

From Variational to Deterministic Autoencoders

Partha Ghosh*, Mehdi S. M. Sajjadi*, Antonio Vergari, Michael Black, Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Tübingen, Germany

{pghosh, msajjadi, avergari, black, bs}@tue.mpg.de

Abstract

Variational Autoencoders (VAEs) provide a theoretically-backed framework for deep generative models. However, they often produce “blurry” images, which is linked to their training objective. Sampling in the most popular implementation, the Gaussian VAE, can be interpreted as simply injecting noise to the input of a deterministic decoder. In practice, this simply enforces a smooth latent space structure. We challenge the adoption of the full VAE framework on this specific point in favor of a simpler, deterministic one. Specifically, we investigate how substituting stochasticity with other explicit and implicit regularization schemes can lead to a meaningful latent space without having to force it to conform to an arbitrarily chosen prior. To retrieve a generative mechanism for sampling new data points, we propose to employ an efficient ex-post density estimation step that can be readily adopted both for the proposed deterministic autoencoders as well as to improve sample quality of existing VAEs. We show in a rigorous empirical study that regularized deterministic autoencoding achieves state-of-the-art sample quality on the common MNIST, CIFAR-10 and CelebA datasets.

1. Introduction

Generative modeling lies at the core of machine learning and computer vision. By capturing the mechanisms behind the data generation process, one can reason about data probabilistically, access and traverse the low-dimensional manifold the data is assumed to live on, and ultimately *generate new data*. It is therefore not surprising that learning generative models has gained momentum in applications like chemistry [16, 25], NLP [8, 46] and computer vision [9, 48].

Variational Autoencoders (VAEs) [27, 38] allow for a principled probabilistic way to model high-dimensional distributions. They do so by casting learning representations as a variational inference problem. Learning a VAE amounts

to the optimization of an objective balancing the quality of autoencoded samples through a stochastic encoder–decoder pair while encouraging the latent space to follow a fixed prior distribution.

Since their introduction, VAEs have become one of the frameworks of choice for generative modeling, promising theoretically well-founded and more stable training than Generative Adversarial Networks (GANs) [17] and more efficient sampling mechanisms than autoregressive models [30]. Much of the recent literature has focused on applying VAEs on image generation tasks [19, 23, 33] and devising new encoder–decoder architectures [25, 56].

Despite this attention from the community, the VAE framework is still far from delivering the promised generative mechanism in many real-world scenarios. In fact, VAEs tend to generate blurry samples, a condition which has been attributed to using overly simplistic distributions for the prior [54]; restrictiveness of the Gaussian assumption for the stochastic architecture [13]; or over-regularization induced by the KL divergence term in the VAE objective [53] (see Fig. 2). Moreover, the VAE objective itself poses several challenges as it admits trivial solutions that decouple the latent space from the input [11, 60], leading to the posterior collapse phenomenon in conjunction with powerful decoders [56]. Training a VAE requires approximating expectations by sampling at the cost of increased variance in gradients [10, 55], making initialization, validation and annealing of hyperparameters fundamental in practice [5, 8, 21]. Lastly, even after a satisfactory convergence of the objective, the learned aggregated posterior distribution rarely matches the assumed latent prior in practice [5, 13, 26], ultimately hurting the quality of generated samples.

In this work, we tackle these shortcomings by reformulating the VAE into a generative modeling scheme that scales better, is simpler to optimize, and most importantly, produces higher-quality samples. We do so based on the observation that under common distributional assumptions made for VAEs, training a stochastic encoder–decoder pair does not differ in practice from training a deterministic architecture where noise is added to the decoder’s input to

*The authors contributed equally to this work.

enforce a smooth latent space. We investigate how to substitute this noise injection mechanism with other regularization schemes in our deterministic *Regularized Autoencoders* (RAEs), and we analyze how we can learn a meaningful latent space without forcing it to conform to a given prior distribution. We equip RAEs with a generative mechanism through a simple and efficient ex-post density estimation step on the learned latent space which leads to improved image quality that surpasses VAEs and stronger alternatives such as Wasserstein Autoencoders (WAEs) [53].

In summary, our contributions are as follows:

1. we introduce the RAE framework for generative modeling,
2. we propose an ex-post density estimation scheme that greatly improves sample quality for the VAE, WAE and RAE without the need for additional training,
3. we conduct a rigorous empirical evaluation on several common image datasets (MNIST, CIFAR-10, CelebA), assaying reconstruction, random samples and interpolation quality for VAE, WAE and RAE,
4. we achieve state-of-the-art FID scores for the above datasets in a non-adversarial setting.

The paper is organized as follows. In Section 2 we introduce the VAE framework and discuss assumptions, practical implementations and limitations, leading to the introduction of our simplified deterministic and regularized framework (Section 3), interpreting explicit regularization as constrained optimization under certain parametric assumptions (Section 4). After discussing ex-post density estimation and related works in Sections 5 and 6, we present experiments in Section 7 before we close with our final conclusions.

2. Variational Autoencoders

For a general discussion, we consider a collection of high-dimensional i.i.d. samples $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ drawn from the true data distribution $p_{\text{data}}(\mathbf{x})$ over a random variable \mathbf{X} taking values in the input space. The aim of generative modeling is to learn from \mathcal{X} a mechanism to draw new samples $\mathbf{x}_{\text{new}} \sim p_{\text{data}}$.

Variational Autoencoders provide a powerful latent variable framework to infer such a mechanism. The generative process of the VAE is defined as

$$\mathbf{z}_{\text{new}} \sim p(\mathbf{Z}), \quad \mathbf{x}_{\text{new}} \sim p_{\theta}(\mathbf{X} | \mathbf{Z} = \mathbf{z}_{\text{new}}) \quad (1)$$

where $p(\mathbf{Z})$ is a fixed prior distribution over a low-dimensional latent variable \mathbf{Z} . A stochastic decoder

$$D_{\theta}(\mathbf{z}) = \mathbf{x} \sim p_{\theta}(\mathbf{x} | \mathbf{z}) = p(\mathbf{X} | g_{\theta}(\mathbf{z})) \quad (2)$$

links the latent space to the input space through the *likelihood*, where g_{θ} is an expressive non-linear function parameterized by θ .¹ As a result, a VAE estimates $p_{\text{data}}(\mathbf{x})$ as the infinite mixture model $p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{x} | \mathbf{z}) p(\mathbf{z}) d\mathbf{z}$. At the same time, the input space is mapped to the latent space via a stochastic encoder

$$E_{\phi}(\mathbf{x}) = \mathbf{z} \sim q_{\phi}(\mathbf{z} | \mathbf{x}) = q(\mathbf{Z} | f_{\phi}(\mathbf{x})) \quad (3)$$

where $q_{\phi}(\mathbf{z} | \mathbf{x})$ is the *posterior* distribution given by a second function f_{ϕ} parameterized by ϕ .

Computing the marginal log-likelihood $\log p_{\theta}(\mathbf{x})$ is generally intractable. We therefore follow a variational approach, maximizing the evidence lower bound (ELBO) for a sample \mathbf{x} :

$$\begin{aligned} \log p_{\theta}(\mathbf{x}) &\geq \text{ELBO}(\phi, \theta, \mathbf{x}) = \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z} | \mathbf{x})} \log p_{\theta}(\mathbf{x} | \mathbf{z}) - \mathbb{KL}(q_{\phi}(\mathbf{z} | \mathbf{x}) || p(\mathbf{z})) \end{aligned} \quad (4)$$

Maximizing Eq. 4 over data \mathcal{X} w.r.t. model parameters ϕ, θ corresponds to minimizing the loss

$$\arg \min_{\phi, \theta} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \mathcal{L}_{\text{ELBO}} = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} \quad (5)$$

where \mathcal{L}_{REC} and \mathcal{L}_{KL} are defined for a sample \mathbf{x} as follows:

$$\mathcal{L}_{\text{REC}} = -\mathbb{E}_{\mathbf{z} \sim q_{\phi}(\mathbf{z} | \mathbf{x})} \log p_{\theta}(\mathbf{x} | \mathbf{z}) \quad (6)$$

$$\mathcal{L}_{\text{KL}} = \mathbb{KL}(q_{\phi}(\mathbf{z} | \mathbf{x}) || p(\mathbf{z})) \quad (7)$$

Intuitively, the reconstruction loss \mathcal{L}_{REC} takes into account the quality of autoencoded samples \mathbf{x} through $D_{\theta}(E_{\phi}(\mathbf{x}))$, while the KL-divergence term \mathcal{L}_{KL} encourages $q_{\phi}(\mathbf{z} | \mathbf{x})$ to match the prior $p(\mathbf{z})$ for each \mathbf{z} which acts as a regularizer during training [22]. For the purpose of generating high-quality samples, a balance between these two loss terms must be found during training, see Fig. 2.

2.1. Practice and shortcomings of VAEs

To fit a VAE to data through Eq. 5 one has to specify the parametric forms for $p(\mathbf{z})$, $q_{\phi}(\mathbf{z} | \mathbf{x})$, $p_{\theta}(\mathbf{x} | \mathbf{z})$, and hence the deterministic mappings f_{ϕ} and g_{θ} . In practice, the choice for the above distributions is guided by trading off computational complexity with model expressiveness.

In the most common formulation of the VAE, $q_{\phi}(\mathbf{z} | \mathbf{x})$ and $p_{\theta}(\mathbf{x} | \mathbf{z})$ are assumed to be Gaussian

$$E_{\phi}(\mathbf{x}) \sim \mathcal{N}(\mathbf{Z} | \boldsymbol{\mu}_{\phi}(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}_{\phi}(\mathbf{x}))) \quad (8)$$

$$D_{\theta}(E_{\phi}(\mathbf{x})) \sim \mathcal{N}(\mathbf{X} | \boldsymbol{\mu}_{\theta}(\mathbf{z}), \text{diag}(\boldsymbol{\sigma}_{\theta}(\mathbf{z}))) \quad (9)$$

with means μ_{ϕ}, μ_{θ} and covariance parameters $\sigma_{\phi}, \sigma_{\theta}$ given by f_{ϕ} and g_{θ} . In practical implementations, the covariance

¹With slight abuse of notation, we use lowercase letters for both random variables and their realizations, e.g. $p_{\theta}(\mathbf{x} | \mathbf{z})$ instead of $p(\mathbf{X} | \mathbf{Z} = \mathbf{z})$, when it is clear to discriminate between the two.

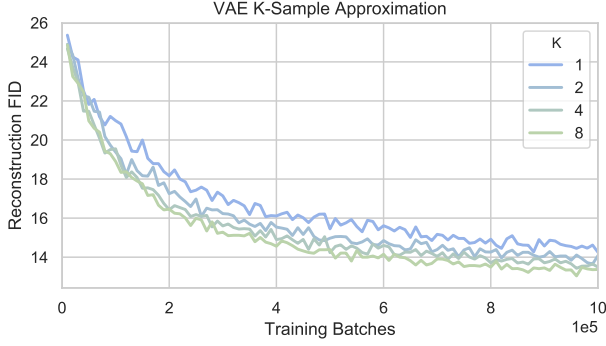


Figure 1. Test reconstruction quality for a VAE trained on MNIST with different numbers of samples in the latent space as in Eq. 8 measured by FID (lower is better). Larger numbers of samples clearly improve training, however, the increased accuracy comes with larger requirements for memory and computation. In practice, the most common choice is therefore $k = 1$.

of the decoder is set to the identity matrix for all \mathbf{z} , *i. e.* $\sigma_\theta(\mathbf{z}) = \mathbf{1}$ [13]. The expectation of \mathcal{L}_{REC} in Eq. 6 is then approximated via k Monte Carlo point estimates. We find clear evidence that larger values lead to improvements in training as shown in Fig. 1. Nevertheless, only a 1-sample approximation is carried out in practice [27] since requirements to memory and computation scale linearly with k . With this approximation, the computation of \mathcal{L}_{REC} is given by the mean squared error between input samples and their mean reconstructions μ_θ through the deterministic decoder:

$$\mathcal{L}_{\text{REC}} = \|\mathbf{x} - \mu_\theta(E_\phi(\mathbf{x}))\|_2^2 \quad (10)$$

Gradients w.r.t. the encoder parameters ϕ are computed through the expectation of \mathcal{L}_{REC} in Eq. 6 via the reparametrization trick [27] where the stochasticity of E_ϕ is relegated to an auxiliary random variable ϵ which does not depend on ϕ :

$$E_\phi(\mathbf{x}) = \mu_\phi(\mathbf{x}) + \sigma_\phi(\mathbf{x}) \odot \epsilon \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (11)$$

where \odot denotes the Hadamard product. An additional simplifying assumption involves fixing the prior $p(\mathbf{z})$ to be a D -dimensional isotropic Gaussian $\mathcal{N}(\mathbf{Z} | \mathbf{0}, \mathbf{I})$. For this choice, the KL-divergence for a sample \mathbf{x} is given in closed form:

$$2\mathcal{L}_{\text{KL}} = \|\mu_\phi(\mathbf{x})\|_2^2 + D + \sum_i^D \sigma_\phi(\mathbf{x})_i - \log \sigma_\phi(\mathbf{x})_i \quad (12)$$

While the chosen assumptions make VAEs easy to implement, the stochasticity in the encoder and decoder has been deemed to be responsible for the “blurriness” in VAE samples [13, 34, 53]. Furthermore, the optimization problem as shown in Eq. 5 presents some further challenges. Imposing a strong weight on the \mathcal{L}_{KL} term during optimization can

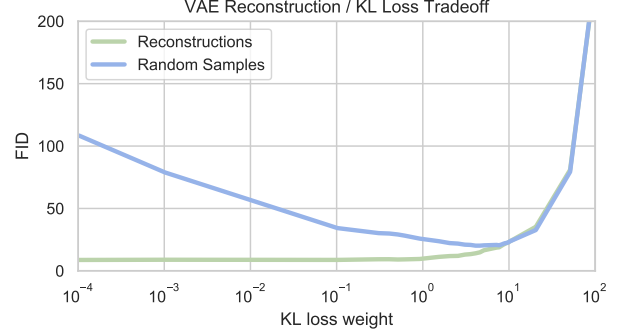


Figure 2. Reconstruction and random sample quality (FID, y-axis, lower is better) of a VAE on MNIST for different tradeoffs between \mathcal{L}_{REC} and \mathcal{L}_{KL} (x-axis, see Eq. 5). Higher weights for \mathcal{L}_{KL} improve random samples but hurt reconstruction. Enforcing structure in the VAE latent space leads to a penalty in quality.

dominate $\mathcal{L}_{\text{ELBO}}$, having the effect of *over-regularization* which leads to blurred samples, see Fig. 2. Heuristics to avoid this include gradually annealing the importance of \mathcal{L}_{KL} during training [5, 8] and manually fine-tuning the balance between the losses.

Even after employing the full array of approximations and “tricks” to reach convergence of Eq. 5 for a satisfactory set of parameters, there is no guarantee that the learned latent space is distributed according to the assumed prior distribution. In other words, the aggregated posterior distribution $q_\phi(\mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} q(\mathbf{z} | \mathbf{x})$ has been shown not to conform well to $p(\mathbf{z})$ after training [5, 13, 53]. This critical issue severely hinders the generative mechanism of a VAE (Eq. 1) since latent codes sampled from $p(\mathbf{z})$ (instead of $q(\mathbf{z})$) might lead to regions of the latent space that are previously unseen to D_θ during training. The result is blurry out-of-distribution samples. We analyze solutions to this problem in Section 5.

2.2. Constant-Variance Encoders

Analogously to what is generally done for decoders, we also investigate fixing the variance of $q_\phi(\mathbf{z} | \mathbf{x})$ to be constant for all \mathbf{x} . This simplifies the computation of E_ϕ from Eq. 11 to

$$E_\phi^{\text{CV}}(\mathbf{x}) = \mu_\phi(\mathbf{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}) \quad (13)$$

where σ is a fixed scalar. At the same time, the KL loss term in Eq. 12 simplifies (up to constants) to

$$2\mathcal{L}_{\text{KL}}^{\text{CV}} = \|\mu_\phi(\mathbf{x})\|_2^2 \quad (14)$$

Constant-Variance VAEs (CV-VAEs) have been previously applied in applications of adversarial robustness [15] and variational image compression [4] but to the best of our knowledge, there is no systematic study of CV-VAEs in the

literature. As noted in [15], treating σ_ϕ as a constant impairs the assumption of $p(\mathbf{z})$ to be an isotropic Gaussian which demands a more complex prior structure over \mathbf{Z} . We address this mismatch in Sec. 5.

3. Deterministic Regularized Autoencoders

As described in Section 2, autoencoding in VAEs is defined in a probabilistic fashion: E_ϕ and D_θ map data points not to a single point, but rather to parameterized distributions as shown in Equations 8 and 9. However, the practical implementation of the VAE admits a deterministic view for this probabilistic mechanism.

A glance at the autoencoding mechanism of the VAE is revealing. The encoder maps a data point \mathbf{x} to a mean $\mu_\phi(\mathbf{x})$ and variance $\sigma_\phi(\mathbf{x})$ in the latent space via the reparametrization trick given in Eq. 11. The input to the decoder is then simply the mean $\mu_\phi(\mathbf{x})$ augmented with random Gaussian noise scaled by $\sigma_\phi(\mathbf{x})$. In the CV-VAE, this relationship is even more obvious, as the magnitude of the noise is fixed for all data points (Eq. 13). In this light, a VAE can be seen as a *deterministic* autoencoder where Gaussian noise is added to the decoder’s input.

Using random noise injection to regularize neural networks during training is a well-known technique that dates back several decades [2, 47]. The addition of noise implicitly smooths the function learned by the network. Since this procedure also adds noise to the gradients, we propose to substitute noise injection with an explicit regularization scheme for the decoder network. Note that from a generative perspective, this is motivated by the goal to learn a smooth latent space where similar data points \mathbf{x} are mapped to similar latent codes \mathbf{z} , and small variations in \mathbf{Z} lead to reconstructions by D_θ that vary only slightly.

By removing the noise injection mechanism from the CV-VAE, we are effectively left with a deterministic *Regularized Autoencoder* (RAE) that can be coupled with any type of explicit regularization for the decoder to enforce a smooth latent space. Training a RAE thus involves minimizing the simplified loss

$$\mathcal{L}_{\text{RAE}} = \mathcal{L}_{\text{REC}} + \beta \mathcal{L}_{\text{KL}}^{\text{RAE}} + \lambda \mathcal{L}_{\text{REG}} \quad (15)$$

where \mathcal{L}_{REG} represents the explicit regularizer for D_θ (see Sec. 3.1) and $\mathcal{L}_{\text{KL}}^{\text{RAE}} = 1/2 \|\mathbf{z}\|_2^2$ from Eq. 14 is equivalent to constraining the size of the learned space. Note that for RAEs, no sampling approximation of \mathcal{L}_{REC} is required, thus relieving the need for more samples from $q_\phi(\mathbf{z} | \mathbf{x})$ to achieve better image quality (see Fig. 1).

3.1. Regularization Schemes for RAEs

Among possible choices for a mechanism to use for \mathcal{L}_{REG} , a first obvious candidate is Tikhonov regularization [52] since is known to be related to the addition of

low-magnitude input noise [7]. Training a RAE within this framework thus amounts to adopting

$$\mathcal{L}_{\text{REG}} = \mathcal{L}_{L_2} = \|\theta\|_2^2 \quad (16)$$

where \mathcal{L}_{L_2} effectively applies weight decay on the decoder parameters θ .

Another avenue comes from the recent GAN literature where regularization is a hot topic [29] and where injecting noise to the input of the adversarial discriminator has led to improved performance in a technique called *instance noise* [49]. To enforce Lipschitz continuity on adversarial discriminators, weight clipping has been proposed [3], which is however known to significantly slow down training. More successfully, a *gradient penalty* on the discriminator can be used similar to [18, 35], yielding the objective

$$\mathcal{L}_{\text{REG}} = \mathcal{L}_{\text{GP}} = \|\nabla D_\theta(E_\phi(\mathbf{x}))\|_2^2 \quad (17)$$

which encourages small L_2 norm of the gradient of the decoder w. r. t. its input.

Additionally, spectral normalization (SN) has been proposed as an alternative way to bound the Lipschitz norm of an adversarial discriminator, showing very promising results for GAN training [36]. SN normalizes the weight matrix θ_ℓ for each layer in the decoder by dividing it by an estimate of its largest singular value:

$$\theta_\ell^{\text{SN}} = \theta_\ell / s(\theta_\ell) \quad (18)$$

where $s(\theta_\ell)$ is the current estimate obtained through the power method.

Lastly, in light of recent success stories of deep neural networks *without* explicit regularization achieving state-of-the-art results [58, 59], it is intriguing to question the need to explicitly regularize the decoder in order to obtain a meaningful latent space. The assumption here is that techniques such as dropout [50], batch normalization [24], adding noise during training [2], or early stopping in conjunction with novel architectural developments implicitly regularize the networks enough to smoothen the latent space. Therefore, as a natural baseline to the \mathcal{L}_{RAE} objectives introduced above, we also consider the RAE framework without \mathcal{L}_{REG} and $\mathcal{L}_{\text{KL}}^{\text{RAE}}$, *i. e.* a standard deterministic autoencoder optimizing \mathcal{L}_{REC} only.

4. A Probabilistic Derivation of Smoothing

In this section, we propose an alternative view on enforcing smoothness on the output of D_θ by augmenting the ELBO optimization problem for VAEs with an explicit constraint. While we keep the Gaussianity assumptions over a stochastic D_θ and $p(\mathbf{z})$ for convenience, we are not fixing a parametric form for $q_\phi(\mathbf{z} | \mathbf{x})$ yet. We will then discuss how some parametric restrictions over $q_\phi(\mathbf{z} | \mathbf{x})$ indeed lead to a

variation of the RAE framework in Eq. 15, specifically as the introduction of \mathcal{L}_{GP} as a regularizer of a deterministic version of the CV-VAE.

To start, we can augment the minimization in Eq. 5 as:

$$\begin{aligned} \arg \min_{\phi, \theta} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{X})} \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} \\ \text{s.t. } \|D_{\theta}(\mathbf{z}_1) - D_{\theta}(\mathbf{z}_2)\|_p < \epsilon \\ \forall \mathbf{z}_1, \mathbf{z}_2 \sim q_{\phi}(\mathbf{z} | \mathbf{x}) \quad \forall \mathbf{x} \sim p_{\text{data}}(\mathbf{X}) \end{aligned} \quad (19)$$

where $D_{\theta}(\mathbf{z}) = \boldsymbol{\mu}_{\theta}(E_{\phi}(\mathbf{x}))$ and the constraint on the decoder encodes that the output has to vary, in the sense of an L_p norm, only by a small amount ϵ for any two possible draws from the encoder. Using the mean value theorem, there exists a $\tilde{\mathbf{z}} \sim q_{\phi}(\mathbf{z} | \mathbf{x})$ such that the left term in the constraint can be bounded as:

$$\begin{aligned} \|D_{\theta}(\mathbf{z}_1) - D_{\theta}(\mathbf{z}_2)\|_p &= \nabla D_{\theta}(\tilde{\mathbf{z}}) \cdot \|\mathbf{z}_1 - \mathbf{z}_2\|_p \\ &\leq \sup\{\|\nabla D_{\theta}(\mathbf{z})\|_p\} \cdot \sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p\} \end{aligned} \quad (20)$$

where we take the supremum of possible gradients of D_{θ} as well as the supremum of a measure of the support of $q_{\phi}(\mathbf{z} | \mathbf{x})$. From this form of the smoothness constraint, it is apparent why the choice of a parametric form for $q_{\phi}(\mathbf{z} | \mathbf{x})$ can be impactful during training. For a compactly supported isotropic PDF $q_{\phi}(\mathbf{z} | \mathbf{x})$, the extension of the support $\sup\{\|\mathbf{z}_1 - \mathbf{z}_2\|_p\}$ would be dependent on $\mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))$, the entropy of $q_{\phi}(\mathbf{z} | \mathbf{x})$, through some functional r . For instance, a uniform posterior over a hypersphere in \mathbf{z} would ascertain $r(\mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))) \cong e^{\mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))/n}$ where n is the dimensionality of the latent space.

Intuitively, one would look for parametric distributions that do not favor overfitting, *e.g.* degenerating in Dirac-deltas (minimal entropy and support) along any dimensions. To this end, an isotropic nature of $q_{\phi}(\mathbf{z} | \mathbf{x})$ would favor such a robustness against decoder over-fitting. We can now rewrite the constraint as

$$r(\mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))) \cdot \sup\{\|\nabla D_{\theta}(\mathbf{z})\|_p\} < \epsilon \quad (21)$$

The \mathcal{L}_{KL} term can be expressed in terms of $\mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))$, by decomposing it as $\mathcal{L}_{\text{KL}} = \mathcal{L}_{\text{CE}} - \mathcal{L}_{\text{H}}$, where $\mathcal{L}_{\text{H}} = \mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))$ and $\mathcal{L}_{\text{CE}} = \mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}), p(\mathbf{z}))$ represents a cross-entropy term. Therefore, the constrained problem in Eq. 19 can be written in a Lagrangian formulation by including Eq. 21:

$$\arg \min_{\phi, \theta} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{X})} \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{CE}} - \mathcal{L}_{\text{H}} + \lambda \mathcal{L}_{\text{LANG}} \quad (22)$$

where $\mathcal{L}_{\text{LANG}} = r(\mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))) * \|\nabla D_{\theta}(\mathbf{z})\|_p$. We argue that a reasonable simplifying assumption for $q_{\phi}(\mathbf{z} | \mathbf{x})$ is to fix $\mathbb{H}(q_{\phi}(\mathbf{z} | \mathbf{x}))$ to a single constant for all samples \mathbf{x} . Intuitively, this can be understood as fixing the variance in $q_{\phi}(\mathbf{z} | \mathbf{x})$ as we did for the CV-VAE in Sec. 2.2. With this simplification, Eq. 22 further reduces to

$$\arg \min_{\phi, \theta} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{X})} \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{CE}} + \lambda \|\nabla D_{\theta}(\mathbf{z})\|_p \quad (23)$$

We can see that $\|\nabla D_{\theta}(\mathbf{z})\|_p$ results to be the gradient penalty \mathcal{L}_{GP} and $\mathcal{L}_{\text{CE}} = \|\mathbf{z}\|_2^2$ corresponds to $\mathcal{L}_{\text{KL}}^{\text{RAE}}$, thus recovering our RAE framework as presented in Eq. 15.

5. Ex-Post Density Estimation

By removing stochasticity and ultimately, the KL divergence term \mathcal{L}_{KL} from RAEs, we have simplified the original VAE objective at the cost of detaching the encoder from the prior $p(\mathbf{z})$ over the latent space. This implies i) we cannot ensure that the latent space \mathbf{Z} is distributed according to a simple distribution anymore (*e.g.* isotropic Gaussian) and consequently, ii) we lose the simple mechanism provided by $p(\mathbf{z})$ to sample from \mathbf{Z} as in Eq. 1.

As discussed in Section 2.1, issue i) is compromising the VAE framework in any case in practice as reported in several works [14, 22, 42]. To fix this, some works extend the VAE objective by encouraging the aggregated posterior to match $p(\mathbf{z})$ [53] or by utilizing more complex priors [5, 26, 54].

To overcome both i) and ii), we instead propose to employ *ex-post density estimation* over \mathbf{Z} . We fit a density estimator denoted as $q_{\delta}(\mathbf{z})$ to $\{\mathbf{z} = E_{\phi}(\mathbf{x}) | \mathbf{x} \in \mathcal{X}\}$. This simple approach not only fits our RAE framework well, but it can also be readily adopted for any VAE or variants thereof such as the WAE as a practical remedy to the aggregated posterior mismatch without adding any computational overhead to the costly training phase.

The choice of $q_{\delta}(\mathbf{z})$ needs to trade-off *expressiveness* – to provide a good fit of an arbitrary space for \mathbf{Z} – with *simplicity* – to improve generalization. Indeed, placing a Dirac distribution on each latent point \mathbf{z} would allow the decoder to output only training sample reconstructions. Striving for simplicity and in order to show the effectiveness of the proposed ex-post density estimation scheme, we compare a full covariance multivariate Gaussian with a 10-component Gaussian mixture model (GMM) in our experiments.

6. Related works

Many works have focused on diagnosing the VAE framework, the terms in its objective [1, 22, 60], and ultimately augmenting it to solve optimization issues [13, 39]. With RAE, we argue that a simpler deterministic framework can be competitive for generative modeling.

Deterministic denoising [57] and contractive autoencoders (CAEs) [41] have received attention in the past for their ability to capture a smooth data manifold. Heuristic attempts to equip them with a generative mechanism include MCMC schemes [6, 40]. However, they are hard to diagnose for convergence, require a considerable effort in tuning [12], and have not scaled beyond MNIST, leading to them being superseded by VAEs. While in spirit the proposed RAE is similar, \mathcal{L}_{GP} requires much less computational effort than computing the Jacobian for CAEs [41].

Approaches to cope with the aggregated posterior mismatch involve fixing a more expressive form for $p(\mathbf{z})$ [5, 26] therefore altering the VAE objective and requiring considerable additional computational efforts. Estimating the latent space of a VAE with a second VAE [13] reintroduces many of the optimization shortcomings discussed for VAEs and is much more expensive in practice compared to fitting a simple $q_\delta(\mathbf{z})$ after training.

GANs [17] have received widespread attention for their ability to produce sharp samples. Despite theoretical and practical advances [3], the training procedure of GANs is still unstable, sensitive to hyperparameters, and prone to the *mode collapse* problem [33, 43, 44].

Adversarial Autoencoders (AAE) [34] add a discriminator to a deterministic encoder-decoder pair, leading to sharper samples at the expense of higher computational overhead and the introduction of instabilities caused by the adversarial nature of the training process. Wasserstein Autoencoders (WAE) [53] have been introduced as a generalization of AAEs by casting autoencoding as an optimal transport (OT) problem. Both stochastic and deterministic models can be trained by minimizing a relaxed OT cost function employing either an adversarial loss term or the maximum mean discrepancy score between $p(\mathbf{z})$ and $q_\phi(\mathbf{z})$ as a regularizer in place of \mathcal{L}_{KL} .

Within the RAE framework, we look at this problem from a different perspective: instead of explicitly imposing a simple structure on \mathbf{Z} that might impair the ability to fit high-dimensional data during training, we propose to model the latent space by an ex-post density estimation step.

7. Experiments

In this Section, we investigate the performance of several VAE and RAE variants on MNIST [31], CIFAR10 [28] and CelebA [32]. We measure three qualities for each model: held-out sample reconstruction quality, random sample quality, and interpolation quality. While reconstructions give us a lower bound on the best quality achievable by the generative model, random sample quality indicates how well the model generalizes. Finally, interpolation quality sheds light on the structure of the learned latent space.

7.1. Models

We compare the the proposed RAE model with the gradient penalty (RAE-GP), with weight decay (RAE-L2), and with spectral normalization (RAE-SN). Additionally, we consider two models for which we either add only the latent code regularizer $\mathcal{L}_{\text{KL}}^{\text{RAE}}$ to \mathcal{L}_{REC} (RAE), or no explicit regularization at all (AE). As baselines, we further employ a regular VAE, the constant-variance VAE (CV-VAE) for comparison, and finally, a Wasserstein Autoencoder (WAE) with the MMD loss as a state-of-the-art alternative.

Aiming for a fair comparison, we employ the same net-

work architecture for all models. We largely follow the models adopted in [53] with the difference that we consistently apply batch normalization [24] for all models as we found it to improve performance across the range. The latent space dimension is 16 for MNIST, 128 for CIFAR-10 and 64 for CelebA. Further details about the network architecture and training procedure are given in Appendix A.

7.2. Evaluation

The evaluation of generative models is a nontrivial research question [33, 45, 51]. Since we are interested in the quality of samples, the ubiquitous Fréchet Inception Distance (FID) [20] is a reasonable choice for comparing different models. More recently, a notion of precision and recall for distributions (PRD) has been proposed [43], separating sample quality from diversity. We report the popular FID scores in this section and we provide PRD scores in Appendix C.

We compute the FID of the reconstructions of random validation samples against the test set to evaluate reconstruction quality. For evaluating generative modeling capabilities, we compute the FID between the test data and randomly drawn samples from a single Gaussian that is either the isotropic $p(\mathbf{z})$ available for VAEs and WAEs, or a single Gaussian fit to $q_\delta(\mathbf{z})$ for CV-VAEs and RAEs. For all models, we also evaluate random samples from a 10-component Gaussian Mixture model (GMM) fit to $q_\delta(\mathbf{z})$. Using only 10 components prevents us from overfitting (which would indeed give good FIDs when compared with the test set). We note that fitting GMMs with up to 100 components, only improved results marginally. Additionally, we provide nearest-neighbours from the training set in Appendix E to show that the models are not overfitting. For interpolations, we report the FID for the furthest interpolation points resulted by applying spherical interpolation to randomly selected validation reconstruction pairs.

7.3. Results

Table 1 summarizes our main results. All of the proposed RAE variants are competitive with the VAE and WAE w. r. t. generated image quality in all settings. Sampling RAEs achieve the best FIDs across all datasets when a modest 10-component GMM is employed for ex-post density estimation. Furthermore, even when \mathcal{N} is considered as $q_\delta(\mathbf{z})$, RAEs rank first with the exception of MNIST, though the best FID achieved there by the VAE is very close to the FID of RAE-SN.

Moreover, our best RAE FIDs are lower than the best results reported for VAEs in the large scale comparison of [33], challenging even the best scores reported for GANs. While we are employing a slightly different architecture than theirs, our models underwent only modest finetuning instead of an extensive hyperparameter search. By looking



Figure 3. Qualitative evaluation of sample quality for VAEs, WAEs and RAEs on CelebA. Left: reconstructed samples (top row is ground truth). Middle: randomly generated samples. Right: interpolations in the latent space between a pair of test images (first and last column). RAE models provide overall sharper samples and reconstructions while interpolating smoothly in the latent space. Corresponding qualitative overviews for MNIST and CIFAR-10 are provided in Appendix D.

at the differently regularized RAEs, there is no clear winner across all settings as all perform equally well. For practical reasons of implementation simplicity, one may prefer RAE-L2 over the GP and SN variants.

Surprisingly, the implicitly regularized RAE and AE models are shown to be able to score impressive FIDs when $q_\delta(\mathbf{z})$ is fit through GMMs. FIDs for AEs decrease from 58.73 to 10.66 on MNIST and from 127.85 to 45.10 on CelebA – a value close to the state of the art. This is a remarkable result that follows a long series of recent con-

firmations that neural networks are surprisingly smooth *by design* [37]. It is also surprising that the lack of an explicitly fixed structure on the latent space of the RAE does not impede interpolation quality. This is further confirmed by the qualitative evaluation on CelebA as reported in Fig. 3 and for the other datasets in Appendix D, where RAE interpolated samples seem sharper than competitors and transitions smoother.

We would like to note that our extensive study confirms and quantifies the effect of the aggregated posterior mis-

	MNIST				CIFAR				CELEBA			
	REC.	SAMPLES			REC.	SAMPLES			REC.	SAMPLES		
		\mathcal{N}	GMM	Interp.		\mathcal{N}	GMM	Interp.		\mathcal{N}	GMM	Interp.
VAE	18.26	19.21	17.66	18.21	57.94	106.37	103.78	88.62	39.12	48.12	45.52	44.49
CV-VAE	15.15	33.79	17.87	25.12	37.74	94.75	86.64	69.71	40.41	48.87	49.30	44.96
WAE	10.03	20.42	9.39	14.34	35.97	117.44	93.53	76.89	34.81	53.67	42.73	40.93
RAE-GP	14.04	22.21	11.54	15.32	32.17	83.05	76.33	64.08	39.71	116.30	45.63	47.00
RAE-L2	10.53	22.22	8.69	14.54	32.24	80.80	74.16	62.54	43.52	51.13	47.97	45.98
RAE-SN	15.65	19.67	11.74	15.15	27.61	84.25	75.30	63.62	36.01	44.74	40.95	39.53
RAE	11.67	23.92	9.81	14.67	29.05	83.87	76.28	63.27	40.18	48.20	44.68	43.67
AE	12.95	58.73	10.66	17.12	30.52	84.74	76.47	61.57	40.79	127.85	45.10	50.94

Table 1. Evaluation of all models by FID (lower numbers are better, best model in bold). We evaluate each model by REC.: test sample reconstruction; \mathcal{N} : random samples generated according to the prior distribution $p(\mathbf{z})$ (for VAE / WAE) or by fitting a Gaussian to $q_\delta(\mathbf{z})$ (for the remaining models); GMM: random samples generated by fitting a mixture of 10 Gaussians in the latent space; Interp.: mid-point interpolation between random pairs of test reconstructions. Note that our (less constrained) RAE models are competitive with or outperform the VAE and WAE throughout the evaluation. Surprisingly, interpolations do not suffer from the lack of explicit prior on the latent space in our models. Furthermore, the unregularized AE achieves very good FID scores when combined with the proposed ex-post density estimation technique.

match as well as the effectivity of our proposed solution to it. Indeed, if we consider the effect of applying ex-post density estimation to each model in Table 1, we see that it consistently improves sample quality across all settings considerably. A 10-component GMM trained to fit \mathbf{Z} seems to be enough to half the FID scores from ~ 20 to ~ 10 for WAE and RAE models on MNIST and from 116 to 46 on CelebA. This is striking since this very cheap additional step to any VAE-like generative model can be employed to boost the quality of generated samples.

All in all, the results strongly support our conjecture that the simple deterministic RAE framework can challenge the VAE and WAE.

8. Conclusion

While the theoretical derivation of the VAE has helped popularize the framework for generative modeling, recent works have started to expose some discrepancies between theory and practice. In this work, viewing sampling in VAEs as noise injection to enforce smoothness has enabled us to distill a deterministic autoencoding framework that is compatible with several regularization techniques to learn a meaningful latent space. We have demonstrated that such a deterministic autoencoding framework can generate comparable or better samples than VAEs, while getting around the practical drawbacks tied to a stochastic framework. Furthermore, we have shown that our solution to fit a simple density estimator such as a Gaussian Mixture Model on the learned latent space is able to consistently improve sample quality both for the proposed RAE framework as well as for the VAE and WAE, acting as a solution for the known mismatch between the prior and the aggregated posterior. The RAE

framework opens interesting future research venues such as learning the density estimator in an end-to-end fashion with the autoencoding network and devising more sophisticated autoencoders that can access the full range of recent structural neural network advancements to scale generative modeling without being bound to the VAE’s restrictions.

Acknowledgements

We would like to thank Anant Raj, Matthias Bauer, Paul Rubenstein and Soubhik Sanyal for fruitful discussions.

References

- [1] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a Broken ELBO. In *ICML*, 2018.
- [2] G. An. The effects of adding noise during backpropagation training on a generalization performance. In *Neural computation*, 1996.
- [3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *ICML*, 2017.
- [4] J. Ballé, V. Laparra, and E. P. Simoncelli. End-to-end optimized image compression. In *ICLR*, 2017.
- [5] M. Bauer and A. Mnih. Resampled Priors for Variational Autoencoders. In *AISTATS*, 2019.
- [6] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *NeurIPS*, 2013.
- [7] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [8] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *CoNLL*, 2016.

- [9] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.
- [10] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.
- [11] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. In *ICLR*, 2017.
- [12] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. In *Journal of the American Statistical Association*, 1996.
- [13] B. Dai and D. Wipf. Diagnosing and Enhancing VAE Models. In *ICLR*, 2019.
- [14] B. Dai and D. Wipf. Diagnosing and Enhancing VAE Models. In *ICLR*, 2019.
- [15] P. Ghosh, A. Losalka, and M. J. Black. Resisting Adversarial Attacks using Gaussian Mixture Variational Autoencoders. In *AAAI*, 2019.
- [16] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. In *ACS central science*, 2018.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
- [18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.
- [19] C. Ham, A. Raj, V. Cartillier, and I. Essa. Variational Image Inpainting. In *Bayesian Deep Learning Workshop, NeurIPS*, 2018.
- [20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. In *NeurIPS*, 2017.
- [21] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.
- [22] M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NeurIPS*, 2016.
- [23] H. Huang, R. He, Z. Sun, T. Tan, et al. Introvae: Introspective Variational Autoencoders for Photographic Image Synthesis. In *NeurIPS*, 2018.
- [24] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *ICML*, 2015.
- [25] W. Jin, R. Barzilay, and T. Jaakkola. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364*, 2018.
- [26] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improving variational inference with inverse autoregressive flow. In *NeurIPS*, 2016.
- [27] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.
- [28] A. Krizhevsky and G. Hinton. Learning Multiple Layers of Features from Tiny Images, 2009.
- [29] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly. The GAN landscape: Losses, architectures, regularization, and normalization. *arXiv preprint arXiv:1807.04720*, 2018.
- [30] H. Larochelle and I. Murray. The neural autoregressive distribution estimator. In *AISTATS*, 2011.
- [31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *IEEE*, 1998.
- [32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *ICCV*, 2015.
- [33] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs Created Equal? A Large-Scale Study. In *NeurIPS*, 2018.
- [34] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. In *ICLR*, 2016.
- [35] L. Mescheder, A. Geiger, and S. Nowozin. Which Training Methods for GANs do actually Converge? In *ICML*, 2018.
- [36] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.
- [37] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.
- [38] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.
- [39] D. J. Rezende and F. Viola. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.
- [40] S. Rifai, Y. Bengio, Y. Dauphin, and P. Vincent. A generative process for sampling contractive auto-encoders. In *ICML*, 2012.
- [41] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML*, 2011.
- [42] M. Rosca, B. Lakshminarayanan, and S. Mohamed. Distribution Matching in Variational Inference. *arXiv preprint arXiv:1802.06847*, 2018.
- [43] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing Generative Models via Precision and Recall. In *NeurIPS*, 2018.
- [44] M. S. M. Sajjadi, G. Parascandolo, A. Mehrjou, and B. Schölkopf. Tempered Adversarial Networks. In *ICML*, 2018.
- [45] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. In *ICCV*, 2017.
- [46] A. Severyn, E. Barth, and S. Semeniuta. A Hybrid Convolutional Variational Autoencoder for Text Generation. 2017.
- [47] J. Sietsma and R. J. Dow. Creating artificial neural networks that generalize. In *Neural networks*. Elsevier, 1991.

- [48] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015.
- [49] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised MAP Inference for Image Super-resolution. In *ICLR*, 2017.
- [50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. In *JMLR*, 2014.
- [51] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. In *ICLR*, 2016.
- [52] A. N. Tikhonov and V. I. Arsenin. *Solutions of Ill Posed Problems*. Vh Winston, 1977.
- [53] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein Auto-Encoders. In *ICLR*, 2017.
- [54] J. Tomczak and M. Welling. VAE with a VampPrior. In *AISTATS*, 2018.
- [55] G. Tucker, A. Mnih, C. J. Maddison, J. Lawson, and J. Sohl-Dickstein. REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models. In *NeurIPS*, 2017.
- [56] A. van den Oord, O. Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.
- [57] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.
- [58] S. Zagoruyko and N. Komodakis. Wide residual networks. In *BMVC*, 2016.
- [59] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.
- [60] S. Zhao, J. Song, and S. Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

Appendix

A. Network architecture and Training Details

For all experiments, we use the Adam optimizer with a starting learning rate of 10^{-3} which is cut in half every time the validation loss plateaus. All models are trained for a maximum of 100 epochs on MNIST and CIFAR and 70 epochs on CelebA. We use mini-batch size of 100 and pad MNIST digits with zeros to make the size 32×32 .

We use official train, validation and test splits of CelebA. For MNIST and CIFAR, we set aside 10k train samples for validation. For random sample evaluation, we draw samples from $\mathcal{N}(0, I)$ for VAE and WAE-MMD and for all remaining models, samples are drawn from a multivariate Gaussian whose mean and covariance are estimated using training set embeddings. For the GMM density estimation, we also utilize the training set embeddings for fitting and validation set embeddings to verify that GMM models are not over fitting to training embeddings. However, due to the very low number of mixture components, we did not encounter overfitting at this step. The GMM parameters are estimated by running EM for at most 100 iterations.

For Figures 1 and 2, we used smaller networks due to computational limitations and since only relative performance is of interest in these experiments. It should be noted that as a result of this, the absolute values of the reported FID scores in these figures cannot be directly compared with the numbers reported in Section 7.

	MNIST	CIFAR_10	CELEBA
ENCODER:	$x \in \mathcal{R}^{32 \times 32}$ $\rightarrow \text{CONV}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{1024} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{FLATTEN} \rightarrow \text{FC}_{16 \times M}$	$x \in \mathcal{R}^{32 \times 32}$ $\rightarrow \text{CONV}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{1024} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{FLATTEN} \rightarrow \text{FC}_{128 \times M}$	$x \in \mathcal{R}^{64 \times 64}$ $\rightarrow \text{CONV}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONV}_{1024} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{FLATTEN} \rightarrow \text{FC}_{64 \times M}$
DECODER:	$z \in \mathcal{R}^{16} \rightarrow \text{FC}_{8 \times 8 \times 1024}$ $\rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_1$	$z \in \mathcal{R}^{128} \rightarrow \text{FC}_{8 \times 8 \times 1024}$ $\rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_1$	$z \in \mathcal{R}^{64} \rightarrow \text{FC}_{8 \times 8 \times 1024}$ $\rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_{512} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_{256} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_{128} \rightarrow \text{BN} \rightarrow \text{ReLU}$ $\rightarrow \text{CONVT}_1$

Conv_n represents a convolutional layer with n filters. All convolutions Conv_n and transposed convolutions ConvT_n have a filter size of 4×4 for MNIST and CIFAR-10 and 5×5 for CELEBA. They all have a stride of size 2 except for the last convolutional layer in the decoder. Finally, $M = 1$ for all models except for the VAE which has $M = 2$ as the encoder has to produce both mean and variance for each input.

B. Evaluation Setup

We use 10k samples for all FID and PRD evaluations. Scores for random samples are evaluated against the test set. Reconstruction scores are computed from validation set reconstructions against the respective test set. Interpolation scores are computed by interpolating latent codes of a pair of randomly chosen validation embeddings vs test set samples. The visualized interpolation samples are interpolations between two randomly chosen test set images.

C. Evaluation by Precision and Recall

	MNIST		CIFAR-10		CELEBA	
	\mathcal{N}	GMM	\mathcal{N}	GMM	\mathcal{N}	GMM
VAE	0.96 / 0.92	0.95 / 0.96	0.25 / 0.55	0.37 / 0.56	0.54 / 0.66	0.50 / 0.66
CV-VAE	0.84 / 0.73	0.96 / 0.89	0.31 / 0.64	0.42 / 0.68	0.25 / 0.43	0.32 / 0.55
WAE	0.93 / 0.88	0.98 / 0.95	0.38 / 0.68	0.51 / 0.81	0.59 / 0.68	0.69 / 0.77
RAE-GP	0.93 / 0.87	0.97 / 0.98	0.36 / 0.70	0.46 / 0.77	0.38 / 0.55	0.44 / 0.67
RAE-L2	0.92 / 0.87	0.98 / 0.98	0.41 / 0.77	0.57 / 0.81	0.36 / 0.64	0.44 / 0.65
RAE-SN	0.89 / 0.95	0.98 / 0.97	0.36 / 0.73	0.52 / 0.81	0.54 / 0.68	0.55 / 0.74
RAE	0.92 / 0.85	0.98 / 0.98	0.45 / 0.73	0.53 / 0.80	0.46 / 0.59	0.52 / 0.69
AE	0.90 / 0.90	0.98 / 0.97	0.37 / 0.73	0.50 / 0.80	0.45 / 0.66	0.47 / 0.71

Table 2. Evaluation of random sample quality by precision / recall [43] (higher numbers are better, best value for each dataset in bold). It is notable that the proposed ex-post density estimation improves not only precision, but also recall throughout the experiment. For example, WAE seems to have a comparably low recall of only 0.88 on MNIST which is raised considerably to 0.95 by fitting a GMM. In all cases, GMM gives the best results. Another interesting point is the low precision but high recall of all models on CIFAR-10 – this is also visible upon inspection of the samples in Fig. 5.

D. More Qualitative Results

	RECONSTRUCTIONS	RANDOM SAMPLES	INTERPOLATIONS
GT			
VAE			
CV-VAE			
WAE			
RAE-GP			
RAE-L2			
RAE-SN			
RAE			
AE			
GT			
VAE			
CV-VAE			
WAE			
RAE-GP			
RAE-L2			
RAE-SN			
RAE			
AE			

Figure 4. Qualitative evaluation for sample quality for VAEs, WAEs and RAEs on MNIST. Left: reconstructed samples (top row is ground truth). Middle: randomly generated samples. Right: spherical interpolations between two images (first and last column).

	RECONSTRUCTIONS	RANDOM SAMPLES	INTERPOLATIONS
GT			
VAE			
CV-VAE			
WAE			
RAE-GP			
RAE-L2			
RAE-SN			
RAE			
AE			
GT			
VAE			
CV-VAE			
WAE			
RAE-GP			
RAE-L2			
RAE-SN			
RAE			
AE			

Figure 5. Qualitative evaluation for sample quality for VAEs, WAEs and RAEs on CIFAR-10. Left: reconstructed samples (top row is ground truth). Middle: randomly generated samples. Right: spherical interpolations between two images (first and last column).

E. Investigating Overfitting

	MNIST		CIFAR-10		CELEBA	
VAE						
CV-VAE						
WAE						
RAE-GP						
RAE-L2						
RAE-SN						
RAE						
AE						

Figure 6. Nearest neighbors to generated samples (leftmost image, red box) from training set. It seems that the models have generalized well and fitting only 10 Gaussians to the latent space prevents overfitting.