse (right first row), LSUN Cars (right second row), and CelebA (right third row) datasets.

5.3 Ablation Study and More Analysis

Effectiveness of Different Elements. To analyze the effectiveness of the proposed parts in our closed-loop disentanglement framework, we perform an ablation to study the influence of Semantics Navigation, Knowledge Distillation $\mathcal{L}_{dt}$, and Diffusion Feedback $\mathcal{L}_{fd}$. We take FFHQ as the dataset to conduct these ablation studies.

As shown in Figure 7, due to the baseline of Diff-AE itself not having a disentanglement ability, when shifting its latent we get generation results with background artifacts. If we directly apply the semantics navigation [76] to the baseline latent, the generation quality has been improved, but the disentanglement is still not guaranteed. For example, when shifting along the semantic direction of "Hair" (The 1^{st} row of Figure 7), global features such as face shape, skin texture, and hairstyle are variously entangled, resulting in a younger appearance for the man. The last two schemes of adding knowledge distillation and diffusion feedback have achieved better performance in disentanglement, particularly for the final closed-loop scheme. We can see that the final scheme together with $\mathcal{L}_{dt}$ & $\mathcal{L}_{fd}$ achieves a better disentanglement effect on the semantic direction of "Hair", compared to that only with $\mathcal{L}_{dt}$ which still changes face characteristics in the traversal generations.

Analysis of Auto-Driven Feedback Mechanism. As introduced, the core contribution of this work lies in an automatically disentangled learning mechanism, achieved by a dynamic C_{dyn} whose value depends on the increasing information capacity in the reverse diffusion process. To prove that, we visualize the changing curves of C_{dyn} during the CL-Dis training. As illustrated in Figure 8(a), the value of C_{dyn} gradually increases with the diffusion feedback, so as to highlight the penalty for KL-divergence item in the co-pilot β-VAE optimization and thus improve disentanglement performance of the entire framework.

5.4 New Metric & Discrimination Task Validation

New Metric Evaluation. As our work belongs to the track of unsupervised representation disentanglement, not image generation or editing. The core of this task is to fully understand the latent factors of the model to enhance its fine-grained controllability. The qualitative results shown above have illustrated that traversing along the disentangled representations results in an "property isolated" generation effect. To measure such effect quantitatively and show our method could progressively change the disentangled properties while keeping

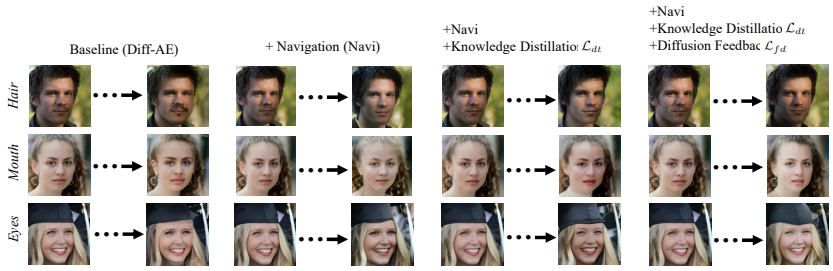
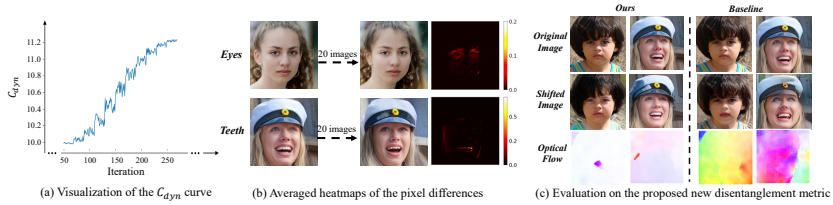


Fig. 7: Ablation study on the FFHQ dataset. We evaluate the effectiveness of different components in our method by shifting different attributes, including "Hair", "Mouth" and "Eyes". We can see that as more elements are added, both the disentanglement performance and the generation quality improve: the first column exists artifacts; the second and third columns change other contents except shift attributes.

others frozen, we propose a new metric (Sec. 4.3) through the lens of optical flow [73]. As illustrated in Figure 8, the optical flow visualization between the original and shifted image pairs on FFHQ dataset [37], shows that our proposed CL-Dis method demonstrates effective disentanglement, where shifting a single dimension of the latent space results in isolated property changes in the generated output. In contrast, the baseline approach (original Diff-AE) changes the entire background content of the image.



(a) Visualization of the C_{dyn} curve (b) Averaged heatmaps of the pixel differences (c) Evaluation on the proposed new disentanglement metric

Fig. 8: (a) Visualization of the curve of changeable C_{dyn} at different iteration steps. C_{dyn} gradually increases with the diffusion feedback to improve the latent disentangling. (b) Averaged heatmaps of the pixel differences between the original and the shifted images for the "eyes shift" (left) and the "teeth shift" (right), revealing a good disentanglement of ours. (c) Evaluation on the proposed new disentanglement metric with the 'hair', 'eyes', and 'smile' attributes. From the optical flow tracking visualization, we find traversing along our disentangled representations results in an "property isolated" generation effect, while the baseline exhibits a bad performance with a fully changed image content.

Furthermore, we perform normalization on the optical flow map and utilize a dynamic threshold to classify pixels as either exceeding or not exceeding it, thereby obtaining the ratio. We make statistics based on the 100 sets, and calculate the average ratio using a static threshold of 0.5, and get the average value. The baseline quantitatively achieves 0.1606 in this new metric, while our method is only 0.0177 (the lower the better), indicating that our method has better disentanglement performance that only isolated property changes when shifting across certain dimensions.

Locality of visual effects. To illustrate that the navigation along the semantic directions results in an isolated and "disentangled" effect, for a particular direction, we compute per-pixel differences averaged over the generation results with

20 shift magnitudes from a uniform grid. Figure 8(b) shows the averaged heatmap for the "Eyes" and "Mouth" directions. We can see that our method achieves a good disentanglement, where shifting "Eyes" just affects the eyes-related image content, and shifting "Mouth" just affects the mouth-related image content.

Table 4: Discrimination task (face recognition) validation. Our method with good disentangled representations achieves >2.0% gains even with fewer (5) training epochs compared to baseline.

Dataset	CASIA-Webface (%)	VGGFace2 (%)
Baseline (50 epochs)	89.45 ± 0.015	81.82 ± 0.019
Ours (5 epochs)	91.72 ± 0.011	87.32 ± 0.014
Ours (50 epochs)	**94.00 ± 0.785**	**87.97 ± 0.012**

Discrimination Task Validation. To validate whether the good disentanglement achieved by our CL-Dis could benefit the discrimination task, we conducted face recognition experiments on two large well-known datasets, CASIA-Webface [91] and VGGFace2 [5]. We use the semantics encoder E_{sem} of CL-Dis as an off-the-shelf feature extractor and introduce two additional learnable linear layers for fine-tuning. Following the same protocol used in [65], we perform a fair comparison in Table 4 with mean classification accuracy as a metric. We observe that the semantic encoder of our CL-Dis performs better than the baseline of learning from scratch (LFS) on the widely adopted test benchmark LFW [34], and even our E_{sem} trained with only 5 epochs still outperforms the LFS model with 50 epochs, which demonstrates that the well-disentangled representations learned by CL-Dis indeed benefit discrimination tasks.

6 Conclusion

In this paper, we present CL-Dis, an unsupervised closed-loop disentanglement framework that achieves a strong DRL ability without relying on heavy supervision or heuristic loss constraints. CL-Dis embeds a closed-loop interconnection, where a pre-trained β-VAE empowers a preliminary disentanglement ability to diffusion model with a distillation loss, while the increasing information capability in the diffusion is taken as feedback, in turn, to progressively improve β-VAE disentangling, forming a core mutually-promoting mechanism. Besides, we also design a self-supervised strategy to navigate semantic directions in the well-disentangled latent for a controllable generation. A new metric is further introduced to measure the disentanglement effect. Experimental results indicate that CL-Dis achieves a state-of-the-art DRL performance, and its strong disentangled ability benefits both generation and discrimination tasks, especially on real-world images.

Acknowledgments

This work was supported in part by NSFC 62302246 and ZJNSFC under Grant LQ23F010008, and supported by High Performance Computing Center at Eastern Institute of Technology, Ningbo, and Ningbo Institute of Digital Twin.

References

1. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: CVPR. pp. 9163–9171 (2019)
2. Bao, Z., Hebert, M., Wang, Y.X.: Generative modeling for multi-task visual learning. In: ICML. pp. 1537–1554. PMLR (2022)
3. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. TPAMI **35**(8), 1798–1828 (2013)
4. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in β-vae. NeurIPS Workshop on Learning Disentangled Representations (2017)
5. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face and Gesture Recognition (2018)
6. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: CVPR (2021)
7. Chen, H., Zhang, Y., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation. arXiv preprint arXiv:2305.03374 (2023)
8. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. NeurIPS **31** (2018)
9. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. NeurIPS **29** (2016)
10. Chen, Z., Jiang, R., Duke, B., Zhao, H., Aarabi, P.: Exploring gradient-based multi-directional controls in gans. In: ECCV. pp. 104–119. Springer (2022)
11. Cheng, P., Min, M.R., Shen, D., Malon, C., Zhang, Y., Li, Y., Carin, L.: Improving disentangled text representation learning with information-theoretic guidance. arXiv preprint arXiv:2006.00693 (2020)
12. Cheng, Y., Wang, D., Zhou, P., Zhang, T.: A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282 (2017)
13. Cherepkov, A., Voynov, A., Babenko, A.: Navigating the gan parameter space for semantic image editing. In: CVPR. pp. 3671–3680 (2021)
14. Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., Yang, Y.: Safe rlhf: Safe reinforcement learning from human feedback. arXiv preprint arXiv:2310.12773 (2023)
15. Dasgupta, S., Dey, A., Roberts, N., Sabato, S.: Learning from discriminative feature feedback. Advances in Neural Information Processing Systems **31** (2018)
16. Dasgupta, S., Sabato, S.: Robust learning from discriminative feature feedback. In: International Conference on Artificial Intelligence and Statistics. pp. 973–982. PMLR (2020)
17. Denton, E.L., et al.: Unsupervised learning of disentangled representations from video. NeurIPS (2017)
18. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
19. Eastwood, C., Williams, C.K.: A framework for the quantitative evaluation of disentangled representations. In: International conference on learning representations (2018)

20. Egan, K.: Memory, imagination, and learning: Connected by the story. Phi Delta Kappan **70**(6), 455–459 (1989)
21. Fleming, R.W.: Visual perception of materials and their properties. Vision research **94**, 62–75 (2014)
22. Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: Ganalyze: Toward visual definitions of cognitive image properties. In: ICCV. pp. 5744–5753 (2019)
23. Gonzalez-Garcia, A., Van De Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. NeurIPS **31** (2018)
24. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS **27** (2014)
25. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. IJCV **129**, 1789–1819 (2021)
26. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. NeurIPS **33**, 9841–9850 (2020)
27. He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: An unsupervised neural attention model for aspect extraction. In: ACL. pp. 388–397 (2017)
28. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
29. Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., Lerchner, A.: Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230 (2018)
30. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2016)
31. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
32. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
33. Hsu, W.N., Glass, J.: Disentangling by partitioning: A representation learning framework for multimodal sensory data. arXiv preprint arXiv:1805.11264 (2018)
34. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
35. Ji, M., Heo, B., Park, S.: Show, attend and distill: Knowledge distillation via attention-based feature matching. In: AAAI. pp. 7945–7952 (2021)
36. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
37. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
38. Katakkar, A., Yoo, C.H., Wang, W., Lipton, Z.C., Kaushik, D.: Practical benefits of feature feedback under distribution shift. arXiv preprint arXiv:2110.07566 (2021)
39. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML. pp. 2649–2658. PMLR (2018)
40. Kim, M., Wang, Y., Sahu, P., Pavlovic, V.: Relevance factor vae: Learning and identifying disentangled factors. arXiv preprint arXiv:1902.01568 (2019)
41. Kim, Y., Kim, D., Lee, H., Moon, I.c.: Unsupervised controllable generation with score-based diffusion models: Disentangled latent code guidance. In: NeurIPS 2022 Workshop on Score-Based Methods (2022)

42. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
43. Knox, W.B., Stone, P.: Augmenting reinforcement learning with human feedback. In: ICML 2011 Workshop on New Developments in Imitation Learning (July 2011). vol. 855, p. 3 (2011)
44. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. ICLR (2018)
45. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
46. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, e253 (2017)
47. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: ECCV. pp. 35–51 (2018)
48. Lee, P.S., Sewell, D.K.: A revised diffusion model for conflict tasks. Psychonomic Bulletin & Review pp. 1–31 (2023)
49. Lee, W., Kim, D., Hong, S., Lee, H.: High-fidelity synthesis with disentangled representation. In: ECCV. pp. 157–174. Springer (2020)
50. Li, M., Jin, Y., Zhu, H.: Surrogate gradient field for latent space manipulation. In: CVPR. pp. 6529–6538 (2021)
51. Lin, Z., Thekumparampil, K.K., Fanti, G., Oh, S.: Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. arXiv preprint arXiv:1906.06034 p. 60 (2019)
52. Liu, L., Li, J., Niu, L., Xu, R., Zhang, L.: Activity image-to-video retrieval by disentangling appearance and motion. AAAI (2021)
53. Liu, Y., Sangineto, E., Chen, Y., Bao, L., Zhang, H., Sebe, N., Lepri, B., Wang, W., De Nadai, M.: Smoothing the disentangled latent style space for unsupervised image-to-image translation. In: CVPR. pp. 10785–10794 (2021)
54. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)
55. Luo, Z., He, Y., Xue, Y., Wang, H., Hauskrecht, M., Li, T.: Hierarchical active learning with qualitative feedback on regions. IEEE Transactions on Human-Machine Systems (2023)
56. Materzyńska, J., Torralba, A., Bau, D.: Disentangling visual and written concepts in clip. In: CVPR. pp. 16410–16419 (2022)
57. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: CVPR (2021)
58. Pearson, J.: The human imagination: the cognitive neuroscience of visual mental imagery. Nature reviews neuroscience **20**(10), 624–634 (2019)
59. Pierobon, M., Akyildiz, I.F.: Capacity of a diffusion-based molecular communication system with channel memory and molecular noise. IEEE Transactions on Information Theory **59**(2), 942–954 (2012)
60. Plumerault, A., Borgne, H.L., Hudelot, C.: Controlling generative models with continuous factors of variations. ICLR (2020)
61. Poulis, S., Dasgupta, S.: Learning with feature feedback: from theory to practice. In: Artificial Intelligence and Statistics. pp. 1104–1113. PMLR (2017)
62. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: CVPR. pp. 10619–10629 (2022)

63. Ren, X., Yang, T., Wang, Y., Zeng, W.: Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. ICLR (2021)
64. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: ICCV. pp. 9339–9347 (2019)
65. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
66. Shao, H., Yao, S., Sun, D., Zhang, A., Liu, S., Liu, D., Wang, J., Abdelzaher, T.: Controlvae: Controllable variational autoencoder. In: International Conference on Machine Learning. pp. 8655–8664. PMLR (2020)
67. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: CVPR. pp. 9243–9252 (2020)
68. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: CVPR. pp. 1532–1540 (2021)
69. Shi, Y., Yang, X., Wan, Y., Shen, X.: Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In: CVPR. pp. 11254–11264 (2022)
70. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2020)
71. Song, Y., Keller, A., Sebe, N., Welling, M.: Latent traversals in generative models as potential flows. ICML (2023)
72. Song, Y., Zhang, J., Sebe, N., Wang, W.: Householder projector for unsupervised latent semantics discovery. In: ICCV. pp. 7712–7722 (2023)
73. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020)
74. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: CVPR. pp. 1415–1424 (2017)
75. Tsai, Y.H.H., Liang, P.P., Zadeh, A., Morency, L.P., Salakhutdinov, R.: Learning factorized multimodal representations. ICLR (2018)
76. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the gan latent space. In: ICML. pp. 9786–9796. PMLR (2020)
77. Wang, H., Zhao, H., Li, X., Tan, X.: Progressive blockwise knowledge distillation for neural network acceleration. In: IJCAI. pp. 2769–2775 (2018)
78. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. In: CVPR. pp. 11379–11388 (2022)
79. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled representation learning. arXiv preprint arXiv:2211.11695 (2022)
80. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: ICCV. pp. 7677–7689 (2023)
81. Wu, J., Li, X., Ao, X., Meng, Y., Wu, F., Li, J.: Improving robustness and generality of nlp models using disentangled representations. arXiv preprint arXiv:2009.09587 (2020)
82. Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., Lin, Z., Zhang, Y., Chang, S.: Uncovering the disentanglement capability in text-to-image diffusion models. In: CVPR. pp. 1900–1910 (2023)
83. Wu, Y., Nakashima, Y., Garcia, N.: Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. pp. 199–208 (2023)
84. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: CVPR. pp. 12863–12872 (2021)

85. Xie, B., Li, B., Zhang, Z., Dong, J., Jin, X., Yang, J., Zeng, W.: Navinerf: Nerf-based 3d representation disentanglement by latent semantic navigation. arXiv preprint arXiv:2304.11342 (2023)
86. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: Text, images and variations all in one diffusion model. In: ICCV. pp. 7754–7765 (2023)
87. Xu, Z., Lin, T., Tang, H., Li, F., He, D., Sebe, N., Timofte, R., Van Gool, L., Ding, E.: Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In: CVPR. pp. 18229–18238 (2022)
88. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: ECCV. pp. 776–791. Springer (2016)
89. Yang, T., Wang, Y., Lan, C., Lu, Y., Zheng, N.: Vector-based representation is the key: A study on disentanglement and compositional generalization. arXiv preprint arXiv:2305.18063 (2023)
90. Yang, T., Wang, Y., Lv, Y., Zh, N.: Disdiff: Unsupervised disentanglement of diffusion probabilistic models. arXiv preprint arXiv:2301.13721 (2023)
91. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
92. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
93. Zhang, Y., Zhang, Y., Guo, W., Cai, X., Yuan, X.: Learning disentangled representation for multimodal cross-domain sentiment analysis. TNNLS (2022)
94. Zhu, X., Xu, C., Tao, D.: Where and what? examining interpretable disentangled representations. In: CVPR. pp. 5861–5870 (2021)