*dp*VAEs: Fixing Sample Generation for Regularized VAEs

Riddhish Bhalodia[1], Iain Lee[1], and Shireen Elhabian[1]

Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, Utah, USA
{riddhishb, iclee, shireen}@sci.utah.edu

Abstract. Unsupervised representation learning via generative modeling is a staple to many computer vision applications in the absence of labeled data. Variational Autoencoders (VAEs) are powerful generative models that learn representations useful for data generation. However, due to inherent challenges in the training objective, VAEs fail to learn useful representations amenable for downstream tasks. Regularization-based methods that attempt to improve the representation learning aspect of VAEs come at a price: poor sample generation. In this paper, we explore this representation-generation trade-off for regularized VAEs and introduce a new family of priors, namely *decoupled priors*, or *dp*VAEs, that decouple the representation space from the generation space. This decoupling enables the use of VAE regularizers on the representation space without impacting the distribution used for sample generation, and thereby reaping the representation learning benefits of the regularizations without sacrificing the sample generation. *dp*VAE leverages invertible networks to learn a bijective mapping from an arbitrarily complex representation distribution to a simple, tractable, generative distribution. Decoupled priors can be adapted to the state-of-the-art VAE regularizers without additional hyperparameter tuning. We showcase the use of *dp*VAEs with different regularizers. Experiments on MNIST, SVHN, and CelebA demonstrate, quantitatively and qualitatively, that *dp*VAE fixes sample generation and improves representation for regularized VAEs.

1 Introduction

Is it possible to learn a *powerful generative model* that matches the true data distribution with *useful data representations* amenable to downstream tasks in an unsupervised way? —This question is the driving force behind most unsupervised representation learning via state-of-the-art (SOTA) generative modeling methods (*e.g.,* [1–4]), with applications in artificial creativity [5, 6], reinforcement learning [7], few-shot learning [8], and semi-supervised learning [9]. A common theme behind such works is learning the data generation process using *latent variable models* [10, 11] that seek to learn representations useful for data generation; an approach known as *analysis-by-synthesis* [12, 13].

Variational autoencoders (VAEs) [14, 15] marry latent variable models and deep learning by having independent, network-parameterized *generative* and *inference* models that are trained jointly to maximize the marginal log-likelihood of the training data. VAE introduces a variational posterior distribution that approximates the true posterior to derive a tractable lower bound on the marginal log-likelihood, a.k.a. the evidence lower bound (ELBO). The ELBO is then maximized using stochastic gradient descent by virtue of the reparameterization trick [14, 15]. Among the many successes of VAEs in representation learning tasks, VAE-based methods have demonstrated SOTA performance for semi-supervised image and text classification tasks [16, 9, 17, 8, 18, 19].

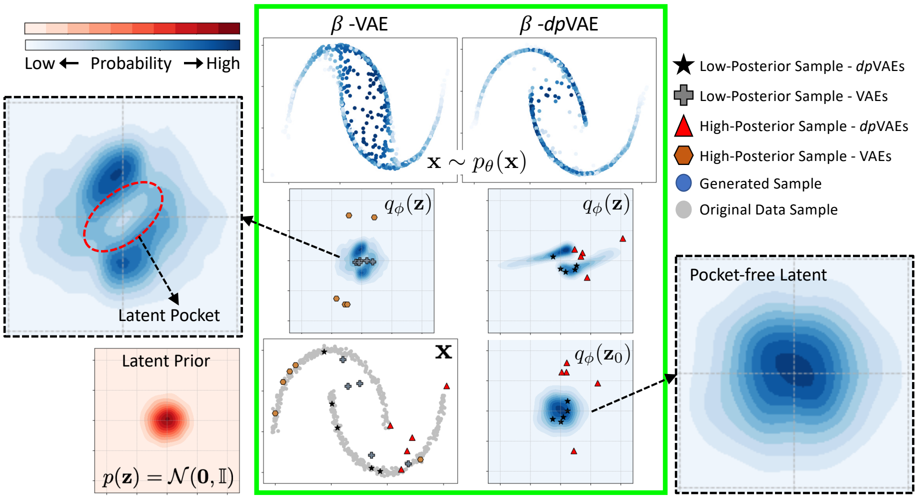

Figure 1. *dp*VAE fixes sample generation for a regularized VAE. The green box shows β-VAE [2] (left column) and β-VAE with the proposed *decoupled prior* (right column), each trained on the two moons dataset. β-VAE: Top to bottom shows the generated samples (colors reflect probability of generation), the aggregate posterior $q_\phi(\mathbf{z})$ and the training samples. The low-posterior samples lie in the latent pockets of $q_\phi(\mathbf{z})$ (shown in enlarged section on the left) and correspond to off-manifold samples in the data space, and high-posterior samples correspond to latent leaks. The β-*dp*VAE decouples the representation $\mathbf{z}$ and generation $\mathbf{z}_0$ spaces. The generation space is pocket-free and very close to standard normal, resulting in generating samples on the data manifold. Furthermore, the representation learning is well established in the representation space (see section 4.1 for more discussion).

Representation learning via VAEs is ill-posed due to the disconnect between the ELBO and the downstream task [20]. Specifically, optimizing the marginal log-likelihood is not always sufficient for good representation learning due to the inherent challenges rooted in the ELBO that result in the tendency to *ignore latent variables* and *not encode information about the data in the latent space* [1, 11, 20–22]. To improve the representations learned by VAEs, a slew of regularizations have been proposed. Many of these regularizers act on the VAE latent space to promote specific characteristics in the learned representations, such as disentanglement [1–3, 6, 23, 24] and informative latent codes [25, 26]. However, better representation learning usually sacrifices sample generation, which is manifested by a distribution mismatch between the marginal (a.k.a. aggregate) latent posterior and the latent prior. This mismatch results in *latent pockets and leaks*; a *submanifold* structure in the latent space (a phenomena demonstrated in Figure 1 and explored in more detail in section 4.1). Latent pockets contain samples that are highly supported under the prior but not covered by the aggregate posterior (*i.e., low-posterior samples* [27]), while latent leaks contain samples supported under the aggregate posterior but less likely to be generated under the prior (*i.e., high-posterior samples*). This behavior has been reported for vanilla VAE [27, 28] but it is substantiated by augmenting the ELBO with regularizers (see Figure 1).

To address this representation-generation trade-off for regularized VAEs, we introduce the idea of decoupling the latent space for representation (*representation space*) from the space that drives sample generation (*generation space*); presenting a general framework for VAE regularization. To this end, we propose a new family of latent priors for VAEs — *decoupled priors* or *dp*VAEs — that leverages the merits of invertible deep networks. In particular, *dp*VAE transforms a tractable, simple base prior distribution in the generation space to a more expressive prior in the representation space that reflects the submanifold structure dictated by the regularizer. This is done using an invertible mapping that is jointly trained with the VAE's inference and generative models. SOTA VAE regularizers can thus be directly plugged in to promote specific characteristics in the representation space without impacting the distribution used for sample generation. We showcase, quantitatively and qualitatively, that *dp*VAE with different SOTA regularizers improve sample generation, without sacrificing their representation learning benefits.

It is worth emphasizing that, being likelihood-based models, VAEs are trained to put probability mass on all training samples, forcing the model to *over-generalize* [29], and generating blurry samples (*i.e.,* off data manifold). This is in contrast to generative adversarial networks (GANs) [30] that generate outstanding image quality but could lack the full data support [31]. *dp*VAE is not expected to resolve the over-generalization problem in VAEs, but to mitigate poor sample quality resulting from regularization.

The contribution of this paper is fourfold:

- Analyze the latent submanifold structure induced by VAE regularizers.
- Introduce a decoupled prior family for VAEs as a general regularization framework that improves both sample generation and representation learning, without sacrificing the ELBO of the vanilla VAE.
- Derive the *dp*VAE ELBO of SOTA regularized VAEs; β-*dp*VAE, β-TC-*dp*VAE, Factor-*dp*VAE, and Info-*dp*VAE.
- Demonstrate empirically on three benchmark datasets the improved generation performance and the preservation of representation characteristics promoted via regularizers without additional hyperparameter tuning.

2 Related Work

To improve sample quality, a family of approaches exist that combine the inference capability of VAEs and the outstanding sample quality of GANs [30]. Leveraging the density ratio trick [30, 32] that only requires samples, VAE-GAN hybrids in the latent (*e.g.,* [33, 34]), data (*e.g.,* [27, 29]), and joint (both latent and data *e.g.,* [35]) spaces avoid restrictions to explicit posterior and/or likelihood distribution families, paving the way for marginals matching [27]. However, such hybrids scale poorly with latent dimensions, lack accurate likelihood bound estimates, and do not provide better quality samples than GAN variants [27]. For instance, VAE variants, such as adversarial [36] and Wasserstein [37] autoencoders, introduce matching penalties (*e.g.,* adversarial or maximum mean discrepancy regularizers) to match distributions in the latent space. Nonetheless, such matching penalties, in contrast to *dp*VAE, modify the likelihood lower bound and use looser bounds for training, and hence introduce a trade off with sample reconstruction [38]. Expressive posterior distributions can lead to better sample quality [33, 39] and are essential to prevent latent variables from being ignored in case of powerful generative models [21]. But results in [27] suggest that the posterior distribution is not the main learning roadblock for VAEs.

More recently, the key role of the prior distribution family in VAE training has been investigated [22, 27, 28]; poor latent representations are often attributed to restricting the latent prior to an overly simplistic distribution. Furthermore, Xu *et al.* presented a formal proof of the necessity and effectiveness of learning the latent prior and theoretically analyzed the failure of the aggregate posterior to match the unit Gaussian prior [28]. This motivates several works to enrich VAEs with more expressive priors. Bauer and Mnih addressed the distribution mismatch between the aggregate posterior and the latent prior by learning a sampling function, parameterized by a neural network, in the latent space [40]. However, this resampled prior requires the estimation of the normalization constant and dictates an inefficient iterative sampling, where a truncated sampling could be used at the price of a less expressive prior due to smoothing. Tomczak and Welling proposed the variational mixture of posteriors prior (VampPrior), which is a parameterized mixture distribution in the latent space given by a fixed number of learnable pseudo (*i.e.,* virtual) data points [41]. VampPrior sampling is non-iterative and is therefore fast. However, density evaluation is expensive due to the requirement of a large number of pseudo points, typically in the order of hundreds, to match the aggregate posterior [40]. A cheaper version is a mixture of Gaussian prior proposed in [42], which gives an inferior performance compared to VampPrior and is more challenging to optimize [40]. Autoregressive priors (*e.g.,* [43, 44]) come with fast density evaluation but a slow, sequential sampling process. VQ-VAE [45, 46] learns VAE prior using PixelCNN [47, 48] to improve sample quality. Yet, unlike *dp*VAE, VQ-VAE is not trained end-to-end and modify the underlying assumption of latent Gaussian models.

The proposed decoupled prior is inspired by flow-based generative models [39, 49–51], which have shown their efficacy in generating images (*e.g.,* GLOW [52]). Such methods hinge on architectural designs that make the model invertible. However, the strict invertibility of these architectures dictate very high-dimensional latent spaces, which are not condusive to representation learning and lead to computationally expensive and oftentimes prohibitively long training. In the context of VAEs, learning latent prior using invertible networks has been proposed by several works with the potential of generating high quality samples [38, 53–55]. Nonetheless, the inherent trade-off between sample representation and generation has not been explored. Such a trade-off is substantiated with regularizers that promote predefined characteristics in the latent space, providing looser bounds for training. Here, we showcase the impact of these looser bounds on sample generation and how *dp*VAE fixes sample generation.

With differences between expressiveness and efficiency, none of these methods address the fundamental challenge of VAE training in concert with existing representation-driven regularization frameworks. The proposed decoupled family of priors addresses the mismatch between the latent prior and the aggregate posterior, which improves sample generation performance and is easy to integrate with existing VAE regularizers that endow representation learning properties to VAEs. Further, the decoupled prior by itself solves the fundamental problem of representation learning in VAEs without using ad-hoc regularizers (see Table 1).

3 Background

In this section, we briefly lay down the foundations and motivations essential for the proposed VAE formulation.

3.1 Variational Autoencoders

VAE seeks to match the learned model distribution $p_\theta(\mathbf{x})$ to the true data distribution $p(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^D$ is the observed variable in the data space. The generative and inference models in VAEs are thus jointly trained to maximize a tractable lower bound $\mathcal{L}(\theta, \phi)$ on the marginal log-likelihood $\mathbb{E}_{p(\mathbf{x})}\left[\log p_\theta(\mathbf{x})\right]$ of the training data, where $\mathbf{z} \in \mathbb{R}^L$ is an unobserved latent variable in the latent space with a prior distribution $p(\mathbf{z})$, such as $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbb{I})$.

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

where θ denotes the generative model parameters, ϕ denotes the inference model parameters, and $q_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}(\mathbf{x}), \boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x}))$ is the variational posterior distribution that approximates the true posterior $p(\mathbf{z}|\mathbf{x})$, where $\boldsymbol{\mu}_\mathbf{z}(\mathbf{x}) \in \mathbb{R}^L$, $\boldsymbol{\Sigma}_\mathbf{z}(\mathbf{x}) = \mathrm{diag}(\boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}))$, and $\boldsymbol{\sigma}_\mathbf{z}(\mathbf{x}) \in \mathbb{R}^L_+$.

Since the ELBO seeks to match the marginal data distribution without penalizing the poor quality of latent representation, VAE can easily ignore latent variables if a sufficiently expressive generative model $p_\theta(\mathbf{x}|\mathbf{z})$ is used (*e.g.,* PixelCNN [47]) and still maximize the ELBO [11, 56, 21], a property known as *information preference* [21, 1]. Furthermore, VAE has the tendency to not encode information about the observed data in the latent codes since maximizing the ELBO is inherently minimizing the mutual information between $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ and $\mathbf{x}$ [22]. Without further assumptions or inductive biases, these failure modes hinder learning useful representations for downstream tasks.

3.2 Invertible Deep Networks

The proposed decoupled prior family for VAEs leverages flow-based generative models that are formed by a sequence of *invertible* blocks (*i.e.,* transformations), parameterized by deep networks. Consider two random variables $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^L$ and $\mathbf{z}_0 \in \mathcal{Z}_0 \subset \mathbb{R}^L$. There exist a bijective mapping between $\mathcal{Z}$ and $\mathcal{Z}_0$ defined by a function $g : \mathcal{Z} \rightarrow \mathcal{Z}_0$, where $g(\mathbf{z}) = \mathbf{z}_0$, and its inverse $g^{-1} : \mathcal{Z}_0 \rightarrow \mathcal{Z}$ such that $\mathbf{z} = g^{-1}(\mathbf{z}_0)$. Given the above condition, we can define the *change of variable formula* for mapping probability distribution on $\mathbf{z}$ to $\mathbf{z}_0$ as follows:

$$p(\mathbf{z}) = p(\mathbf{z}_0)\left|\frac{\partial \mathbf{z}_0}{\partial \mathbf{z}}\right| = p(g(\mathbf{z}))\left|\frac{\partial g(\mathbf{z})}{\partial \mathbf{z}}\right| \tag{2}$$

By maximizing the log-likelihood and parameterizing the invertible blocks with deep networks, flow-based methods learn to transform a simple, tractable base distribution (*e.g.,* standard normal) into a more expressive one. To model distributions with arbitrary dimensions, the $g-$bijection needs to be defined such that the Jacobian determinant can be computed in a closed form. Dinh *et al.* [50] proposed the *affine coupling layers* to build a flexible bijection function g by stacking a sequence of K simple bijection blocks $\mathbf{z}_{k-1} = g_\eta^{(k)}(\mathbf{z}_k)$ of the form,

$$g_\eta^{(k)}(\mathbf{z}_k) = \mathbf{b}_k \odot \mathbf{z}_k + (1 - \mathbf{b}_k) \odot \left[\mathbf{z}_k \odot \exp\left(s_k(\mathbf{b}_k \odot \mathbf{z}_k)\right) + t_k\left(\mathbf{b}_k \odot \mathbf{z}_k\right)\right] \tag{3}$$

$$g_\eta(\mathbf{z}) = \mathbf{z}_0 = g_\eta^{(1)} \circ \cdots \circ g_\eta^{(K-1)} \circ g_\eta^{(K)}(\mathbf{z}) \tag{4}$$

where $\mathbf{z} = \mathbf{z}_K$, $\odot$ is the Hadamard (element-wise) product, $\mathbf{b}_k \in \{0, 1\}^L$ is a binary mask used for partitioning the $k-$th block input, and $\eta = \{s_1, ..., s_K, t_1, ..., t_K\}$ are the deep networks parameters of the scaling s_k and translation t_k functions of the K blocks (see the supplementary material for network architectures).

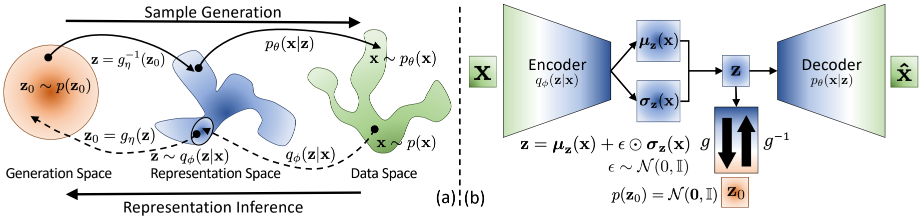
Sample Generation
$\mathbf{z} = g_\eta^{-1}(\mathbf{z}_0)$
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x} \sim p_\theta(\mathbf{x})$
$\mathbf{z}_0 \sim p(\mathbf{z}_0)$
$\mathbf{z}_0 = g_\eta(\mathbf{z})$
$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$
$q_\phi(\mathbf{z}|\mathbf{x})$
$\mathbf{x} \sim p(\mathbf{x})$
Generation Space
Representation Space
Data Space
Representation Inference
(a)
(b)
$\mathbf{X}$
Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
$\boldsymbol{\mu}_\mathbf{z}(\mathbf{x})$
$\boldsymbol{\sigma}_\mathbf{z}(\mathbf{x})$
$\mathbf{z}$
Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\hat{\mathbf{X}}$
$\mathbf{z} = \boldsymbol{\mu}_\mathbf{z}(\mathbf{x}) + \epsilon \odot \boldsymbol{\sigma}_\mathbf{z}(\mathbf{x})$
g
g^{-1}
$\epsilon \sim \mathcal{N}(0, \mathbb{I})$
$p(\mathbf{z}_0) = \mathcal{N}(\mathbf{0}, \mathbb{I})$
$\mathbf{z}_0$

Figure 2. ***dp*VAE:** (a) The latent space is decoupled into a *generation space* with a simple, tractable distribution (*e.g.,* standard normal) and a *representation space* whose distribution can be arbitrarily complex and is learned via a bijective mapping to the generation space. (b) VAE with the decoupled prior. The $g-$bijection is jointly trained with the VAE generative (*i.e.,* decoder) and inference (*i.e.,* encoder) models.

4 General Framework for VAE Regularization

In this section, we formally define and analyze how VAE regularizations affect the generative property of VAE. We also present the decoupled prior family for VAEs (see Figure 2) and analyze its utility to solve the submanifold problem of SOTA regularization-based VAEs.

4.1 VAE Regularizers: Latent Pockets and Leaks

ELBO regularization is a conventional mechanism that enforces inductive biases (*e.g.,* disentanglement [1–3, 6, 23, 24] and informative latent codes [25, 26]) to improve the representation learning aspect of VAEs [20]. These methods have shown their efficacy in learning good representations but neglect the generative property. Empirically, these regularizations improve the learned latent representation but inherently cause a mismatch between the aggregate posterior $q_\phi(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})}\left[q_\phi(\mathbf{z}|\mathbf{x})\right]$ and the prior $p(\mathbf{z})$. This mismatch leads to *latent pockets and leaks*, or a *submanifold* in the aggregate posterior that results in poor generative capabilities. Specifically, if a sample $\mathbf{z} \sim p(\mathbf{z})$ (*i.e.,* likely to be generated under the prior) lies in a pocket, (*i.e.,* $q_\phi(\mathbf{z})$ is low), then its corresponding decoded sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$ will not lie on the data manifold. This problem, caused by VAE regularizations, we call the *submanifold problem.*

To better understand this phenomena, we define two different types of samples in the VAE latent space that corresponds to two VAE failure modes.

Low-Posterior (LP) samples are highly likely to be generated under the prior (*i.e.,* $p(\mathbf{z})$ is high) but are not covered by the aggregate posterior (*i.e.,* $q_\phi(\mathbf{z})$ is low). The low-posterior samples are typically generated from the *latent pockets* dictated by the regularizer(s) used and are of poor quality since they lie off the data manifold. To generate low-posterior samples, we follow the logic of [27], where we sample $\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbb{I})$, rank them according to their aggregate posterior support, *i.e.,* values of $q_\phi(\mathbf{z})$, and choose the samples with lowest aggregate posterior values. In the case of *dp*VAEs, samples are generated from $\mathbf{z}_0 \sim p(\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_0; \mathbf{0}, \mathbb{I})$, which is a standard normal, and then transformed by $\mathbf{z} = g_\eta^{-1}(\mathbf{z}_0)$ before plugging it into the aggregate posterior.

High-Posterior (HP) samples are samples supported under the aggregate posterior (*i.e.,* $q_\phi(\mathbf{z})$ is high) but are less likely to be generated under the prior (*i.e.,* $p(\mathbf{z})$ is low). Specifically, these are samples in the latent space that can produce good generated samples but are unlikely to be sampled due to the low support of the prior, and thereby they are samples that are in the *latent leaks.* To generate high-posterior samples, we sample from $\mathbf{z} \sim q_\phi(\mathbf{z}) = \mathbb{E}_{p(\mathbf{x})}\left[q_\phi(\mathbf{z}|\mathbf{x})\right]$, rank them according to their prior support, *i.e.,* values of $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbb{I})$, and choose the samples with lowest prior support values. For *dp*VAEs, sampled $\mathbf{z}$ are first mapped to the $\mathbf{z}_0-$space by $\mathbf{z}_0 = g_\eta(\mathbf{z})$ before computing prior probabilities.

VAE performs well in the generative sense if the latent space is free of pockets and leaks. A pocket-free latent space is manifested by low-posterior samples that lie on the data manifold when mapped to the data space via the decoder $p_\theta(\mathbf{x}|\mathbf{z})$. In a leak-free latent space, high-posterior samples are supported by the aggregate posterior, yet with a tiny probability under the prior, and thereby these samples fall off the data manifold. This submanifold problem is demonstrated using four SOTA VAE regularizers (see Figure 1 and Figure 3). With β-VAE [2], FactorVAE [3] and β-TCVAE [23], we can clearly see that the low-posterior samples lie in the latent pockets formed in the aggregate posterior (see Figure 3b) and they lie outside the data manifold (see Figure 3c), causing the sample generation to be very noisy (see Figure 3a). In the case of InfoVAE, the low-posterior samples

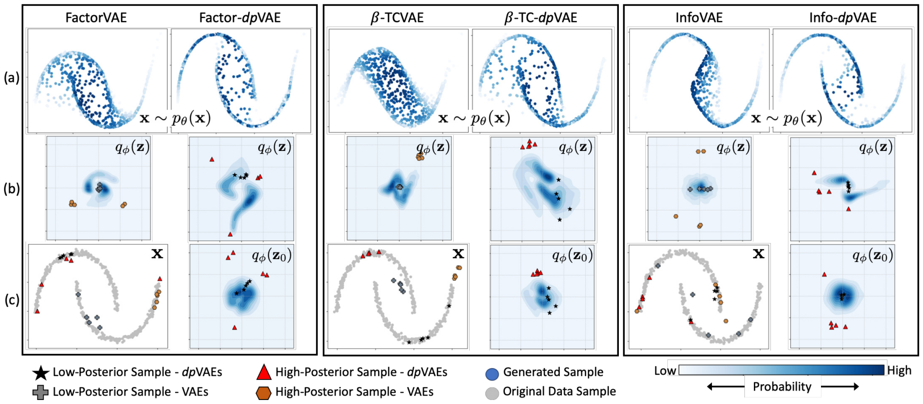

Figure 3. Sample generation and latent spaces for regularized VAEs: Each block is a VAE trained with a different regularizer on the moons dataset, with and without the decoupled prior. In each block, (a) showcases the sample quality, (b) shows the aggregate posterior $q_\phi(\mathbf{z})$ with top five low- and high-posterior samples marked, and (c) shows the generation space for the decoupled prior and the training samples in the data space with corresponding low- and high-posterior samples are marked.

lie in regions with not much aggregate posterior support causing a slightly noisy sample generation (see Figure 3a). More importantly, there are high-posterior samples that come from $q_\phi(\mathbf{z})$ but can very rarely be captured by a standard normal prior distribution. With the InfoVAE, for instance, the model fails to generate samples that lie on the tail-end of the top moon.

Although VAE regularizers improve latent representations, they sacrifice sample generation through the introduction of latent pockets and leaks. To fix sample generation, we propose a decoupling of the representation and generation spaces (see Figure 2a for illustration). This is demonstrated for β-VAE with and without decoupled prior in Figure 1, where the decoupled generation space $p(\mathbf{z}_0) \sim \mathcal{N}(\mathbf{z}_0; \mathbf{0}, \mathbb{I})$ is used for generation and all the low-posterior samples lie on the data manifold. We formulate this prior in detail in following section.

4.2 *dp*VAE: Decoupled Prior for VAE

Decoupled prior family, as the name suggests, decouples the latent space that performs the representation and the space that drives sample generation. For this decoupling to be meaningful, the representation and generation spaces should be related by a functional mapping. The decoupled prior effectively learns the latent space distribution $p(\mathbf{z})$ by simultaneously learning the functional mapping g_η together with the generative and inference models during optimization.

Specifically, the latent variables $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^L$ and $\mathbf{z}_0 \in \mathcal{Z}_0 \subset \mathbb{R}^L$ are the random variables of the *representation* and *generation* spaces, respectively, where $p(\mathbf{z}_0) \sim \mathcal{N}(\mathbf{z}_0; \mathbf{0}, \mathbb{I})$. The bijective mapping between the representation space $\mathcal{Z}$ and the generation space $\mathcal{Z}_0$ is defined by an invertible function $g_\eta(\mathbf{z}) = \mathbf{z}_0$, parameterized by the network parameters η. VAE regularizers still act on the posteriors in the representation space, *i.e.*, $q_\phi(\mathbf{z}|\mathbf{x})$ and/or $q_\phi(\mathbf{z})$, without affecting the latent distribution of the generation space. Sample generation starts by sampling $\mathbf{z}_0 \sim p(\mathbf{z}_0)$, passing through the inverse mapping to obtain $\mathbf{z} = g_\eta^{-1}(\mathbf{z}_0)$, which is then decoded by the generative model $p_\theta(\mathbf{x}|\mathbf{z})$ (see Figure 2a). These decoupled spaces can allow any modifications in the representation space dictated by the regularizer to infuse its submanifold structure in that space (see Figure 3b) without significantly impacting the generation space (see Figure 3c), and thereby improving sample generation for regularized VAEs (see Figure 3a). Moreover, the decoupled prior $p(\mathbf{z})$ is an expressive prior that is learned jointly with the VAE, and thereby it can match an arbitrarily complex aggregate posterior $q_\phi(\mathbf{z})$, thanks to the flexibility of deep networks to model complex mappings. Additionally, due to the bijective mapping g_η, we have a one-to-one correspondence between samples in $p(\mathbf{z}_0) \sim \mathcal{N}(\mathbf{z}_0; \mathbf{0}, \mathbb{I})$ and those in $p(\mathbf{z})$.

To derive the ELBO for *dp*VAE, we replace the standard normal prior in (1) with the decoupled prior defined in (2). Using the change of variable formula, the KL divergence term in (1) can be simplified into[1]:

$$\begin{aligned}\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right] = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\sum_{k=1}^{K}\sum_{l=1}^{L} b_k^l s_k\left(b_k^l z_k^l\right)\right] \\ -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{z}}(\mathbf{x})| + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[g_\eta(\mathbf{z})^T g_\eta(\mathbf{z})\right] \qquad (5)\end{aligned}$$

where L is the latent dimension, K is number of invertible blocks that defines the decoupled prior in (4), s_k is the scaling network of the $k-$th block, $\boldsymbol{\Sigma}_{\mathbf{z}}(\mathbf{x})$ is the covariance matrix of the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ (typically assumed to be diagonal), and b_k^l and z_k^l are the $l-$th element in $\mathbf{b}_k$ and $\mathbf{z}_k$ vectors, respectively.

4.3 *dp*VAE in Concert with VAE Regularizers

The KL divergence in (5) can be directly used for any regularized ELBO. However, there are some regularized models such as β-TCVAE [23], and InfoVAE [1] that introduce additional terms other than $\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right]$ with $p(\mathbf{z})$. These regularizers need to be modified when used with decoupled priors[2].

β-*dp*VAE: For β-VAE (both β-VAE-H [2] and β-VAE-B [57] versions), the only difference in the ELBO (1) is reweighting the KL-term and the addition of certain constraints without introducing any additional terms. Hence, β-*dp*VAE will retain the same reweighting and constraints, and only modify the KL divergence term according to (5).

Factor-*dp*VAE: FactorVAE [3] introduces a total correlation term $\mathrm{KL}\left[q_\phi(\mathbf{z}) \| q_\phi(\bar{\mathbf{z}})\right]$ to the ELBO in (1), where $q_\phi(\bar{\mathbf{z}}) = \prod_{l=1}^{L} q_\phi(z^l)$ and z^l is the $l-$th element of $\mathbf{z}$. This term promotes disentanglement of the latent dimensions of $\mathbf{z}$, impacting the representation learning aspect of VAE. Hence, in the case of the decoupled prior, the total correlation term should be applied to the *representation space*. In this sense, the decoupled prior only affects the KL divergence term as described in (5) for the Factor-*dp*VAE model.

[1] Complete derivation can be found in the supplementary material.

[2] The ELBOs for these regularizers can be found in the supplementary material.

β-TC-*dp*VAE: Regularization provided by β-TCVAE [23] factorizes the ELBO into the individual latent dimensions based on the decomposition given in [22]. The only term that includes $p(\mathbf{z})$ is the KL divergence between marginals, *i.e.,* KL $[q_\phi(\mathbf{z}) \| p(\mathbf{z})]$. This term in β-TCVAE is assumed to be factorized and is evaluated via sampling, facilitating the direct incorporation of the decoupled prior. In particular, we can just sample from the base distribution $\mathbf{z}_0 \sim p(\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_0; \mathbf{0}, \mathbb{I})$ and compute the corresponding sample $\mathbf{z} \sim p(\mathbf{z})$ using $\mathbf{z} = g_\eta^{-1}(\mathbf{z}_0)$.

Info-*dp*VAE: In InfoVAE [1], the additional term in the ELBO is again the divergence between aggregate posterior and the prior, *i.e.,* KL $[q_\phi(\mathbf{z}) \| p(\mathbf{z})]$. This KL divergence term is replaced by different divergence families; adversarial training [34], Stein variational gradient [58], and maximum-mean discrepancy MMD [59–61]. However, adversarial-based divergences can have unstable training and Stein variational gradient scales poorly with high dimensions [1]. Motivated by the MMD-based results in [1], we focus here on the MMD divergence to evaluate this marginal divergence. For Info-*dp*VAE, we start with the ELBO of InfoVAE and modify the standard KL divergence term using (5). In addition, we compute the marginal KL divergence using MMD, which quantifies the divergence between two distributions by comparing their moments through sampling. Similar to β-TC-*dp*VAE, we can sample from $p(\mathbf{z}_0)$ and use the inverse mapping to compute samples in the $\mathbf{z}-$space.

Table 1. Generative metrics (***lower** is better*) for vanilla VAE and regularized VAEs using standard normal and decoupled priors. **FID** = Frchet Inception Distance. **LP** = Low-Posterior FID score. **sKL** = symmetric KL divergence. **NLL** = Negative Log-Likelihood ($\times 10^3$)

Methods	**MNIST**				**SVHN**				**CelebA**			
	FID	LP	sKL	NLL	FID	LP	sKL	NLL	FID	LP	sKL	NLL
VAE [14, 15]	137.4	165.0	1.26	3.56	78.9	83.8	53.67	0.386	81.4	79.0	59.3	9.26
*dp*VAE	**129.0**	**153.1**	**0.88**	**3.53**	**50.8**	**55.2**	**13.02**	**0.318**	**71.5**	**74.3**	**10.4**	**4.91**
β-VAE-H [2]	144.2	163.1	4.49	4.12	96.7	97.6	10.35	0.611	80.3	79.9	39.7	**6.93**
β-*dp*VAE-H	**127.1**	**127.4**	**1.07**	**2.98**	**65.2**	**67.7**	**4.05**	**0.592**	**67.2**	**72.5**	**33.5**	10.6
β-VAE-B [57]	130.8	163.6	2.74	3.11	61.7	68.5	2.62	0.606	75.7	79.6	25.8	12.5
β-*dp*VAE-B	**113.3**	**114.1**	**1.32**	**2.80**	**51.1**	**50.3**	**2.47**	**0.550**	**67.9**	**72.0**	**19.1**	**10.4**
β-TCVAE [23]	149.8	200.3	4.48	2.91	69.2	70.5	7.76	9.86	83.8	83.0	93.6	**9.33**
β-TC-*dp*VAE	**133.3**	**133.1**	**2.07**	**2.70**	**50.3**	**53.8**	**2.94**	**4.52**	**80.3**	**81.4**	**90.3**	10.0
FactorVAE [3]	130.5	191.2	1.04	**3.50**	97.2	108.5	1.91	**2.13**	82.6	86.8	71.3	**9.89**
Factor-*dp*VAE	**120.8**	**121.3**	**0.85**	3.60	**86.3**	**86.9**	**1.57**	2.36	**65.0**	**73.4**	**51.3**	12.2
InfoVAE [1]	128.7	133.2	2.89	2.88	81.3	83.2	4.91	**1.55**	76.5	79.1	30.6	**11.1**
Info-*dp*VAE	**110.1**	**110.5**	**1.70**	**2.81**	**62.9**	**67.7**	**2.67**	1.56	**68.9**	**72.9**	**20.3**	12.1

5 Experiments

We experiment with three benchmark image datasets, namely MNIST [62], SVHN [63], CelebA (cropped version) [64] to provide a fair comparison with SOTA regularized VAEs, which used the same datasets. We train these datasets with VAE [14, 15] and five regularized VAEs, namely β-VAE-H [2], β-VAE-B [57], β-TCVAE) [23], FactorVAE [3] and InfoVAE [1]. We showcase, qualitatively and quantitatively, that *dp*VAEs improve sample generation while retaining the benefits of representation learning provided by the regularizers[3]

5.1 Generative Metrics

We use the following quantitative metrics to assess the generative performance of the regularized VAEs with and without the decoupled prior.

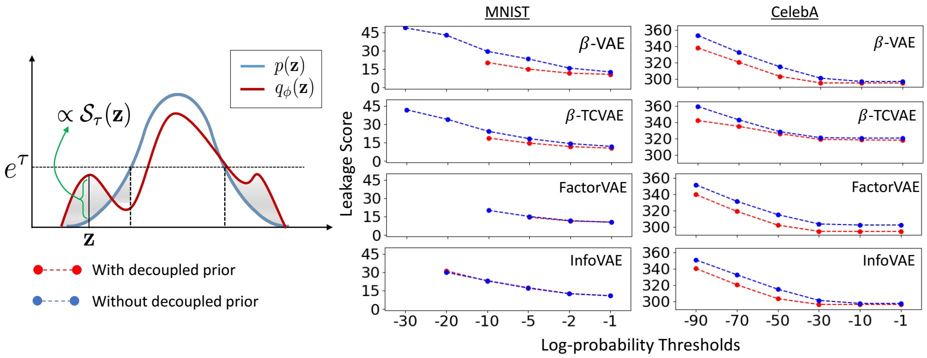

Figure 4. *dp*VAEs have less latent leaks: Leakage scores for regularized VAEs on MNIST (a) and CelebA (b) data (missing values mean there are no samples with $\log(p(\mathbf{z})) < \tau$, implying zero leakage at that threshold). The illustration on the left represents the intuition behind the lekage score. For a probability threshold τ, the leakage score is proportional to the probaility difference at a sample in latent sapce, this area is marked in grey : $\mathbb{E}_{q_\phi(\mathbf{z})}\left[\mathcal{S}_\tau(\mathbf{z})\right]$

.

Frchet Inception Distance (FID): The FID score is based on the statistics, assuming Gaussian distribution, computed in the feature space defined using the inception network features [65]. FID score quantifies both the sample diversity and quality. Lower FID means better sample generation.

Symmetric KL Divergence (sKL): To quantify the overlap between $p(\mathbf{z})$ and $q_\phi(\mathbf{z})$ in the representation space ($p(\mathbf{z})$ being the decoupled prior for *dp*VAEs or the standard normal), we compute sKL$=$ KL$\left[p(\mathbf{z})\|q_\phi(\mathbf{z})\right]$ + KL$\left[q_\phi(\mathbf{z})\|p(\mathbf{z})\right]$ through sampling (using 5,000 samples). sKL also captures the existence of pockets and leaks in $q_\phi(\mathbf{z})$. Lower sKL implies there is a better overlap between $p(\mathbf{z})$ and $q_\phi(\mathbf{z})$, indicating better generative capabilities.

Negative Log-likelihood: We estimate the likelihood of held-out samples under a trained model using importance sampling (with 21,000 samples) as in [15], where NLL $= \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})/q_\phi(\mathbf{z}|\mathbf{x})\right]$. Lower NLL means better sam-

[3] Architectures and hyperparameters are described in the supplementary material. Additionally, results showcasing that representation learning (specifically disentanglement) is not adversely affected by the introduction of decoupled priors are also presented in the supplementary material.

ple generation since the learned model supports unseen samples drawn from the data distribution.

Leakage Score: To assess the effect of decoupled priors on latent leaks (as manifested by high posterior samples), we devise a new metric based on log-probability differences. We sample from the aggregate posterior $\mathbf{z} \sim q_\phi(\mathbf{z})$. If $\log(p(\mathbf{z})) < \tau$, where $\tau \in \mathbb{R}$ is a chosen threshold value, then we consider the sample to lie in a "leakage region" defined by τ. This sample is considered a high-posterior sample at the $\tau-$level since the sample is better supported under the aggregate posterior than the prior (see Figure 4). Based on the threshold value, these leakage regions are less likely to be sampled from. In order to not lose significant regions from the data manifold, we want the aggregate posterior corresponding to these samples to attain low values as well. To quantify latent leakage for a trained model, we propose a *leakage score* as $\text{LS}(\tau) = \mathbb{E}_{q_\phi(\mathbf{z})}\left[\mathcal{S}_\tau(\mathbf{z})\right]$, where for a given $\mathbf{z} \sim q_\phi(\mathbf{z})$ at a particular threshold τ, $\mathcal{S}_\tau(\mathbf{z})$ is defined in (6), where h is the identity function for VAEs and g_η for *dp*VAEs.

$$\mathcal{S}_\tau(\mathbf{z}) = \begin{cases} \log\left(\frac{q_\phi(\mathbf{z})}{p(h(\mathbf{z}))}\right) & \log(p(\mathbf{z})) < \tau \\ 0 & \log(p(\mathbf{z})) \geq \tau \end{cases} \quad (6)$$

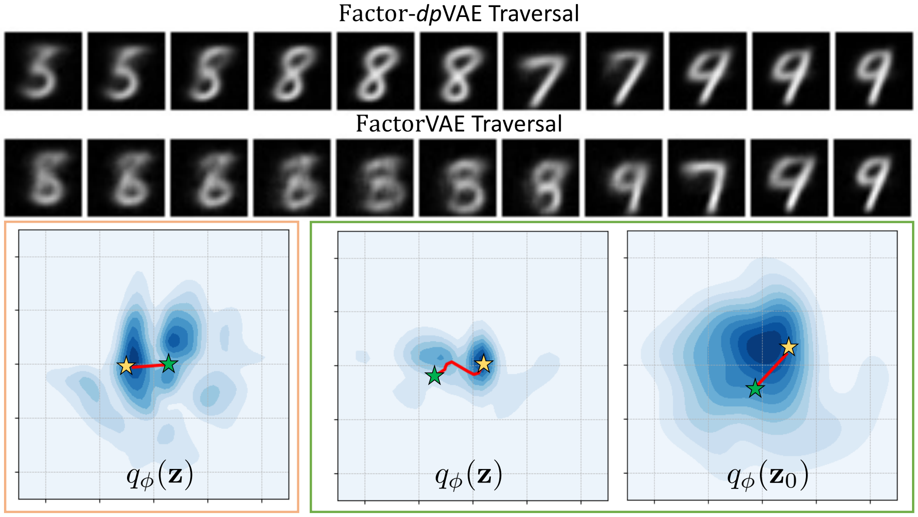

Figure 5. Latent traversal for *dp*VAEs does not path through latent pockets: The top rows showcases latent traversal for FactorVAE and Factor-*dp*VAE on MNIST data. The orange box is the $q_\phi(\mathbf{z})$ for FactorVAE and the red line shows the traversal between starting and ending points (green and yellow stars, respectively). The green box shows the *same* traversal in $q_\phi(\mathbf{z}_0)$ that is mapped using g_η^{-1} to the representation space, demonstrated using $q_\phi(\mathbf{z})$. We see that the traversal path in $q_\phi(\mathbf{z})$ tries to avoid low probability regions, which correspond to better image quality.

5.2 Generation Results and Analysis

In Table 1, we observe that *dp*VAEs perform better than their corresponding regularized VAEs without the decoupled prior.When comparing VAEs with and without decoupled priors (*e.g.,* InfoVAE and Info-*dp*VAE), we use the same hyperparameters and perform no additional tuning. This showcases the robustness of the decoupled prior wrt hyperparameters, facilitating its direct use with any regularized VAE. We report the FID scores on both the randomly generated samples from the prior and the low-posterior samples. As analyzed in section 4.1, if the low-posterior samples lie on the data manifold, then the learned latent space is pocket-free. Results in Table 1 suggest that for all *dp*VAEs, the FID scores for the randomly generated samples and low-posterior ones are comparable, suggesting that all the pockets in the latent space are filled. Qualitative results of sample generation for CelebA and MNIST are shown in the supplementary material (due to space constraints). We show both the random prior and low-posterior sample generation with and without the decoupled prior for three different regularizers. Sample quality of *dp*VAEs is better or on par with those without the decoupled prior. But more importantly, one can observe a significant quality improvement in the low-posterior samples, which aligns with the quantitative results in Table 1. In Figure 4, we report the leakage score LS(τ) as a function of log-probability thresholds for different regularizers with and without the decoupled priors. We observe that *dp*VAEs result in models with lower latent leakage. This is especially true at lower thresholds, which suggests that even when $p(\mathbf{z})$ is small, the $q_\phi(\mathbf{z})$ is small as well, preventing the loss of significant regions from the data manifold.

5.3 Latent Traversals Results

We perform latent traversals between samples in the latent space. We expect that in VAEs, there will be instances of the traversal path crossing the latent pockets resulting in poor intermediate samples. In contrast, we expect *dp*VAEs will map the linear traversal in $\mathbf{z}_0$ (generation space) to a non-linear traversal in $\mathbf{z}$ (representation space), while avoiding low probability regions. This is observed for MNIST data traversal ($L = 2$) and is depicted in Figure 5.

We also qualitatively observe similar occurrences in CelabA traversals (see supplementary material). Finally, we want to attest that the addition of the decoupled prior to a regularizer does not affect it's ability to improve the latent representation. We demonstrate this by observing latent factor traversals for CelebA trained on Factor-*dp*VAE, where we vary one dimension of the latent space while fixing the others. One can observe that Factor-*dp*VAE is able to isolate different attributes of variation in the data, as shown in Figure 6.

6 Conclusion

In this paper, we define and analyze the submanifold problem for regularized VAEs, or the tendency of a regularizer to accentuate the creation of pockets

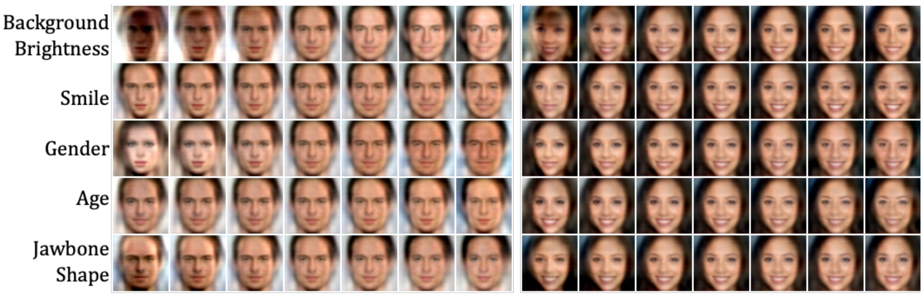

Figure 6. Factor-*dp*VAE latent traversals across the top 5 latent dimensions: Traversals start with the reconstructed image of a given sample and move ±5 standard deviations along a latent dimension. Results from other *dp*VAEs similarly retain the latent space disentanglement.

and leaks in the latent space. This submanifold structure manifests the mismatch between the aggregate posterior and the latent prior which in turn causes degradation in generation quality. To overcome this trade-off between sample generation and latent representation, we propose the decoupled prior family as a general regularization framework for VAE and demonstrate its efficacy on SOTA VAE regularizers. *dp*VAE does not modify the ELBO of the vanilla VAE, rather it leverages learnable priors that are optimized jointly with the inference and generation models to match the aggregate posterior and the latent prior. We demonstrate that *dp*VAEs generate better quality samples as compared with their standard normal prior based counterparts, via qualitative and quantitative results. Additionally, we qualitatively observe that the representation learning (as improved by the regularizer) is not adversely affected by *dp*VAEs. Decoupled priors can act as a pathway to realizing the true potential of VAEs as both a representation learning and a generative modeling framework. Further work in this direction will include exploring more expressive inference and generative models (*e.g.,* PixelCNN [47]) in conjuction with decoupled priors. We also believe more sophisticated invertible architectures (*e.g.,*. RAD [66]) and base distributions will provide further improvements.

References

1. Zhao, S., Song, J., Ermon, S.: Infovae: Balancing learning and inference in variational autoencoders. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 5885–5892
2. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR **2** (2017) 6
3. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. (2018) 2654–2663
4. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems. (2016) 2172–2180

5. Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J.: Plug & play generative networks: Conditional iterative generation of images in latent space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4467–4477
6. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: Advances in Neural Information Processing Systems. (2016) 5040–5048
7. Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., Lerchner, A.: Darla: Improving zero-shot transfer in reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 1480–1490
8. Rezende, D., Danihelka, I., Gregor, K., Wierstra, D., et al.: One-shot generalization in deep generative models. In: International Conference on Machine Learning. (2016) 1521–1529
9. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in neural information processing systems. (2014) 3581–3589
10. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35** (2013) 1798–1828
11. Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R.A., Murphy, K.: Fixing a broken elbo. In: International Conference on Machine Learning. (2018) 159–168
12. Yuille, A., Kersten, D.: Vision as bayesian inference: analysis by synthesis? Trends in cognitive sciences **10** (2006) 301–308
13. Nair, V., Susskind, J., Hinton, G.E.: Analysis-by-synthesis by learning to invert generative black boxes. In: International Conference on Artificial Neural Networks, Springer (2008) 971–981
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ICLR (2014)
15. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International Conference on Machine Learning. (2014) 1278–1286
16. Maaløe, L., Sønderby, C.K., Sønderby, S.K., Winther, O.: Auxiliary deep generative models. In: International Conference on Machine Learning. (2016) 1445–1453
17. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: How to train deep variational autoencoders and probabilistic ladder networks. In: 33rd International Conference on Machine Learning (ICML 2016). (2016)
18. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. In: Advances in neural information processing systems. (2016) 2352–2360
19. Xu, W., Sun, H., Deng, C., Tan, Y.: Variational autoencoder for semi-supervised text classification. In: Thirty-First AAAI Conference on Artificial Intelligence. (2017)
20. Tschannen, M., Bachem, O., Lucic, M.: Recent advances in autoencoder-based representation learning. Third workshop on Bayesian Deep Learning (NeurIPS 2018) (2018)
21. Chen, X., Kingma, D.P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., Abbeel, P.: Variational lossy autoencoder. ICLR (2017)
22. Hoffman, M.D., Johnson, M.J.: Elbo surgery: yet another way to carve up the variational evidence lower bound. In: Workshop in Advances in Approximate Bayesian Inference, NIPS. Volume 1. (2016)

23. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. (2018) 2610–2620
24. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. ICLR (2018)
25. Makhzani, A., Frey, B.J.: Pixelgan autoencoders. In: Advances in Neural Information Processing Systems. (2017) 1975–1985
26. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. ICLR (2016)
27. Rosca, M., Lakshminarayanan, B., Mohamed, S.: Distribution matching in variational inference. arXiv preprint arXiv:1802.06847 (2018)
28. Xu, H., Chen, W., Lai, J., Li, Z., Zhao, Y., Pei, D.: On the necessity and effectiveness of learning the prior of variational auto-encoder. arXiv preprint arXiv:1905.13452 (2019)
29. Shmelkov, K., Lucas, T., Alahari, K., Schmid, C., Verbeek, J.: Coverage and quality driven training of generative image models. arXiv preprint arXiv:1901.01091 (2019)
30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
31. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. (2017) 214–223
32. Sugiyama, M., Suzuki, T., Kanamori, T.: Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. Annals of the Institute of Statistical Mathematics **64** (2012) 1009–1044
33. Mescheder, L., Nowozin, S., Geiger, A.: Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 2391–2400
34. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)
35. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning. In: Advances in Neural Information Processing Systems. (2017) 3308–3318
36. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. In: International Conference on Learning Representations. (2016)
37. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. arXiv preprint arXiv:1711.01558 (2017)
38. Xiao, Z., Yan, Q., Amit, Y.: Generative latent flow. arXiv preprint arXiv:1905.10485 (2019)
39. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. In: Advances in neural information processing systems. (2016) 4743–4751
40. Bauer, M., Mnih, A.: Resampled priors for variational autoencoders. In: The 22nd International Conference on Artificial Intelligence and Statistics. (2019) 66–75
41. Tomczak, J., Welling, M.: Vae with a vampprior. In: International Conference on Artificial Intelligence and Statistics. (2018) 1214–1223
42. Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648 (2016)

43. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: Draw: A recurrent neural network for image generation. In: International Conference on Machine Learning. (2015) 1462–1471
44. Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A.A., Visin, F., Vazquez, D., Courville, A.: Pixelvae: A latent variable model for natural images. ICLR (2017)
45. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems. (2017) 6306–6315
46. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: Advances in Neural Information Processing Systems. (2019) 14866–14876
47. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: Advances in neural information processing systems. (2016) 4790–4798
48. Oord, A.v.d., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759 (2016)
49. Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation (2014)
50. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. ICLR (2017)
51. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: Proceedings of the 32nd International Conference on Machine Learning. Volume 37 of Proceedings of Machine Learning Research., Lille, France, PMLR (2015) 1530–1538
52. Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., eds.: Advances in Neural Information Processing Systems 31. Curran Associates, Inc. (2018) 10215–10224
53. Huang, C.W., Touati, A., Dinh, L., Drozdzal, M., Havaei, M., Charlin, L., Courville, A.: Learnable explicit density for continuous latent space and variational inference. arXiv preprint arXiv:1710.02248 (2017)
54. Das, H.P., Abbeel, P., Spanos, C.J.: Dimensionality reduction flows. arXiv preprint arXiv:1908.01686 (2019)
55. Gritsenko, A.A., Snoek, J., Salimans, T.: On the relationship between normalising flows and variational-and denoising autoencoders. (2019)
56. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. (2016) 10–21
57. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in beta-vae. arXiv preprint arXiv:1804.03599 (2018)
58. Liu, Q., Wang, D.: Stein variational gradient descent: A general purpose bayesian inference algorithm. In: Advances in neural information processing systems. (2016) 2378–2386
59. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.J.: A kernel method for the two-sample-problem. In: Advances in neural information processing systems. (2007) 513–520
60. Li, Y., Swersky, K., Zemel, R.: Generative moment matching networks. In: International Conference on Machine Learning. (2015) 1718–1727
61. Dziugaite, G.K., Roy, D.M., Ghahramani, Z.: Training generative neural networks via maximum mean discrepancy optimization. In: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, AUAI Press (2015) 258–267

62. LeCun, Y., Cortes, C.: MNIST handwritten digit database. (2010)
63. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning. (2011)
64. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. (2015) 3730–3738
65. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,. (2016)
66. Dinh, L., Sohl-Dickstein, J., Pascanu, R., Larochelle, H.: A RAD approach to deep mixture models. CoRR **abs/1903.07714** (2019)