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, data mining etc. In this article, we comprehensively review DRL from various aspects including motivations, definitions, methodologies, evaluations, applications and model designs. We discuss works on DRL based on two well-recognized definitions, i.e., Intuitive Definition and Group Theory Definition. We further categorize the methodologies for DRL into four groups, i.e., Traditional Statistical Approaches, Variational Auto-encoder Based Approaches, Generative Adversarial Networks Based Approaches, Hierarchical Approaches and Other Approaches. 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 number of attentions 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 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

- *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 is supported by the National Key Research and Development Program of China No.2020AAA0106300 and National Natural Science Foundation of China No. 62222209, 62250008, 62102222.*

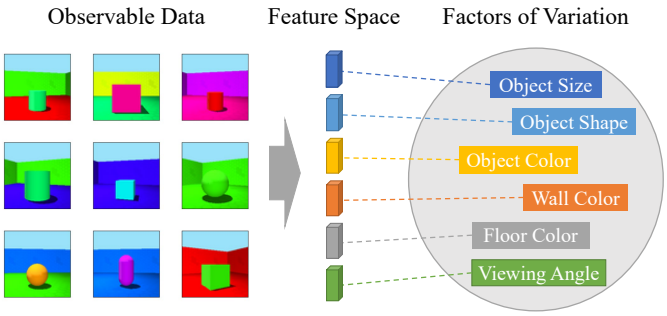

Fig. 1. The scene of Shape3D [4], where the six rectangles in the gray circle represent the six factors of variation in the Shape3D respectively. DRL is expected to encode these distinct factors with independent latent variables in the latent feature space.

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. In particular, we present the definitions of DRL in Section 2 and comprehensively review DRL approaches in Section 3. In Section 4, we discuss popular evaluation metrics for DRL implementation. We discuss the applications of DRL for various downstream tasks in Section 5, followed by our insights in designing proper DRL models for different tasks in Section 6. Last but not least, we summarize several open questions and future directions for DRL in Section 7. 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 (Section 3.2.2).

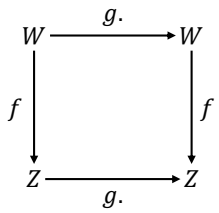

Fig. 2. The illustration of condition (ii).

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 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 via considering that there might be an underlying causal structure which renders generative factors dependent. 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. Nevertheless, this topic is not the focus of this paper and we leave it as an open question for further explorations.

3 DRL Taxonomy

In this section, we in detail present DRL approaches through categorizing them into five groups: i.e., traditional

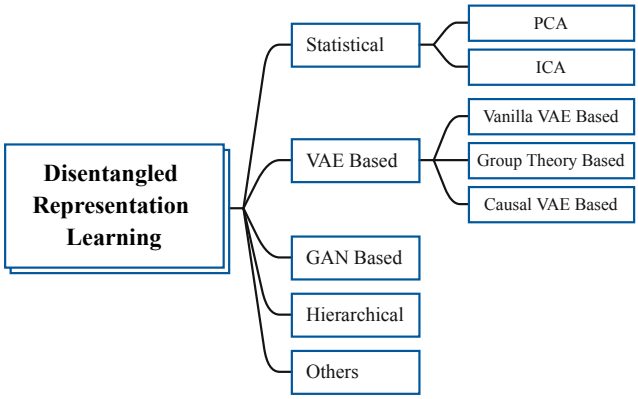

Fig. 4. A categorization of DRL approaches.

statistical approaches, VAE based approaches, GAN based approaches, hierarchical approaches and other methods.

3.1 Traditional Statistical Approaches

Though not being equipped with deep architectures, several traditional statistical approaches have always been effective in disentangling latent factors in the vector space, among which *Principal Component Analysis* and *Independent Component Analysis* are the most two representative algorithms. Although these shallow models are not the focus of this paper, we still provide brief descriptions for completeness. Interested readers may refer to more statistics literature for details.

3.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [44] is a well-established method for dimension reduction. PCA linearly transforms the original data into linearly independent representation in each dimension to extract the principle feature components of the given data. The idea of PCA to obtain linearly independent representation resembles the independence assumption of DRL. Though PCA can be applied to learn disentangled representation, there exist some bottlenecks. For example, PCA is non-parametric so that the optimization procedure can not be tailored for specific tasks. Moreover, PCA is only effective for data samples that are generated following Gaussian distribution. There is some research (e.g., [45]) pointing out that the similarities between VAE and PCA could possibly be one explanation for the reason why VAE to some extent is capable of learning disentangled representations.

3.1.2 Independent Component Analysis (ICA)

Independent Component Analysis (ICA) [46] looks for latent factors or components that are statistically independent and non-Gaussian in multivariate statistics. The underlying assumption is that the observed signals X are generated by the combinations of several statistically independent non-Gaussian components S which are required to be recovered. The independence assumption of ICA also resembles the idea of DRL. For DRL, under the common assumption that high dimension observations are generated by a number of independent latent factors through certain non-linear functions, we always consider non-linear ICA. In non-linear ICA, the generation process can be written as $X = f(S \mid \theta) + n$, where f is a non-linear mixing function and n is noise. The main bottleneck of non-linear ICA Lies in its incapability of identifying disentangled solution with entangled ones without any assumed conditions [47]. To tackle this identifiability problem, Horan et al. [48] propose a constraint of local isometry about the mapping from latent space to observation space.

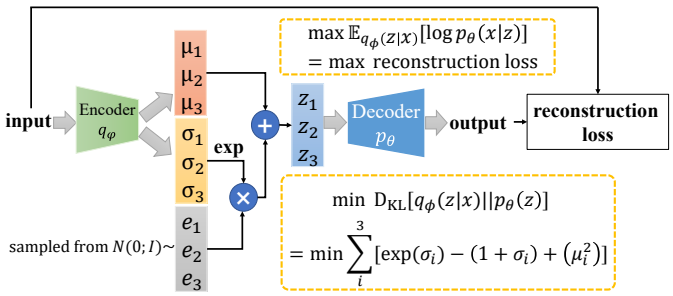

Fig. 5. The general framework of variational auto-encoder (VAE).

3.2 Variational Auto-encoder (VAE) Based Approaches

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 [49]). To obtain better disentanglement performance, researchers design various extra regularizers to combine with the original VAE loss function, resulting in the family of VAE Based Approaches.

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,

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

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,

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

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 constraint on the representations learned though neural network [5], which may be the reason that VAE has the potential ability of disentanglement.

3.2.1 Vanilla VAE Based Methods

Although the essential design of VAE provides the potential ability to disentangle, we observe that VAE shows poor disentanglement capability on complex datasets such as CelebA [50] and 3D Chairs [51] 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., β-VAE [6], DIP-VAE [35], and β-TCVAE [5] etc.). Specifically, to strengthen the independence constraint of the variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$, β-VAE [6] introduces a β penalty coefficient before the KL term in ELBO, where the updated objective function is shown in Eq.(4).

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

When β=1, β-VAE degenerates to the original VAE formulation. The experimental results of β-VAE [6] show that larger values of β encourage learning more disentangled representations while harming the performance of reconstruction. Therefore, it is important to select an appropriate β 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 β 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 β-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 β-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 $\max \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$. $D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})\big)$ 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,

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

where γ 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,

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

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,

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

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

where λ_d and λ_{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),

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

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 informa-

tion 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.

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

From Eq.(12), we can straightforwardly obtain the explanation of the trade-off in β-VAE, i.e., higher β 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 β, we can penalize these three terms with three different coefficients respectively, which is referred as β-TCVAE and is shown in Eq.(13) as follows.

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

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. They separate the latent variables into two subsets, i.e., $\mathbf{R}$ (relevant variables) and $\mathbf{N}$ (nuisance variables). Different from the original FactorVAE, RF-VAE only focuses on the relevant part when computing the total correlation. Moreover, the KL loss between posterior $q_\phi(\mathbf{z})$ and prior $p_\theta(\mathbf{z})$ is handled differently by penalizing less for relevant dimensions and more for nuisance dimensions, which follows the intuition that the posterior of the nuisance part should be independent of input sample $\mathbf{x}$. When $\mathbf{R}$ and $\mathbf{N}$ are known in advance, the objective function can be formulated in Eq.(14)) as follows.

$$\begin{aligned} \mathcal{L} =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - \sum_{j=1}^{d}\lambda_j D_{KL}\big(q_\phi\left(z_j|\mathbf{x}\right)\|p_\theta\left(z_j\right)\big) \\ &- \gamma D_{KL}\Big(q_\phi(\mathbf{z_R})||\prod_{j\in\mathbf{R}} q_\phi(z_j))\Big), \\ \text{where} \quad \lambda_j =& \begin{cases} \lambda_{\min} & \text{if } j\in\mathbf{R} \\ \lambda_{\max} & \text{if } j\in\mathbf{N} \end{cases} \quad (\lambda_{\min}<\lambda_{\max}), \end{aligned} \tag{14}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are predefined hyperparameters. When $\mathbf{R}$ and $\mathbf{N}$ are not accessible, a learnable relevance vector $\mathbf{r}$ is employed, where r_j = 1 indicates that z_j is a relevant factor and r_j = 0 indicates that z_j is a nuisance factor. The objective function is shown in Eq.(15) as follows,

$$\begin{aligned} \mathcal{L} =& \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - \sum_{j=1}^{d}\lambda\left(r_j\right) D_{KL}\big(q_\phi\left(z_j|\mathbf{x}\right)\|p_\theta\left(z_j\right)\big) \\ &- \gamma D_{KL}\big(q_\phi(\mathbf{r}\circ\mathbf{z})\|\prod_{j=1}^{d} q_\phi(r_j\circ z_j)\big) - \eta\|\mathbf{r}\|_1, \end{aligned} \tag{15}$$

where $\circ$ denotes element-wise product and $\lambda(\cdot)$ is a monotone decreasing function with $\lambda(0)=\lambda_{max}>\lambda_{min}=\lambda(1)$. $\eta\|\mathbf{r}\|_1$ is the L1 regularizer which penalizes the situations where too many dimensions are chosen as relevant, thus encouraging minimal redundancy.

The aforementioned VAE based methods are designed for continuous latent variables, failing to model the discrete variables. Dupont et al. [32] propose a β-VAE based framework, JointVAE, which is capable of disentangling both continuous and discrete representations in an unsupervised manner. They separate latent variables into continuous $\mathbf{z}$ and discrete $\mathbf{c}$, assuming the continuous and discrete latent variables are conditionally independent. Therefore, the objective function Eq.(16) can be extended from the modified β-VAE function Eq.(6), where C_z and C_c are gradually increased during training.

$$\begin{aligned} \mathcal{L}(\theta,\phi) =& \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{c}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{c})\big] \\ &- \gamma\big|D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big) - C_z\big| \\ &- \gamma\big|D_{KL}\big(q_\phi(\mathbf{c}|\mathbf{x})\|p(\mathbf{c})\big) - C_c\big|. \end{aligned} \tag{16}$$

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.

It is worth noting that we can also utilize supervised signals to obtain more disentangled and nicely aligned latent representations if applicable. For example, DC-IGN [52] 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. For another example, 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.

To further conduct DRL on sequential data such as video or audio, Li [53] et al. modify the original VAE model to adapt sequential data. Considering the temporal nature of sequences, they separate latent representation into time-invariant and time-varying part. The probabilistic generative model is shown in Eq.(17) as follows,

TABLE 1
The summary of VAE based approaches.

Method	Regularizer	Description
β-VAE	$-\beta D_{KL}\big(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\big)$	β controls the trade-off between reconstruction fidelity and the quality of disentanglement in latent representations.
Understanding disentangling in β-vae	$-\gamma\left\|D_{KL}\big(q_\phi(\mathbf{z}\|\mathbf{x})\|p(z)\big)-C\right\|$	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	$-\lambda D_{KL}\big(q_\phi(\mathbf{z})\|p(\mathbf{z})\big)$	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	$-\gamma D_{KL}\big(q_\phi(\mathbf{z})\|\prod_j q_\phi(z_j)\big)$	Directly impose independence constraint on $q_\phi(\mathbf{z})$ in the form of total correlation.
β-TCVAE	$-\alpha I_q(\mathbf{z};\mathbf{x})-\beta D_{KL}\big(q(\mathbf{z})\|\prod_j q(z_j)\big)-\gamma\sum_j D_{KL}\big(q(z_j)\|p(z_j)\big)$	Decompose $D_{KL}\big(q(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\big)$ into three terms: i) mutual information, ii) total correlation, iii) dimension-wise KL divergence and then penalize them respectively.
JointVAE	$-\gamma\left\|D_{KL}\big(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\big)-C_z\right\|-\gamma\left\|D_{KL}\big(q_\phi(\mathbf{c}\|\mathbf{x})\|p(\mathbf{c})\big)-C_c\right\|$	Separate latent variables into continuous $\mathbf{z}$ and discrete $\mathbf{c}$, then modify the objective function of β-VAE to capture discrete generative factors.
RF-VAE	$-\sum_{j=1}^{d}\lambda(r_j)D_{KL}\big(q(z_j\|\mathbf{x})\|p(z_j)\big)-\gamma D_{KL}\big(q(\mathbf{r}\circ\mathbf{z})\|\prod_{j=1}^{d}q(r_j\circ z_j)\big)-\eta\|\mathbf{r}\|_1$	Introduce relevance indicator variables $\mathbf{r}$ by only focusing on relevant part when computing the total correlation, penalize $D_{KL}\big(q(z_j\|\mathbf{x})\|p(z_j)\big)$ less for relevant dimensions and more for nuisance (noisy) dimensions.

$$p_\theta\left(\mathbf{x_{1:T}},\mathbf{z_{1:T}},\mathbf{f}\right)=p_\theta(\mathbf{f})\prod_{t=1}^{T}p_\theta\left(\mathbf{z_t}\mid\mathbf{z_{<t}}\right)p_\theta\left(\mathbf{x_t}\mid\mathbf{z_t},\mathbf{f}\right). \quad (17)$$

The inference model is shown in Eq.(18) for full q and Eq.(19) for factorized q,

$$q_\phi\left(\mathbf{z_{1:T}},\mathbf{f}\mid\mathbf{x_{1:T}}\right)=q_\phi\left(\mathbf{f}\mid\mathbf{x_{1:T}}\right)q_\phi\left(\mathbf{z_{1:T}}\mid\mathbf{f},\mathbf{x_{1:T}}\right), \quad (18)$$

$$q_\phi\left(\mathbf{z_{1:T}},\mathbf{f}\mid\mathbf{x_{1:T}}\right)=q_\phi\left(\mathbf{f}\mid\mathbf{x_{1:T}}\right)\prod_{t=1}^{T}q_\phi\left(\mathbf{z_t}\mid\mathbf{x_t}\right), \quad (19)$$

where $\mathbf{x_{1:T}}=(\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_T})$ denote a high dimensional sequence, $\mathbf{f}$ is a latent variable which can model global aspects of the whole sequence which are time-invariant and $\mathbf{z_i}$ 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.(20) as follows,

$$\begin{aligned}\mathcal{L}(\theta,\phi,\mathbf{x},\mathbf{z},\mathbf{f})=&\mathbb{E}_{q_\phi(\mathbf{z_{1:T}},\mathbf{f}|\mathbf{x_{1:T}})}\big[\log p_\theta(\mathbf{x_{1:T}}|\mathbf{z_{1:T}},\mathbf{f})\big]\\&-D_{KL}\big(q_\phi(\mathbf{z_{1:T}},\mathbf{f}|\mathbf{x_{1:T}})\|p_\theta(\mathbf{z_{1:T}},\mathbf{f})\big)\end{aligned} \quad (20)$$

3.2.2 Group Theory Based VAE Methods

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). They consider the data space O and latent representation space V, where a dataset of trajectories $(o_0,g_0,o_1,g_1,...)$ with o_i denoting the observation of data and $g_i\in G$ denoting the transformation that transforms o_i to o_{i+1}. They map G to a group of matrices belonging to the special orthogonal group $SO(n)$, i.e., mapping g_i to an element of general linear group $GL(V)$, which is shown in Eq.(22)). $R_{i,j}$ denotes the rotation in the (i,j) plane. For instance, in the case of 3-dimensional space:

$$R_{i,i}\left(\theta_{i,j}\right)=\begin{pmatrix}\cos\theta_{i,j}&0&\sin\theta_{i,j}\\0&1&0\\-\sin\theta_{i,j}&0&\cos\theta_{i,j}\end{pmatrix}. \quad (21)$$

In the training procedure, they first randomly select an observation o_i in the trajectories, then generate a series of reconstructions $\{\hat{o}_k\}_{k=i+1,\ldots,i+m}$ through Eq.(23), where f_ϕ is the encoder mapping the observations to the n-dimensional latent space V and d_ψ is the decoder. The first objective is to minimize the reconstruction loss $\mathcal{L}_{\text{rec}}(\phi,\psi,\theta)$ between the true observations $\{o_k\}_{k=i+1,\ldots,i+m}$ generated by the transformations in the environment and the reconstructed observations $\{\hat{o}_k\}_{k=i+1,\ldots,i+m}$ generated by the transformations in the latent space. Furthermore, to enforce disentanglement, they propose another loss function $\mathcal{L}_{ent}(\theta)$ which penalizes the number of rotations that a transformation $g_a(\theta^a_{i,j})$ involves, which is shown in Eq.(24). Lower $\mathcal{L}_{ent}$ indicates that g_a involves fewer rotations and thus g_a acts on fewer dimensions, which means better disentanglement.

$$g\left(\theta_{1,2},\theta_{1,3}...,\theta_{1,n},\theta_{2,3}...,\theta_{2,n},.....\theta_{n-1,n}\right)=\prod_{i=1}^{n-1}\prod_{j=i+1}^{n}R_{i,j}\left(\theta_{i,j}\right), \quad (22)$$

$$\hat{o}_{i+m}(\phi,\psi,\theta)=d_\psi\big(g_{i+m}(\theta)\cdot g_{i+m-1}(\theta)\ldots.g_{i+1}(\theta)\cdot f_\phi(o_i)\big), \quad (23)$$

$$\mathcal{L}_{\text{ent}}(\theta)=\sum_a\sum_{(i,j)\neq(\alpha,\beta)}\left|\theta^a_{i,j}\right|^2\quad\text{with}\quad\theta^a_{\alpha,\beta}=\max_{i,j}\left(\left|\theta^a_{i,j}\right|\right). \quad (24)$$

Different from environment-based methods [38], [39] which leverage 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. They propose three sufficient conditions in the framework, namely model constraint, data constraint and group structure constraint, together with a specific implementation of the framework based on the existing VAE-based models through integrating additional loss. The authors assume that G is a direct product of m rings of integers modulo n, i.e., $G = (\mathbb{Z}/n\mathbb{Z})^m = \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \times \cdots \times \mathbb{Z}/n\mathbb{Z}$, where n denotes the number of possible values for a factor and m denotes the number of all factors. They assume Z has the same elements as G and further assume the group action of G on Z is element-wise addition, i.e., $g \cdot z = \overline{g+z}, \forall z \in Z, g \in G$. In order not to involve the group action on world state space W for the unsupervised setting, they construct the permutation group Φ, then use the group action of Φ on the data space O to replace the group action of G on W, which can be formulated in Eq.(25) as follows,

$$f(g \cdot w) = h\big(\varphi_g \cdot b(w)\big) = h\left(\varphi_g \cdot o\right), \forall w \in W, g \in G, \tag{25}$$

where h represents the mapping from O to Z and b represents the mapping from W to O.

Here the Φ satisfying Eq.(25) exists if and only if: (i) Φ is isomorphic to G; (ii) For each generator of dimension i of G, i.e., g_i, there exists a generator of Φ, i.e., φ_i, such that $\varphi_i \cdot b(w) = b\left(g_i \cdot w\right), \forall w \in W$; (iii) $\varphi_g \cdot b(w) = h^{-1}\big(g \cdot f(w)\big), \forall w \in W, \varphi_g \in \Phi$, where φ_g is the corresponding element of g under the isomorphism. Condition (i) and (ii) are respectively referred as group structure constraint and data constraint. Condition (iii) is a model constraint which further guarantees that group Φ can be achieved by encoder, decoder and the group action of G on Z. When these three conditions are satisfied, it can be derived that Z is disentangled with respect to G. However, given that condition (ii) directly involves the world states, a learning method named GROUPIFIED VAE utilizing a necessary condition to substitute (ii) is proposed to satisfy the unsupervised setting. Thus under the architecture of VAE, the model constraint in condition (iii) can be formulated by Eq.(26) as follows,

$$\varphi_g \cdot o = h^{-1}\big(g \cdot h(o)\big) \triangleq d\big(g \cdot h(o)\big), \forall o \in O, g \in G, \tag{26}$$

where h is the encoder and d is the decoder. Moreover, the data constraint can be satisfied to some extent by VAE based models for the unsupervised setting because of the intuition that VAE based models can generate the data from statistical independent latent variables which are similar to generators of Φ. To satisfy the group structure constraint, GROUPIFIED VAE proposes *Abel Loss* and *Order Loss* to guarantee that ϕ is isomorphic to G, which are formulated in Eq.(27) and Eq.(28) respectively as follows,

$$\mathcal{L}_a = \sum_{o \in O} \sum_{(i,j)} \big\| \varphi_i \cdot (\varphi_j \cdot o) - \varphi_j \cdot (\varphi_i \cdot o) \big\|, \tag{27}$$

$$\mathcal{L}_o = \sum_{o \in O} \sum_{1 \le i \le m} \big(\big\| \varphi_i \cdot (\varphi_i^{n-1} \cdot o) - o \big\| + \big\| \varphi_i^{n-1} \cdot (\varphi_i \cdot o) - o \big\| \big), \tag{28}$$

where φ is the generator of Φ.

Beyond learning the homomorphism from a group to group action, Wang et al. [41] propose Iterative Partition-based Invariant Risk Minimization (IP-IRM), an iterative algorithm based on the self-supervised learning fashion, to specifically learn a mapping between observation space $\mathcal{I}$ and feature space $\mathcal{X}$, i.e., a disentangled feature extractor ϕ such that $\mathbf{x} = \phi(I)$ under the group-theoretical disentanglement conditions. They first argue that most existing self-supervised learning approaches only disentangle the augmentation related features, thus failing to modularize the global semantics. In contrast, IP-IRM is able to ground the abstract semantics and the group actions successfully. Specifically, IP-IRM partitions the training data into disjoint subsets with a partition matrix $\mathbf{P}$, and defines a pretext task with contrastive loss $\mathcal{L}(\phi, \theta = 1, k, \mathbf{P})$ on the samples in the k-th subset, where θ is a constant parameter. At each iteration, it finds a new partition $\mathbf{P}^*$ through maximizing the variance across the group orbits by Eq.(29), which reveals an entangled group element g_i.

$$\begin{aligned} \mathbf{P}^* = \arg\max_{\mathbf{P}} \sum_k \big[& \mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) \\ & + \lambda_2 \left\| \nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) \right\|^2 \big]. \end{aligned} \tag{29}$$

Then the Invariant Risk Minimization (IRM) [54] approach is adopted to update ϕ by Eq.(30), which disentangles the representation w.r.t g_i. It sets $\mathcal{P} = \{\mathbf{P}\}$ at beginning and update $\mathcal{P} \leftarrow \mathcal{P} \cup \mathbf{P}^*$ each time.

$$\min_{\phi} \sum_{\mathbf{P} \in \mathcal{P}} \sum_k \big[\mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) + \lambda_1 \left\| \nabla_{\theta=1} \mathcal{L}(\phi, \theta = 1, k, \mathbf{P}) \right\|^2 \big] \tag{30}$$

It is theoretically proved that iterating the above two steps eventually converges to a fully disentangled representation w.r.t. $\prod_{i=1}^m g_i$. IP-IRM is devised to delay the group action learning to downstream tasks on demand so that it learns a disentangled representation with an inference process, which provides wide feasibility and availability on large-scale tasks.

Moreover, Zhu et al. [55] propose an unsupervised DRL framework, named Commutative Lie Group VAE. They introduce a matrix Lie group G and corresponding Lie algebra $\mathfrak{g}$ which satisfies Eq.(31),

$$\begin{aligned} & g(t) = \exp(B(t)), g \in G, B \in \mathfrak{g}, \\ & B(t) = t_1 B_1 + t_2 B_2 + \ldots + t_m B_m, \forall t_i \in \mathbb{R}, \end{aligned} \tag{31}$$

where $\exp(\cdot)$ denotes the matrix exponential map and $\{B_i\}_{i=1}^m$ is a basis of the Lie algebra. In this case, every sample have a group representation $\mathbf{z}$ and can also be identified by coordinate t in the Lie algebra. The objective function is written in Eq.(32) as follows,

$$\begin{aligned}\log p(\mathbf{x}) \geq & \mathcal{L}_{\text{bottleneck}}(\mathbf{x}, \mathbf{z}, \mathbf{t}) \\ = & \mathbb{E}_{q(\mathbf{z}|\mathbf{x})q(\mathbf{t}|\mathbf{z})} \log p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{t}) \\ & - \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} D_{KL}\big(q(\mathbf{t}|\mathbf{z})||p(\mathbf{t})\big) - \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log q(\mathbf{z}|\mathbf{x}),\end{aligned} \tag{32}$$

where $q(\mathbf{z}|\mathbf{x})$ is implemented as a deterministic encoder, while $q(\mathbf{t}|\mathbf{z})$ is implemented as a stochastic encoder. $p(\mathbf{z}|\mathbf{t})$ is implemented through $\mathbf{z} = g(\mathbf{t})$, and $p(\mathbf{x}|\mathbf{z})$ is implemented as an image decoder. The first and second term can be regarded as reconstruction loss on data space and representation space respectively. The third term is the conditional entropy, which is constant. Moreover, a one-parameter decomposition constraint and a Hessian penalty constraint on $\{B_i\}_{i=1}^m$ are proposed to encourage disentanglement as well.

3.2.3 Causal VAE Based Methods

Based on the statement from Suter et al. [14], Reddy et al. [56] 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 [57]. 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.

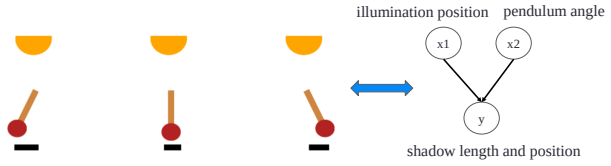

Fig. 6. The position of the illumination source and the angle of the pendulum are causes of the position and the length of the shadow.

Since Locatello et al. [58] 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. 6. 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. (33),

$$\boldsymbol{\epsilon} = h(\mathbf{x}, \mathbf{u}) + \zeta, \tag{33}$$

where h is the encoder and ζ 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.(34),

$$\mathbf{z} = \mathbf{A}^\top \mathbf{z} + \boldsymbol{\epsilon} = \left(\mathbf{I} - \mathbf{A}^\top\right)^{-1} \boldsymbol{\epsilon}, \tag{34}$$

where $\mathbf{A}$ is the learnable adjacency matrix of the causal directed acyclic graph (DAG). Before being fed into the decoder, $\mathbf{z}$ is passed through a *Mask Layer* to reconstruct itself, as is illustrated in Eq.(35), for the i-th latent dimension of $\mathbf{z}$, z_i,

$$z_i = g_i\left(\mathbf{A}_i \circ \mathbf{z}; \eta_i\right) + \epsilon_i, \tag{35}$$

where $\circ$ represents element-wise product and g_i is a mild nonlinear function with the learnable parameter η_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 ξ is also a noise.

Bengio et al. [59] 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. [60] in which the previously calculated $\mathbf{z}$ is employed to define $F_\beta(\boldsymbol{\epsilon})$, as is shown in Eq.(36),

$$\left[\mathbf{z} = f_1\big((\mathbf{I} - \mathbf{A}^\top)^{-1} f_2(\boldsymbol{\epsilon})\big)\right] := F_\beta(\boldsymbol{\epsilon}), \tag{36}$$

where f_1 and f_2 are element-wise transformations, which are usually non-linear. A is the same learnable adjacency matrix in Eq.(34) and Eq.(35). β denotes the parameters of f_1, f_2 and A. When f_1 is invertible, Eq.(36) will be equivalent to Eq.(37) in the following:

$$f_1^{-1}(\mathbf{z}) = \mathbf{A}^\top f_1^{-1}(\mathbf{z}) + f_2(\boldsymbol{\epsilon}). \tag{37}$$

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.(37) 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 $E(\mathbf{x}) \sim \mathcal{N}\big(m(\mathbf{x}), \Sigma(\mathbf{x})\big)$, where $m(\mathbf{x})$ and $\Sigma(\mathbf{x})$ are the mean and variance output by the encoder, respectively.

3.3 Generative Adversarial Networks (GAN) Based Approaches

GAN (Generative Adversarial Nets) [17], as another important generative model proposed by Goodfellow et al., has drawn a lot of attentions from researchers. Instead of adopting conventional Bayesian statistical methods, GAN directly sample 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.(38),

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim P_{data}}\big[\log D(\mathbf{x})\big] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\Big[\log\Big(1-D\big(G(\mathbf{z})\big)\Big)\Big], \tag{38}$$

where P_{data} represents the real dataset and $p(\mathbf{z})$ represents the prior distribution of the latent representation $\mathbf{z}$. Based on GAN, researchers has also proposed a number of methods for DRL.

InfoGAN [9] is one of the earliest works using the GAN paradigm to conduct DRL. 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 λ, such that the adversarial loss of InfoGAN is written in Eq. (39) as follows,

$$\min_G \max_D V_I(D,G) = V^{'}(D,G) - \lambda I\big(\mathbf{c};G(\mathbf{z},\mathbf{c})\big), \tag{39}$$

where $V^{'}(D,G)$ is defined in Eq.(40), taking $\mathbf{c}$ into account.

$$V'(D,G) = \mathbb{E}_{\mathbf{x}\sim P_{data}}\big[\log D(\mathbf{x})\big] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\Big[\log\Big(1-D\big(G(\mathbf{z},\mathbf{c})\big)\Big)\Big]. \tag{40}$$

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.(41),

$$\begin{aligned} I\big(\mathbf{c};G(\mathbf{z},\mathbf{c})\big) &= H(\mathbf{c}) - H\big(\mathbf{c}\mid G(\mathbf{z},\mathbf{c})\big) \\ &= \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\Big[\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})}\big[\log p\left(\mathbf{c}'\mid\mathbf{x}\right)\big]\Big] + H(\mathbf{c}) \\ &= \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\Big[\underbrace{D_{KL}\big(p(\cdot\mid\mathbf{x})\|q(\cdot\mid\mathbf{x})\big)}_{\geq 0} \\ &\quad + \mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})}\big[\log q\left(\mathbf{c}'\mid\mathbf{x}\right)\big]\Big] + H(\mathbf{c}) \\ &\geq \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\Big[\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})}\big[\log q\left(\mathbf{c}'\mid\mathbf{x}\right)\big]\Big] + H(\mathbf{c}) \\ &= \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}),\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}[\log q(\mathbf{c}\mid\mathbf{x})] + H(\mathbf{c}), \end{aligned} \tag{41}$$

where $H(.)$ denotes the entropy of the random variable and $q(\mathbf{c}|\mathbf{x})$ is the auxiliary posterior distribution approximating the true posterior $p(\mathbf{c}|\mathbf{x})$. Actually, q is implemented as a neural network. The overall framework of InfoGAN is shown in Figure. 7.

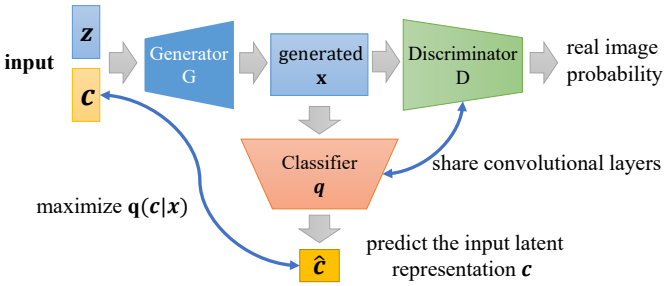

Fig. 7. The overall framework of InfoGAN.

Nevertheless, the performance of InfoGAN for disentanglement is constantly reported to be lower than VAE-based models. To enhance disentanglement, Jeon et al. [61] 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. [62] propose InfoGAN-CR, which is a self-supervised variant of InfoGAN with contrastive regularizer. 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. [63] 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 ϵ 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. [64] 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. (42),

$$\mathcal{L}_{\text{Jacob}}(G) = \sum_{d=1}^{D} \sum_{i=1}^{m} \sum_{j \neq i} \left| \left[\frac{\partial G_d}{\partial z_i} \right]^T \frac{\partial G_d}{\partial z_j} \right|^2, \tag{42}$$

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

On the other hand, we can also introduce supervision to further facilitate disentanglement. For instance, Tran et al. [65] propose DR-GAN which uses the class labels of input images as supervision signals, where the manually preset one-hot latent representation $\mathbf{c}$ is forced to align with the class label during the training stage.

Xiao et al. [66] 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 8

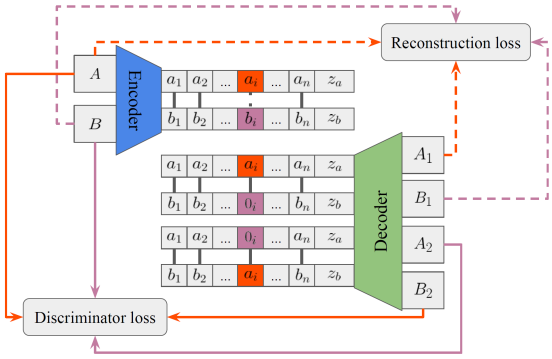

Fig. 8. The architecture of DNA-GAN, figure from [66].

The powerful representations obtained from GAN based approaches also promote research on cross-domain DRL. Liu et al. [67] study the challenge of image translation through learning a joint distribution from two marginal distributions of different domains. They propose UNIT, a model adopting the assumption that a shared latent space can be obtained via mapping images from two domains to a common space. Directly reconstructing images or translating images to the other domains based on the shared latent space can be accomplished in an unsupervised manner, with a design implying the idea of cycle-consistency constraint [68]. The extension of UNIT, MUNIT [69] is explicitly inspired by the idea of disentanglement, following the assumption that the representation space can be decomposed into content space and style space, capturing domain-invariant and domain-specific properties respectively. MUNIT introduces multimodal and diverse translation through combining a content representation with a style representation sampled from the style space of an alternative target domain. Specifically, MUNIT obtains content representation and style representation of each domain, reconstructing the samples through two pairs of auto-encoders for within-domain generation. Afterwards, the content encoders from two different domains will be swapped to generate translated samples for cross-domain translation, ensuring that the translated images are indistinguishable from real images by the discriminator in the target domain with adversarial objectives.

3.4 Hierarchical Approaches

In practice, many generative processes naturally involve hierarchical structures [70] where the factors of variation have different levels of semantic abstraction, either dependent or independent across levels. For example, the factor controlling *gender* has higher level of abstraction than the independent factor controlling *eye-shadow* in CelebA dataset [50], while there exist dependencies between factors controlling *shape* and *phase* in Spaceshapes dataset [70], 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.

Li et al. [71] propose a VAE-based model which learns hierarchical disentangled representations through formulating the hierarchical generative probability model in Eq. (43),

$$p(\mathbf{x}, \mathbf{z}) = p\left(\mathbf{x} \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L\right) \prod_{l=1}^{L} p\left(\mathbf{z}_l\right), \tag{43}$$

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., deeper network layer is responsible for outputting representations with higher abstraction level. It is worth noting that Eq.(43) assumes that there is no dependency among latent representations with different abstraction levels. In other words, each latent representation tends to capture the factors in one single abstraction level, which will not be covered in other levels. The corresponding inference model is formulated in Eq.(44) as follows,

$$q\left(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L \mid \mathbf{x}\right) = \prod_{l=1}^{L} q\big(\mathbf{z}_l \mid \mathbf{h}_l(\mathbf{x})\big), \tag{44}$$

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 9, where h_i and g_i are a set of encoders and decoders at different abstraction levels.

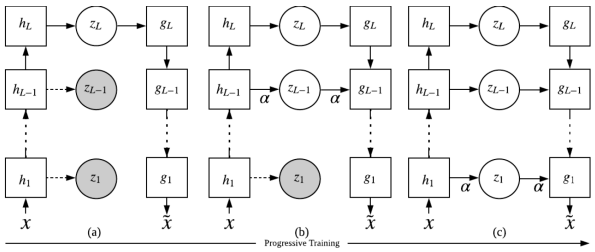

Fig. 9. The architecture of DNA-GAN, figure from [71].

Tong et al. [72] propose to learn a set of hierarchical disentangled representations $\mathbf{z} = \left\{\mathbf{z}_l^i\right\}_{i=1}^{c_l}$, where $\mathbf{z}_l^i$ is the 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 every hierarchical level, they design a loss function shown in Eq.(45),

$$\mathcal{L}_{\text{disentangle}} = \sum_l \frac{2}{c_l\left(c_l-1\right)} \sum_{i \neq j}^{c_l} \mathrm{dCov}^2\left(\mathbf{z}_l^i, \mathbf{z}_l^j\right), \qquad (45)$$

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

Singh et al. [73] 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. [74] 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. [70] 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 learn the hierarchical structure based on which an autoencoder is trained for hierarchical disentanglement.

Hsu et al. [75] propose a hierarchical conditional VAE-based framework with two levels of hierarchical latent variables: i)categorical variable and ii)multivariate Gaussian variable. The first level represents attribute groups (clusters) and the second level characterizes specific attribute configurations conditioned on the first level, with the distribution of these two latent variables following a Gaussian mixture model (GMM).

3.5 Other Methods

Pretrained Generator as Prior. Most encoder-decoder based methods such as VAE train the encoder and decoder (or generator) simultaneously. However, recent works [76], [77], [78] have shown semantically meaningful variations when traversing along different 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. Based on this, Ren et al. [79] claim that training the encoder and generator simultaneously may not be the best choice and then propose a framework, DisCo, which optimizes the encoder with the pretrained generator fixed. They discover the traversal directions of the fixed generator as factors for disentanglement and further encode traversed images into the variation space, where contrastive learning is utilized to enforce disentanglement.

Distilling Unknown Factors with Weak Supervision. Most unsupervised 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 [80]. However, in some cases there exist intractable factors which are unclear or difficult for labeling, where these factors are usually regarded as noises unrelated to the target task. Xiang et al. [80] 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 the 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 which predict the probability distribution of factor labels to enforce disentanglement, similar to the idea of InfoGAN [9].

Incorporating Supervisions with Few Labels. Locatello et al. [58] claim that "pure unsupervised DRL is theoretically impossible without inductive bias on methods and data sets". Given that the amount of supervisions are quite limited in practice, they point out that using a few labeled and even imprecise data for training can be sufficient and beneficial both in terms of disentanglement and downstream performance [81] . From the perspective on correlation of observation, Träuble et al. [82] demonstrate that systematically induced correlations in the dataset remain statistically dependent and entangled in the latent representations, which

can be resolved either through weak supervision during training or by post-hoc correcting a pre-trained model with a small number of labels.

3.6 Discussions

Dimension-wise v.s. Vector-wise Categorization. According to the structure of disentangled representations, we can categorize DRL methods into two groups, i.e., dimension-wise and vector-wise methods. For dimension-wise methods, generative factors are fine-grained and different dimensions represent different types of semantic meanings. For vector-wise methods, generative factors are coarse-grained and different vectors represent different types of semantic meanings. The comparisons of dimension-wise and vector-wise methods are shown as Table 2 where most approaches discussed in Section 3 belong to dimension-wise methods, e.g., various VAE-based methods, InfoGAN and IB-GAN. Dimension-wise methods are always experimented on synthetic and simple datasets, while vector-wise methods are always tested in real-world scenes such as image translation.

4 METRICS

Many works [5], [6], [7], [9] qualitatively evaluate the performance of disentanglement by inspecting the change in reconstructions when traversing one variable in the latent space. Qualitative observation is straightforward, but not precise or mathematically rigorous. In order to promote the research of learning disentangled representations, it is important to design reliable metrics which can quantitatively measure disentanglement. We review and divide a series of quantitative metrics into two categories: supervised metrics and unsupervised metrics. As for a deeper understanding, discussion and taxonomy for metrics, we refer interested readers to Zaidi et al.'s work [85].

4.1 Supervised Metrics

Supervised metrics assume that we have access to the ground truth generative factors.

Z-diff. Higgins et al. [6] propose a supervised disentanglement metric based on a low capacity linear classifier network to measure both the independence and explainability. They conduct inference on a number of image pairs that are generated by fixing the value of one data generative factor while randomly sampling all others. Taking a batch of image pairs as input, the classifier is expected to identity which factor is fixed and report the accuracy value as the disentanglement metric score.

Z-min Variance. Kim et al. [7] point out that the aforementioned method using linear network has several weaknesses, such as being sensitive to hyperparameters of the linear classifier optimization. Most importantly, the metric has a failure mode: giving 100% accuracy even when only $K-1$ factors out of K have been disentangled. The authors propose a metric based on a majority-vote classifier with no optimization hyperparameters. They also generate a number of images with one factor k fixed and all others varying randomly. After obtaining representations with normalization, they take the index of the dimension with the lowest empirical variance and the label k as the input & output for the majority-vote classifier. The accuracy of the classifier is regarded as the disentanglement metric score.

Z-max Variance. Kim et al. [33] propose a metric which is almost the same as Z-min Variance. The main difference lays that they generate samples with one factor k varying and all others fixed. Consequently, they choose the index of the dimension with the highest empirical variance as the input of majority-vote classifier. They claim that this metric shows better consistency with qualitative assessment than Z-min Variance.

Mutual Information Gap (MIG). Chen et al. [5] propose a classifier-free information-theoretic metric named MIG. The key insight of MIG is to evaluate the empirical mutual information between a latent variable z_j and a ground truth factor k. For each factor k, MIG computes the gap between the top two latent variables with the highest mutual information. The average gap over all factors is used as the disentanglement metric score. Higher MIG score means better disentanglement performance because it indicates that each generative factor is principally captured by only one latent dimension.

SAP Score. Kumar et al. [35] propose a metric referred as Separated Attribute Predictability (SAP) score. They construct a score matrix $S \in R^{d\times k}$ and the $(i,j)-$th element represents the linear regression or classification score of predicting $j-$th factor using only $i-$th latent variable distribution. Then for each column of the score matrix, they compute the difference between the top two elements and take the average of these differences as the SAP score. Higher SAP score means better disentanglement performance because it also indicates that each generative factor is principally corresponding to only one latent dimension, just like MIG.

DCI. Eastwood et al. [86] design a framework which evaluates disentangled models from three aspects, i.e., disentanglement (D), completeness (C) and informativeness (I). Specifically, disentanglement denotes the degree of capturing at most one generative factor for each latent variable. Completeness denotes the degree to which each generative factor is captured by only one latent variable. Informativeness denotes the amount of information that latent variables captures about the generative factors. It is worth noting that the disentanglement and the completeness together quantify the deviation between bijection and the actual mapping.

Modularity and Explicitness. Ridgeway et al. [87] evaluate disentanglement from two aspects, i.e., modularity and explicitness. They claim a latent dimension is ideally modular only when it has high mutual information with only one factor and zero with all others. They obtain the modularity score by computing the deviation between the empirical case and the desired case. Explicitness focuses on the coverage of latent representation with respect to generative factors. Assuming factors have discrete values, they fit a one-versus-rest logistic-regression factor classifier and record the ROC area-under-the-curve (AUC). They then take the mean of AUC values over all classes for all factors as the final explicitness score.

UNIBOUND. Tokui et al. [88] propose UNIBOUND to evaluate disentanglement by lower bounding the unique information in the term of Partial Information Decomposition (PID). PID decomposes the information between a

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

Methods	Dimension of Each Latent Factor	Representative Works	Semantic Alignment	Applicability
Vector-wise	two or more	MAP-IVR [83], DRNET [84], DR-GAN [65], DRANet [21], Lee et al. [8], Liu et al. [22], Singh et al. [73]	each latent variable aligns to one coarse-grained semantic meaning	real scenes
Dimension-wise	one	VAE-based methods, InfoGAN [9], IB-GAN [61], Zhu et al. [19], InfoGAN-CR [62], PS-SC GAN [63], Wei et al. [64], DNA-GAN [66]	each dimension aligns to one fine-grained semantic meaning	synthetic and simple datasets

latent variable z_l and a generative factor y_k into three parts: redundant information, unique information and complementary information. Let $\mathcal{U}(y_k; z_\ell \backslash \mathbf{z}_{\setminus \ell})$ denote the unique information which is held by z_ℓ and not held by remaining variables, they then lower bound this unique information term. Similar to MIG, for each generative factor, the difference between the top two latent variables with largest lower bound value is computed. The average value over all factors is taken as the final score.

UC and GC. Reddy [56] propose Unconfoundedness (UC) Metric and Counterfactual Generativeness (CG) Metric from the causal perspective. As mentioned in Section 2, they leverage an SCM to describe the data generation process. UC metric evaluates the degree how the mapping from G_i to Z_I is unique and unconfounded with respect to a set of confounders C. UC is defined as $UC := 1 - \mathbb{E}_{x\sim p_X}\left[\frac{1}{S}\sum_{I,J}\frac{|\mathbf{Z}_I^x \cap \mathbf{Z}_J^x|}{|\mathbf{Z}_I^x \cup \mathbf{Z}_J^x|}\right]$. CG evaluates whether or not any causal intervention on Z_I influence the generated aspects about G_i. This means only the intervention on Z_I can influence G_i for the generation process. CG is defined as $CG = \mathbb{E}_I[|ACE^{X_I^{cf}}_{\mathbf{Z}_I^X} - ACE^{X_{\setminus I}^{cf}}_{\mathbf{Z}_{\setminus I}^X}|]$, where $ACE^X_{do(Z=\alpha)} = \mathbb{E}[X|do(Z=\alpha)] - \mathbb{E}[X|do(Z=\alpha^*)]$.

4.2 Unsupervised Metrics

When we do not have access to the ground truth factors, unsupervised metrics then become important and useful.

ISI. Do et al. [89] suggest three important properties of disentanglement from the perspective of mutual information, i.e., informativeness (I), separability (S) and interpretability (I). Furthermore, they propose a series of metrics to conduct the evaluation based on the three aspects respectively. Specifically, informativeness denotes the mutual information between original data x and latent variable z_i, formulated as $I(x, z_i)$ Separability means that any two latent variables z_i, z_j do not share common information about the data x, which denotes the ability to separate two latent variables with respect to the data x, formulated as $I(x, z_i, z_j)$. Explainability means a one-one mapping (or bijection) between latent variables z_i and the data generative factors y_k, formulated as $I(z_i, y_k) = H(z_i) = H(y_k)$. They further propose specific methods of estimating these mutual information terms, which are applicable to both supervised and unsupervised scenarios.

5 DRL Applications

In this section, we discuss the broad applications of DRL for various downstream tasks.

5.1 Image

Images, as one of the most widely investigated visual data type, can benefit a lot from DRL in terms of generation, translation and explanation etc.

5.1.1 Generation

By taking advantages of DRL, independent factors in generation objectives can be learned and aligned with latent representation through disentanglement, hence capable of controlling the generation process.

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 β-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 widely 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 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. [90] analyze disentanglement generation operation in StyleGAN [91], especially in *StyleSpace*, to manipulate semantically meaningful attributes in generation. Zeng et al. [92] 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 [93], which disentangles identity-relevant and identity-irrelevant information through optimizing information bottleneck to generate more identity-discriminative swapped faces.

5.1.2 Translation

In addition to generation, image translation is also a hot topic in image processing and understanding. Disentangled factors contribute to coherent and robust performance for cross-domain scenarios, ultimately enhancing and expanding the controllability and applicability of image translation.

Gonzalez et al. [20] present cross-domain disentanglement, disentangling the internal representations into shared and exclusive parts through bidirectional image translation based on GAN and cross-domain autoencoders with only paired images as input. This design achieves satisfactory performance on various tasks such as diverse sample generation, cross-domain retrieval, domain-specific image transfer and interpolation. Lee et al. [8] disentangle latent representations into domain-invariant content space and domain-specific attribute space by introducing a content discriminator and cross-cycle consistency loss on GAN-based framework, achieving diverse multimodal translation without using pre-aligned image pairs for training. Later, DRANet [21] is proposed to disentangle content and style factors, and synthesize images by transferring visual attributes for unsupervised multi-directional domain adaption. Liu et al. [22] point out the lack of graduality for existing image translation models in semantic interpolations both within domains and across domains. As such, they propose a new training protocol, which learns a smooth and disentangled latent style space to perform gradual changes and better preserve the content of the source image.

5.1.3 Others

The idea of DRL has also been employed in other image-related fields and tasks. Sanchez et al. [94] disentangle shared and exclusive representations in paired images through optimizing mutual information, which is well applied to image classification and image retrieval tasks without relying on image reconstruction or image generation. Hamaguchi et al. [95] propose a VAE-based network to disentangle variant and invariant factors for rare event detection on imbalanced datasets, requiring only pairs of observations. Gidaris et al. [96] propose a self-supervised semantic feature learning method through predicting rotated images with ConvNet model to achieve comparable performances with supervised methods. Inspired by Gidaris et al.'s work, Feng et al. [97] later disentangle feature representations relevant to semantic rotation and irrelevant ones through joint training on image rotating prediction and instance discrimination, which benefits in the generalization ability in image classification, retrieval, segmentation and other tasks. Ghandeharioun et al. [12] propose DISSECT, which enforces the disentanglement of latent concepts by encouraging the distinctness across different concepts and the proximity within a same concept. They achieve multiple counterfactual image explanations which can intervene the output of model by changing disentangled concepts.

To summarize, representations learned through various image representation models can always be structured based on DRL strategy to separate variant events from the inherent attributes. Therefore, as an appropriate learning strategy for image-related tasks, DRL particularly contributes to significant improvement in image generation and translation, facilitating more comprehensive and diverse implementations for various image applications.

5.2 Video

Besides static images, DRL also promotes dynamic videos analysis, including video prediction, video retrieval and motion retargeting etc.

5.2.1 Video Prediction

Video prediction is a challenging yet interesting task of predicting future frames given a set of contextual frames. Denton et al. propose DRNET [84], 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 lays in the high dimensional representation space of visual data. To tackle this problem, Sreekar et al. propose mutual information predictive auto-encoder (MIPAE) [98], 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 [99], 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.

5.2.2 Activity Image-to-Video Retrieval

Activity image-to-video retrieval (AIVR) aims to retrieve videos containing a similar activity as the query image, which focuses on both static appearance and dynamic motion, making the problem much more challenging. Considering the asymmetric relationship between images and videos, Liu et al [83] propose a disentangled framework named MAP-IVR to separate video representation into appearance and motion, transforming image query to video query

through the motion features extracted from the candidate video. As such, the retrieval performance can be dramatically increased through the direct matching between images and videos.

5.2.3 Motion Retargeting

Motion retargeting aims at transferring the human motion from a source video to a target video. Ma et al. [100] point out that previous methods neglect the subject-dependent motion features in the transferring process, which leads to unnatural synthesis. To tackle this problem, they propose to disentangle subject-dependent motion features and subject-independent motion features, generating target motion features through combination of the source subject-independent features and the target subject-dependent features. To achieve this purpose, they design triplet loss for both subject-dependent and subject-independent features to ensure the disentanglement.

5.2.4 Others

To deal with the large mode variations in the real-world applications, Kim et al. [101] propose a DRL framework for robust facial authentication, which disentangles identity and mode (e.g., illumination, pose) features. They first use two encoders to encode identity and mode, respectively. To ensure disentanglement, they design an exclusion-based strategy which encourages the two encoders to remove the characteristics of the peer from their own representations. Moreover, they design a reconstruction-based strategy to reinforce the disentanglement, which uses a decoder to reconstruct the original features by exchanging identity features for an image-pair before re-disentangling the identity and the mode features. Xing et al. [102] propose a disentangled generative framework for video sequences, where two independent latent vectors are employed to represent appearance features and geometric features respectively. They utilize an appearance generator taking the appearance vector as input to generate the original image as well as a geometric generator taking the geometric vector as input to generate the coordinate residual. Taking the original image and coordinate residual as input, a warping function is designed to transform the original image to the target image.

5.3 Natural Language Processing

DRL has also been widely used in natural language processing (NLP) tasks, such as text generation, style transfer, semantic understanding etc.

5.3.1 Text Representation

The initial DRL applications in NLP aims at learning disentangled text representations w.r.t. various criteria, primarily by encoding different aspects of representations into distinct spaces. 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. [103] 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.

5.3.2 Style Transfer

Several works are motivated by employing DRL to disentangle style information from text representations in the practice of text style transfer tasks. Hu et al. [104] 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. [105] 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.

5.3.3 Others

There also exist specific tasks in NLP community where DRL serves as an effective approach. Zou et al. [106] propose to address the fundamental task, i.e., text semantic matching, by disentangling factual keywords from abstract to learn the fundamental way of content matching under different levels of granularity. Dougrez-Lewis et al. [107] disentangle the latent topics of social media messages through an adversarial learning setting, to achieve rumour veracity classification. Zhu et al. [108] disentangle the content and style in latent space by diluting sentence-level information in style representations to generate stylistic conversational responses. Other works [109], [110] also propose to exploit the large pretrained language models (PLM) using DRL. Zhang et al. [109] try to uncover disentangled representations from pretrained models such as BERT [111] by identifying existing subnetworks within them, aiming to extract representations that can factorize into distinct, complementary properties of input. Zeng et al. [110] propose task-guided disentangled tuning for PLMs, which enhances the generalization of representations by disentangling task-relevant signals from the entangled representations.

5.4 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 [112], [113] study the typical modal-level disentanglement through encouraging independence between modality-specific and multimodal factors. Shi et

al. [114] 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. [115] propose a disentangled sentiment representation adversarial network (DiSRAN) to reduce the domain shift of expressive styles for cross-domain sentiment analysis. Recent works [116], [117], [118], [119], [120] tend to focus on disentangling the rich information among multi modalities and leveraging that to perform various downstream tasks. Alaniz et al. [116] 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 [117] realizes disentangled text-driven image manipulation through exploiting the power of the pretrained vision-language model CLIP [121]. Similarly, Yu et al. [118] achieve counterfactual image manipulation via disentangling and leveraging the semantic in text embedding of CLIP. Materzynska et al. [119] disentangle the spelling capabilities from the visual concept processing of CLIP.

5.5 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 [123], [124] 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], [125] 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 [125].

5.6 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 [126], [127]. 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 [128] is proposed to be one unsupervised disentangled method that can disentangle node and edge features from attributed graphs. FactorGCN [129] 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. [130] first propose to learn disentangled graph representations with self-supervision. Given the input graph, the proposed method DGCL identifies the latent factors of the input graph and derives its factorized representations. Then it conducts factor-wise contrastive learning to encourage the factorized representations to independently reflect the expressive information from different latent factors. They further propose IDGCL [131] that is able 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. Li et al. [132] find that learning disentangled graph representations can improve the out-of-distribution (OOD) generalization ability of GNNs. The proposed OOD-GNN model encourages the graph representation disentanglement by eliminating the statistical dependence among all dimensions of the output representation through iteratively optimizing the sample graph weights and graph encoder.

6 DRL Design for Different Tasks

In this section, we discuss commonly adopted strategies for DRL in practical applications, providing inspirations on designing various DRL models for specific tasks. We summarize two key aspects for designing a DRL model: i) designing an appropriate representation structure according to a specific task, and ii) designing corresponding loss functions which force the representation to be disentangled without losing task-specific information.

6.1 Design of Representation Structure

Two approaches for designing the representation structure include i) dimension-wise: use a whole vector representation $\mathbf{z}$, which is fine-grained and 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, approach i) usually requires that $\mathbf{z}$ is dimension-wise independent, while approach ii) usually requires that $\mathbf{z}_i$ is independent with $\mathbf{z}_j$ where $i \neq j$.

If we choose dimension-wise approach 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 models such as InfoSwap [93] 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.

TABLE 3
Representatives of disentangled representation learning applications

Papers	Method	Paradigm	Application
[5], [6], [7], [16], [32], [33], [34], [35]	VAE-based	Unsupervised	Image generation
[9], [10], [19], [90]	GAN-based		
[36], [52], [81], [82]	VAE-based	Supervised	
[65], [66]	GAN-based		
[11], [43]	Causal-based		
[8], [20], [21], [22], [92]	GAN-based	Unsupervised	Image translation
[95]	VAE-based	Supervised	Image classification, segmentation, etc.
[94], [97]	Others	Unsupervised	
[84], [99]	VAE-based	Unsupervised	Video
[83]	Others	Supervised	
[23]	Others	Unsupervised	Natural language processing
[13], [25]	Others	Supervised	
[112], [114], [116]	VAE-based	Unsupervised	Multimodal Application
[113], [115]	VAE-based	Supervised	
[118]	GAN-based		
[117], [119], [120]	Others		
[26], [28], [123], [124], [125]	VAE-based	Supervised	Recommendation
[27], [29], [122]	Others		
[128], [129], [130], [131]	VAE-based	Supervised	Graph
[29], [30], [132]	Others		

As for vector-wise approach, there are 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 [65] 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. [83] leverage two encoders, namely motion encoder and appearance encoder, to respectively extract motion feature and appearance feature by passing through the original representation. Cheng et al. [133] utilize two encoders E_{cls} and E_{var} to extract class-specific and class-irrelevant features, respectively. DRNET [84] also uses two encoders to extract the pose feature and content feature, respectively. DRANet [21] employs only one encoder to extract content feature and then obtains style feature by subtracting content feature from original feature. Similar to DRANet, Wu et al. [134] 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 can also obtain disentangled representations with clustering-based methods by separating data into several relatively independent parts, then extract feature from these independent parts respectively, and finally fuse the features in some way. For example, DGCF [29], as a model for Collaborative Filtering, obtains independent representations according to different user intentions before concatenating them as the final disentangled representation.

We also 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 data.

6.2 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.

6.2.1 Generative model

The reconstruction loss, which is always essential for generation tasks, ensures that the representation is semantically meaningful. The disentanglement loss, on the other hand, enforces the disentanglement of the representation. Moreover, reconstruction loss can sometimes provide guidance for disentanglement. The task loss is directly related to task objective and also can provide guidance for disentanglement, or in other words, the task loss also plays an important role in ensuring that the disentangled features learned can meet the expectations of objective task. Generative models usually have non-zero λ_1 and λ_2, while having zero λ_3.

For example, 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 [61]. Wu et al. [134] use an orthogonal loss to promote the independence between domain-invariant and domain-specific features. Meanwhile, they also utilize a domain classifier to prompt domain-specific features which capture much more domain-specific information, and further use a detection loss of domain adaptive object detection as task loss. DRANet [21] adopts a L_1 loss, a consistency loss and an adversarial loss as reconstruction loss, in addition to

TABLE 4
The summary of loss functions of several generative and discriminative models.

Methods	reconstruction loss	disentanglement loss	task loss
VAE-based Approaches	L2 loss	extra regularizers added to ELBO	-
InfoGAN [9]	GAN loss	maximizing $\lambda I(c; G(z, c))$	-
DRANet [21]	L1 loss, adversarial loss, consistency loss	perceptual loss	-
DRNET [84]	L2 loss	similarity loss, adversarial loss	-
InfoSwap [93]	cycle-consistency loss, adversarial loss	information-compression loss	-
Hamaguchi et al. [95]	VAE loss	similarity loss	similarity loss, activation loss
MAP-IVR [83]	L2 loss	$\mathcal{L}_{\text{orth}} = \cos(m^v, a^v)$, $\mathcal{L}_{class}$	-
Cheng et al. [133]	L1 loss	discriminative Loss	classification loss

the usage of a perceptual loss to enhance disentanglement. InfoSwap [93] resorts to an information compression loss based on information bottleneck theory as the disentanglement loss, as well as using several reconstruction loss functions such cycle-consistency loss. DRNET [84] adopts a L_2 loss as reconstruction loss and uses a similarity loss together with an adversarial loss to ensure disentanglement. Besides, several works also introduce extra supervisions to enforce disentanglement without explicit disentanglement loss function, such as DR-GAN [65] and DNA-GAN [66]. Table 4 summarizes designs of loss functions.

6.2.2 Discriminative model

In contrast to generative models, discriminative models normally set λ_1 to 0, because there will be no need for reconstruction. Hence, we mainly consider the task loss and the disentanglement loss which enforces the disentanglement of representation for discriminative models. In other words, a discriminative model typically utilizes DRL as an infrastructure to achieve better performance for target task.

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 auxiliary loss required by the target task such as image classification, recommendation, neural architecture search etc., will be added. For example, Hamaguchi et al. [95] add similarity loss and activation loss on the basis of using two pairs of VAEs to encode image pairs. This strategy aims at making common features encode invariant factors in an input image pair and avoiding a trivial solution, which encourages the model to learn common features and specific features of images and thus achieve the goal of rare event detection. In terms of text style transfer, John et al. [105] divide the latent representation of text into two parts: the style space and content space, as well as additionally design a systematic set of auxiliary losses to encourage disentanglement. Specifically, multi-task objectives are utilized so that the desired information is constrained to be encoded in latent space while the adversarial objectives are employed to minimize the predictability of irrelevant information. Cheng et al. [133] use a gradient reverse layer and a class discriminative loss to minimize the class-specific information captured by class-irrelevant encoder. Moreover, the reconstruction loss and classification loss can ensure that the class-specific encoder is capable of capturing the class-specific information. MAP-IVR [83] 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 has no task loss since it uses the trained motion and appearance features to tackle the downstream task, i.e., activity image-to-video retrieval.

7 Future Directions

Last but not least, we conclude this paper by pointing out some potential interesting directions that deserve future investigations. **i) Diverse scenes.** Existing works on theoretic DRL methodology and benchmark mostly focus on image generation tasks over simple synthetic datasets. It will be interesting to conduct more analysis on DRL in diverse scenes over more complicated datasets. **ii) Diverse learning paradigms.** Existing DRL methods mostly start from VAE-based and GAN-based models. It will be promising to conduct more research on other potential models, e.g., diffusion model, which may open new ways for DRL. **iii) Explainability and generalization.** Although DRL has achieved several success in explainability and generalization, future works should continue focusing on these two advantages of DRL, e.g., exploring generalization in few-shot or zero-shot learning, and demonstrating explainability in more types of practical tasks.

References

[1] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, pp. 665–673, 2020.
[2] Y. Bengio et al., "Representation learning: A review and new perspectives," *IEEE TPAMI*, 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 *NeurIPS*, 2019, pp. 2615–2625.
[6] I. Higgins et al., "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2016.
[7] H. Kim and A. Mnih, "Disentangling by factorising," in *ICML*, 2018, pp. 2649–2658.
[8] H.-Y. Lee et al., "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018.
[9] X. Chen et al., "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NeurIPS*, 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*, 2021.
[13] J. Wu et al., "Improving robustness and generality of nlp models using disentangled representations," *arXiv*, 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*, pp. 25 123–25 133, 2021.
[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv*, 2013.
[17] I. Goodfellow et al., "Generative adversarial nets," in *NeurIPS*, 2014.
[18] I. Higgins et al., "Towards a definition of disentangled representations," *arXiv*, 2018.
[19] J.-Y. Zhu et al., "Visual object networks: Image generation with disentangled 3d representations," in *NeurIPS*, 2018, pp. 118–129.
[20] A. Gonzalez-Garcia et al., "Image-to-image translation for cross-domain disentanglement," in *NeurIPS*, 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 *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*, 2019.
[27] Y. Zhang et al., "Content-collaborative disentanglement representation learning for enhanced recommendation," in *ACM RecSys*, 2020, pp. 43–52.
[28] X. Wang et al., "Multimodal disentangled representation for recommendation," in *ICME*, 2021, pp. 1–6.
[29] X. Wang et al., "Disentangled graph collaborative filtering," in *ACM SIGIR*, 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*, 2018.
[33] M. Kim et al., "Relevance factor vae: Learning and identifying disentangled factors," *arXiv*, 2019.
[34] C. P. Burgess et al., "Understanding disentangling in β-vae," *arXiv*, 2018.
[35] A. Kumar et al., "Variational inference of disentangled latent concepts from unlabeled observations," in *ICLR*, 2018.
[36] D. Bouchacourt et al., "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *AAAI*, 2018.
[37] S. Bing et al., "On disentanglement in gaussian process variational autoencoders," *arXiv*, 2021.
[38] H. Caselles-Dupré et al., "Symmetry-based disentangled representation learning requires interaction with environments," *NeurIPS*, 2019.
[39] R. Quessard et al., "Learning disentangled representations and group structure of dynamical environments," *NeurIPS*, pp. 19 727–19 737, 2020.
[40] T. Yang et al., "Towards building a group-based unsupervised representation disentanglement framework," in *ICLR*, 2022.
[41] T. Wang et al., "Self-supervised learning disentangled group representation as feature," *NeurIPS*, 2021.
[42] J. Pearl, *Causality*. Cambridge university press, 2009.
[43] X. Shen et al., "Disentangled generative causal representation learning," *arXiv*, 2020.
[44] S. Wold et al., "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
[45] M. Rolinek et al., "Variational autoencoders pursue pca directions (by accident)," in *CVPR*, 2019, pp. 12 406–12 415.
[46] J. V. Stone, "Independent component analysis: an introduction," *Trends in cognitive sciences*, vol. 6, no. 2, pp. 59–64, 2002.
[47] A. Hyvarinen et al., "Nonlinear ica using auxiliary variables and generalized contrastive learning," in *AISTATS*, 2019, pp. 859–868.
[48] D. Horan et al., "When is unsupervised disentanglement possible?" *NeurIPS*, pp. 5150–5161, 2021.
[49] Y. Lecun et al., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, no. 11, pp. 2278–2324, 1998.
[50] Z. Liu et al., "Deep learning face attributes in the wild," in *ICCV*, December 2015.
[51] M. Aubry et al., "Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models," in *CVPR*, 2014.
[52] T. D. Kulkarni et al., "Deep convolutional inverse graphics network," *arXiv*, 2015.
[53] Y. Li and S. Mandt, "Disentangled sequential autoencoder," *arXiv*, 2018.
[54] M. Arjovsky et al., "Invariant risk minimization," *arXiv*, 2019.
[55] X. Zhu et al., "Commutative lie group vae for disentanglement learning," in *ICML*, 2021, pp. 12 924–12 934.
[56] A. G. Reddy et al., "On causally disentangled representations," *arXiv*, 2021.
[57] S. Greenland et al., "Confounding and collapsibility in causal inference," *Statistical science*, pp. 29–46, 1999.
[58] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in *ICML*, 2019, pp. 4114–4124.
[59] Y. Bengio et al., "A meta-transfer objective for learning to disentangle causal mechanisms," in *ICLR*, 2020.
[60] Y. Yu et al., "Dag-gnn: Dag structure learning with graph neural networks," in *ICML*, 2019, pp. 7154–7163.
[61] I. Jeon et al., "Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks," in *AAAI*, 2021, pp. 7926–7934.
[62] Z. Lin et al., "Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers," 2019.
[63] X. Zhu et al., "Where and what? examining interpretable disentangled representations," in *CVPR*, 2021, pp. 5861–5870.
[64] Y. Wei et al., "Orthogonal jacobian regularization for unsupervised disentanglement in image generation," in *CVPR*, 2021, pp. 6721–6730.
[65] L. Tran et al., "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, 2017, pp. 1415–1424.
[66] T. Xiao et al., "Dna-gan: Learning disentangled representations from multi-attribute images," *arXiv*, 2017.
[67] M.-Y. Liu et al., "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017.
[68] J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.
[69] X. Huang et al., "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189.
[70] A. Ross and F. Doshi-Velez, "Benchmarks, algorithms, and metrics for hierarchical disentanglement," in *ICML*, 2021, pp. 9084–9094.
[71] Z. Li et al., "Progressive learning and disentanglement of hierarchical representations," *arXiv*, 2020.
[72] B. Tong et al., "Hierarchical disentanglement of discriminative latent features for zero-shot learning," in *CVPR*, 2019, pp. 11 467–11 476.
[73] K. K. Singh et al., "Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," in *CVPR*, 2019, pp. 6490–6499.
[74] X. Li et al., "Image-to-image translation via hierarchical style disentanglement," in *CVPR*, 2021, pp. 8639–8648.
[75] W.-N. Hsu et al., "Hierarchical generative modeling for controllable speech synthesis," *arXiv*, 2018.
[76] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in gans," in *CVPR*, 2021, pp. 1532–1540.
[77] V. Khrulkov et al., "Disentangled representations from non-disentangled models," *arXiv*, 2021.
[78] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the gan latent space," in *ICML*, 2020, pp. 9786–9796.
[79] X. Ren et al., "Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view," in *ICLR*, 2021.
[80] S. Xiang et al., "Disunknown: Distilling unknown factors for disentanglement learning," in *CVPR*, 2021, pp. 14 810–14 819.
[81] F. Locatello et al., "Disentangling factors of variations using few labels," in *ICLR*, 2020.
[82] F. Träuble et al., "On disentangled representations learned from correlated data," in *ICML*, 2021, pp. 10 401–10 412.
[83] L. Liu et al., "Activity image-to-video retrieval by disentangling appearance and motion," in *AAAI*, 2021, pp. 1–9.
[84] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *NeurIPS*, 2017, pp. 4417–4426.
[85] J. Zaidi et al., "Measuring disentanglement: A review of metrics," *arXiv*, 2020.
[86] C. Eastwood and C. K. Williams, "A framework for the quantitative evaluation of disentangled representations," in *ICLR*, 2018.
[87] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss," *NeurIPS*, 2018.
[88] S. Tokui and I. Sato, "Disentanglement analysis with partial information decomposition," in *ICLR*, 2022.
[89] K. Do and T. Tran, "Theory and evaluation metrics for learning disentangled representations," *arXiv*, 2019.
[90] Z. Wu et al., "Stylespace analysis: Disentangled controls for stylegan image generation," in *CVPR*, 2021, pp. 12 863–12 872.

[91] T. Karras et al., "A style-based generator architecturanetre for generative adversarial networks," in *CVPR*, 2019.
[92] X. Zeng et al., "Realistic face reenactment via self-supervised disentangling of identity and pose," in *AAAI*, 2020, pp. 12 757–12 764.
[93] G. Gao et al., "Information bottleneck disentanglement for identity swapping," in *CVPR*, 2021, pp. 3404–3413.
[94] E. H. Sanchez et al., "Learning disentangled representations via mutual information estimation," in *ECCV*, 2020, pp. 205–221.
[95] R. Hamaguchi et al., "Rare event detection using disentangled representation learning," in *CVPR*, 2019.
[96] S. Gidaris et al., "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.
[97] Z. Feng et al., "Self-supervised representation learning by rotation feature decoupling," in *CVPR*, 2019.
[98] P. A. Sreekar et al., "Mutual information based method for unsupervised disentanglement of video representation," in *ICPR*, 2021, pp. 6396–6403.
[99] J.-T. Hsieh et al., "Learning to decompose and disentangle representations for video prediction," *NeurIPS*, 2018.
[100] J. Ma and S. Yu, "Human identity-preserved motion retargeting in video synthesis by feature disentanglement," *arXiv*, 2022.
[101] M. Kim et al., "Robust video facial authentication with unsupervised mode disentanglement," in *ICIP*, 2020, pp. 1321–1325.
[102] X. Xing et al., "Deformable generator networks: unsupervised disentanglement of appearance and geometry," *IEEE TPAMI*, 2020.
[103] P. Colombo et al., "A novel estimator of mutual information for learning to disentangle textual representations," in *ACL*, 2021, pp. 6539–6550.
[104] Z. Hu et al., "Toward controlled generation of text," in *ICML*, 2017, pp. 1587–1596.
[105] V. John et al., "Disentangled representation learning for non-parallel text style transfer," in *ACL*, 2019, pp. 424–434.
[106] Y. Zou et al., "Divide and conquer: Text semantic matching with disentangled keywords and intents," in *ACL*, 2022, pp. 3622–3632.
[107] J. Dougrez-Lewis et al., "Learning disentangled latent topics for twitter rumour veracity classification," in *ACL*, 2021, pp. 3902–3908.
[108] Q. Zhu et al., "Neural stylistic response generation with disentangled latent variables," in *ACL*, 2021, pp. 4391–4401.
[109] X. Zhang et al., "Disentangling representations of text by masking transformers," in *EMNLP*, 2021, pp. 778–791.
[110] J. Zeng et al., "Task-guided disentangled tuning for pretrained language models," in *ACL*, 2022, pp. 3126–3137.
[111] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.
[112] W.-N. Hsu and J. Glass, "Disentangling by partitioning: A representation learning framework for multimodal sensory data," *arXiv*, 2018.
[113] Y.-H. H. Tsai et al., "Learning factorized multimodal representations," in *ICLR*, 2018.
[114] Y. Shi et al., "Variational mixture-of-experts autoencoders for multi-modal deep generative models," *NeurIPS*, 2019.
[115] Y. Zhang et al., "Learning disentangled representation for multimodal cross-domain sentiment analysis," *IEEE TNNLS*, 2022.
[116] S. Alaniz et al., "Compositional mixture representations for vision and text," in *CVPR Workshops*, 2022, pp. 4202–4211.
[117] 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.
[118] Y. Yu et al., "Towards counterfactual image manipulation via clip," in *ACM Multimedia*, 2022, pp. 3637–3645.
[119] J. Materzyńska et al., "Disentangling visual and written concepts in clip," in *CVPR*, 2022, pp. 16 410–16 419.
[120] W. Zou et al., "Utilizing bert intermediate layers for multimodal sentiment analysis," in *ICME*, 2022, pp. 1–6.
[121] A. Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.
[122] L. Hu et al., "Graph neural news recommendation with unsupervised preference disentanglement," in *ACL*, 2020, pp. 4255–4264.
[123] J. Ma et al., "Disentangled self-supervision in sequential recommenders," in *ACM SIGKDD*, 2020, pp. 483–491.
[124] H. Chen et al., "Curriculum disentangled recommendation with noisy multi-feedback," *NeurIPS*, pp. 26 924–26 936, 2021.
[125] X. Wang et al., "Disentangled representation learning for recommendation," *IEEE TPAMI*, 2022.
[126] Z. Zhang et al., "Deep learning on graphs: A survey," *IEEE TKDE*, 2020.
[127] J. Zhou et al., "Graph neural networks: A review of methods and applications," *AI Open*, pp. 57–81, 2020.
[128] X. Guo et al., "Interpretable deep graph generation with node-edge co-disentanglement," in *ACM SIGKDD*, 2020, pp. 1697–1707.
[129] Y. Yang et al., "Factorizable graph convolutional networks," *NeurIPS*, pp. 20 286–20 296, 2020.
[130] H. Li et al., "Disentangled contrastive learning on graphs," *NeurIPS*, pp. 21 872–21 884, 2021.
[131] H. Li et al., "Disentangled graph contrastive learning with independence promotion," *IEEE TKDE*, 2022.
[132] H. Li et al., "Ood-gnn: Out-of-distribution generalized graph neural network," *IEEE TKDE*, 2022.
[133] H. Cheng et al., "Disentangled feature representation for few-shot image classification," *arXiv*, 2021.
[134] A. Wu et al., "Vector-decomposed disentanglement for domain-invariant object detection," in *CVPR*, 2021, pp. 9342–9351.

Xin Wang is currently an Assistant 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 and recommendation in social media. He has published high-quality research papers in ICML, NeurIPS, IEEE TPAMI, IEEE TKDE, ACM KDD, WWW, ACM SIGIR, ACM Multimedia etc. He is the recipient of 2020 ACM China Rising Star Award and 2022 IEEE TCMC Rising Star Award.

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. His research interests are in the area of data-driven multimedia networking and Cross-media big data computing. He has published over 380 referred papers and is the inventor or co-inventor of over 80 patents. He received eight Best Paper Awards, including ACM Multimedia 2012 and IEEE Transactions on Circuits and Systems for Video Technology in 2001 and 2019.

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).