Cascade Variational Auto-Encoder for Hierarchical Disentanglement

Fudong Lin
University of Louisiana at Lafayette
Lafayette, LA, USA
fudong.lin1@louisiana.edu

Xu Yuan*
University of Louisiana at Lafayette
Lafayette, LA, USA
xu.yuan@louisiana.edu

Lu Peng
Tulane University
New Orleans, LA, USA
lpeng3@tulane.edu

Nian-Feng Tzeng
University of Louisiana at Lafayette
Lafayette, LA, USA
tzeng@louisiana.edu

ABSTRACT

While deep generative models pave the way for many emerging applications, decreased interpretability for larger model sizes and complexities hinders their generalizability to wide domains such as economy, security, healthcare, *etc.* Considering this obstacle, a common practice is to learn interpretable representations through latent feature disentanglement, aiming for exposing a set of mutually independent factors of data variations. However, existing methods either fail to catch the trade-off between the synthetic data quality and model interpretability, or consider the first-order feature disentangling only, overlooking the fact that a subset of salient features can carry decomposable semantic meanings and hence be of *high-order* in nature. Hence, we in this paper propose a novel generative modeling paradigm by introducing a Bayesian network-based regularizer on a cascade Variational Auto-Encoder (VAE). Specifically, this regularizer guides the learner to discover a representation space that comprises both first-order disentangled features and high-order salient features, with the feature interplay captured by the Bayesian structure. Experiments demonstrate that this regularizer gives us free control over the representation space and can guide the learner to discover decomposable semantic meanings by capturing the interplay among independent factors. Meanwhile, we benchmark extensive experiments on six widely-used vision datasets, and the results exhibit that our approach outperforms the state-of-the-art VAE competitors in terms of the trade-off between the synthetic data quality and model interpretability. Although our design is framed in the VAE regime, it in effect is generic and can be better amenable to both GANs and VAEs in terms of letting them concurrently enjoy both high model interpretability and high synthesis quality.

*Corresponding author.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA.

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557254

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; *Neural Networks*; *Regularization*; *Unsupervised Learning*.

KEYWORDS

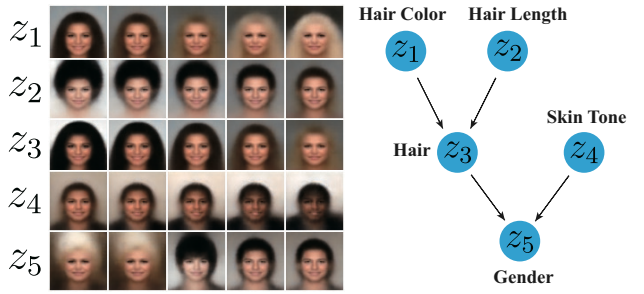
Interpretable Machine Learning; Deep Generative Models; Representation Learning; Bayesian Network

ACM Reference Format:
Fudong Lin, Xu Yuan, Lu Peng, and Nian-Feng Tzeng. 2022. Cascade Variational Auto-Encoder for Hierarchical Disentanglement. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557254

1 INTRODUCTION

Generative modeling has resulted in a plethora of computer vision applications, such as DeepFake [30], image-style translation [13], 3D-object generation [31], *etc.* Its common focus is on creating synthetic vision contents that are seemingly authentic in the view of human beings. Such modeling tends to require a large size and higher complexity for producing better synthesized data. However, a large and complex model suffers from diminished model interpretability, thus hindering its wide applicability to essential domains, like economy, security, and healthcare. For sound model interpretability, disentangled representation learning has become the *de facto* practice, popularized by InfoGAN [4] and β-VAE [10]. In essence, such learning results are derived from interpretable generative modeling, which extracts a set of latent features from the vision contents with each feature carrying one salient or meaningful characteristic of data variation independently. Generative modeling is based on either a Generative Adversarial Network (GAN) [6] or Variational Auto-Encoder (VAE) [15] as the backbone, with diverse recent studies, including [3, 5, 12, 22, 23]. All studies so far aim at optimizing latent representations jointly for i) reconstruction fidelity, which ensures the synthetic data to be human-indistinguishable, and ii) model interpretability, which encourages disentanglement and independency among the latent features.

Despite being reasonably effective, existing studies exhibit two drawbacks. First, prior arts focus on the first-order disentangling only, where each latent feature has to be purely independent from



(a) Latent feature traversals. **(b) Hierarchical structure.**

Figure 1: Qualitative results of running our approach on CelebA. (a) Illustration of the latent traversals, where each latent feature is traversed between $[-5, 5]$. (b) Semantic meanings of the latent features and the structure that characterizes their hierarchical dependencies.

the rest, overlooking the fact that the dependency structure underlying the latent space could be hierarchical and of *high-order* in nature. For example, as shown in Figure 1, the feature capturing semantic of "gender" is resulted from an interplay of two salient features corresponding to the variations of "hair" and "skin tone", where the "hair" can be further disentangled into two independent features describing the "hair color" and "hair length". We argue that only by capturing both the independent factors of data variations and the correlation among them, can a model be deemed as fully interpretable. So far, existing studies simply treat all features that do not show salient independent tendency as nuisance features; yet, no study has explored how to characterize the hierarchical structure for deeply explaining the potential interactions among latent features.

Second, prior studies have no provision for trading off synthetic data quality against interpretability satisfactorily. In particular, the GAN variants model the reconstruction process from the latent representations to the synthetic data in an implicit means, thereby having weak control over the relations among extracted features. This, along with the game-playing nature of training a GAN, makes the discovery of an equilibrium that suffices both the high-quality synthetic data and salient interpretable features at once extremely hard. On the other hand, the VAE variants that construct the representation space with a mixture of prior densities, albeit explicitly, have to use variational approximations to make the inference process tractable. Thus, extensive expert knowledge is entailed to dissect and decompose the VAE objectives into two blocks, respectively for handling the reconstruction fidelity and the latent feature disentanglement. Often, a good balance ratio between the two blocks has to be tuned ad-hoc, being a costly and time-consuming process.

To overcome the aforementioned challenges, we propose a novel generative modeling paradigm, aiming to better learn interpretable representations with an aid of high-order disentangling patterns while retaining data reconstruction fidelity. Our key idea is to model the dependency structure among the latent features with a Bayesian network, guiding the learner to extract a set of independent factors and their correlation. A cascade VAE architecture is tailored to extract two *conjugate* latent feature sets respectively from the raw input and the reconstructed data, by using two consecutive encoders that share parameters. By encouraging the Bayesian structure to form the two conjugate feature sets, an intermediate regularization is imposed on the VAE during learning. As such, this generative modeling paradigm enriches the latent space by capturing the independent factors of data variations and the dependency among them. Meanwhile, it guides the learner to discover the interactions among latent features in accordance with the hierarchical dependency structure, permitting our free control over the representation space.

Our specific contributions are summarized as follows.

(1) A novel generative model with better interpretability than those under the prior studies is realized by discovering the representation space that contains both first-order disentangled features and high-order salient features.

(2) A Bayesian network-based regularizer is crafted to pattern the high-order latent feature disentanglement, enabling our free control over the representation space as the regularizer guides the learner to extract independent factors and the interplay among them in accordance with the hierarchical dependency structure.

(3) Although the VAE architecture is used as the backbone of our design, this regularizer is generic and can be readily applicable to other generative models based on GANs to better the trade-off between reconstruction fidelity and interpretability. Our empirical experiments evidence this point.

(4) Extensive experiments are carried out, and the results substantiate that our approach outperforms the state-of-the-art VAE competitors in that it suffers from lower data reconstruction errors while discovering a more informative and interpretable latent representation space.

2 RELATED WORK

This section reviews the existing generative models that learn interpretable representations with deep neural networks, grouped into the GAN-based and the VAE-based methods.

Generative Adversarial Networks (GANs). Since the pioneer InfoGAN [4], learning disentangled latent representations has become a common practice of interpreting deep generative models. Basically, InfoGAN splits the input variables into two sets, with one for fake data synthesis via running a traditional GAN and the other for salient visual characteristics captured by maximizing mutual information between input variables and latent representations. Subsequent works include Elastic-InfoGAN [23] which generalizes InfoGAN into long-tail data distributions, Casual-GAN [18] which deals with sequential input, and Style-GAN [22] which allows the input image and synthetic data to have high-resolution. All these works inherit the two-player game-playing paradigm from GAN [6], where a generator strives to fool a discriminator learning toward classifying real and synthetic data. Such an adversarial training nature not only makes it very difficult to discover an equilibrium having both high reconstruction fidelity and latent feature disentanglement, but also leaves the data generating mechanism inside the generator in a black box, thereby undermining full model interpretability.

Variational Auto-Encoders (VAEs). VAE [15] constructs the latent representation space with a mixture of prior densities via minimizing the KL-divergence between the prior and the variational posteriors in an explicit means. Thus, VAE naturally encourages disentanglement among latent features to possess interpretability. The follow-up of β-VAE [10] claims that disentangling performance could be further improved by assigning a larger weight to the KL-divergence term. However, subsequent studies [11] argue that β-VAE, on the other hand, sacrifices the data reconstruction quality. To remedy this, [11] decomposes the KL-divergence term into two blocks: index-code mutual information (IMI) block and total correction (TC) block, with a more significant weight imposed on TC only while fixing that on IMI to yield better disentanglement without losing the reconstruction fidelity. Follow-up studies, including Factor-VAE [12], β-TCVAE [3], Bayes-Factor-VAE [14], and Guided-VAE [5], achieve better performance than β-VAE by adding newly crafted penalty terms to VAE objective (or substituting the TC block). Unfortunately, prior arts all focus only on the first-order latent feature disentanglement, missing out the high-order structure underlying the representation space where a rich set of salient features exists to represent an informative correlation of multiple independent factors. Disregarding this hierarchical structure makes it impossible to explore a full interpretation of the learned features, especially their interactions. The most relevant study to our work is Casual-VAE [33], where the authors model causal relationships via the masked directed acyclic graph (DAG). However, its design lacks flexibility, unable to be used to decompose causally-related latent factors. Also, its generalizability to other regimes, such as GAN, is unknown.

3 OUR PROPOSED APPROACH

We start by formulating our learning problem in Section 3.1, followed by highlighting the deficiencies hidden behind existing VAE variants from a mathematical perspective in Section 3.2. Then, we give the overview of our idea for learning hierarchical disentanglement in Section 3.3. We end by presenting the details of our Bayesian regularizer design in Section 3.4.

3.1 Problem Statement

Assume a set of raw inputs $\boldsymbol{x} \in \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ are represented by a set of latent variables $\boldsymbol{z} \in \{\boldsymbol{z}^{(i)}\}_{i=1}^{N}$, where $\boldsymbol{z}^{(i)} \in \mathbb{R}^d$. We express the generative model by a standard Gaussian distribution $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The encoder $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ are parameterized by two neural networks. The encoder produces the mean and variance of the variational posterior for raw data, where $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{d} \mathcal{N}(z_j \mid \mu_j(\boldsymbol{x}), \sigma_j^2(\boldsymbol{x}))$. The VAE aims to learn the marginal likelihood of raw inputs by maximizing the log evidence lower-bound (ELBO) $\mathcal{L}$:

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

where its first term can be interpreted as the negative synthetic error and the second term as a complex regularizer.

3.2 Challenges

Two challenges impede VAE variants' further advances in learning interpretable representations. First, β-VAE [10] claims that a larger weight ($\beta > 1$) on the second term in Eq. (1) can improve the interpretability by guiding VAE to learn disentangled representations, with the learning objective for β-VAE equal to:

$$\mathcal{L}_{\beta} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})\right] - \beta\, KL\left(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{z})\right). \quad (2)$$

We further break down the KL term in Eq. (1) as [11, 19]:

$$\begin{aligned} &\mathbb{E}_{p_{data}(\boldsymbol{x})}\left[KL(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{z}))\right] \\ &= I(\boldsymbol{x}; \boldsymbol{z}) + KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}) \,||\, \prod_{j=1}^{d} q_{\boldsymbol{\phi}}(z_j)), \end{aligned} \quad (3)$$

where $I(\boldsymbol{x}; \boldsymbol{z})$ denotes mutual information between $\boldsymbol{x}$ and $\boldsymbol{z}$. The penalty on the KL term encourages independence among latent dimensions and hence promotes disentanglement. However, the penalty on $I(\boldsymbol{x}; \boldsymbol{z})$ reduces the amount of information about the raw data $\boldsymbol{x}$ stored in the latent variable $\boldsymbol{z}$ and hence lowers synthetic data quality. Therefore, high values of β on the KL term result in better disentanglement but poor synthetic data quality [19]. We interpret this issue as *the trade-off between synthetic data quality and model interpretability*. Subsequent studies (*i.e.*, Factor-VAE and β-TCVAE) achieve a better trade-off by introducing the *total correlation* (TC) [29] penalty to vanilla VAE. According to Eq. (3), we rewrite the objective of β-VAE, Factor-VAE, and β-TCVAE to better understand this trade-off, arriving at:

$$\mathcal{L}_{qi} = \mathcal{L}_{\backslash qi} - \alpha\, I(\boldsymbol{z}; \boldsymbol{x}) - \beta\ KL(q_{\boldsymbol{\phi}}(\boldsymbol{z}) \,||\, \prod_{j=1}^{d} q_{\boldsymbol{\phi}}(z_j)),$$

where $\mathcal{L}_{\backslash qi}$ denotes the part of objective function unrelated to the trade-off between synthetic data quality and model interpretability. Factor-VAE and β-TCVAE can achieve a better trade-off (compared to β-VAE) by tuning a good balance ratio between the second and third terms. Unfortunately, tuning the ratio in an ad-hoc manner is a costly and time-consuming process, limiting their generalizability.

Second, previous arts [3, 10, 12, 14] are effective to capture independent factors of data variations, but their representation spaces all comprise first-order disentangled features only, unable to discover high-order salient features that convey decomposable semantic meanings. To learn high-order salient features, a learner should capture both independent factors and the correlation among them. However, previous work cannot characterize the hierarchical structure among latent features, making high-order disentangling unattainable. Meanwhile, having control over the representation space sheds the light on the generative process as it will i) become easier to discover what the independent factors of data variations represent and ii) be more flexible to explore the potential correlation among independent factors.

3.3 Our Idea

To overcome the two challenges, we follow conventional VAE for learning interpretations but with two new designs: i) developing cascade architecture to extract two sets of latent features and ii) introducing Bayesian network-based regularizer on vanilla VAE to explore independent factors of data variations and the interplay among them, making the discovery of high-order salient features that convey decomposable semantic meanings feasible. As illustrated in Figure 2, we frame our design in the VAE regime to involve

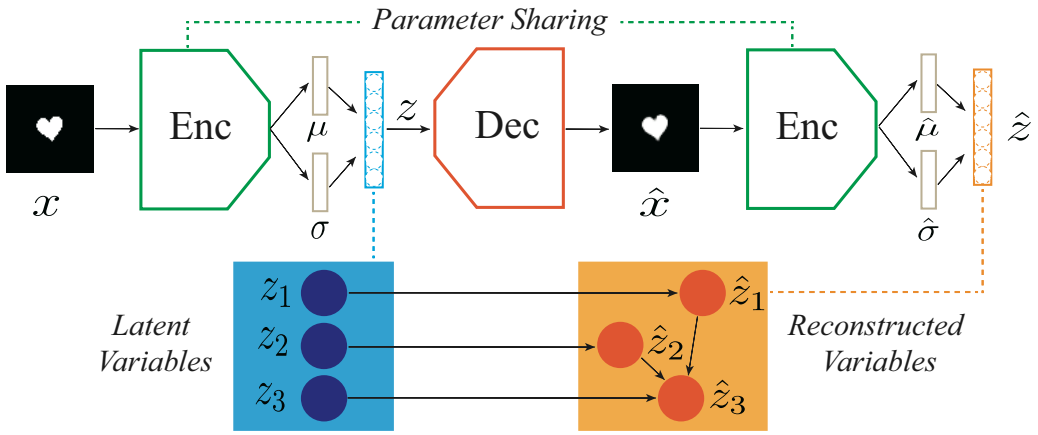


Figure 2: Architecture of the proposed cascade VAE, where the first encoder receives the raw data and the second encoder takes in the output of the decoder. A Bayesian network models the dependency structure formed between the two sets of salient features.

two parameter-sharing encoders that learn latent representations from the raw inputs and from the reconstructed data, respectively. First, we employ the Bayesian network to characterize the hierarchical structure among latent features. Two reasons motivate this: i) Bayesian network naturally encodes the correlation among latent features and ii) Bayesian network allows us to learn the causal relationship [8]. Second, similar to standard VAE, the latent variables in the first encoder learn the latent features underlying data. We model the hierarchical structure on the second encoder, aiming to guide the reconstructed variables to capture both latent features and the interactions among them in accordance with the dependency structure.

3.4 Bayesian Network-based VAE Regularizer

In this section, we present the details of the Bayesian network-based regularizer. Let *Enc* and *Dec* denote the encoder and the decoder, respectively, and $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ denote their respective parameters. Note that *Enc* includes two encoders that share the same parameters. Following the standard VAE [15], we sample the latent variable $\boldsymbol{z}$ from the standard Gaussian distribution. We assume that a set of N generated samples $\hat{\boldsymbol{x}} \in \{\hat{\boldsymbol{x}}^{(i)}\}_{i=1}^{N}$ are decoded by a set of latent variables $\boldsymbol{z} \in \{\boldsymbol{z}^{(i)}\}_{i=1}^{N}$, where $\boldsymbol{z}^{(i)} \in \mathbb{R}^d$ and $\hat{\boldsymbol{x}} = Dec(\boldsymbol{z})$. We assume that a set of reconstructed variables $\hat{\boldsymbol{z}} \in \{\hat{\boldsymbol{z}}^{(i)}\}_{i=1}^{N}$ are encoded by $\hat{\boldsymbol{x}}$, where $\hat{\boldsymbol{z}} = Enc(\hat{\boldsymbol{x}}) = Enc(Dec(\boldsymbol{z}))$.

We let a Bayesian network $\mathcal{B}$ model the joint distribution of latent variables $\boldsymbol{z}$ and reconstructed variables $\hat{\boldsymbol{z}}$, *w.r.t.* a given connectivity pattern, as a production of local distribution probabilities. Let $Pa(\hat{\boldsymbol{z}}_j)$ denote the parents of the j-th dimension of reconstructed latent variables $\hat{\boldsymbol{z}}$. Note that i) the latent variables $\boldsymbol{z}$ do not have parents as they are sampled directly from a standard Gaussian distribution; ii) any reconstructed variables $\hat{\boldsymbol{z}}_j$ can be parented by either latent or reconstructed variables according to the dependency structure $\mathcal{B}$. As such, the joint distribution of $\hat{\boldsymbol{z}}$ is defined as:

$$p_{\mathcal{B}}(\hat{\boldsymbol{z}}) = \prod_{j=1}^{d} p(\hat{\boldsymbol{z}}_j \mid Pa(\hat{\boldsymbol{z}}_j)). \tag{4}$$

Minimizing $KL(p(\hat{\boldsymbol{z}}|\hat{\boldsymbol{x}}) \,||\, p_{\mathcal{B}}(\hat{\boldsymbol{z}}))$, we are able to regularize VAE to discover latent features and the interactions among them in accordance with the dependency structure $\mathcal{B}$. Unfortunately, $p(\hat{\boldsymbol{z}}|\hat{\boldsymbol{x}})$ is intractable. As $\hat{\boldsymbol{z}}$ is sampled from $\hat{\boldsymbol{x}}$ by the encoder, we use $q_{\phi}(\hat{\boldsymbol{z}} \mid \hat{\boldsymbol{x}})$ to approximate $p(\hat{\boldsymbol{z}}|\hat{\boldsymbol{x}})$. Hence, the learning objective is then given by:

$$\mathcal{L}_{\mathcal{B}} = \mathcal{L}_{vae} - KL(q_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}} \mid \hat{\boldsymbol{x}}) \,||\, p_{\mathcal{B}}(\hat{\boldsymbol{z}})), \tag{5}$$

where $\mathcal{L}_{vae}$ denotes the learning objective of vanilla VAE, *i.e.*, Eq. (1). As we have mentioned before, the cascade architecture is utilized to avert the conflict between VAE and the Bayesian network-based regularizer.

As such, our approach enjoys both high synthetic data quality and model interpretability since i) modeling the joint distribution of the latent variables $\boldsymbol{z}$ and the reconstructed variable $\hat{\boldsymbol{z}}$ via the Bayesian network perseveres more information from raw inputs, contributing to high data reconstruction fidelity and ii) the hierarchical structure guides a learner to discover high-order salient features conveying decomposable semantic meanings, thereby promoting model interpretability.

Density-Ratio Trick. The prerequisite of using the *density-ratio trick* [21, 27] to minimize the KL divergence is to have access to the samples from both distributions. We first randomly choose a latent variable $\boldsymbol{z}^{(i)}$ from the standard Gaussian distribution $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and generate a sample $\hat{\boldsymbol{x}}^{(i)}$ by the decoder, where $\hat{\boldsymbol{x}}^{(i)} = Dec(\boldsymbol{z}^{(i)}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}^{(i)})$. Then we fetch a sample and the reconstructed variable $\hat{\boldsymbol{z}}^{(i)}$ from $q_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}} \mid \hat{\boldsymbol{x}})$, by the encoder, where $\hat{\boldsymbol{z}}^{(i)} = Enc(\hat{\boldsymbol{x}}^{(i)}) = q_{\phi}(\hat{\boldsymbol{z}}^{(i)} \mid \hat{\boldsymbol{x}}^{(i)})$. Since we have the latent variable $\boldsymbol{z}^{(i)}$ and the reconstructed variable $\hat{\boldsymbol{z}}^{(i)}$, we can access to a sample from $p_{\mathcal{B}}(\hat{\boldsymbol{z}})$ by using the dependency structure $\mathcal{B}$.

We follow Factor-VAE [12] to approximate the density ratio by a discriminator. Specifically, we train a discriminator D (e.g., an MLP, multi-layer perceptron) to estimate the probability $D(\hat{\boldsymbol{z}})$ with its input being a sample from $q_{\boldsymbol{\phi}}(\hat{\boldsymbol{z}} \mid \hat{\boldsymbol{x}})$ instead of $p_{\mathcal{B}}(\hat{\boldsymbol{z}})$, yielding:

$$KL\left(q_{\boldsymbol{\phi}}\ (\hat{\boldsymbol{z}} \mid \hat{\boldsymbol{x}}) \mid\mid p_{\mathcal{B}}(\hat{\boldsymbol{z}})\right) = \mathbb{E}_{q_{\boldsymbol{\phi}}\ (\hat{\boldsymbol{z}}|\hat{\boldsymbol{x}})}\left[\log \frac{q_{\boldsymbol{\phi}}\ (\hat{\boldsymbol{z}} \mid \hat{\boldsymbol{x}})}{p_{\mathcal{B}}\ (\hat{\boldsymbol{z}})}\right] \approx \mathbb{E}_{q_{\boldsymbol{\phi}}\ (\hat{\boldsymbol{z}}|\hat{\boldsymbol{x}})}\left[\log \frac{D(\hat{\boldsymbol{z}})}{1 - D(\hat{\boldsymbol{z}})}\right]. \tag{6}$$

We train the VAE and the discriminator jointly, where the $\mathcal{L}_{vae}$ term in Eq. (5) is updated using the learning objective of standard VAE, with the KL term replaced by the discriminator-based approximation from Eq. (6).

4 EXPERIMENTS

We conduct extensive experiments to show the performance of our solution. Our goal is threefold. First, to compare with state-of-the-art counterparts, we show our solution can regularize the learner towards a better trade-off between synthetic data quality and model interpretability (in Section 4.1) via the non-hierarchical structure. Second, we exhibit that our solution can effectively guide the learner to discover the high-order salient features via the hierarchical structures (in Sections 4.2 and 4.3). In the end, we validate that our method is generic in augmenting state-of-the-art generative models (in Section 4.4).

Datasets. We experiment on six synthetic or real-world datasets: i) **MNIST** [16]: 70,000 greyscale 28×28 examples of handwritten digits with 10 distinct categories; ii) **Fashion-MNIST** [32]: 70,000 greyscale 28×28 examples with 10 distinct categories; iii) **dSprites** [20]: 737,280 binary 64×64 examples of 2D shapes; iv) **3D Faces** [24]: 239,840 greyscale 64×64 examples of 3D faces; v) **3D Chairs** [1]: 86,366 RGB $64 \times 64 \times 3$ examples of 3D chair models; vi) **CelebA** [17]: 202,599 RGB $64 \times 64 \times 3$ examples of celebrity faces.

Compared Models. We compare to state-of-the-art generative models, *i.e.,* β-**VAE**, **Factor-VAE**, and β-**TCVAE**. For a fair comparison, we build VAE models in identical architecture and hyperparameters are set as reported in their original literatures. The proposed regularizer is implemented on the same encoder/decoder architecture as in [2], and the discriminator for the density-ratio trick is implemented on the same Multi-layer Perceptron (MLP) architecture as in Factor-VAE [12].

Metrics. We use four common metrics to evaluate the trade-off between synthetic data quality and model interpretability.

Inception Score (IS) [26] is effective to quantitatively evaluate the synthetic data quality, where the score is based on the following considerations: i) evaluated by the classifier, the class distribution of high-quality synthetic data should have low entropy and ii) high-quality synthetic data should enjoy high diversity, *i.e.*, the predictions on generated samples should cover all classes. [26] combines the two considerations into one score, arriving at:

$$IS = \exp\left(\mathbb{E}_{\boldsymbol{x} \sim p_g}\left[d_{KL}(p(y \mid \boldsymbol{x}), p(\boldsymbol{x}))\right]\right), \tag{7}$$

where p_g is the distribution of synthetic data, y is the prediciton on synthetic data made by the Inception Net [28] which was trained on ImageNet [25].

Frechet Inception Distance (FID) [9] is another way to quantify the synthetic data quality. The authors exploit a specific layer of Inception Net to embed data into a feature space and regard the embedding layer as a continuous multivariate Gaussian. The mean and covariance for both the real data and the synthetic data are estimated by the embedding layer. As such, the quality of synthetic data is measured by the Frechet distance between two Gaussians, yielding:

$$FID = \|\mu_x - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}}\right), \tag{8}$$

where (μ_x, Σ_x) and (μ_g, Σ_g) are the mean and covariance of the real data distribution and the synthetic data distribution, respectively.

Mutual Information Gap (MIG) [3] is to quantitatively evaluate the first-order feature disentanglement. We let $V = \{v_1, v_2, \ldots, v_k\}$ denote the set of ground-truth generative factors, $Z = \{z_1, z_2, \ldots, z_d\}$ denote the set of latent variables, and $I(z_i; v_j)$ denotes the mutual information between a latent variable z_i and a generative factor v_j. Given a generative factor v_j, MIG measures the difference between the maximal mutual information $I(z_*; v_j)$ and the second highest mutual information $I(z_o; v_j)$. The first-order feature disentanglement is evaluated by the averaged difference of all generative factors, *i.e.*,

$$MIG = \frac{1}{K}\sum_{j=1}^{K} \frac{1}{H(v_j)}\left(I(z_*; v_j) - I(z_o; v_j)\right), \tag{9}$$

where $H(v_j)$ is the entropy of v_j, serving to normalize the difference.

Factor Score [12] quantifies the first-order disentanglement in another aspect. Following the above notation, let $V = \{v_1, v_2, \ldots, v_k\}$ denote the set of ground-truth generative factors. Each time, the authors generate data by fixing the value of a generative factor v_j but randomly varying the values for other generative factors v_{-j}. Let $Z = \{z_1, z_2, \ldots, z_d\}$ denote the set of latent variables extracted from the generated data. Each dimension in latent variables $z_i \in Z$ is normalized by its standard deviation over the entire dataset. The index of dimension with lowest variations $d^* \in \{1, 2, \ldots, d\}$ and the target index $j \in \{1, 2, \ldots, k\}$ serve as a pair of data point (d^*, j) for a marjority-vote classifier. The factor score is measured by the accuracy of the classifier.

4.1 Trade-off between Synthetic Data Quality and Model Interpretability

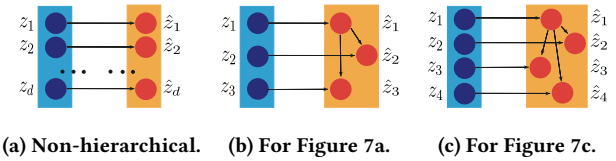
In this section, we exhibit the performance of our approach via the non-hierarchical dependency structure, so that we can compare our method with the state-of-the-arts in terms of synthetic data quality and model interpretability. The non-hierarchical dependency structure shown in Figure 3a is utilized for all the experiments in this section.

Synthetic Data Quality. We conduct experiments on the 6 datasets. The *Inception Score* (IS) [26] and *Frechet Inception Distance* (FID) [9] are used for quantitatively evaluating the synthetic data quality. Table 1 presents the experimental results. Note that a large IS or a small FID indicates high synthetic data quality.

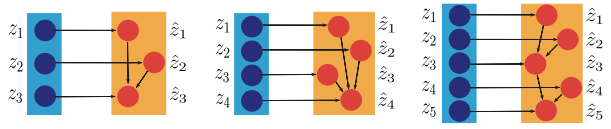
Following observations can be made from Table 1. First, β-VAE achieves the worst synthetic data quality with the IS of 1.89 and the FID of 88.8 on average. The reason is that the large weight ($\beta > 1$) on the $KL(q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}) \mid\mid p(\boldsymbol{z}))$ term reduces the amount of information about the raw data stored in latent variables. Hence, it suffers from a huge synthetic data quality loss. Second, Factor-VAE

Table 1: Experimental results (Mean Score ± Standard Deviation) for benchmark datasets. We utilize the Inception Score (IS) and Frechet Inception Distance (FID) to quantitatively evaluate the synthetic data quality. The best results are bold.

Score	Method	MNIST	Fashion-MNIST	3D Faces	3D Chairs	dSprites	CelebA
IS (↑)	β-VAE	2.00 ± .04	2.34 ± .06	1.65 ± .03	2.37 ± .07	1.24 ± .02	1.71 ± .03
	Factor-VAE	2.01 ± .04	2.51 ± .04	1.75 ± .04	2.67 ± .09	1.36 ± .03	1.73 ± .05
	β-TCVAE	2.02 ± .04	2.28 ± .06	1.79 ± .05	2.63 ± .13	1.53 ± .07	1.63 ± .04
	Ours.	**2.05 ± .04**	**2.99 ± .09**	**1.83 ± .05**	**3.23 ± .10**	**1.64 ± .06**	**2.50 ± .09**
FID (↓)	β-VAE	39.4 ± .17	74.5 ± .15	91.4 ± .68	86.1 ± .70	121.9 ± .12	119.4 ± .80
	Factor-VAE	**26.8 ± .15**	50.8 ± .09	61.9 ± .15	72.6 ± .58	76.9 ± .52	113.7 ± .53
	β-TCVAE	44.4 ± .14	82.6 ± .24	67.1 ± .72	74.7 ± .61	83.9 ± .68	111.6 ± .93
	Ours.	26.9 ± .14	**46.2 ± .09**	**52.4 ± .54**	**67.6 ± .69**	**64.7 ± .09**	**97.0 ± .54**



(a) Non-hierarchical. (b) For Figure 7a. (c) For Figure 7c.



(d) For Figures 8a and 11a. (e) For Figure 8c. (f) For Figures 1a and 9a.

Figure 3: Illustration of 6 different Bayesian network structures used in our study. (a) refers to a non-hierarchical structure, encouraging the salient features to be mutually independent; (b) and (c) correspond to hierarchical structures, where the high-order salient feature at the top is decomposed into respective two and three first-order disentangled ones; (d), (e) and (f) profile another type of hierarchical structures, where the high-order salient features at the bottom are synthesized from several first-order disentangled ones.

and β-TCVAE achieve better synthetic data quality levels, with 2.01 and 1.98 for their respective IS and 67.2 and 77.4 for their respective FID, on average. This is because tuning a good balance ratio lowers the synthetic data quality loss. Third, our approach achieves the best synthetic data quality by exhibiting the averaged IS and FID of 2.37 and 59.0, respectively. The statistical evidence exhibits our approach to outperform β-VAE, Factor-VAE and β-TCVAE across 6 datasets, with respective 20.3%, 15.2%, and 16.5% IS improvement, and respective 35.6%, 12.2%, and 23.7% FID reduction on average. The reason is that the Bayesian structure provides additional information per se, *i.e.*, the dependency structure penalizes difference between the latent variables z and the reconstructed variables $\hat{z}$, therefore in turn yielding improved synthetic data quality.

Interpretability. We conduct experiments to qualitatively and quantitatively evaluate the model interpretability. In particular, we focus on the first-order disentanglement, which conveys non-decomposable semantic meanings. To qualitatively evaluate the

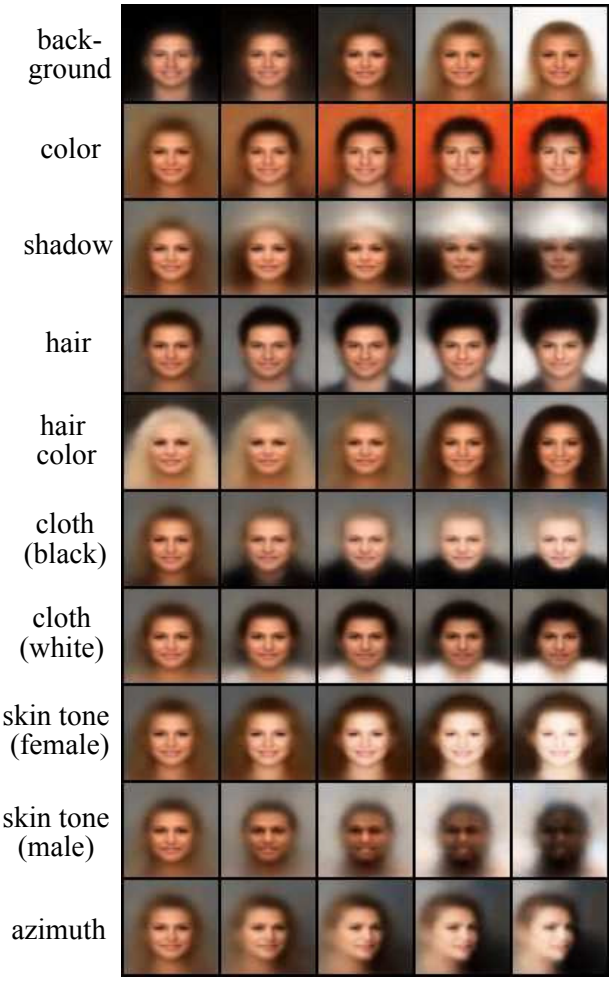
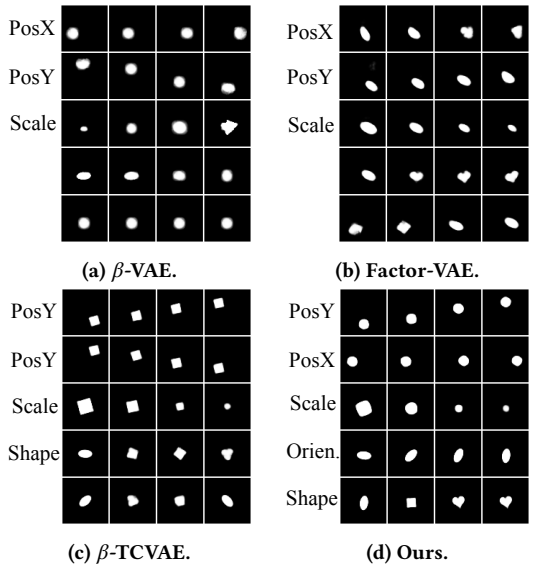


Figure 4: Illustration of first-order salient feature disentanglement under our approach on CelebA.

first-order disentangling, we generate samples by varying all z's from -2 to 2 for dSprites and from -5 to 5 for CelebA in evenly spaced intervals, respectively. The qualitative result of our method on CelebA is shown in Figure 4, where our approach learns 10 disentangled features on CelebA. This shows that our approach achieves



(a) β-VAE. (b) Factor-VAE.

(c) β-TCVAE. (d) Ours.

Figure 5: Comparison of latent traversals on dSprites. The competitors (a), (b), and (c) discover 3 distinct ground truth factors, while our approach (d) captures all 5 distinct factors: *Position Y, Position X, Scale, Orientation,* and *Shape.*

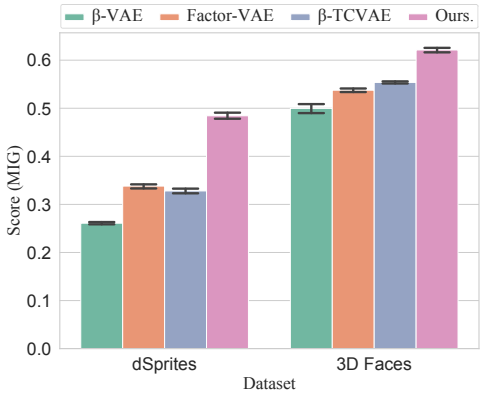
high interpretability on the RGB dataset. The comparative results between our approach and its every competitor under dSprites are shown in Figure 5. Comparing to Figure 5a, 5b, and 5c (for β-VAE, Factor-VAE, and β-TCVAE, respectively), Figure 5d (under our approach) clearly achieves the best synthetic data quality and learns the most first-order disentangled features. This demonstrates that our approach outperforms all competitors in terms of the synthetic data quality and model interpretability.

We next quantitatively compare our method to VAE variants in terms of two disentanglement metrics, namely, Mutual Information Gap (MIG) [3] and Factor score [12]. Note that MIG and Factor score measure the first-order disentanglement only. We train our model on dSprites and 3D Faces. Figures 6a and 6b present the comparative results, clearly demonstrating that our approach achieves the best first-order disentangling. Our approach achieves the best averaged MIG (or Factor score) of 0.484 (or 0.744) for dsprites and of 0.621 (or 0.861) for 3D faces. The statistical evidence exhibits that our approach outperforms β-VAE, Factor-VAE, and β-TCVAE, with respective 31.3% (or 20.2%), 20.8% (or 6.4%), and 20.3% (or 6.3%) improvement on average in terms of MIG (or Factor score).

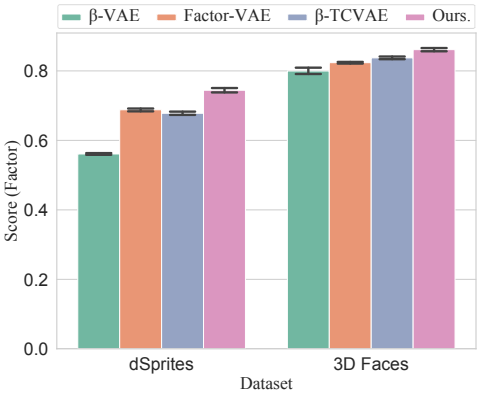
Notably, better first-order feature disentangling allows a model to convey more non-decomposable semantic meanings, thereby improving model interpretability.

4.2 Decomposing High-order Salient Feature into First-order Disentangled ones

We next present how the hierarchical dependency structures guide the learner to discover high-order salient features that convey decomposable semantic meanings, by decomposing them into first-order disentangled ones. We employ the dependency structure of Figure 3b to guide the learner to decompose the higher-order salient

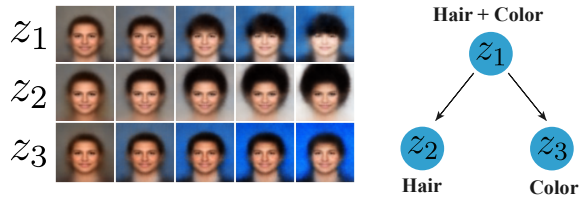


(a) MIG.

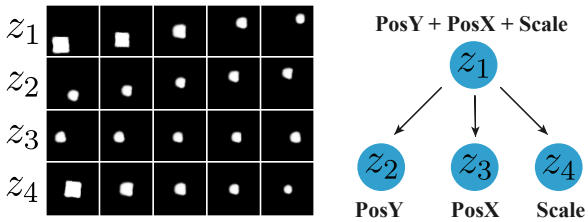


(b) Factor score.

Figure 6: Disentanglement scores on dSprites and 3D Faces. (a) MIG and (b) Factor score.



(a) Latent traversals on CelebA. (b) Structure.



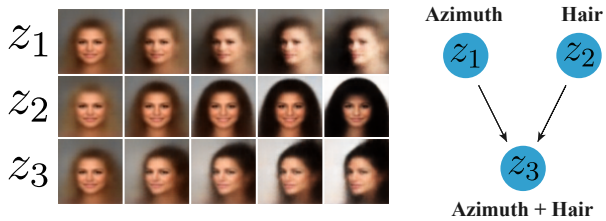
(c) Latent traversals on dSprites. (d) Structure.

Figure 7: Illustration of our dependency structure guiding the learner to decompose a high-order salient feature into (a) two and (c) three first-order disentangled ones. (b) and (d) illustrate corresponding hierarchical structures.

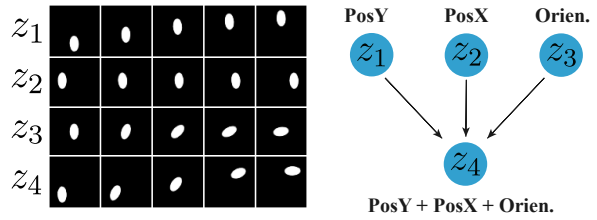
feature into two first-order disentangled ones. We train our model on CelebA and generate samples by varying all z's from -5 to 5 in evenly spaced intervals. Figures 7a and 7b respectively present latent feature traversals and their corresponding hierarchical structure. From Figure 7a, we observe that z_1 captures decomposable semantic meanings, *i.e.*, an interplay of *Hair* and *Color*, while z_2 and z_3 capture corresponding non-decomposable semantic meanings, *i.e.*, *Hair* and *Color*, respectively. This demonstrates that the hierarchical dependency structure guides the learner to decompose a high-order salient feature z_1 into two first-order disentangled ones (z_2 and z_3), as Figure 7b shows.

Similar experiments are conducted on dSprites where we employ the dependency structure of Figure 3c. We generate samples by varying all z's from -2 to 2 in evenly spaced intervals after training, and Figures 7c and 7d depict the qualitative results. We observe that our approach guides the learner to decompose the high-order salient feature z_1, conveying composed of *Position Y*, *Position X*, and *Scale*, into three first-order disentangled ones z_2, z_3, and z_4, representing *Position Y*, *Position X*, and *Scale*, respectively.

4.3 Synthesize High-order Salient Feature from Multiple First-order Disentangled Ones



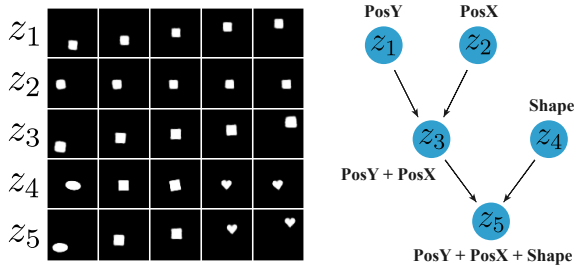
(a) Latent traversals on CelebA. (b) Structure.



(c) Latent traversals on dSprites. (d) Structure.

Figure 8: Illustration of our dependency structure guiding the learner to synthesize first-order disentangled features to form a second-order salient feature. (a) and (c) refer to latent feature traversals on CelebA and dSprites, respectively; and (b) and (d) illustrate corresponding hierarchical structures.

We next signify how the dependency structure guides the learner to capture decomposable semantic meanings by synthesizing a high-order salient feature from multiple first-order disentangled ones. We first employ the dependency structure of Figure 3d to guide the learner to synthesize a second-order salient one from two first-order disentangled ones. After training the model on CelebA, we generate samples by varying all z's from -5 to 5 in evenly spaced intervals. Figure 8a shows the resulting latent feature traversals. We observe that the latent variables z_1 and z_2 capture non-decomposable semantic meanings, respectively representing the *Azimuth* and *Hair*, while z_3 captures a decomposable semantic meaning, conveying an interplay of *Azimuth* and *Hair*. This affirms that the dependency structure guides the learner to compose two first-order disentangled features (*i.e.*, z_1 and z_2) into a second-order salient feature (*i.e.*, z_3), as shown in Figure 8b. Then, we conduct experiments on dSprites by employing the dependency structure in Figure 3e and generate samples by varying all z's from -2 to 2 in evenly spaced intervals. Figure 8c shows the qualitative result, where z_4 captures the interplay of *Position Y*, *Position X*, and *Orientation*, conveying the semantic meaning of z_1 (*i.e.*, *Position Y*), z_2 (*i.e.*, *Position X*), and z_3 (*i.e.*, *Orientation*) jointly.



(a) Latent traversals on dSprites. (b) Structure.

Figure 9: Illustration of our dependency structure guiding the learner to synthesize first-order disentangled features to form a third-order salient feature.

We next employ the dependency structure of Figure 3f to guide the learner to capture a complex interplay of latent features under dSprites. The samples are generated by varying z's from -2 to 2. Figure 9a presents the latent feature traversals. It is seen that z_1, z_2, and z_4 capture non-decomposable semantic meanings, respectively conveying *Position Y*, *Position X*, and *Shape*. The latent variables z_3 and z_5 capture decomposable semantic meanings, where z_3 represents composed of *Position Y* and *Position X*, and z_5 conveys an interplay of *Position Y*, *Position X*, and *Shape*. This further demonstrates the effectiveness of our dependency structure to guide the learner to synthesize a third-order salient feature from multiple first-order disentangled ones.

4.4 Generalizing to Other Generative Models

In this section, we conduct experiments to validate that our Bayesian-network-based regularizer is generic and can be readily applicable to GAN-based and VAE-based generative models. For the GAN-based generative models, the model architecture is similar to InfoGAN [4, 7], where we exploit the Bayesian network to model the joint distribution of the latent variables z in the generator and the reconstructed variable $\hat{z}$ in the discriminator. Aiming to apply the proposed regularizer on the VAE-based generative models, we can simply replace the $\mathcal{L}_{vae}$ term in Eq. (5) with the learning objective of VAE variants, *i.e.*, β-VAE, Factor-VAE, and β-TCVAE.

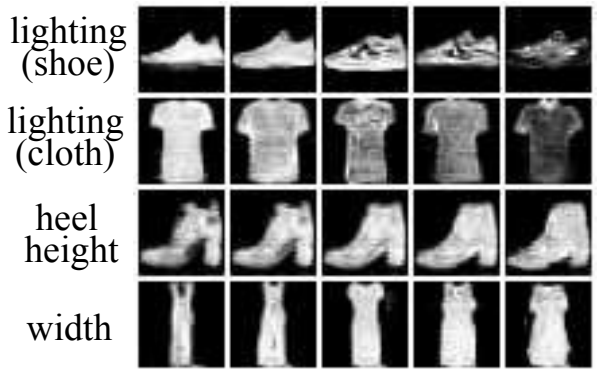


Figure 10: Illustration of first-order feature disentanglement on Fashion-MNIST.

(a) Latent traversals on MNIST. **(b) Structure.**

Figure 11: Illustration of our dependency structure guiding InfoGAN to discover high-order salient features.

We shall qualitatively evaluate the performance of the Bayesian-network-based regularizer on the GAN-based model. First, we employ the non-hierarchical dependency structure of Figure 3a to guide InfoGAN to learn mutually independent factors of data variations. After training, we generate samples by varying all z's from -10 to 10 in evenly spaced intervals. Figure 10 shows the qualitative results of Fashion-MNIST, whereas our approach generates the *sneaker*, *T-shirt*, *boot*, and *dress* with a high quality, with each latent dimension learning a distinct first-order disentangled feature. This demonstrates our approach can guide InfoGAN to achieve a high generated data quality and high model interpretability at once.

Then, we utilize the hierarchical dependency structure of Figure 3d to guide InfoGAN to capture decomposable semantic meanings by synthesizing first-order disentangled features to a high-order salient one. After training on MNIST, we vary all z's from -12 to 12, with the generated result depicted in Figure 11a, where the second-order salient feature (*i.e.*, z_3), representing the interactions among *size* and *rotate*, is synthesized from two first-order disentangled features *size* (*i.e.*, z_1) and *rotate* (*i.e.*, z_2). The corresponding hierarchical structure is illustrated in Figure 11b. This confirms that our approach can guide the GAN-based model to learn high-order salient features.

Next, we verify whether the proposed regularizer can improve VAE-based generative models in terms of first-order feature disentanglement. We exploit the dependency structure of Figure 3a to guide β-VAE, Factor-VAE, and β-TCVAE to learn first-order disentangled features. After training on dSprites, we utilize *mutual information gap* (MIG) [3] to quantitatively evaluate the disentanglement score. Figure 12 presents the MIG score of VAE variants with and without the Bayesian-network-based regularizer. The GAN-based model InfoGAN is also included in this experiment. As shown in Figure 12, our approach improves the performance of InfoGAN, β-VAE, Factor-VAE, and β-TCVAE, with 26.1%, 48.0%, 30.5%, and 30.6% disentanglement improvement, respectively.

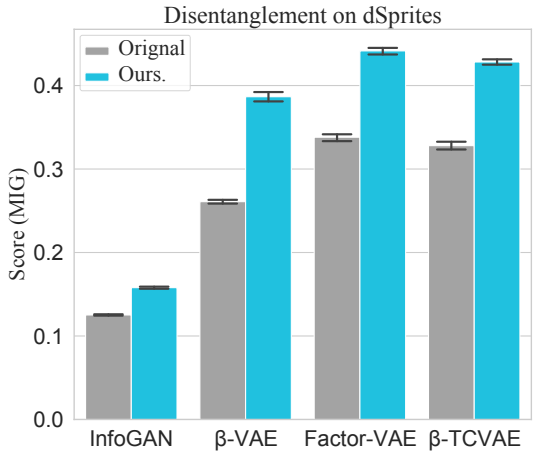


Figure 12: Illustration of disentanglement score (MIG) for different modeling algorithms.

In summary, the aforementioned experimental results validate that our method is generic and can be applicable to both GAN-based and VAE-based generative models, improving their performance in terms of the synthetic data quality and model interpretability.

5 CONCLUSION

This paper has proposed a novel generative modeling paradigm that can synthesize human-indistinguishable vision contents while possessing strong interpretability. Our key idea lies in characterizing a hierarchical dependency structure, comprising the first-order disentangled features and high-order salient features carrying interactions among first-order disentangled features. Such a hierarchical structure results from imposing a Bayesian-network-based regularizer on a cascade variational auto-encoder (VAE) to arrive at a novel generative modeling paradigm. With this paradigm, we have bettered known modern deep-learning-based generative models, including GANs and VAEs, in terms of both interpretability and reconstruction fidelity. Meanwhile, this paradigm guides the learner to capture independent factors of data variations and their correlation in accordance with the hierarchical dependency structure, offering free control over the representation space. Extensive experiments have been carried out, with their results evidencing the effectiveness of our approach. We hope our work can shed the light on extracting more informative semantics from vision contents by learning interpretable representations that involve dependencies.

ACKNOWLEDGMENTS

This work was supported in part by NSF under Grants 1763620, 1948374, 2019511, and 2146447. Any opinion and findings expressed in the paper are those of the authors and do not necessarily reflect the view of funding agency.

REFERENCES

[1] Mathieu Aubry, Daniel Maturana, Alexei A. Efros, Bryan C. Russell, and Josef Sivic. 2014. Seeing 3D Chairs: Exemplar Part-Based 2D-3D Alignment Using a Large Dataset of CAD Models. In *CVPR*. 3762–3769.

[2] Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in β-VAE. *CoRR* abs/1804.03599 (2018).

[3] Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *NIPS*. 2615–2625.

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *NIPS*. 2172–2180.

[5] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. 2020. Guided Variational Autoencoder for Disentanglement Learning. In *CVPR*. 7917–7926.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *NIPS* 27 (2014), 2672–2680.

[7] Yi He, Fudong Lin, Xu Yuan, and Nian-Feng Tzeng. 2021. Interpretable Minority Synthesis for Imbalanced Classification. In *IJCAI*. 2542–2548.

[8] David Heckerman. 1998. A Tutorial on Learning with Bayesian Networks. In *Learning in Graphical Models*. Vol. 89. 301–354.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NIPS*. 6626–6637.

[10] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

[11] Matthew D Hoffman and Matthew J Johnson. 2016. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, Vol. 1. 2.

[12] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by Factorising. In *ICML*, Vol. 80. 2654–2663.

[13] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. 2020. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*.

[14] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. 2019. Bayes-Factor-VAE: Hierarchical Bayesian Deep Auto-Encoder Models for Factor Disentanglement. In *ICCV*. 2979–2987.

[15] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

[16] Yann LeCun and Corinna Cortes. 2010. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/. (2010). http://yann.lecun.com/exdb/mnist/

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.

[18] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled Self-Supervision in Sequential Recommenders. In *SIGKDD*. 483–491.

[19] Alireza Makhzani and Brendan J. Frey. 2017. PixelGAN Autoencoders. In *NIPS*. 1975–1985.

[20] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. 2017. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/.

[21] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. 2010. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Trans. Inf. Theory* 56, 11 (2010), 5847–5861.

[22] Weili Nie, Tero Karras, Animesh Garg, Shoubhik Debnath, Anjul Patney, Ankit Patel, and Animashree Anandkumar. 2020. Semi-supervised stylegan for disentanglement learning. In *ICML*. 7360–7369.

[23] Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, and Yong Jae Lee. 2020. Elastic-InfoGAN: Unsupervised Disentangled Representation Learning in Class-Imbalanced Data. In *NIPS*.

[24] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *AVSS*. 296–301.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* (2015).

[26] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *NIPS*. 2226–2234.

[27] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. 2012. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics* 64, 5 (2012), 1009–1044.

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*.

[29] Michael Satosi Watanabe. 1960. Information Theoretical Analysis of Multivariate Correlation. *IBM J. Res. Dev.* 4, 1 (1960), 66–82.

[30] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology Innovation Management Review* 9, 11 (2019).

[31] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*.

[32] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[33] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *CVPR*.