# Disentangled Representation Learning

Xin Wang, *Member, IEEE*, Hong Chen, Si'ao Tang, Zihao Wu and Wenwu Zhu, *Fellow, IEEE*

**Abstract**—Disentangled Representation Learning (DRL) aims to learn a model capable of identifying and disentangling the underlying factors hidden in the observable data in representation form. The process of separating underlying factors of variation into variables with semantic meaning benefits in learning explainable representations of data, which imitates the meaningful understanding process of humans when observing an object or relation. As a general learning strategy, DRL has demonstrated its power in improving the model explainability, controlability, robustness, as well as generalization capacity in a wide range of scenarios such as computer vision, natural language processing, and data mining. In this article, we comprehensively investigate DRL from various aspects including motivations, definitions, methodologies, evaluations, applications, and model designs. We first present two well-recognized definitions, i.e., Intuitive Definition and Group Theory Definition for disentangled representation learning. We further categorize the methodologies for DRL into four groups from the following perspectives, the model type, representation structure, supervision signal, and independence assumption. We also analyze principles to design different DRL models that may benefit different tasks in practical applications. Finally, we point out challenges in DRL as well as potential research directions deserving future investigations. We believe this work may provide insights for promoting the DRL research in the community.

**Index Terms**—Disentangled Representation Learning, Representation Learning, Computer Vision, Pattern Recognition.

## 1 Introduction

When humans observe an object, we seek to understand the various properties of this object (e.g., shape, size and color etc.) with certain prior knowledge. However, existing end-to-end black-box deep learning models take a shortcut strategy through directly learning representations of the object to fit the data distribution and discrimination criteria [1], failing to extract the hidden attributes carried in representations with human-like generalization ability. To fill this gap, an important representation learning paradigm, *Disentangled Representation Learning* (DRL) is proposed [2] and has attracted an increasing amount of attention in the research community.

DRL is a learning paradigm where machine learning models are designed to obtain representations capable of identifying and disentangling the underlying factors hidden in the observed data. DRL always benefits in learning explainable representations of the observed data that carry semantic meanings. Existing literature [2], [3] demonstrates the potential of DRL in learning and understanding the world as humans do, where the understanding towards real-world observations can be reflected in disentangling the semantics in the form of disjoint factors. The disentanglement in the feature space encourages the learned representation to carry explainable semantics with independent factors, showing great potential to improve various machine learning tasks from the three aspects: i) Explainability: DRL learns semantically meaningful and separate representations which are aligned with latent generative factors. ii) Generalizability: DRL separates the representations that our tasks are interested in from the original entangled input and thus has better generalization ability. iii) Controllability: DRL achieves controllable generation by manipulating the learned disentangled representations in latent space.

Then a natural question arises, *what are disentangled representations supposed to learn?* The answer may lie in the concept of disentangled representation proposed by Bengio et al. [2], which refers to *factor of variations* in brief. As shown by the example illustrated in Figure 1, Shape3D [4] is a frequently used dataset in DRL with six distinct factors of variation, i.e., object size, object shape, object color, wall color, floor color and viewing angle. DRL aims at separating these factors and encoding them into independent and distinct latent variables in the representation space. In this case, the latent variables controlling object shape will change only with the variation of object shape and be constant over other factors. Analogously, it is the same for variables controlling other factors including size, color etc.

Through both theoretical and empirical explorations, DRL benefits in the following three perspectives: i) Invariance: an element of the disentangled representations is invariant to the change of external semantics [5], [6], [7], [8], ii) Integrity: all the disentangled representations are aligned with real semantics respectively and are capable of generating the observed, undiscovered and even counterfactual samples [9], [10], [11], [12], and iii) Generalization: representations are intrinsic and robust instead of capturing confounded or biased semantics, thus being able to generalize for downstream tasks [13], [14], [15].

Following the motivation and requirement of DRL, there have been numerous works on DRL and its applications over various tasks. Most typical methods for DRL are based on generative models [6], [9], [16], [17], which initially show great potential in learning explainable representations for visual images. In addition, approaches based on causal

- *Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu and Wenwu Zhu are with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. E-mail: {xin_wang,wwzhu}@tsinghua.edu.cn, {h-chen20,tsa22,wuzh22}@mails.tsinghua.edu.cn.*

- *This work was supported by the National Key Research and Development Program of China No.2023YFF1205001, National Natural Science Foundation of China (No. 62222209, 62250008, 62102222), Beijing National Research Center for Information Science and Technology under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.*

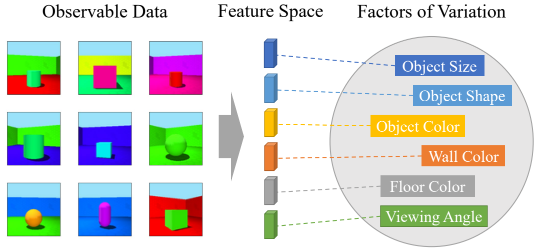

Fig. 1. The scene of Shape3D [4], where the six rectangles in the gray circle represent the six factors of variation in the Shape3D respectively. DRL is expected to encode these distinct factors with independent latent variables in the latent feature space.

inference [14] and group theory [18] are widely adopted in DRL as well. The core concept of designing DRL architecture lies in encouraging the latent factors to learn disentangled representations while optimizing the inherent task objective, e.g., generation or discrimination objective. Given the efficacy of DRL at capturing explainable, controllable and robust representations, it has been widely used in many fields such as computer vision [8], [19], [20], [21], [22], natural language processing [23], [24], [25], recommender systems [26], [27], [28], [29] and graph learning [29], [30] etc., boosting the performances of various downstream tasks.

**Contributions.** In this paper, we comprehensively review DRL through summarizing the theories, methodologies, evaluations, applications and design schemes, to the best of our knowledge, for the first time. Existing work most related to this paper is Liu et al.'s work [31], which only focuses on imaging domain and applications in medical imaging. In comparison, our work discusses DRL from a general perspective, taking full coverage of definitions, taxonomies, applications and design scheme.

## 2 DRL DEFINITIONS

**Intuitive Definition**. Bengio et al. [2] propose an intuitive definition about disentangled representation:

**Definition 1.** *Disentangled representation should separate the distinct, independent and informative generative factors of variation in the data. Single latent variables are sensitive to changes in single underlying generative factors, while being relatively invariant to changes in other factors.*

The definition also indicates that latent variables are statistically independent. Following this intuitive definition, early DRL methods can be traced back to independent component analysis (ICA) and principal component analysis (PCA). Numerous Deep Neural Network (DNN) based methods also follow this definition [5], [6], [7], [9], [32], [33], [34], [35], [36], [37]. Most models and metrics hold the view that generative factors and latent variables are statistically independent.

Definition 1 is widely adopted in the literature, and is followed by the majority of DRL approaches discussed in Section 3.

**Group Theory Definition**. For a more rigorous mathematical definition, Higgins et al. [18] propose to define DRL from the perspective of group theory, which is later adopted by a series of works [38], [39], [40], [41]. We briefly review the group theory-based definition as follows:

**Definition 2.** *Consider a symmetry group $G$, world state space $W$ (i.e., ground truth factors which generate observations), data space $O$, and representation space $Z$. Assume $G$ can be decomposed as a direct product $G = G_1 \times G_2 \times \cdots \times G_n$. Representation $Z$ is disentangled with respect to $G$ if:*

*(i) There is an action of $G$ on $Z$: $G \times Z \to Z$.*

*(ii) There exists a mapping from $W$ to $Z$, i.e., $f : W \to Z$ which is equivariant between the action of $G$ on $W$ and $Z$. This condition can be formulated as follows:*

$$g \cdot f(w) = f(g \cdot w), \forall g \in G, \forall w \in W \tag{1}$$

*which can be illustrated as Figure. 2.*

*(iii) The action of $G$ on $Z$ is disentangled with respect to the decomposition of $G$. In other words, there is a decomposition $Z = Z_1 \times \ldots \times Z_n$ or $Z = Z_1 \oplus \ldots \oplus Z_n$ such that each $Z_i$ is affected only by $G_i$ and invariant to $G_j, \forall j \neq i$.*

Definition 2 is mainly adopted by DRL approaches originating from the perspective of group theory in VAE (Group theory based VAEs in Section 3.1.1).

$$\begin{tikzcd} W \arrow[r, "g\cdot"] \arrow[d, "f"'] & W \arrow[d, "f"] \\ Z \arrow[r, "g\cdot"'] & Z \end{tikzcd}$$

Fig. 2. The illustration of condition (ii).

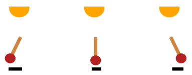

Fig. 3. Swinging pendulum, light and shadow, figure from [11].

**Discussions**. All the two definitions hold the assumption that generative factors are naturally independent. However, Suter et al. [14] propose to define DRL from the perspective of the structural causal model (SCM) [42], where they additionally introduce a set of confounders which causally influence the generative factors of observable data. Yang et al. [11] and Shen et al. [43] further discard the independence assumption by considering that there might be an underlying causal structure which renders generative factors. For example, in Figure 3, the position of the light source and the angle of the pendulum are both responsible for the position and length of the shadow. Consequently, instead of the independence assumption, they use SCM which characterizes the causal relationship of generative factors as prior. We refer to these works holding the assumption of causal factors as causal disentanglement methods, which will be discussed in detail in Section 3.4.

## 3 DRL TAXONOMY

As shown in Figure 4, we categorize DRL approaches from i) the perspective of base model type, ii) the perspective of representation structure, i.e., dimension-wise v.s. vector-wise and flat v.s. hierarchical, iii) the perspective of available supervision signal, i.e., unsupervised v.s. supervised v.s. weakly supervised, and iv) from the independence assumption of generative factors, i.e., independent v.s. causal.

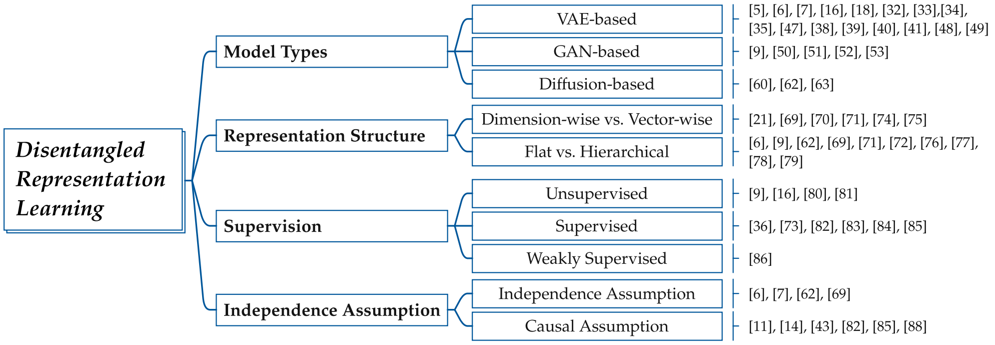

Fig. 4. A categorization of DRL approaches.

## 3.1 Model Type

### 3.1.1 VAE-based DRL Methods

**Vanilla VAE-based Methods.** Variational auto-encoder (VAE) [16] is a variant of the auto-encoder, which adopts the idea of variational inference. VAE is originally proposed as a deep generative probabilistic model for image generation. Later researchers find that VAE also has the potential ability to learn disentangled representation on simple datasets (e.g., FreyFaces [16], MNIST [44]). The general VAE model structure is shown in Figure 5. The fundamental idea of VAE is to model data distributions from the perspective of maximum likelihood using variational inference, i.e., to maximize $\log p_\theta(\mathbf{x})$. This objective can be written as Eq.(2) in the following,

$$\log p_\theta(\mathbf{x}) = D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}|\mathbf{x})\right) + \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}), \tag{2}$$

where $q$ represents variational posterior distribution and $z$ represents the latent representation in hidden space. The key point of Eq.(2) is leveraging variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ to approximate true posterior distribution $p_\phi(\mathbf{z}|\mathbf{x})$, which is generally intractable in practice. The detailed derivation of Eq.(2) can be found in the original paper [16]. The first term of Eq.(2) is the KL divergence between variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$, and the second term is denoted as the (variational) evidence lower bound (ELBO) given that the KL divergence term is always non-negative. In practice, we usually maximize the ELBO to provide a tight lower bound for the original $\log(p_\theta(\mathbf{x}))$. The ELBO can also be rewritten as Eq.(3) in the following,

$$\begin{aligned}\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = &- D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\right) \\ &+ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right],\end{aligned} \tag{3}$$

where the conditional logarithmic likelihood $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$ is in charge of the reconstruction, and the KL divergence reflects the distance between the variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior distribution $p_\theta(\mathbf{z})$. Generally, a standard Gaussian distribution $N(0, I)$ is chosen for $p_\theta(\mathbf{z})$ so that the KL term actually imposes independent constraints on the representations learned through neural network [5], which may be the reason that VAE has the potential ability of disentanglement.

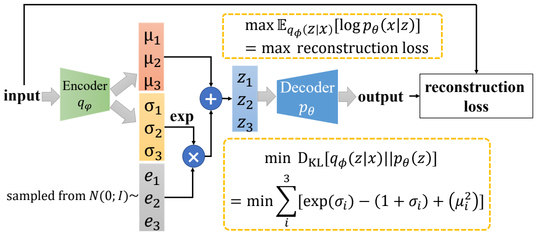
input
Encoder
$q_\varphi$
$\mu_1$
$\mu_2$
$\mu_3$
$\sigma_1$
$\sigma_2$
$\sigma_3$
exp
$e_1$
$e_2$
$e_3$
sampled from $N(0; I)\sim$
$z_1$
$z_2$
$z_3$
Decoder
$p_\theta$
output
reconstruction loss
$\max \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$
$= \max$ reconstruction loss
$\min \mathrm{D_{KL}}[q_\phi(z|x)||p_\theta(z)]$
$= \min \sum_i^3 [\exp(\sigma_i) - (1+\sigma_i) + (\mu_i^2)]$

Fig. 5. The general framework of variational auto-encoder (VAE).

Although having the potential ability to disentangle, it has been observed that the vanilla VAE shows poor disentanglement capability on relatively complex datasets such as CelebA [45] and 3D Chairs [46] etc. To tackle this problem, a large amount of improvement has been proposed through adding implicit or explicit inductive bias to enhance disentanglement ability, resorting to various regularizers (e.g., $\beta$-VAE [6], DIP-VAE [35], and $\beta$-TCVAE [5] etc.). Specifically, to strengthen the independence constraint of the variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$, $\beta$-VAE [6] introduces a $\beta$ penalty coefficient before the KL term in ELBO, where the updated objective function is shown in Eq.(4).

$$\mathcal{L}(\theta, \phi, \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z})\right). \tag{4}$$

When $\beta$=1, $\beta$-VAE degenerates to the original VAE formulation. The experimental results of $\beta$-VAE [6] show that larger values of $\beta$ encourage learning more disentangled representations while harming the performance of reconstruction. Therefore, it is important to select an appropriate $\beta$ to control the trade-off between reconstruction accuracy and the quality of disentangling latent representations. To further investigate this trade-off phenomenon, Chen et al. [5] gives a more straightforward explanation from the perspective of ELBO decomposition. They prove that the penalty tends to increase dimension-wise independence of representation $\mathbf{z}$ but decrease the ability of $\mathbf{z}$ in preserving the information from input $\mathbf{x}$.

However, it is practically intractable to obtain the optimal $\beta$ that balances the trade-off between reconstruction and disentanglement. To handle this problem, Burgess et al. [34] propose a simple modification, such that the quality of disentanglement can be improved as much as possible without losing too much information of the original data. They regard $\beta$-VAE objective as an optimization problem from the perspective of information bottleneck theory, whose objective function is shown in Eq.(5) as follows,

$$\max[I(Z;Y)-\beta I(X;Z)], \tag{5}$$

where $X$ represents the original input to be compressed, $Y$ represents the objective task, $Z$ is the compressed representations for $X$, and $I(;)$ stands for mutual information. Recall the $\beta$-VAE framework, we can regard the first term in Eq.(4), $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$ as $I(Z;Y)$, and approximately treat the second term, $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$ as $I(X;Z)$. To be specific, $q_\phi(\mathbf{z}|\mathbf{x})$ can be considered as the information bottleneck of the reconstruction task $\max \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$. $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$ can be seen as an upper bound over the amount of information that $q_\phi(\mathbf{z}|\mathbf{x})$ can extract and preserve for original data $\mathbf{x}$. The strategy is to gradually increase the information capacity of the latent channel, and the modified objective function is shown in Eq.(6) as follows,

$$\begin{aligned}\mathcal{L}(\theta,\phi,C;\mathbf{x},\mathbf{z})=&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}|\mathbf{z})-\\&\gamma\left|D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))-C\right|,\end{aligned} \tag{6}$$

where $\gamma$ and $C$ are hyperparameters. During the training process, $C$ will gradually increase from 0 to a value large enough to guarantee the expressiveness of latent representations, or in other words, to guarantee satisfactory reconstruction quality when achieving good disentanglement quality.

Furthermore, DIP-VAE [35] proposes an extra regularizer to improve the ability to disentangle, with objective function shown in Eq.(7) as follows,

$$\begin{aligned}\max_{\theta,\phi}\mathbb{E}_{\mathbf{x}}\Big[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]-D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})\right)\Big]\\-\lambda D\left(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z})\right),\end{aligned} \tag{7}$$

where $D(\cdot\|\cdot)$ represents distance function between $q_\phi(\mathbf{z})$ and $p_\theta(\mathbf{z})$. The authors point out that $q_\phi(\mathbf{z})$ should equal to $\prod_j q_j(\mathbf{z}_j)$ to guarantee the disentanglement. Given the assumption that $p_\theta(\mathbf{z})$ follows the standard Gaussian distribution $N(0,I)$, the objective imposes independence constraint on the variational posterior cumulative distribution $q_\phi(\mathbf{z})$. In order to minimize the distance term, Kumar et al. match the covariance of $q_\phi(\mathbf{z})$ and $p_\theta(\mathbf{z})$ by decorrelating the dimensions of $\mathbf{z}\sim q_\phi(\mathbf{z})$ given $p_\theta(\mathbf{z})\sim N(0,I)$, i.e., they force Eq.(8) to be close to the identity matrix,

$$\mathrm{Cov}_{q_\phi(\mathbf{z})}[\mathbf{z}]=\mathbb{E}_{p(\mathbf{x})}\left[\boldsymbol{\Sigma}_\phi(\mathbf{x})\right]+\mathrm{Cov}_{p(\mathbf{x})}\left[\boldsymbol{\mu}_\phi(\mathbf{x})\right], \tag{8}$$

where $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\Sigma}_\phi(\mathbf{x})$ denote the prediction of VAE model for posterior $q_\phi(\mathbf{z}|\mathbf{x})$, i.e., $q_\phi(\mathbf{z}|\mathbf{x})\sim N(\boldsymbol{\mu}_\phi(\mathbf{x}),\boldsymbol{\Sigma}_\phi(\mathbf{x}))$. Finally, they propose two variants, DIP-VAE-I and DIP-VAE-II, whose objective functions are shown in Eq.(9) and Eq.(10) respectively as follows,

$$\begin{aligned}\max_{\theta,\phi}\mathrm{ELBO}(\theta,\phi)-\lambda_{od}\sum_{i\neq j}\left[\mathrm{Cov}_{p(\mathbf{x})}[\boldsymbol{\mu}_\phi(\mathbf{x})]\right]_{ij}^2\\-\lambda_d\sum_i\Big(\left[\mathrm{Cov}_{p(\mathbf{x})}[\boldsymbol{\mu}_\phi(\mathbf{x})]\right]_{ii}-1\Big)^2,\end{aligned} \tag{9}$$

$$\begin{aligned}\max_{\theta,\phi}\mathrm{ELBO}(\theta,\phi)-\lambda_{od}\sum_{i\neq j}\left[\mathrm{Cov}_{q_\phi(\mathbf{z})}[\mathbf{z}]\right]_{ij}^2\\-\lambda_d\sum_i\Big(\left[\mathrm{Cov}_{q_\phi(\mathbf{z})}[\mathbf{z}]\right]_{ii}-1\Big)^2,\end{aligned} \tag{10}$$

where $\lambda_d$ and $\lambda_{od}$ are hyperparameters. DIP-VAE-I regularizes $\mathrm{Cov}_{p(\mathbf{x})}\left[\boldsymbol{\mu}_\phi(\mathbf{x})\right]$, while DIP-VAE-II directly regularizes $\mathrm{Cov}_{q_\phi(\mathbf{z})}[\mathbf{z}]$.

Kim et al. [7] propose FactorVAE which imposes independence constraint according to the definition of independence, as shown in Eq.(11),

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\Big[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})\right]-D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}^{(i)})\|p_\theta(\mathbf{z}))\Big]\\-\gamma D_{KL}\left(q_\phi(\mathbf{z})\|\bar{q}_\phi(\mathbf{z})\right),\end{aligned} \tag{11}$$

where $\bar{q}_\phi(\mathbf{z})=\prod_j q_\phi(\mathbf{z}_j)$ and $\mathbf{x}^{(i)}$ represents $i$-th sample. $D_{KL}\big(q_\phi(\mathbf{z})\|\prod_j q_\phi(\mathbf{z}_j)\big)$ is called *Total Correlation* which evaluates the degree of dimension-wise independence in $\mathbf{z}$.

Chen et al. [5] propose to elaborately decompose $D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\big)$ into three terms, as is shown in Eq.(12). i) The first term demonstrates the mutual information which can be rewritten as $I_q(\mathbf{z};\mathbf{x})$, ii) the second term denotes the total correlation and iii) the third term is the dimension-wise KL divergence.

$$\begin{aligned}D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))=&\underbrace{D_{KL}\left(q_\phi(\mathbf{z},\mathbf{x})\|q_\phi(\mathbf{z})p_\theta(\mathbf{x})\right)}_{\text{(i) Mutual Information}}\\&+\underbrace{D_{KL}\Big(q_\phi(\mathbf{z})\|\prod_j q_\phi(z_j)\Big)}_{\text{(ii) Total Correlation}}\\&+\underbrace{\sum_j D_{KL}\left(q_\phi(z_j)\|p_\theta(z_j)\right)}_{\text{(iii) Dimension-wise K L Divergence}}.\end{aligned} \tag{12}$$

From Eq.(12), we can straightforwardly obtain the explanation of the trade-off in $\beta$-VAE, i.e., higher $\beta$ tends to decrease $I_q(\mathbf{z};\mathbf{x})$ which is related to the reconstruction quality, while increasing the independence in $q_\phi(\mathbf{z})$ which is related to disentanglement. As such, instead of penalizing $D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\big)$ as a whole with coefficient $\beta$, we can penalize these three terms with three different coefficients respectively, which is referred as $\beta$-TCVAE and is shown in Eq.(13) as follows.

$$\begin{aligned}\mathcal{L}=&\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]-\alpha I_q(\mathbf{z};\mathbf{x})\\&-\beta D_{KL}\Big(q_\phi(\mathbf{z})\|\prod_j q_\phi(z_j)\Big)-\gamma\sum_j D_{KL}\left(q_\phi(z_j)\|p_\theta(z_j)\right).\end{aligned} \tag{13}$$

To further distinguish between meaningful and noisy factors of variation, Kim et al. [33] propose Relevance Factor VAE (RF-VAE) through introducing relevance indicator variables that are endowed with the ability to identify all meaningful factors of variation as well as the cardinality. The aforementioned VAE based methods are designed for continuous latent variables, failing to model the discrete variables. Dupont et al. [32] propose a $\beta$-VAE based framework, JointVAE, capable of disentangling both continuous and discrete representations in an unsupervised manner. The formulas of the two methods are shown in Tabel 1.

We conclude that all the above VAE based approaches are unsupervised, with the common characteristic of adding extra regularizer(s), e.g., $D_{KL}\big(q_\phi(\mathbf{z})||p(\mathbf{z})\big)$ [35] and Total Correlation [7], in addition to ELBO such that the disentanglement ability can be guaranteed. The summary of these unsupervised VAE based approaches is illustrated in Table 1. Besides, there are also several works that incorporate supervised signals into VAE-based models, which we will discuss in the Section 3.3.

VAE-based methods can also be modified to process sequential data, e.g., video and audio. Li et al. [47] separate latent representations of video frames into time-invariant and time-varying parts. They use $\mathbf{f}$ to model the global time-invariant aspects of the video frames, and use $\mathbf{z_i}$ to represent the time-varying feature of the $i$-th frame. The training procedure conforms to the VAE algorithm [16] with the objective of maximizing ELBO in Eq.(14) as follows,

$$\begin{aligned}\mathcal{L}(\theta, \phi, \mathbf{x}, \mathbf{z}, \mathbf{f}) =& \mathbb{E}_{q_\phi(\mathbf{z_{1:T}}, \mathbf{f} | \mathbf{x_{1:T}})}\left[\log p_\theta(\mathbf{x_{1:T}} | \mathbf{z_{1:T}}, \mathbf{f})\right] \\ & - D_{KL}\left(q_\phi(\mathbf{z_{1:T}}, \mathbf{f} | \mathbf{x_{1:T}}) \| p_\theta(\mathbf{z_{1:T}}, \mathbf{f})\right)\end{aligned} \tag{14}$$

**VAE-based Methods with Group Theory.** Besides the intuitive definition from Definition 1, Higgins et al. [18] propose a mathematically rigorous group theory definition of DRL in Definition 2, which is followed by a series of works [38], [39], [40], [41] on group-based DRL. Quessard et al. [39] propose a method for learning disentangled representations of dynamical environments (which returns observations) from the trajectories of transformations (which act on the environment). Different from environment-based methods [38], [39] which leverage the environment to provide world states, Yang et al. [40] propose a theoretical framework to make Definition 2 feasible in the setting of unsupervised DRL without relying on the environment. Additionally, Wang et al. [41] propose Iterative Partition-based Invariant Risk Minimization (IP-IRM) to learn a mapping between observation space and feature space in a self-supervised manner based on the Invariant Risk Minimization (IRM) [48]. Moreover, Zhu et al. [49] propose an unsupervised DRL framework, named Commutative Lie Group VAE with a matrix Lie group and corresponding Lie algebra.

### 3.1.2 GAN-based DRL Methods

Instead of adopting conventional Bayesian statistical methods, GAN (Generative Adversarial Nets) [17] directly samples latent representations $\mathbf{z}$ from a prior distribution $p(\mathbf{z})$. Specifically, GAN has a generative network (generator) $G$ and a discriminative network (discriminator) $D$ where the generator $G$ simulates a complex unknown generative system which transforms latent representation $\mathbf{z}$ to a generated image, while the discriminator $D$ receives an image (real or generated by $G$) as input and then outputs the probability of the input image being real. In the training process, the goal of generator $G$ is to generate images which can deceive discriminator $D$ into believing the generated images are real. Meanwhile, the goal of discriminator $D$ is to distinguish the images generated by generator $G$ from the real ones. Thus, generator $G$ and discriminator $D$ constitute a dynamic adversarial *minimax* game. Ideally, generator $G$ can finally generate an image that looks like a real one so that discriminator $D$ fails to determine whether the image generated by generator $G$ is real or not. The objective function is shown as Eq.(15),

$$\begin{aligned}\min_G \max_D V(D, G) = & \mathbb{E}_{\mathbf{x} \sim P_{data}}\left[\log D(\mathbf{x})\right] \\ & + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}\Big[\log\Big(1 - D\big(G(\mathbf{z})\big)\Big)\Big],\end{aligned} \tag{15}$$

where $P_{data}$ represents the real dataset and $p(\mathbf{z})$ represents the prior distribution of the latent representation $\mathbf{z}$.

Similar to VAE-based methods, researchers have explored a mass of GAN-based methods to achieve DRL.

InfoGAN [9] is one of the earliest works using the GAN paradigm to conduct dimension-wise DRL, whose framework is shown in Figure. 6. The generator takes two latent variables as input, where one is the incompressible noise $\mathbf{z}$, and the other is the target latent variable $\mathbf{c}$ which captures the latent generative factors. To encourage the disentanglement in $\mathbf{c}$, InfoGAN designs an extra variational regularization of mutual information, i.e., $I(\mathbf{c}; G(\mathbf{z}, \mathbf{c}))$ controlled by hyperparameter $\lambda$, such that the adversarial loss of InfoGAN is written in Eq. (16) as follows,

$$\min_G \max_D V_I(D, G) = V'(D, G) - \lambda I\big(\mathbf{c}; G(\mathbf{z}, \mathbf{c})\big), \tag{16}$$

where $V'(D, G)$ is defined in Eq.(17), taking $\mathbf{c}$ into account.

$$\begin{aligned}V'(D, G) = & \mathbb{E}_{\mathbf{x} \sim P_{data}}\left[\log D(\mathbf{x})\right] \\ & + \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}\Big[\log\Big(1 - D\big(G(\mathbf{z}, \mathbf{c})\big)\Big)\Big].\end{aligned} \tag{17}$$

However, it is intractable to directly optimize $I\big(\mathbf{c}; G(\mathbf{z}, \mathbf{c})\big)$ because of the inaccessibility of posterior $p(\mathbf{c}|\mathbf{x})$. Therefore, InfoGAN derives a lower bound of $I\big(\mathbf{c}; G(\mathbf{z}, \mathbf{c})\big)$ with variational inference in Eq.(18),

$$\begin{aligned}I\big(\mathbf{c}; G(\mathbf{z}, \mathbf{c})\big) &= H(\mathbf{c}) - H\big(\mathbf{c} \mid G(\mathbf{z}, \mathbf{c})\big) \\ &= \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})}\Big[\mathbb{E}_{\mathbf{c}' \sim p(\mathbf{c}|\mathbf{x})}\left[\log p\left(\mathbf{c}' \mid \mathbf{x}\right)\right]\Big] + H(\mathbf{c}) \\ &= \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})}\Big[\underbrace{D_{KL}\big(p(\cdot \mid \mathbf{x}) \| q(\cdot \mid \mathbf{x})\big)}_{\geq 0} \\ &\qquad + \mathbb{E}_{\mathbf{c}' \sim p(\mathbf{c}|\mathbf{x})}\left[\log q\left(\mathbf{c}' \mid \mathbf{x}\right)\right]\Big] + H(\mathbf{c}) \\ &\geq \mathbb{E}_{\mathbf{x} \sim G(\mathbf{z}, \mathbf{c})}\Big[\mathbb{E}_{\mathbf{c}' \sim p(\mathbf{c}|\mathbf{x})}\left[\log q\left(\mathbf{c}' \mid \mathbf{x}\right)\right]\Big] + H(\mathbf{c}) \\ &= \mathbb{E}_{\mathbf{c} \sim p(\mathbf{c}), \mathbf{x} \sim G(\mathbf{z}, \mathbf{c})}[\log q(\mathbf{c} \mid \mathbf{x})] + H(\mathbf{c}),\end{aligned} \tag{18}$$

where $H(.)$ denotes the entropy of the random variable and $q(\mathbf{c}|\mathbf{x})$ is the auxiliary posterior distribution approximating the true posterior $p(\mathbf{c}|\mathbf{x})$.

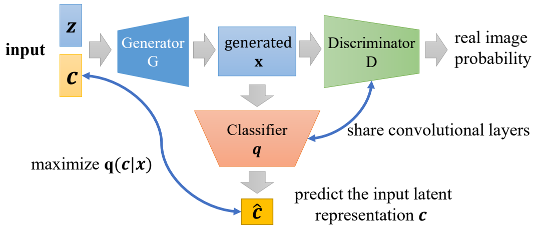

Fig. 6. The overall framework of InfoGAN.

Nevertheless, the performance of InfoGAN for disentanglement is constantly reported to be lower than VAE-based models. To enhance disentanglement, Jeon et al. [50] propose IB-GAN which compresses the representation by adding a constraint on the maximization of mutual information between latent representation $\mathbf{z}$ and $G(\mathbf{z})$, which is actually a kind of application for information bottleneck. The hypothesis behind IB-GAN is that the compressed representations usually tend to be more disentangled.

Lin et al. [51] propose InfoGAN-CR, which is a self-supervised variant of InfoGAN with contrastive regularizer.

TABLE 1
The summary of VAE based approaches.

| Method | Regularizer | Description |
|---|---|---|
| $\beta$-VAE | $-\beta D_{KL}\left(q_\phi(\mathbf{z}\|\mathbf{x})\\|p(\mathbf{z})\right)$ | $\beta$ controls the trade-off between reconstruction fidelity and the quality of disentanglement in latent representations. |
| Understanding disentangling in $\beta$-vae | $-\gamma \left\| D_{KL}\left(q_\phi(\mathbf{z}\|\mathbf{x})\\|p(z)\right) - C \right\|$ | The quality of disentanglement can be improved as much as possible without losing too much information from original data by linearly increasing $C$ during training. |
| DIP-VAE | $-\lambda D_{KL}\left(q_\phi(\mathbf{z})\\|p(\mathbf{z})\right)$ | Enhance disentanglement by minimizing the distance between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$. In practice, we can match the moments between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$. |
| FactorVAE | $-\gamma D_{KL}\left(q_\phi(\mathbf{z})\\| \prod_j q_\phi(z_j)\right)$ | Directly impose independence constraint on $q_\phi(\mathbf{z})$ in the form of total correlation. |
| $\beta$-TCVAE | $-\alpha I_q(\mathbf{z};\mathbf{x}) - \beta D_{KL}\left(q(\mathbf{z})\\| \prod_j q(z_j)\right) - \gamma \sum_j D_{KL}\left(q(z_j)\\|p(z_j)\right)$ | Decompose $D_{KL}\left(q(\mathbf{z}\|\mathbf{x})\|\|p(\mathbf{z})\right)$ into three terms: i) mutual information, ii) total correlation, iii) dimension-wise KL divergence and then penalize them respectively. |
| JointVAE | $-\gamma \left\|D_{KL}\left(q_\phi(\mathbf{z}\|\mathbf{x})\\|p(\mathbf{z})\right) - C_z\right\| - \gamma \left\|D_{KL}\left(q_\phi(\mathbf{c}\|\mathbf{x})\\|p(\mathbf{c})\right) - C_c\right\|$ | Separate latent variables into continuous $\mathbf{z}$ and discrete $\mathbf{c}$, then modify the objective function of $\beta$-VAE to capture discrete generative factors. |
| RF-VAE | $-\sum_{j=1}^{d} \lambda(r_j)\, D_{KL}\left(q(z_j\|\mathbf{x})\\|p(z_j)\right) - \gamma D_{KL}\left(q(\mathbf{r}\circ\mathbf{z})\\| \prod_{j=1}^{d} q(r_j \circ z_j)\right) - \eta\\|\mathbf{r}\\|_1$ | Introduce relevance indicator variables $\mathbf{r}$ by only focusing on relevant part when computing the total correlation, penalize $D_{KL}\left(q(z_j\|\mathbf{x})\|\|p(z_j)\right)$ less for relevant dimensions and more for nuisance (noisy) dimensions. |

They generate multiple images by keeping one dimension of the latent representation, i.e., $c_i$, fixed and randomly sampling others, i.e., $c_j$ where $j \neq i$. Then a classifier which takes these images as input will be trained to determine which dimension is fixed. The contrastive regularizer encourages distinctness across different dimensions in the latent representation, thus being capable of promoting disentanglement.

Zhu et al. [52] propose PS-SC GAN based on InfoGAN which employs a Spatial Constriction (SC) design to obtain the focused areas of each latent dimension and utilizes Perceptual Simplicity (PS) design to encourage the factors of variation captured by latent representations to be simpler and purer. The Spatial Constriction design is implemented as a spatial mask with constricted modification. Moreover, PS-SC GAN imposes a perturbation $\epsilon$ on a certain latent dimension $c_i$ (i.e., $c'_i = c_i + \epsilon$) and then computes the reconstruction loss between $\mathbf{c}$ and $\hat{\mathbf{c}}$ with $\hat{\mathbf{c}} = q\big(G(\mathbf{c},\mathbf{z})\big)$, as well as the reconstruction loss between $\mathbf{c}'$ and $\hat{\mathbf{c}}'$ with $\hat{\mathbf{c}}' = q\big(G(\mathbf{c}',\mathbf{z})\big)$, where $q$ is a classifier same in InfoGAN. The principle of Perceptual Simplicity is to punish more on the reconstruction errors for the perturbed dimensions and give more tolerance for the misalignment of the remaining dimensions.

Wei et al. [53] propose an orthogonal Jacobian regularization (OroJaR) to enforce disentanglement for generative models. They employ the Jacobian matrix of the output with respect to the input (i.e., latent variables for representation) to measure the output changes caused by the variations in the input. Assuming that the output changes caused by different dimensions of latent representations are independent with each other, then the Jacobian vectors are expected to be orthogonal with each other, i.e., minimizing Eq. (19),

$$\mathcal{L}_{\text{Jacob}}(G) = \sum_{d=1}^{D} \sum_{i=1}^{m} \sum_{j \neq i} \left| \left[ \frac{\partial G_d}{\partial z_i} \right]^T \frac{\partial G_d}{\partial z_j} \right|^2, \qquad (19)$$

where $G_d$ denotes the $d$-th layer of the generative models and $z_i$ denotes $i$-th dimension in the latent representation.

### 3.1.3 Diffusion-based DRL Methods

DRL can be utilized to disentangle the latent code of the diffusion models. Yang et al. [57] propose an unsupervised disentanglement framework DisDiff whose network structure is shown in Figure 7. They learn a disentangled representation and a disentangled gradient field for each generative factor. Specifically, they assign a separate encoder $E^i_\phi$ for each factor $f^i$ to extract the disentangled representations $z^i$, i.e., $\{z^1, z^2, \ldots, z^N\} = \{E^1_\phi(x_0), E^2_\phi(x_0), \ldots, E^N_\phi(x_0)\}$. Motivated by class-guided [58] sampling of diffusion models, they employ a decoder $G_\psi(x_t, z, t)$ to estimate the score function, i.e., gradient field $\nabla_{x_t} \log p(z^i|x_t)$. Then the sampling process can be written as follows,

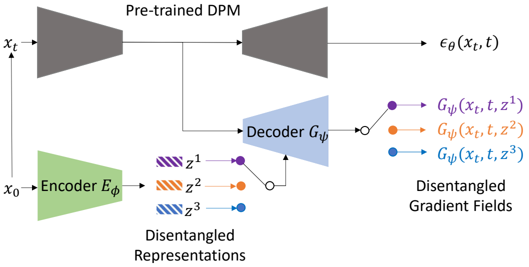

Fig. 7. The network structure of DisDiff, figure from [57].

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t) + \sigma_t \sum_i \nabla_{x_t} \log p(z^i|x_t), \sigma_t). \qquad (20)$$

The authors propose the following disentanglement loss that minimizes the upper bound for the mutual information between $z^i$ and $z^j$, where $i \neq j$:

$$\mathcal{L}_{dis} = \min_{i,j,x_0,\hat{x}^j_0} \|\hat{z}^{i|j} - \hat{z}^i\| - \|\hat{z}^{i|j} - z^i\| + \|\hat{z}^{j|j} - z^j\|, \qquad (21)$$

where $\hat{x}^j_0$ is conditioned on $z^j$ through the sampling process $p_\theta(x_{t-1}|x_t, z^j)$, $\hat{z}^i = E^i_\phi(\hat{x}_0)$, and $\hat{z}^{i|j} = E^i_\phi(\hat{x}^j_0)$. Note that $\hat{x}_0$ denotes the sampled data, while $x_0$ denotes the groundtruth data.

Other works [59], [60], [61] focus on disentangling the conditions of the latent diffusion models for more controllable customized text-to-image and text-to-video generation, with the help of various disentangled losses.

### 3.1.4 Disentangled Latent Directions in Pretrained Generative Models

Most GAN-based and VAE-based DRL methods conform to the paradigm of training from scratch with certain disentanglement loss functions. However,  Recent works [62], [63], [64] have shown semantically meaningful variations when traversing along certain directions in the latent space of pretrained generative models. The phenomenon indicates that there exist certain properties of disentanglement in the latent space of the pretrained generator. Each meaningful and interpretable latent direction aligns with one generative factor. Discovering the disentangled latent directions is a more efficient way to achieve DRL than conventional methods, which can leverage the power of pretrained models and also save resources.

Voynov et al. [64] propose the first unsupervised framework for the discovery of disentangled latent directions in the pretrained GAN latent space. Specifically, they learn a matrix $A \in \mathbb{R}^{d \times K}$ where $d$ denotes the dimension of latent space and $K$ is the total number of generative factors. The $k$-th column represents the direction vector of the $k$-th factor. The direction vectors can be added to a latent code $z$ to transform corresponding generative factors in the image space. To learn $A$, they obtain image pairs in the form of $(G(z), G(z + A(\varepsilon e_k)))$, where $z \sim \mathcal{N}(0, I)$, G denotes the generator, $e_k$ denotes an axis-aligned unit vector, and $\varepsilon$ denotes a shift scalar. A reconstructor R is employed to predict the index $k$ and the shift $\varepsilon$ given the generated pairs $(G(z), G(z + A(\varepsilon e_k)))$. This loss function is shown as follows,

$$\min_{A,R} \mathop{\mathbb{E}}_{z,k,\varepsilon} L(A, R) = \min_{A,R} \mathop{\mathbb{E}}_{z,k,\varepsilon} \left[ L_{cl}(k, \widehat{k}) + \lambda L_r(\varepsilon, \widehat{\varepsilon}) \right], \tag{22}$$

where $\hat{k}, \hat{\varepsilon} = R\Big(G(z), G\big(z + A(\varepsilon e_k)\big)\Big)$. Minimizing this loss will force $A$ to form disentangled direction vectors that are easier to be distinguished by the reconstructor R.

Ren et al. [65] propose an unsupervised framework Disco with contrastive learning for the discovery of latent directions. Similar to Voynov et al. [64], they also utilize a learnable matrix $A \in \mathbb{R}^{d \times K}$ to represent latent directions, with each column representing a candidate direction. Besides, they employ an encoder $E$ to explicitly extract disentangled representations. Note that "disentangled representations" here are not in the original latent space of the pretrained generator $G$, but in a separate space extracted by the encoder. Therefore, the discovered disentangled latent directions can be used to conduct disentangled controllable generation, while the extracted disentangled representations can be applied in downstream tasks. The authors adopt contrastive learning to jointly optimize the latent directions $A$ and the encoder $E$. Specifically, they first construct a variation space by $\boldsymbol{v}(z, k, \varepsilon) = \big|E\big(G(z + A(\varepsilon e_k))\big) - E(G(z))\big|$, where $k$ denotes the direction index and $\varepsilon$ denotes the shift. Then the contrastive loss is employed to pull together the variation samples with the same $k$ and push away the ones from different $k$.

Kwon et al. [66] propose a framework called Asyrp for the image-editing task, by discovering the semantic directions in the space of the deepest feature maps of a pretrained diffusion model. Specifically, they align the direction $\Delta h_i$ to the $i$-th attribute such as "smiling", where $h$ is the deepest feature maps of the pretrained UNet. Then they can edit the attribute by shifting along the direction, i.e., $\tilde{h} = h + \alpha \Delta h_i$. Then this shift will change the UNet output, i.e., the predicted noise $\epsilon_\theta$, and finally change the $i$-th attribute of generated images. They employ a CLIP-based directional loss with cosine distance to optimize $\Delta h_i$.

Besides, DisDiff [57] introduced in Section 3.1.3 can also be regarded as a direction discovery method, as it obtains disentangled latent directions $\nabla_{x_t} \log p(z^i|x_t)$.

#### 3.1.5 Comparison

As we mention in previous sections, VAE-based methods usually suffer from the trade-off between disentanglement performance (i.e., explanation ability) and generative ability (i.e., reconstruction performance). Although GAN-based models usually have a remarkable generative ability, they lack the reversibility property, making them less flexible than VAEs. It seems that the more recent generative models, e.g., diffusion models, are able to combine the advantages of GANs and VAEs. Diffusion models not only possess powerful generative abilities, but also naturally have a friendly reversibility property, demonstrating huge potential to become the mainstream in future DRL research.

### 3.2 Representation Structure

#### 3.2.1 Dimension-wise DRL v.s. Vector-wise DRL

Based on the structure of disentangled representations, we can categorize DRL methods into two groups, i.e., dimension-wise and vector-wise methods. Let $k_1$ and $k_2$ denote the total number of generative factors for dimension-wise and vector-wise methods. Dimension-wise methods adopt a set of disentangled dimensions $\mathbf{z} = \{z_i, i = 0, ..k_1 - 1\}$, where $\mathbf{z}$ is the whole representation and each single dimension $z_i$ (a 1-dimension scalar) represents one fine-grained generative factor. In contrast, vector-wise methods employ $k_2$ disentangled vectors $\mathbf{z}_i, i = 0, .., k_2 - 1$ to represent $k_2$ coarse-grained factors, where the dimension of each vector $\mathbf{z}_i$ equals to or is larger than 2. The comparisons of dimension-wise and vector-wise methods are shown in Figure 8 and Table 2. Dimension-wise methods are always experimented on synthetic and simple datasets, while vector-wise methods are always used in real-world tasks such as identity swapping, image classification, subject-driven generation, and video understanding etc. Synthetic and simple datasets usually have fine-grained latent factors, leading to the applicability of dimension-wise disentanglement. In contrast, for real-world datasets and applications, we usually concentrate on coarse-grained factors (e.g., identity and pose), making it more suitable for vector-wise disentanglement. Dimension-wise methods are mostly early theoretical explorations of DRL, and usually rely on certain model architectures, e.g., VAE or GAN. Vector-wise methods are more flexible, for example, we can design task-specific encoders to extract disentangled vectors and design appropriate loss functions to ensure disentanglement.

The fundamental difference between dimension-wise and vector-wise methods lies in the information capacity determined by the number of dimensions to represent a factor. Specifically, dimension-wise methods usually allocate 1-dimension scalar to represent a generative factor, while vector-wise methods employ more dimensions in the form of vectors to represent generative factors. With more dimensions, vector-wise representations naturally are more powerful than dimension-wise representations to represent complex information. Therefore, we believe this is the main reason why vector-wise DRL is suitable for coarse-grained factors capturing more information while dimension-wise DRL is suitable for fine-grained factors capturing relatively less information. However, we note that granularity is sometimes a relative concept. For example, on the one hand "object size" in simple synthetic datasets is usually a fine-grained factor, which is enough to be represented via a single dimension scalar. On the other hand, "object size" can be a coarse-grained concept in complex real-world scenes by decomposing "object size" into more fine-grained concepts such "width" and "height", which needs vectors with more dimensions to represent. Therefore, we remark that in practice the choice of dimension-wise or vector-wise heavily relies on the complexity of target scenes and tasks.

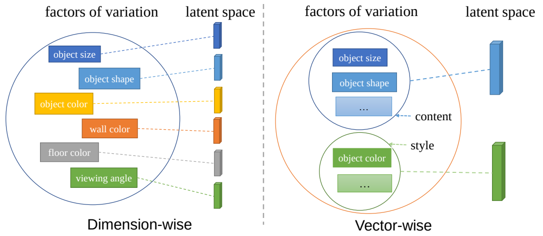

Fig. 8. The illustration of the comparison between dimension-wise and vector-wise DRL.

**Dimension-wise Methods.** A typical architecture that can be used to achieve dimension-wise disentanglement is the VAE-based method, on which we have spent a lot of words in Section 3.1.1. Besides, GAN-based methods are another important paradigm that can achieve dimension-wise DRL, as discussed in Section 3.1.2. Next, we will focus more on various vector-wise methods.

**Vector-wise Methods.** Besides dimension-wise DRL, GAN-based models can also be used to achieve vector-wise DRL on more real-world tasks. Tran et al. [69] propose DR-GAN for pose-invariant face recognition. They use two vectors to represent identity and pose respectively. Specifically, they explicitly set a one-hot latent vector to represent the pose and use an encoder to extract the identity vector from input images. Two Discriminators for pose and identity respectively are used to ensure that the latent vectors can align with the corresponding semantics (i.e., generative factors). Finally, the learned identity vector can be used to conduct pose-invariant face recognition or synthesize identity-preserving faces.

Liu et al. [67] propose a disentangled framework MAP-IVR for activity image-to-video retrieval, which separates video representation into appearance and motion parts. Specifically, they use two video encoders to extract the video motion feature $m^v$ and video appearance feature $a^v$ from the video feature $\bar{v}$. They also use an image encoder to extract the appearance feature of the reference image. They design several objectives to ensure the disentanglement:

$$\mathcal{L}_{orth} = \cos(m^v, a^v), \tag{23}$$

$$\mathcal{L}_{class} = -\log(p(a^v)_y) - \log(p(a^u)_y), \tag{24}$$

$$\mathcal{L}_{re} = \|\bar{v} - \hat{v}\|_2^2. \tag{25}$$

$\mathcal{L}_{orth}$ is a cosine distance loss that facilitates the orthogonality between the video motion and appearance feature. $\mathcal{L}_{class}$ leverages an activity classifier $p$ to ensure that the video appearance feature $a^v$ and the image appearance feature $a^u$ can both capture the activity information. The reconstruction loss $\mathcal{L}_{re}$ also helps the disentanglement, where $\bar{v}$ is the original video feature and $\hat{v}$ is reconstructed by combining $a^u$ and the motion information from $m^v$. The disentangled features can be used for better activity image-to-video retrieval by translating the image reference to a video reference.

Denton et al. propose an autoencoder-based model DR-NET [68] that disentangles each video frame into a time-invariant (content) and a time-varying (pose) component. They use two encoders to extract the content feature and the pose feature respectively. Let $h_c^t$ and $h_p^t$ denote the content feature and the pose feature of the $i$-th frame. The disentanglement objectives are:

$$\mathcal{L}_{re}(D) = ||D(h_c^t, h_p^{t+k}) - x^{t+k}||_2^2, \tag{26}$$

$$\mathcal{L}_{sim}(E_c) = ||E_c(x^t) - E_c(x^{t+k})||_2^2. \tag{27}$$

$\mathcal{L}_{re}(D)$ is the reconstruction loss that aims to ensure combining $h_c^t$ and $h_p^{t+k}$ can reconstruct the frame $x^{t+k}$. $\mathcal{L}_{sim}$ means $h_t^c = E_c(x^t)$ should be invariant across t. The two losses expect $h_c$ to capture the time-invariant content and $h_p$ to capture the time-varying pose. Besides, they also use an adversarial loss to help $h_p$ not carry information about the content. The disentangled representations can be applied in future frames prediction or classification tasks.

Cheng et al. [72] propose DFR that leverages disentangled features to achieve few-shot image classification. They utilize two encoders $E_{cls}$ and $E_{var}$ to extract class-specific and class-irrelevant features, respectively. The disentangled objectives are:

$$\mathcal{L}_{dis} = -\sum_{i=1}^{P} \left(l_i \cdot log(s_i) + (1 - l_i) \cdot log(1 - s_i)\right), \tag{28}$$

$$\mathcal{L}_{cls} = -\sum_{i=1} y_i \log P\left(\hat{y}_i = y_i \mid \mathcal{T}_{FS}\right), \tag{29}$$

$\mathcal{L}_{dis}$ is a discriminative loss that removes the class-specific information in the class-irrelevant feature. $s_i = r_\varphi\left(E_{var}(x_{i1}), E_{var}(x_{i2})\right)$, where $r_\varphi$ is the discriminator which outputs the probability that the pair $x_{i1}$ and $x_{i2}$ are from the same class. They employ a gradient reversal layer to encourage $E_{var}$ to remove the class-specific information. $\mathcal{L}_{cls}$ is the cross-entropy loss for classification loss, where the prediction $\hat{y}_i$ is obtained on the basis of the class-specific feature $E_{cls}(x_i)$. $\mathcal{L}_{cls}$ encourages $E_{cls}$ to capture class-specific information. Besides, they employ a reconstruction loss and a translation loss to further promote disentanglement.

Lee et al. propose a disentangled cross-domain adaptation framework DRANet [21]. They use one encoder to

TABLE 2
The comparisons of dimension-wise and vector-wise methods.

| Methods | Dimension of Each Latent Factor | Representative Works | Semantic Alignment | Applicability |
|---|---|---|---|---|
| Vector-wise | multiple | MAP-IVR [67], DRNET [68], DR-GAN [69], DRANet [21], Lee et al. [8], Liu et al. [22], Singh et al. [70] | each latent variable aligns to one coarse-grained semantic meaning | real scenes |
| Dimension-wise | one | VAE-based methods, InfoGAN [9], IB-GAN [50], Zhu et al. [19], InfoGAN-CR [51], PS-SC GAN [52], Wei et al. [53], DNA-GAN [71] | each dimension aligns to one fine-grained semantic meaning | synthetic and simple datasets |

extract the content feature while obtaining the style feature by subtracting the content feature from the original feature. They also design several losses, e.g., a perceptual loss to enhance disentanglement. Domain adaption can be achieved by combining the content feature with the style feature of the target domain. Gao et al. propose a disentangled identity-swapping framework InfoSwap [73]. They achieve disentanglement by optimizing a loss objective based on the information bottleneck theory. Identity-swapping can be achieved by combining the identity-relevant feature with the target identity-irrelevant feature.

**Discussion.** We have introduced a variety of dimension-wise and vector-wise disentangled representation learning methods. Dimension-wise DRL methods use a single dimension (or several dimensions) to represent one fine-grained generative factor, while vector-wise DRL methods use a single vector to represent one coarse-grained generative factor. A common key point of them is how to enforce disentanglement by designing certain loss objectives, e.g., various regularizers or specifically-designed supervised signals. We will discuss in depth how to design loss objectives for DRL tasks in Sec 5. As for how to determine which kind of structure (i.e., dimension-wise or vector-wise) to use in different tasks, it depends on the number and the granularity of the generative factors we hope to take into account. For example, in many real-world applications, we only need to consider two or several coarse-grained factors. On the other hand, for some specific generative tasks on simple datasets, we need to consider multiple fine-grained factors. The dimension-wise methods are mostly early theoretical exploration for DRL in simple scenes, while in recent years, researchers focus more on how to incorporate vector-wise DRL to tackle real-world applications. It might be a trend to explore the power of vector-wise DRL in realistic tasks.

### 3.2.2 Flat DRL vs. Hierarchical DRL

The aforementioned DRL methods hold an assumption that the architecture of generative factors is flat, i.e., all the factors are parallel and at the same abstraction level. For example, as for dimension-wise DRL, $\beta$-VAE [6] disentangles face rotation, smile, skin color, fringe, etc. on CelebA dataset. InfoGAN [9] disentangles azimuth, elevation, lighting, etc. on 3D Faces dataset. As for vector-wise DRL, DR-GAN [69] disentangles face identity and pose. MAP-IVR [67] disentangles motion and appearance features for video. DisenBooth [59] disentangles the identity-preserved and the identity-irrelevant features. In summary, there doesn't exist a hierarchical structure among these disentangled factors.

However, in practice, generative processes might naturally involve hierarchical structures [74], [75] where the factors of variation have different levels of semantic abstraction, either dependent [74] or independent [75] across levels. For example, the factor controlling *gender* has a higher level of abstraction than the independent factor controlling *eyeshadow* on CelebA dataset [75], while there exist dependencies between factors controlling *shape* (higher level) and *phase* (lower level) on Spaceshapes dataset [74], e.g., the dimension of "phase" is active only when the object shape equals to "moon". To capture these hierarchical structures, a series of works have been proposed to achieve hierarchical disentanglement. Figure 9 demonstrates the paradigm of hierarchical DRL.

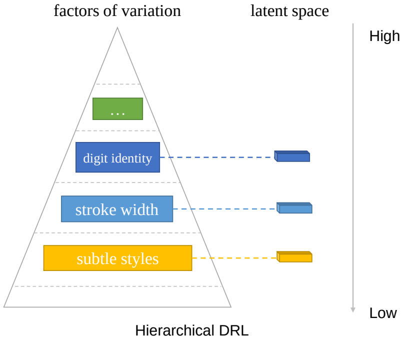

Fig. 9. The illustration of hierarchical DRL. There exists a hierarchical structure among generative factors, i.e., the factors belong to different abstraction levels, resembling a pyramid.

Li et al. [75] propose a VAE-based model which learns hierarchical disentangled representations through formulating the hierarchical generative probability model in Eq. (30),

$$p(\mathbf{x}, \mathbf{z}) = p\left(\mathbf{x} \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L\right) \prod_{l=1}^{L} p\left(\mathbf{z}_l\right), \tag{30}$$

where $\mathbf{z}_l$ denotes the latent representation of the $l$-th level abstraction, and a larger value of $l$ indicates a higher level of abstraction. The authors estimate the level of abstraction with the network depth, i.e., the deeper network layer is responsible for outputting representations with higher abstraction level. It is worth noting that Eq.(30) assumes that there is no dependency among latent representations with different abstraction levels. In other words, each latent representation tends to capture the factors that belong to a single abstraction level, which will not be covered in other levels. The corresponding inference model is formulated in Eq.(31) as follows,

$$q\left(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L \mid \mathbf{x}\right)=\prod_{l=1}^{L} q\left(\mathbf{z}_l \mid \mathbf{h}_l(\mathbf{x})\right), \tag{31}$$

where $\mathbf{h}_l(\mathbf{x})$ represents the abstraction of $l$-th level. In the training stage, the authors design a progressive strategy of learning representations from high to low abstraction levels with modified ELBO objectives. The hierarchical progressive learning is shown in Figure 10, where $h_i$ and $g_i$ are a set of encoders and decoders at different abstraction levels. The framework can disentangle digit identity, stroke width, and subtle digit styles on MNIST dataset, from high to low abstraction levels. It can also disentangle gender, smile, wavy-hair, and eye-shadow on CelebA.

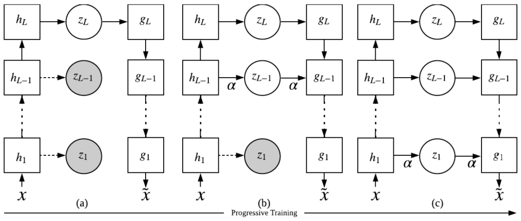

Fig. 10. The architecture of the hierarchical framework proposed by Li et al. [75]. The figure is from the original paper.

Tong et al. [76] propose to learn a set of hierarchical disentangled representations $\mathbf{z}=\{\mathbf{z}_l^i\}_{i=1}^{c_l}$, where $\mathbf{z}_l^i$ is the $i$-th latent variable of the $l$-th layer in the hierarchical structure and $c_l$ is the total number of latent variables of the $l$-th layer. To ensure disentanglement at each hierarchical level, they design a loss function shown in Eq.(32),

$$\mathcal{L}_{\text{disentangle}}=\sum_l \frac{2}{c_l\left(c_l-1\right)} \sum_{i \neq j}^{c_l} \operatorname{dCov}^2\left(\mathbf{z}_l^i, \mathbf{z}_l^j\right), \tag{32}$$

where $\operatorname{dCov}^2(\cdot,\cdot)$ denotes the distance covariance.

Singh et al. [70] propose an unsupervised hierarchical disentanglement framework FineGAN for fine-grained object generation. They design three latent representations for different hierarchical levels, i.e., background code $\mathbf{b}$, parent code $\mathbf{p}$ and child code $\mathbf{c}$, which represent background, object shape and object appearance respectively. Background is the lowest level, followed by shape and appearance. In the generation process, FineGAN first generates a realistic background image by taking $\mathbf{b}$ and noise $\mathbf{z}$ as input. Then it generates the shape and stitches it on top of the background image through taking $\mathbf{p}$ and noise $\mathbf{z}$ as input. Finally, by taking $\mathbf{c}$ as input conditioned on $\mathbf{p}$, the model fills in the shape (parent) outline with appearance (child) details. The authors further employ information theory (similar to InfoGAN) to disentangle the parent (shape) and child (appearance), and use an adversarial loss together with an auxiliary background classification loss to constrain the background generation.

Li et al. [77] propose a hierarchical disentanglement framework for image-to-image translation. They manually organize the labels into a hierarchical tree structure from root to leaves and from high to low level of abstraction, for example, *tags* (e.g., glasses), *attributes* (e.g., with or without), *styles* (e.g., myopic glasses, sunglasses). It is worth noting that the tree hierarchical structure indicates that the child nodes depend on their parents. The authors train a translator module to deal with tags and train an encoder to extract style features.

Ross et al. [74] propose a hierarchical disentanglement framework, which assumes that a group of dimensions may only be active in some cases. Specifically, they organize generative factors as a hierarchical structure (e.g., tree) such that whether a child node can be active depends on the value of its parent node. Take the Spaceshapes dataset as an example, the dimension representing *phase* will only be active when the value of its parent *shape* equals to "moon". They design an algorithm named MIMOSA to train an autoencoder to learn the hierarchical disentangled representations.

**Discussion.** In summary, we can choose the flat or hierarchical DRL methods, depending on whether the generative factors have a hierarchical structure. Although flat DRL methods might also have the potential to disentangle factors from different levels of abstraction, the hierarchical DRL methods with particular designs for the hierarchical structure perform much better. In specific applications, we can consider if there is an underlying hierarchical structure that we can leverage to facilitate disentanglement.

## 3.3 Supervision Signal

### 3.3.1 Unsupervised Methods

The original VAE [16] model demonstrates the possibility of learning latent space in unsupervised manner using Bayesian inference. A class of adversarial generative network models represented by InfoGAN [9] strive to learn explainable representations in an unsupervised manner. Additionally, research grounded in information theory also introduces methods such as mutual information estimation [78] and DeepVIB [79] to disentangle underlying factors and achieve better robustness and generalization ability.

As a primitive stage of development in DRL, unsupervised learning paradigm is built as the original vision for most researchers, which represents a class of intuitive and effective implementation methods of DRL.

### 3.3.2 Supervised Methods

Locatello et al. [80] prove that "pure unsupervised DRL is theoretically impossible without inductive bias on methods and datasets" recently. In other words, disentanglement itself does not occur naturally, which breaks the situation that researchers have been focusing on "unsupervised disentanglement". Locatello et al. [81] later propose that using some of the labeled data for training is beneficial both in terms of disentanglement and downstream performance.

DC-IGN [82] restricts only one factor to be variant and others to be invariant in each mini-batch. One dimension of latent representation $\mathbf{z}$ is chosen as $z_{train}$ which is trained to explain all the variances within the batch and through supervision, thus aligns to the selected variant factor. ML-VAE [36] divides samples into groups according to one selected factor $f_s$, where samples in each group share the same value of $f_s$. This setting is more applicable for some applications such as image-to-image translation, where images in each group share the same label as well as the same posterior of latent variables with respect to $f_s$, which depends on all the samples in the group. While as for other factors except $f_s$, the posterior may be dependent on each individual sample.

Besides, Bengio et al. [83] claim that adaptation speed can evaluate how well a model fits the underlying causal structure from the view of causal inference, and then propose exploiting a meta-learning objective to learn well-represented, disentangled and structured causal representations given an unknown mixture of causal variables.

Xiao et al. [71] propose DNA-GAN, a supervised model whose training procedure similar to gene swap. In concrete, DR-GAN takes a pair of multi-labeled images $I_a$ and $I_b$ with different labels as the input of the encoder. After obtaining the original representations $\mathbf{a}$ and $\mathbf{b}$ of $I_a$ and $I_b$ through an encoder, the swapped representations $\mathbf{a}'$ and $\mathbf{b}'$ are constructed by swapping the value of a particular dimension in the attribute-relevant part of the original representations. After decoding, the reconstruction and the adversarial loss are applied to ensure that each dimension of attribute-relevant representations can align with the corresponding labels. The architecture is shown in Figure 11

Moreover, the guidance from reconstruction loss and task loss can also be regarded as supervised signals, which are widely used in real-world applications. We will elaborately discuss this in the section "Designs of DRL " (Sec. 5).

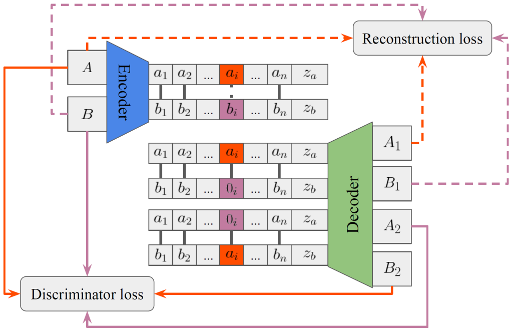
Reconstruction loss
A
B
Encoder
$a_1$ $a_2$ ... $a_i$ ... $a_n$ $z_a$
$b_1$ $b_2$ ... $b_i$ ... $b_n$ $z_b$
$a_1$ $a_2$ ... $a_i$ ... $a_n$ $z_a$
$b_1$ $b_2$ ... $0_i$ ... $b_n$ $z_b$
$a_1$ $a_2$ ... $0_i$ ... $a_n$ $z_a$
$b_1$ $b_2$ ... $a_i$ ... $b_n$ $z_b$
Decoder
$A_1$
$B_1$
$A_2$
$B_2$
Discriminator loss

Fig. 11. The architecture of DNA-GAN, figure from [71].

### 3.3.3 Weakly Supervised Methods

Supervised DRL methods hold the assumption that the target dataset is semantically clear and well-structured to be disentangled into explanatory, independent, and recoverable generative factors [84]. However, in some cases there exist intractable factors which are unclear or difficult to annotate, where these factors are usually regarded as noises unrelated to the target task.

Xiang et al. [84] propose a weakly-supervised DRL framework, DisUnknown, with the setting of $N-1$ factors labeled and 1 factor unknown out of totally $N$ factors. As such, all the intractable factors or task-irrelevant factors can be covered in a single unknown factor. The DisUnknown model is a two-stage method including i) unknown factor distillation and ii) multi-conditional generation, where the first stage extracts the unknown factor by adversarial training and the second stage embeds all labeled factors for reconstruction. They use a set of discriminative classifiers that predict the probability distribution of factor labels to enforce disentanglement, similar to the idea of InfoGAN [9].

### 3.3.4 The Identifiability of DRL

One of the most significant concerns of disentangled representation learning is the identifiability [11], [80], [85]. It is mainly discussed in the context of unsupervised DRL, which focuses on the feasibility of unsupervised DRL. The identifiability indicates whether we can distinguish the disentangled model that we expect to obtain from other entangled ones. Locatello et al. [80] claim that it is impossible to identify the disentangled model by unsupervised learning without inductive biases both on the learning approaches and the data sets, or in other words, unsupervised DRL is impossible without inductive biases. Specifically, let us consider a generative paradigm as follows:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \tag{33}$$

The unsupervised DRL method has access to observations $\mathbf{x}$, i.e., $p(\mathbf{x})$, but it can not identify the true prior of latent variables $\mathbf{z}$ according to the marginal distribution $p(\mathbf{z})$. Theoretically, there is an infinite number of different distributions having the same $p(\mathbf{z})$ [80], [85]. For example, if $p(\mathbf{z})$ is a multivariate Gaussian distribution, which will be invariant to rotation. Therefore, according to Eq.(33), we will obtain infinite equivalent generative models which have the same $p(\mathbf{z})$. Let $\hat{\mathbf{z}}$ denote the latent variable of another generative model, so we have:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}, \tag{34}$$

where $p(\mathbf{z})$ equals to $p(\hat{\mathbf{z}})$. In summary, we can not ensure we actually obtain the disentangled model rather than other equivalent ones.

Therefore, we need extra inductive biases to identify the disentangled model, or we can leverage supervision to help find the target model [80]. As for the aforementioned unsupervised methods, the reason why they can achieve DRL to some extent is that they also have inductive biases, e.g., the regularizers and their regularization strength. Locatello et al. [80] also point out that the disentanglement scores of unsupervised DRL can be easily influenced by randomness and hyper-parameters. As such, although designing appropriate inductive biases is important for unsupervised DRL, it might be more effective to use implicit and explicit supervision.

## 3.4 Independence Assumption

Intuitively, typical DRL models discussed so far hold the assumption that latent factors are statistically independent, so that they are supposed to be independently disentangled through independent or factorial regularization [6], [7] or various disentanglement losses [59], [67]. However, in some cases, underlying generative factors are not independent and hold certain causal relations. In this section, we discuss causal DRL methods that can capture the underlying causal mechanism of the data generation process and potentially achieve more interpretable and robust representations via disentangling causal factors.

Based on the statement from Suter et al. [14], Reddy et al. [86] propose two essential properties that a generative latent variable models (e.g., VAE) should fulfill to achieve causal disentanglement. Consider a latent variable model $M(e, g, p_X)$, where $e$ denotes an encoder, $g$ denotes a generator and $p_x$ denotes a data distribution. Let $G_i$ denote the $i$-th generative factor and $C$ be the confounders in the causal learning literature [87]. The two properties with respect to encoder and generator are presented in the following:

**Property 1**. Encoder $e$ can learn the mapping from $G_i$ to unique $Z_I$, where $I$ is a set of indices and $Z_I$ is a set of latent dimensions indexed by $I$ . The unique $Z_I$ means that $Z_I \cap Z_J = \emptyset, \ \forall\ I \neq J, |I|, |J| \geq 0$. In this case, we assert that $Z$ is unconfounded with respect to $C$, i.e., there is no spurious correlation between $Z_I$ and $Z_J$, $\forall\ I \neq J$.

**Property 2**. For a generative process by $g$ , only $Z_I$ can influence the aspects of generated output controlled by $G_i$, while the others, denoted as $Z_{I^-}$, can not.

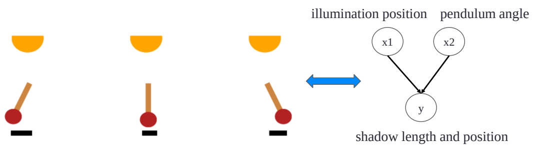

Fig. 12. The position of the illumination source and the angle of the pendulum are causes of the position and the length of the shadow.

Since Locatello et al. [80] challenge the common assumption in the vanilla VAE based DRL approaches that latent variables need to be independent, some following works also attempt to discard the independence assumptions. Yang et al. [11] propose CausalVAE which first introduces structural causal model (SCM) as prior. CausalVAE considers the relationships between the factors of variation in the data from the perspective of causality, describing these relationships with SCM, as is illustrated in Figure. 12. CausalVAE employs an encoder to map the input $\mathbf{x}$ and supervision signal $\mathbf{u}$ associated with the true causal concepts to an independent exogenous variable $\boldsymbol{\epsilon}$ whose prior distribution follows a standard Multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This encoding process is illustrated in Eq. (35),

$$\boldsymbol{\epsilon} = h(\mathbf{x}, \mathbf{u}) + \zeta, \tag{35}$$

where $h$ is the encoder and $\zeta$ is a noise. Then a *Causal Layer* is designed to transforms $\boldsymbol{\epsilon}$ to causal representation $\mathbf{z}$ through the linear structural equation in Eq.(36),

$$\mathbf{z} = \mathbf{A}^\top \mathbf{z} + \boldsymbol{\epsilon} = \left(\mathbf{I} - \mathbf{A}^\top\right)^{-1} \boldsymbol{\epsilon}, \tag{36}$$

where $\mathbf{A}$ is the learnable adjacency matrix of the causal directed acyclic graph (DAG). Before fed into the decoder, $\mathbf{z}$ is passed through a *Mask Layer* to reconstruct itself, as illustrated in Eq.(37), for the $i$-th latent dimension of $\mathbf{z}$, $z_i$,

$$z_i = g_i\left(\mathbf{A}_i \circ \mathbf{z}; \eta_i\right) + \epsilon_i, \tag{37}$$

where $\circ$ represents element-wise product and $g_i$ is a mild nonlinear function with the learnable parameter $\eta_i$. In this mask stage, causal intervention is conducted in the form of "do operation" by setting $z_i$ to a fixed value. After the *Mask Layer*, $\mathbf{z}$ is passed through the decoder to reconstruct the observation $\mathbf{x}$, i.e., $\hat{\mathbf{x}} = \mathbf{d}(\mathbf{z}) + \xi$, where $\xi$ is also a noise.

Bengio et al. [83] point out adaptation speed can evaluate how well a model fits the underlying causal structure from the view of causal inference, and exploit a meta-learning objective to learn disentangled and structured causal representations given unknown mixtures of causal variables.

Different from the supervised scheme of CausalVAE, Shen et al. [43] propose a weakly supervised framework named DEAR, which also introduces SCM as prior. First, the causal representation $\mathbf{z}$ is obtained by an encoder E (or obtained by sampling from prior $p_z$), taking sample $\mathbf{x}$ as input, i.e., $\mathbf{z} = E(\mathbf{x})$. Second, the exogenous variable $\boldsymbol{\epsilon}$ is computed based on the general non-linear SCM proposed by Yu et al. [88] in which the previously calculated $\mathbf{z}$ is employed to define $F_\beta(\boldsymbol{\epsilon})$, as is shown in Eq.(38),

$$\left[\mathbf{z} = f_1\left((\mathbf{I} - \mathbf{A}^\top)^{-1} f_2(\boldsymbol{\epsilon})\right)\right] := F_\beta(\boldsymbol{\epsilon}), \tag{38}$$

where $f_1$ and $f_2$ are element-wise transformations, which are usually non-linear. $A$ is the same learnable adjacency matrix in Eq.(36) and Eq.(37). $\beta$ denotes the parameters of $f_1$, $f_2$ and $A$. When $f_1$ is invertible, Eq.(38) will be equivalent to Eq.(39) in the following:

$$f_1^{-1}(\mathbf{z}) = \mathbf{A}^\top f_1^{-1}(\mathbf{z}) + f_2(\boldsymbol{\epsilon}). \tag{39}$$

Third, we can carry out "do operation" on $\mathbf{z}$ by setting $z_i$ to a fixed value and then reconstruct $\mathbf{z}$ using ancestral sampling by performing Eq.(39) iteratively. Finally, $\mathbf{z}$ is passed through a decoder for reconstruction. To guarantee disentanglement, a weakly supervised loss $L = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[L_s(E; \mathbf{x}, \mathbf{y})\right]$ is applied, only needing a small piece of labeled data, with $L_s = \sum_{i=1}^m \text{CrossEntropy}\left(\bar{E}(x_i), y_i\right)$ when label $y_i$ is binarized or $L_s = \sum_{i=1}^m \left(\bar{E}(x_i) - y_i\right)^2$ when $y_i$ is continuous. Note that $\bar{E}$ is the deterministic part of $E(\mathbf{x})$. When using the VAE structure, $\bar{E}(\mathbf{x}) = m(\mathbf{x})$ is derived with $E(\mathbf{x}) \sim \mathcal{N}(m(\mathbf{x}), \Sigma(\mathbf{x}))$, where $m(\mathbf{x})$ and $\Sigma(\mathbf{x})$ are the mean and variance output by the encoder, respectively.

**Discussion**. Causal models in disentangled representation learning typically involve two main components:

- Structural Causal Models (SCMs): SCMs provide a way to represent causal relationships between variables using a directed acyclic graph (DAG). In an SCM, each node in the graph represents a variable, and the directed edges indicate causal dependencies. The variables can be observed or latent, and the model specifies how the variables interact with each other.
- Interventional Inference: Causal models enable intervention, which means we can manipulate or intervene on specific variables to observe the effects on other variables. Interventions involve changing the value of a specific variable in the model and observing the resulting changes in the other variables. This helps in understanding the causal relationships and in encoding causal mechanisms into disentangled representations.

By incorporating causal models into disentangled representation learning, we can explicitly identify and disentangle the causal factors that influence the observed data. In our opinion, compared to independent DRL, causal DRL methods are better suited for scenarios characterized by the presence of multiple generating factors and potential causal relationships among these factors. Note that learning causal models and disentangled representations is a challenging task. In practice, it may be difficult to accurately specify the underlying causal structure and capture all the causal relationships in the data. Various techniques and algorithms, such as causal inference, causal discovery, and structural equation modeling, are employed in this area to address these challenges.

## 3.5 The Interrelations of DRL

### 3.5.1 The Interrelations with Capsule Nets

Capsule networks [89], [90], [91], [92] introduced by Hinton et al. [90] are an alternative to traditional convolutional neural networks (CNNs), which aim to tackle the limitations of CNNs, e.g., the information loss brought by pooling, and the lack of ability to cope with part-object spatial relationships. Compared to scalar neural units, capsule networks organize the neurons into capsules, each of which is a group of neurons that work together to represent a specific feature, such as pose, color and texture. The capsules can encode not only the presence of an object but also its properties and relationships with other objects. Capsule networks explicitly model the part-whole hierarchical relations through capsules in multiple levels. The lower capsules capture the information at the lower abstraction level, and the higher capsules capture the higher ones. Routing strategies are used to transmit information from the lower capsules to the higher capsules.

The concepts of capsule networks inherently coincide with those of DRL, as they tend to represent various factors of variation as separate capsules and decompose the features of objects into their composing parts. However, it can not be ensured that the representations of each capsule are indeed disentangled [93]. Therefore it is still an open problem to obtain disentangled capsules. Hu et al. [93] propose $beta$-CapsNet to learn disentangled capsules by adding information bottleneck constraints. The results show that $beta$-CapsNet has a better ability to learn disentangled representations than $beta$-VAE and the original CapsNet. In our humble opinions, capsule networks inherently have the potential to achieve DRL for their hierarchical part-whole architecture, providing a suitable framework for learning disentangled representations. Furthermore, it needs extra regularizers or explicit supervisions to ensure the disentanglement of capsules.The combination of capsule networks and DRL has the potential to obtain more interpretable and robust representations of data, which is no doubt a promising research direction.

### 3.5.2 The Interrelations with Object-centric Learning

Different from conventional deep learning models, object-centric learning [94], [95], [96], [97] aims to explicitly model and understand different objects individually, and capture the underlying structure of the scene and relationships of objects. For example, object-centric learning can output the masks of different objects and their object-centric representations that can be used in downstream tasks. Object-centric learning highlights the significance of identifying and reasoning about objects as distinct entities, which can benefit a variety of downstream tasks such as controllable image generation [98], segmentation [99], and visual question answering [100]. Moreover, there are also a series of works that achieve more fine-grained object-centric learning, i.e., further disentangling the representations of each object. For example, Ferraro et al. [101] disentangle shape and pose for each object. Li [102] disentangle several factors for each object such as rotation and color. In summary, object-centric learning can be regarded as a specific instance of DRL which focuses on the disentanglement of individual objects and their properties.

## 3.6 Discussions on Connections over Taxonomy

We present the taxonomy of disentangled representation learning from four aspects: i) model type, ii) representation structure, iii) supervision signal, and iv) independence assumption, together with a lot of related works for each aspect. Although these works may focus on different aspects of DRL, we argue that these four aspects are interrelated, with the common assumption that there exist explainable factors hidden in the data and the final goal is to discover these factors. For instance, the independence assumption is able to influence the model design, where if the factors have causal relationships, we may choose a causal-VAE model instead of vanilla-VAE. Similarly, the representation structure also has an impact on the model type, where the dimension-wise representation structure is more suitable for VAE-based model. Future investigations may simultaneously consider these four aspects and conduct joint optimization over them, to achieve disentanglement on various target tasks.

# 4 DRL APPLICATIONS

## 4.1 Image Generation

On the one hand, the original VAE [16] model learns well-disentangled representations on image generation and reconstruction tasks. Later approaches have achieved more prominent results on image manipulation and intervene through improvement in disentanglement and reconstruction. Representative models such as $\beta$-VAE [6], [34] and FactorVAE [7] can better disentangle independent factors of variation, enabling applicable manipulations of latent variables in the image generation process. JointVAE [32] pays attention to joint continuous and discrete features, which acquires more generalized representations compared with previous methods, thus broadening the scope of image generation to a wider range of fields. CausalVAE [11] introduces causal structure into disentanglement with weak supervision, supporting the generation of images with causal semantics and creation of counterfactual results.

On the other hand, GAN-based disentangled models have also been applied in image generation tasks, benefiting in the high fidelity of GAN. InfoGAN [9], as a typical GAN based model, disentangles latent representation in an unsupervised manner to learn explainable representations and generates images under manipulation, while lacking of stability and sample diversity [6], [7]. Larsen et al. [10] combine VAE and GAN as an unsupervised generative model by i) merging the decoder and the generator into one, ii) using feature-wise similarity measures instead of element-wise errors, which learns high-level visual attributes for image generation and reconstruction in high fidelity, iii) suggesting

TABLE 3
Representatives of disentangled representation learning applications

| Papers | Model | Paradigm | Application |
|---|---|---|---|
| [5], [6], [7], [16], [32], [33], [34], [35] | VAE-based | Unsupervised | Image generation |
| [9], [10], [19], [103] | GAN-based | | |
| [36], [81], [82], [104] | VAE-based | Supervised | |
| [69], [71] | GAN-based | | |
| [11], [43] | Causal-based | | |
| [8], [20], [21], [22], [105] | GAN-based | Unsupervised | Image translation |
| [106] | VAE-based | Supervised | Image classification, segmentation, etc. |
| [107], [108] | Others | Unsupervised | |
| [68], [109] | VAE-based | Unsupervised | Video |
| [67] | Others | Supervised | |
| [23] | Others | Unsupervised | Natural language processing |
| [13], [25] | Others | Supervised | |
| [110], [111], [112] | VAE-based | Unsupervised | Multimodal Application |
| [113], [114] | VAE-based | Supervised | |
| [115] | GAN-based | | |
| [116], [117], [118] | Others | | |
| [26], [28], [119], [120], [121] | VAE-based | Supervised | Recommendation |
| [27], [29], [122] | Others | | |
| [123], [124], [125], [126] | VAE-based | Supervised | Graph |
| [29], [30], [127] | Others | | |

that unsupervised training produces certain disentangled image representations. Zhu et al. [19] utilize GAN architecture to disentangle 3D representations including shape, viewpoint, and texture, to synthesize natural images of objects. Wu et al. [103] analyze disentanglement generation operation in StyleGAN [128], especially in *StyleSpace*, to manipulate semantically meaningful attributes in generation. Zeng et al. [105] propose a hybrid model DAE-GAN, which utilizes a deforming autoencoder and conditional generator to disentangle identity and pose representations from video frames, generating realistic face images of particular poses in a self-supervised manner without manual annotations.

Other works based on information theory also make considerable contributions for long. For example, Gao et al. propose InfoSwap [73], which disentangles identity-relevant and identity-irrelevant information through optimizing information bottleneck to generate more identity-discriminative swapped faces.

## 4.2 Video

Besides static images, DRL also promotes dynamic videos analysis. Denton et al. propose DRNET [68], an autoencoder-based model factorizing each frame into an invariant part and a varying component, which is able to coherently generate future frames in videos. One of the major challenges for video prediction lies in the high-dimensional representation space of visual data. To tackle this problem, Sreekar et al. propose mutual information predictive autoencoder (MIPAE) [129], separating latent representations into time-invariant (content) and time-varying (pose) part, which avoids directly predictions of high dimensional video frames. They use a mutual information loss and a similarity loss to enforce disentanglement, as well as employ LSTM to predict low dimensional pose representations. Latent representations of content and the predicted representations of pose are then decoded to generate future frames. Hsieh et al. later propose DDPAE [109], a framework which also disentangles the content representations and the low-dimensional pose representations. They utilize a pose prediction neural network to predict future pose representations based on the existing pose representations. Based on an inverse spatial transformer parameterized by the predicted pose representations, the invariant content representations can also be used to predict future frames.

## 4.3 Natural Language Processing

### 4.3.1 Text Representation

He et al. [23] apply attention mechanism to an unsupervised neural word embedding model so as to discover meaningful and semantically coherent aspects with strong identification, which improves disentanglement among diverse aspects compared with previous approaches. Bao et al. [24] generate sentences from disentangled syntactic and semantic spaces through modeling syntactic information in the latent space of VAE and regularizing syntactic and semantic spaces via an adversarial reconstruction loss. Cheng et al. [25] propose a disentangled learning framework with partial supervision for NLP, to disentangle the information between style and content of a given text by optimizing the upper bound of mutual information. With the semantic information being preserved, this framework performs well on conditional text generation and text-style transfer. Wu et al. [13] propose a disentangled learning method that optimizes the robustness and generalization ability of NLP models. Colombo et al. [130] propose to learn disentangled representation for text data by minimizing the mutual information between the latent representations of the sentence contents and the attributes. They design a novel variational upper bound based on the Kullback-Leibler and the Renyi divergences to estimate the mutual information.

### 4.3.2 Style Transfer

Hu et al. [131] combine VAE with an attribute discriminator to disentangle content and attributes of the given textual data, for generating texts with desired attributes of sentiment and tenses. John et al. [132] incorporate auxiliary multi-task and adversarial objectives based on VAE to disentangle the latent representations of sentence, achieving high performance in non-parallel text style transfer.

### 4.4 Multimodal Application

With the fast development of multimodal data, there have also been an increasing number of research interests on DRL for multimodal tasks, where DRL is primarily conductive to the separation, alignment and generalization of representations of different modalities.

Early works [110], [113] study the typical modal-level disentanglement through encouraging independence between modality-specific and multimodal factors. Shi et al. [111] posit four criteria for multimodal generative models and propose a multimodal VAE using a mixture-of-experts layer, achieving disentanglement among modalities. Zhang et al. [114] propose a disentangled sentiment representation adversarial network (DiSRAN) to reduce the domain shift of expressive styles for cross-domain sentiment analysis. Recent works [112], [115], [116], [117], [118] tend to focus on disentangling the rich information among multi modalities and leveraging that to perform various downstream tasks. Alaniz et al. [112] propose to use the semantic structure of the text to disentangle the visual data, in order to learn an unified representation between the text and image. The PPE framework [116] realizes disentangled text-driven image manipulation through exploiting the power of the pretrained vision-language model CLIP [133]. Similarly, Yu et al. [115] achieve counterfactual image manipulation via disentangling and leveraging the semantic in text embedding of CLIP. Materzynska et al. [117] disentangle the spelling capabilities from the visual concept processing of CLIP.

### 4.5 Recommendation

Application of DRL in recommendation tasks has also drawn researchers' attention substantially. Latent factors behind user's behaviors can be complicated and entangled in recommender systems. Disentangled factors bring new perspectives, reduce the complexity and improve the efficiency and explainability of recommendation.

DRL in recommendation mostly aims at capturing user's interests of different aspects. Early works [26], [27], [29], [122] focus on learning disentangled representations for collaborative filtering. Specifically, Ma et al. [26] propose MacridVAE to learn the user's macro and micro preference on items, which can be used for controllable recommendation. Wang et al. [29], [122] decomposes the user-item bipartite graph into several disentangled subgraphs, indicating different kinds of user-item relations. Zhang et al. [27] propose to learn users' disentangled interests from both behavioral and content information. More recent works [119], [120], [134] also applied DRL in the sequential recommendations, where the user's future interest are matched with historical behaviors in the disentangled intention space. Additionally, some works [28], [121] also utilize auxiliary information to help the disentangled recommendation. In particular, Wang et al. [28] utilize both visual images and textual descriptions to extract the user interests, providing recommendation explainability from the visual and textual clues. Later they incorporate both visual and categorical information to provide disentangled visual semantics which further boost both recommendation explainability and accuracy [121]. Wang et al. [135] further learn co-disentangled representation across different envrionments for social recommendation.

### 4.6 Graph Representation Learning

Graph representation learning and reasoning methods are being significantly demanded due to increasing applications on various domains dealing with graph structured data, while real-world graph data always carry complex relationships and interdependency between objects [136], [137]. Consequently, research efforts have been devoted to applying DRL to graphs, resulting in beneficial advances in graph analysis tasks.

Ma et al. [30] point out the absence of attention for complex entanglement of latent factors contemporaneously and proposes DisenGCN, which learns disentangled node representations through *neighborhood routing mechanism* iteratively segmenting the neighborhood according to the underlying factors. Later, NED-VAE [123] is proposed to be one unsupervised disentangled method that can disentangle node and edge features from attributed graphs. FactorGCN [124] is then proposed to decompose the input graph into several factor graphs for graph-level disentangled representations. After that, each of the factor graphs is separately fed to the GNN model and then aggregated together for disentangled graph representations. Li et al. [125] first propose to learn disentangled graph representations with self-supervision. They further propose to learn disentangled self-supervised graph representation via explicit enforcing independence between the latent representations so as to improve the quality of disentangled graph representations [126].

### 4.7 Abstract Reasoning

Abstract reasoning in neural networks is proposed by Barrett et al. [138]. Inspired by the human IQ test Raven's Progressive Matrices (RPMs), they explore ways to measure and induce abstract reasoning ability as well as the generalization performance in neural networks. Particularly, Steenkiste et al. [139] investigate the utility of DRL for abstract reasoning tasks. They evaluate the usefulness of the representations learned by DRL models and train abstract reasoning models based on these disentangled representations. They observe that disentangled representations may lead to better downstream performance, e.g., learning faster with fewer samples in tasks similar to RPMs. Locatello et al. [140] learn disentangled representations from pairs of observations, and present adaptive group-based disentanglement methods without requiring annotations of the groups. They demonstrate that disentangled representations can be learned with only weak supervision, which exhibits capabilities beyond statistical correlations and shows effectiveness over abstract visual reasoning tasks. Amizadeh et al. [141] investigate the disentanglement between the reasoning and perception for visual question answering (VQA). They introduce a differentiable first-order logic formalism for VQA that explicitly decouples logical reasoning from visual perception in the process of question answering, which enables the independent evaluation of reasoning and perception.

# 5 DRL Design for Different Tasks

## 5.1 Design of Representation Structure

Given a specific task, we first need to work out the structure of the disentangled representations. To this end, we should consider the structure of the underlying generative factors. Specifically, on the one hand, we should consider whether the factors have a hierarchical or flat structure. If it is the former, we should adopt hierarchical DRL methods, otherwise use the flat DRL. On the other hand, we should consider the number and the granularity of the generative factors that we need to disentangle. As discussed before, dimension-wise DRL methods are suitable for multiple fine-grained factors in simple scenarios, while vector-wise DRL methods are suitable for several coarse-grained factors in more complex and real-world scenarios. In this section, we discuss the designs of the representation structure of DRL according to the taxonomy of "dimension-wise or vector-wise methods".

Consider the two complementary structures: i) dimension-wise: use a whole vector representation $\mathbf{z}$, which is fine-grained. ii) vector-wise: use two or more independent vectors $\mathbf{z}_1, \mathbf{z}_2$... to represent different parts of data features, which is coarse-grained. To guarantee the disentanglement property, method i) usually requires that $\mathbf{z}$ is dimension-wise independent, while method ii) usually requires that $\mathbf{z}_i$ is independent with $\mathbf{z}_j$ where $i \neq j$.

If we choose dimension-wise methods for our application, typical models that we can select are the various VAE-based and GAN-based methods which have been elaborated in Section 3. In this case, we can use VAE or GAN as our backbone and design extra loss functions to adapt to specific tasks. We can also use other model architectures, for example, InfoSwap [73] which uses a multi-layer encoder to extract task-relevant features and compresses the features layer by layer based on information bottleneck to discard task-irrelevant features.

As for vector-wise methods, there are usually two ways of obtaining multiple latent vectors: i) preset these vectors or ii) employ different encoders which take original representations as input to separate the original whole vector into several different vectors. For example, DR-GAN [69] explicitly sets a latent representation to represent pose and uses an encoder to extract identity code from input images, then leverages a supervised loss function to guarantee that the pose code and the identity code can really capture the pose and the identity information correspondingly. Liu et al. [67] leverage two encoders, namely motion encoder and appearance encoder, to respectively extract the motion feature and the appearance feature by passing through the original representation. Cheng et al. [72] utilize two encoders $E_{cls}$ and $E_{var}$ to extract class-specific and class-irrelevant features, respectively. DRNET [68] also uses two encoders to extract the pose feature and content feature, respectively. DRANet [21] employs only one encoder to extract the content feature and then obtains the style feature by subtracting the content feature from the original feature. Similar to DRANet, Wu et al. [142] adopt one encoder to extract domain-invariant features from an image feature map, followed by obtaining domain-specific features through subtracting domain-invariant features.

We have to point out that no matter which model structure is chosen, appropriate loss functions must be designed to guarantee that the representation is disentangled without losing the information carried in the data.

## 5.2 Design of Loss Function

Here, we will discuss the design of loss functions which enforce disentanglement and informativeness according to different model types, i.e., generative model and discriminative model. Overall, we summarize loss functions as $\mathcal{L} = \lambda_1 \mathcal{L}_{re} + \lambda_2 \mathcal{L}_{disen} + \lambda_3 \mathcal{L}_{task}$, where $\mathcal{L}_{re}$ denotes reconstruction loss, $\mathcal{L}_{disen}$ denotes disentanglement loss, and $\mathcal{L}_{task}$ denotes specific task loss.

The reconstruction loss, which is always essential for generation tasks, ensures that the disentangled representation are semantically meaningful and can recover the original data. The disentanglement loss enforces the disentanglement of the representation. Moreover, reconstruction loss can sometimes facilitate disentanglement, as it expects the model to correctly reconstruct the data by these disentangled features. The task loss is directly related to the task objective. In the cases that we jointly optimize the task loss and the disentangled module, the task loss can usually provide guidance for disentanglement. In contrast, if we adopt a two-stage scheme, i.e., firstly training the disentangled module and then applying the disentangled features in downstream tasks, the task loss can't guide the disentangling process.

### 5.2.1 *Generative task*

Various VAE-based models mentioned in Section 3 all have explicit reconstruction loss included in ELBO and also utilize extra regularizers as disentanglement loss. As for GAN-based methods, the adversarial loss can be regarded as reconstruction loss as well, and the disentanglement loss can be mutual information constraints such as those adopted in InfoGAN [9] and IB-GAN [50].

DRANet [21] adopts a $L_1$ loss as the reconstruction loss and uses an adversarial loss as the task loss to ensure the task objective(i.e., image cross-domain adaption). This task loss is computed according to the generated images based on disentangled features, so it also encourages the model to obtain correct disentangled features. It also uses consistency and perceptual losses to enhance the disentanglement.

DRNET [68] adopts a $L_2$ loss as the reconstruction loss and uses a similarity loss together with an adversarial loss to ensure disentanglement. We ignore the task loss, because it is a two-stage scheme that the disentangled representations will be used in downstream tasks such as video prediction.

InfoSwap [73] resorts to an information compression loss based on information bottleneck theory as the disentanglement loss. It uses an adversarial loss as the task loss of face identity swapping, and uses a consistency loss as the reconstruction loss.

MAP-IVR [67] employs a cosine similarity loss to enforce orthogonality between the motion and appearance feature, in addition to the $L_2$ reconstruction loss which ensures the motion feature and the appearance feature capturing the dynamic and static information respectively. MAP-IVR also has no task loss since it is a two-stage scheme. The learned motion and appearance features to tackle the downstream task, e.g., activity image-to-video retrieval.

Besides, several works introduce extra supervision to enforce disentanglement without explicit disentanglement loss function, such as DR-GAN [69] and DNA-GAN [71].

### 5.2.2 Discriminative task

Discriminative tasks sometimes also need reconstruction loss. For example, discriminative tasks can use a VAE backbone to do feature extraction, or leverage reconstruction loss to ensure the disentangled features can correctly recover the data. Discriminative tasks usually do not restrict any specific backbone models, they adopt the latent disentangled representation encoded by appropriate models such as VAE or GAN, based on which the task loss required by the target task such as image classification, recommendation, neural architecture search etc., will be added.

For example, Hamaguchi et al. [106] add similarity loss and activation loss on the basis of using two pairs of VAEs to encode image pairs, which aims to make common features encode invariant factors in an input image pair. The two losses encourage the model to learn common features and specific features of images separately. It uses a classification loss as the task loss to achieve rare event detection, which also guides the training of the disentanglement encoders.

Wu et al. [142] use an orthogonal loss to promote the independence between domain-invariant and domain-specific features. Meanwhile, they also use an adversarial mechanism to encourage the domain-specific features to capture more domain-specific information. They use a detection loss to do domain adaptive object detection.

Cheng et al. [72] use a gradient reverse layer and a class discriminative loss to minimize the class-specific information captured by the class-irrelevant encoder. They use a L1 reconstruction loss and a translation loss to ensure the disentangled features can correctly recover the data. They use a classification to accomplish few-shot image classification.

## 6 Future Directions

**Deriving Better Theoretical Foundations**. Although DRL has been empirically shown to be effective, there still a lack of strict mathematical guarantees for whether particular representations can be completely disentangled, or to which degree can the disentanglement be. Moreover, is it possible to discover any theoretical connections between disentanglement and generalization or robustness? Therefore, the following two points may be a good start for future investigation.

1) Proposing solid mathematical frameworks. Establish a rigorous mathematical foundation that can define and measure disentanglement precisely. This includes understanding the conditions under which disentanglement is possible and beneficial.
2) Connecting to generalization and robustness. Develop theories that explain how disentangled representations improve generalization across different tasks and domains and enhance robustness against adversarial attacks.

**Improving Evaluation Metrics.** It is very important to create more objective, standard, comprehensive and universally accepted benchmarks and metrics for evaluating the quality and effectiveness of disentangled representations. This will help in comparing different approaches more fairly and systematically, pushing forward the future research on DRL.

**Enhancing Explainability towards A New Level**. Indeed one major motivation of DRL is to learn explainable representations of data, however, it is a pity that none of existing DRL approaches are able to fully explain the semantic meaning of latent representations in vector space. As such we believe that continuing to enhance the capability of explanation to next level is still of great importance for further investigations.

1) Human-understandable representations. The ultimate goal of DRL is expected to aim for representations that are not only disentangled but also understandable by humans, facilitating explainable AI. This includes aligning the learned factors with human intuition and semantic meanings.
2) Interactive disentangled representation learning. It may also be interesting to explore interactive learning environments where humans can guide the disentanglement process, ensuring that the learned representations align with user-defined concepts.

**Disentangling Foundation Models**. Foundation models such as *chatGPT* and *Stable Diffusion* are now prevailing in the research community, which are powerful in various downstream tasks. The strength of foundation models mainly comes from large-scale training data and billions of deep neural network parameters. We believe that foundation models may benefit from DRL for the following points.

1) Adaptation to specific tasks. For a specific downstream task, there usually exists redundancy in foundation models. Existing methods always use transfer learning based techniques such as fine-tuning to adapt foundation models to specific tasks. However, transfer learning can not distinguish which part of the knowledge in the pretrained model is relevant to the task, which may result in redundancy as well. DRL is able to disentangle the task-relevant parts and the task-irrelevant parts, having the potential to make the adaption more precise and efficient.
2) Interpretability of Foundation model. Foundation models are powerful and even have the reasoning capability to some extent, however, it is still a black box model that lacks the ability to explain. DRL has the potential to make foundation models more transparent and interpretable. For example, DRL might be able to identify the roles of representations within different network modules and decompose the inference process into human-understandable individual components.

**Exploring the Power of DRL in Advanced Real-world Scenes**. As mentioned before, theoretical research of DRL mainly focuses on simple datasets, which might not keep the path with various advanced applications (e.g., visual question answering, text-to-image generation, and text-to-video generation) and models (e.g., Diffusion Models) which involve real-world datasets and scenes. We believe it is necessary to explore the power of DRL to facilitate these advanced real-world applications as follows.

1) On the one hand, we may need to design more effective model architectures suited for real-world disentanglement that usually involve coarse-grained complex factors. For example, Capsule networks might be a promising direction. Traditional dimension-wise DRL such as VAE-based methods might not be suitable for real-world scenes.
2) On the other hand, we should explore the power of existing ideas of DRL in various advanced tasks, e.g., diffusion text-to-image generation. Considering the factorability, robustness, and generalization brought by DRL might facilitate these tasks.
3) In addition, the interdisciplinary research trending requires us to solve real-world problems across various domains such as healthcare, autonomous vehicles, and finance. This being the case, understanding domain-specific needs can inspire new techniques for DRL.

**Paying Attention to Ethical and Fair AI**. Last but not least, we should always keep in mind to address ethical concerns and biases in AI systems through leveraging disentangled representations to identify and mitigate sources of bias and unfairness as much as we can.

## References

[1] R. Geirhos, et.al., "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[2] Y. Bengio, et.al., "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] B. M. Lake, et.al., "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, 2017.

[4] C. Burgess and H. Kim, "3d shapes dataset," https://github.com/deepmind/3d-shapes, 2018.

[5] R. T. Chen, et.al., "Isolating sources of disentanglement in vaes," in *the 32nd International Conference on Neural Information Processing Systems*, 2019, pp. 2615–2625.

[6] I. Higgins, et.al., "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations* , 2016.

[7] H. Kim and A. Mnih, "Disentangling by factorising," in *ICML*. PMLR, 2018, pp. 2649–2658.

[8] H.-Y. Lee, et.al., "Diverse image-to-image translation via disentangled representations," in *the ECCV (ECCV)*, September 2018.

[9] X. Chen, et.al., "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 2180–2188.

[10] A. B. L. Larsen, et.al., "Autoencoding beyond pixels using a learned similarity metric," in *ICML*, 2016, pp. 1558–1566.

[11] M. Yang, et.al., "Causalvae: Disentangled representation learning via neural structural causal models," in *CVPR* , 2021, pp. 9593–9602.

[12] A. Ghandeharioun, et.al., "Dissect: Disentangled simultaneous explanations via concept traversals," *arXiv preprint arXiv:2105.15164*, 2021.

[13] J. Wu, et.al., "Improving robustness and generality of nlp models using disentangled representations," *arXiv preprint arXiv:2009.09587*, 2020.

[14] R. Suter, et.al., "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *ICML*, 2019, pp. 6056–6065.

[15] J. Lee, et.al., "Learning debiased representation via disentangled feature augmentation," *NeurIPS*, vol. 34, pp. 25 123–25 133, 2021.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] I. Goodfellow, et.al., "Generative adversarial nets," in *NeurIPS*, 2014.

[18] I. Higgins, et.al., "Towards a definition of disentangled representations," *arXiv preprint arXiv:1812.02230*, 2018.

[19] J.-Y. Zhu, et.al., "Visual object networks: Image generation with disentangled 3d representations," in *NeurIPS*, vol. 31, 2018, pp. 118–129.

[20] A. Gonzalez-Garcia, et.al., "Image-to-image translation for cross-domain disentanglement," in *NeurIPS*, vol. 31. 2018.

[21] S. Lee, et.al., "Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation," in *CVPR* ,2021, pp. 15 252–15 261.

[22] Y. Liu, et.al., "Smoothing the disentangled latent style space for unsupervised image-to-image translation," in *CVPR*, 2021, pp. 10 785–10 794.

[23] R. He, et.al., "An unsupervised neural attention model for aspect extraction," in *the 55th ACL*, 2017, pp. 388–397.

[24] Y. Bao, et.al., "Generating sentences from disentangled syntactic and semantic spaces," in *ACL* 2019, pp. 6008–6019.

[25] P. Cheng, et.al., "Improving disentangled text representation learning with information-theoretic guidance," in *ACL*, 2020, pp. 7530–7541.

[26] J. Ma, et.al., "Learning disentangled representations for recommendation," in *NeurIPS*, vol. 32, 2019.

[27] Y. Zhang, et.al., "Content-collaborative disentanglement representation learning for enhanced recommendation," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 43–52.

[28] X. Wang, et.al. "Multimodal disentangled representation for recommendation," in *2021* . IEEE, 2021, pp. 1–6.

[29] X. Wang, et.al., "Disentangled graph collaborative filtering," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1001–1010.

[30] J. Ma, et.al., "Disentangled graph convolutional networks," in *ICML*, 2019, pp. 4212–4221.

[31] X. Liu, et.al., "Learning disentangled representations in the imaging domain," *Medical Image Analysis*, p. 102516, 2022.

[32] E. Dupont, "Learning disentangled joint continuous and discrete representations," in *NeurIPS*, vol. 31. 2018.

[33] M. Kim, et.al., "Relevance factor vae: Learning and identifying disentangled factors," *arXiv preprint arXiv:1902.01568*, 2019.

[34] C. P. Burgess, et.al., "Understanding disentangling in $\beta$-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[35] A. Kumar, et.al., "Variational inference of disentangled latent concepts from unlabeled observations," in *International Conference on Learning Representations*, 2018.

[36] D. Bouchacourt, et.al., "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *AAAI Conference on Artificial Intelligence*, 2018.

[37] S. Bing, et.al., "On disentanglement in gaussian process variational autoencoders," *arXiv preprint arXiv:2102.05507*, 2021.

[38] H. Caselles-Dupré, et.al., "Symmetry-based disentangled representation learning requires interaction with environments," *NeurIPS*, vol. 32, 2019.

[39] R. Quessard, et.al., "Learning disentangled representations and group structure of dynamical environments," *NeurIPS*, vol. 33, pp. 19 727–19 737, 2020.

[40] T. Yang, et.al., "Towards building a group-based unsupervised representation disentanglement framework," in *International Conference on Learning Representations*, 2022.

[41] T. Wang, et.al., "Self-supervised learning disentangled group representation as feature," *NeurIPS*, vol. 34, 2021.

[42] J. Pearl, *Causality*. Cambridge university press, 2009.

[43] X. Shen, et.al., "Disentangled generative causal representation learning," *arXiv preprint arXiv:2010.02637*, 2020.

[44] Y. Lecun, et.al., "Gradient-based learning applied to document recognition," *the IEEE*, no. 11, pp. 2278–2324, 1998.

[45] Z. Liu, et.al., "Deep learning face attributes in the wild," in *International Conference on Computer Vision (ICCV)*, December 2015.

[46] M. Aubry, et.al., "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *CVPR*, 2014.

[47] Y. Li and S. Mandt, "Disentangled sequential autoencoder," *arXiv preprint arXiv:1803.02991*, 2018.

[48] M. Arjovsky, et.al., "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[49] X. Zhu, et.al., "Commutative lie group vae for disentanglement learning," in *ICML*, 2021, pp. 12 924–12 934.

[50] I. Jeon, et.al., "Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks," in *35th AAAI Conference on Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence*. ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE, 2021, pp. 7926–7934.

[51] Z. Lin, et.al., "Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers," 2019.

[52] X. Zhu, et.al., "Where and what? examining interpretable disentangled representations," in *CVPR*, 2021, pp. 5861–5870.

[53] Y. Wei, et.al., "Orthogonal jacobian regularization for unsupervised disentanglement in image generation," in *CVPR*, 2021, pp. 6721–6730.

[54] R. Rombach, et.al., "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[55] B. Kawar, et.al., "Imagic: Text-based real image editing with diffusion models," in *CVPR*, 2023, pp. 6007–6017.

[56] J. Z. Wu, et.al., "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *IEEE International Conference on Computer Vision*, 2023, pp. 7623–7633.

[57] T. Yang, et.al., "Disdiff: Unsupervised disentanglement of diffusion probabilistic models," *arXiv preprint arXiv:2301.13721*, 2023.

[58] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *NeurIPS*, vol. 34, pp. 8780–8794, 2021.

[59] H. Chen, et.al., "Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation," in *International Conference on Learning Representations* , 2024.

[60] H. Chen, et.al., "Disendreamer: Subject-driven text-to-image generation with sample-aware disentangled tuning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[61] H. Chen, et.al., "Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning," *arXiv preprint arXiv:2311.00990*, 2023.

[62] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *CVPR*, 2021, pp. 1532–1540.

[63] V. Khrulkov, et.al., "Disentangled representations from non-disentangled models," *arXiv preprint arXiv:2102.06204*, 2021.

[64] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *ICML*, 2020, pp. 9786–9796.

[65] X. Ren, et.al., "Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view," in *International Conference on Learning Representations*, 2021.

[66] M. Kwon, et.al., "Diffusion models already have a semantic latent space," in *The Eleventh International Conference on Learning Representations*, 2022.

[67] L. Liu, et.al., "Activity image-to-video retrieval by disentangling appearance and motion," in *Proc. AAAI*, 2021, pp. 1–9.

[68] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *NeurIPS*, 2017, pp. 4417–4426.

[69] L. Tran, et.al., "Disentangled representation learning gan for pose-invariant face recognition," in *the CVPR*, 2017, pp. 1415–1424.

[70] K. K. Singh, et.al., "Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," in *CVPR*, 2019, pp. 6490–6499.

[71] T. Xiao, et.al., "Dna-gan: Learning disentangled representations from multi-attribute images," *arXiv preprint arXiv:1711.05415*, 2017.

[72] H. Cheng, et.al., "Disentangled feature representation for few-shot image classification," *arXiv preprint arXiv:2109.12548*, 2021.

[73] G. Gao, et.al., "Information bottleneck disentanglement for identity swapping," in *the CVPR*,2021, pp. 3404–3413.

[74] A. Ross and F. Doshi-Velez, "Benchmarks, algorithms, and metrics for hierarchical disentanglement," in *ICML*, 2021, pp. 9084–9094.

[75] Z. Li, et.al., "Progressive learning and disentanglement of hierarchical representations," *arXiv preprint arXiv:2002.10549*, 2020.

[76] B. Tong, et.al., "Hierarchical disentanglement of discriminative latent features for zero-shot learning," in *CVPR*, 2019, pp. 11 467–11 476.

[77] X. Li, et.al., "Image-to-image translation via hierarchical style disentanglement," in *CVPR*, 2021, pp. 8639–8648.

[78] R. D. Hjelm, et.al., "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations* , 2018.

[79] A. A. Alemi, et.al., "Deep variational information bottleneck," in *International Conference on Learning Representations* , 2017.

[80] F. Locatello, et.al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in *ICML*, 2019, pp. 4114–4124.

[81] F. Locatello, et.al., "Disentangling factors of variations using few labels," in *International Conference on Learning Representations*, 2020.

[82] T. D. Kulkarni, et.al., "Deep convolutional inverse graphics network," *arXiv preprint arXiv:1503.03167*, 2015.

[83] Y. Bengio, et.al., "A meta-transfer objective for learning to disentangle causal mechanisms," in *International Conference on Learning Representations*, 2020.

[84] S. Xiang, et.al., "Disunknown: Distilling unknown factors for disentanglement learning," in *CVPR*, 2021, pp. 14 810–14 819.

[85] I. Khemakhem, et.al., "Variational autoencoders and nonlinear ica: A unifying framework," in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 2207–2217.

[86] A. G. Reddy, et.al., "On causally disentangled representations," *arXiv preprint arXiv:2112.05746*, 2021.

[87] S. Greenland, et.al., "Confounding and collapsibility in causal inference," *Statistical science*, vol. 14, no. 1, pp. 29–46, 1999.

[88] Y. Yu, et.al., "Dag-gnn: Dag structure learning with graph neural networks," in *ICML*, 2019, pp. 7154–7163.

[89] F. D. S. Ribeiro, et.al., "Learning with capsules: A survey," *arXiv preprint arXiv:2206.02664*, 2022.

[90] S. Sabour, et.al., "Dynamic routing between capsules," *NeurIPS*, vol. 30, 2017.

[91] M. K. Patrick, et.al., "Capsule networks–a survey," *Journal of King Saud University-computer and information sciences*, vol. 34, no. 1, pp. 1295–1310, 2022.

[92] V. Mazzia, et.al., "Efficient-capsnet: Capsule network with self-attention routing," *Scientific reports*, vol. 11, no. 1, p. 14634, 2021.

[93] M.-f. Hu and J.-w. Liu, "$\beta$-capsnet: learning disentangled representation for capsnet by information bottleneck," *Neural Computing and Applications*, vol. 35, no. 3, pp. 2503–2525, 2023.

[94] F. Locatello, et.al., "Object-centric learning with slot attention," *NeurIPS*, vol. 33, pp. 11 525–11 538, 2020.

[95] M. Seitzer, et.al., "Bridging the gap to real-world object-centric learning," in *The Eleventh International Conference on Learning Representations*, 2022.

[96] G. Singh, et.al., "Simple unsupervised object-centric learning for complex and naturalistic videos," *NeurIPS*, vol. 35, pp. 18 181–18 196, 2022.

[97] G. Elsayed, et.al., "Savi++: Towards end-to-end object-centric learning from real-world videos," *NeurIPS*, vol. 35, pp. 28 940–28 954, 2022.

[98] T. Sylvain, et.al., "Object-centric image generation from layouts," in *the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2647–2655.

[99] B. Cheng, et.al., "Boundary iou: Improving object-centric image segmentation evaluation," in *CVPR*, 2021, pp. 15 334–15 342.

[100] Z. Wu, et.al., "Slotformer: Unsupervised visual dynamics simulation with object-centric models," *arXiv preprint arXiv:2210.05861*, 2022.

[101] S. Ferraro, et.al., "Disentangling shape and pose for object-centric deep active inference models," in *International Workshop on Active Inference*. Springer, 2022, pp. 32–49.

[102] N. Li, et.al., "Learning object-centric representations of multi-object scenes from multiple views," *NeurIPS*, vol. 33, pp. 5656–5666, 2020.

[103] Z. Wu, et.al., "Stylespace analysis: Disentangled controls for stylegan image generation," in *CVPR* , 2021, pp. 12 863–12 872.

[104] F. Träuble, et.al., "On disentangled representations learned from correlated data," in *ICML*, 2021, pp. 10 401–10 412.

[105] X. Zeng, et.al., "Realistic face reenactment via self-supervised disentangling of identity and pose," in *the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 757–12 764.

[106] R. Hamaguchi, et.al., "Rare event detection using disentangled representation learning," in *CVPR* , 2019.

[107] Z. Feng, et.al., "Self-supervised representation learning by rotation feature decoupling," in *CVPR* , 2019.

[108] E. H. Sanchez, et.al., "Learning disentangled representations via mutual information estimation," in *ECCV*, 2020, pp. 205–221.

[109] J.-T. Hsieh, et.al., "Learning to decompose and disentangle representations for video prediction," *NeurIPS*, vol. 31, 2018.

[110] W.-N. Hsu and J. Glass, "Disentangling by partitioning: A representation learning framework for multimodal sensory data," *arXiv preprint arXiv:1805.11264*, 2018.

[111] Y. Shi, et.al., "Variational mixture-of-experts autoencoders for multi-modal deep generative models," *NeurIPS*, vol. 32, 2019.

[112] S. Alaniz, et.al., "Compositional mixture representations for vision and text," in *CVPR Workshops*, 2022, pp. 4202–4211.

[113] Y.-H. H. Tsai, et.al., "Learning factorized multimodal representations," in *International Conference on Learning Representations*, 2018.

[114] Y. Zhang, et.al., "Learning disentangled representation for multimodal cross-domain sentiment analysis," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[115] Y. Yu, et.al., "Towards counterfactual image manipulation via clip," in *ACM Multimedia*, 2022, pp. 3637–3645.

[116] Z. Xu, et.al., "Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model," in *CVPR*,2022, pp. 18 229–18 238.

[117] J. Materzyńska, et.al., "Disentangling visual and written concepts in clip," in *CVPR*,2022, pp. 16 410–16 419.

[118] W. Zou, et.al., "Utilizing bert intermediate layers for multimodal sentiment analysis," in *2022 ICME*. IEEE, 2022, pp. 1–6.

[119] J. Ma, et.al., "Disentangled self-supervision in sequential recommenders," in *the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 483–491.

[120] H. Chen, et.al., "Curriculum disentangled recommendation with noisy multi-feedback," *NeurIPS*, vol. 34, pp. 26 924–26 936, 2021.

[121] X. Wang, et.al., "Disentangled representation learning for recommendation," *IEEE TPAMI*, 2022.

[122] L. Hu, et.al., "Graph neural news recommendation with unsupervised preference disentanglement," in *ACL*, 2020, pp. 4255–4264.

[123] X. Guo, et.al., "Interpretable deep graph generation with node-edge co-disentanglement," in *the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 1697–1707.

[124] Y. Yang, et.al., "Factorizable graph convolutional networks," *NeurIPS*, vol. 33, pp. 20 286–20 296, 2020.

[125] H. Li, et.al., "Disentangled contrastive learning on graphs," *NeurIPS*, vol. 34, pp. 21 872–21 884, 2021.

[126] H. Li, et.al., "Disentangled graph contrastive learning with independence promotion," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[127] H. Li, et.al., "Ood-gnn: Out-of-distribution generalized graph neural network," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[128] T. Karras, et.al., "A style-based generator architecturanetre for generative adversarial networks," in *CVPR* ,2019.

[129] P. A. Sreekar, et.al., "Mutual information based method for unsupervised disentanglement of video representation," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6396–6403.

[130] P. Colombo, et.al., "A novel estimator of mutual information for learning to disentangle textual representations," in *the 59th ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6539–6550.

[131] Z. Hu, et.al., "Toward controlled generation of text," in *ICML*, 2017, pp. 1587–1596.

[132] V. John, et.al., "Disentangled representation learning for non-parallel text style transfer," in *ACL*, 2019, pp. 424–434.

[133] A. Radford, et.al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[134] Y. Zhang,et.al., "Adaptive disentangled transformer for sequential recommendation," in *the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3434–3445.

[135] X. Wang, et.al., "Curriculum co-disentangled representation learning across multiple environments for social recommendation," in *ICML*, 2023.

[136] Z. Zhang, et.al., "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[137] J. Zhou, et.al., "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[138] D. Barrett, et.al., "Measuring abstract reasoning in neural networks," in *ICML*, 2018, pp. 511–520.

[139] S. Van Steenkiste, et.al., "Are disentangled representations helpful for abstract visual reasoning?" *NeurIPS*, vol. 32, 2019.

[140] F. Locatello, et.al., "Weakly-supervised disentanglement without compromises," in *ICML*, 2020, pp. 6348–6359.

[141] S. Amizadeh, et.al., "Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning"," in *ICML*, 2020, pp. 279–290.

[142] A. Wu, et.al., "Vector-decomposed disentanglement for domain-invariant object detection," in *CVPR*, 2021, pp. 9342–9351.

**Xin Wang** is currently an Associate Professor at the Department of Computer Science and Technology, Tsinghua University. He got both of his Ph.D. and B.E degrees in Computer Science and Technology from Zhejiang University, China. He also holds a Ph.D. degree in Computing Science from Simon Fraser University, Canada. His research interests include multimedia intelligence, machine learning and its applications. He is the recipient of ACM China Rising Star Award, IEEE TCMC Rising Star Award and DAMO Academy Young Fellow.

**Hong Chen** received B.E. from the Department of Electronic Engineering, Tsinghua University, Beijing, China in 2020. He is currently a PH.D. candidate in the Department of Computer Science and Technology of Tsinghua University. His main research interests include auxiliary learning and multi-modal learning. He has published several papers in top conferences and journals including NeurIPS, ICML, IEEE TPAMI, etc.

**Si'ao Tang** is a master candidate at Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, majored in Data Science and Information Technology. His research interests include machine learning, multimedia intelligence, video understanding, etc.

**Zihao Wu** is currently working toward the master's degree in computer science and technology with Tsinghua University, Beijing, China. He recieved his B.E. degree from the Department of Computer Science, Tongji University. His research interests include machine learning, multimedia intelligence, and recommendation.

**Wenwu Zhu** is currently a Professor in the Department of Computer Science and Technology at Tsinghua University. He also serves as the Vice Dean of National Research Center for Information Science and Technology, and the Vice Director of Tsinghua Center for Big Data. He served as EiC for IEEE Transactions on Multimedia from 2017-2019. He served in the steering committee for IEEE Transactions on Multimedia (2015-2016) and IEEE Transactions on Mobile Computing (2007-2010), respectively.He is an AAAS Fellow, IEEE Fellow, SPIE Fellow, and a member of The Academy of Europe (Academia Europaea).