SELF-SUPERVISED VARIATIONAL AUTO-ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Density estimation, compression and data generation are crucial tasks in artificial intelligence. Variational Auto-Encoders (VAEs) constitute a single framework to achieve these goals. Here, we present a novel class of generative models, called *self-supervised Variational Auto-Encoder* (selfVAE), that utilizes deterministic and discrete variational posteriors. This class of models allows to perform both conditional and unconditional sampling, while simplifying the objective function. First, we use a single self-supervised transformation as a latent variable, where a transformation is either downscaling or edge detection. Next, we consider a hierarchical architecture, i.e., multiple transformations, and we show its benefits compared to the VAE. The flexibility of selfVAE in data reconstruction finds a particularly interesting use case in data compression tasks, where we can trade-off memory for better data quality, and vice-versa. We present performance of our approach on three benchmark image data (Cifar10, Imagenette64, and CelebA).

1 INTRODUCTION

The framework of variational autoencoders (VAEs) provides a principled method for jointly learning latent-variable models and corresponding inference models. As it utilizes a meaningful low-dimensional latent space with density estimation capabilities, it forms an attractive solution for generative modelling tasks. However, as its performance in terms of test likelihood and quality of generated samples was far from the desired one, many modifications were proposed in order to improved its performance on high-dimensional data like natural images. In general, one can obtain a

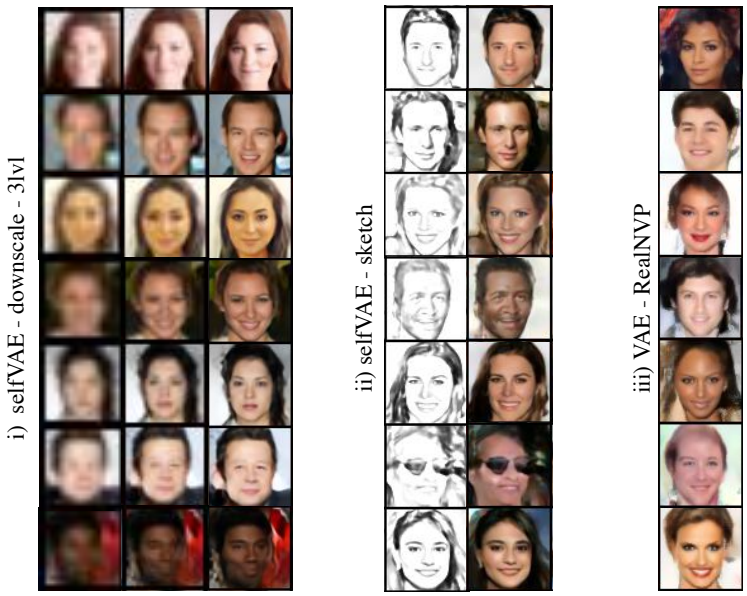

Figure 1: Unconditional CelebA generations from (i) the three-level self-supervised VAE with downscaling, (ii) the self-supervised VAE employed with edge detection (sketches), (iii) the VAE with RealNVP prior.

tighter lower bound, and, thus, a more powerful and flexible model, by advancing over the following three elements: the *encoder* (Rezende et al., 2014; van den Berg et al., 2018; Hoogeboom et al., 2020; Maaløe et al., 2016), the *prior* (or *marginal* over latents) (Chen et al., 2016; Habibian et al., 2019; Lavda et al., 2020; Lin & Clark, 2020; Tomczak & Welling, 2017) and the *decoder* (Gulrajani et al., 2016). Nevertheless, recent studies have shown that by employing deep hierarchical architectures and by carefully designed the building blocks of the neural networks, VAEs can successful model large high-dimensional data and reach state-of-the-art test likelihoods (Zhao et al., 2017; Maaløe et al., 2019; Vahdat & Kautz, 2020).

In this work, we present a novel class of generative models, called *self-supervised Variational Auto-Encoders*, where we introduce additional variables to VAEs that result from discrete and deterministic transformations of the given data samples, and specifically images. Since the transformations are deterministic, namely, non-trainable, and they provide a specific aspect of images (e.g., contextual information through detecting edges or downscaling), we refer to them as *self-supervised representations*. The introduction of the discrete and deterministic variables allows to train deep hierarchical models efficiently and steadily, by breaking down the discovery of a highly complex distribution into smaller, overall simpler ones. In this way, the model allows to integrate the prior knowledge about the data, but still enables to synthesize unconditional samples. Furthermore, the discrete and deterministic variables could be used to conditionally reconstruct data, which could be of great use in data compression and super-resolution tasks.

In summary, we make the following contributions: i) We propose an extension of the VAE framework by employing a bijective prior and by incorporating self-supervised representations of the data. ii) We analyze the impact of modelling natural images with different data transformations as self-supervised representations. iii) This new type of generative model, namely, *self-supervised Variational Auto-Encoders*, which is able to perform both conditional and unconditional sampling, demonstrate improved quantitative performance in terms of density estimation and generative capabilities on image benchmarks.

2 BACKGROUND

2.1 VARIATIONAL AUTO-ENCODERS

Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ with $\mathbf{x}_n \in \mathcal{X}^{\mathrm{D}}$ be independent and identically distributed observable variables, where $\mathcal{X}$ could be a discrete set of values (e.g., $\mathcal{X} = \{0, \ldots, 255\}$). Given that the dimensionality D of real-world data is typically large, thus, resulting in a complex, high-dimensional distributions, calculating the marginal likelihood $p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n)$ becomes a challenging problem. Moreover, a latent variable model with latent variables $\mathbf{z} \in \mathbb{R}^{\mathrm{M}}$, $p_\vartheta(\mathbf{x}) = \int p_\vartheta(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z}$, where ϑ denotes parameters, provides a framework for learning data representations. However, the optimization of $p_\vartheta(\mathbf{x})$ through maximum likelihood estimation (MLE) becomes infeasible due to the intractability of the integration at hand. One possible way of overcoming this issue and obtaining a highly scalable framework is by introducing a variational family of distributions $\mathcal{Q}$. In order to identify its (amortized) member $q_\phi(\mathbf{z}|\mathbf{x})$, we aim at minimizing the Kullback-Leibler divergence (KL) to the *real* posterior $p(\mathbf{z}|\mathbf{x})$. In consequence, we can derive a tractable objective function, namely, the *evidence lower bound* (ELBO) (Jordan et al., 1999):

$$\log p_\vartheta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_\lambda(\mathbf{z})\right] \equiv \mathcal{L}(\theta, \phi, \lambda),$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is the variational posterior (or the *encoder*), $p_\theta(\mathbf{x}|\mathbf{z})$ is the likelihood function (or the *decoder*) and $p_\lambda(\mathbf{z})$ is the *prior* over the latent variables, parameterized by ϕ, θ and λ respectively. The optimization is done efficiently by computing the expectation by Monte Carlo integration while exploiting the *reparameterization trick* in order to obtain an unbiased estimator of the gradients. This generative framework is known as *Variational Auto-Encoder* (VAE) (Kingma & Welling, 2013; Rezende et al., 2014).

2.2 VAEs WITH BIJECTIVE PRIOR

Even though the lower-bound suggests that the prior plays a crucial role in improving the variational bounds, usually it is modelled by a fixed distribution (i.e., a standard multivariate Gaussian). While being relatively simple and computationally cheap, a fixed prior is known to result in over-regularized

models that tend to ignore more of the latent dimensions (Burda et al., 2015; Hoffman & Johnson, 2016; Tomczak & Welling, 2017). Moreover, even with powerful encoders, VAEs may still fail to match the variational posterior to a unit Gaussian prior (Rosca et al., 2018).

However, it is possible to obtain a rich, multi-modal prior distribution $p(\mathbf{z})$ by using a *bijective* (or *flow-based*) model (Dinh et al., 2016). Formally, given a latent code $\mathbf{z} \sim q_Z(\mathbf{z}|\mathbf{x})$, a base distribution $p_V(\mathbf{v})$ over latent variables $\mathbf{v} \in V$, and $f : V \rightarrow Z$ consisting of a sequence of L diffeomorphic transformations[1], where $f_i(\mathbf{v}_{i-1}) = \mathbf{v}_i$, $\mathbf{v}_0 = \mathbf{v}$ and $\mathbf{v}_L = \mathbf{z}$, the *change of variable* can be used sequentially to express the distribution of $\mathbf{z}$ as a function of $\mathbf{v}$ as follows:

$$\log p_Z(\mathbf{z}) = \log p_V(\mathbf{v}) - \sum_{i=1}^{L} \log \left| \frac{\partial f_i(\mathbf{v}_{i-1})}{\partial \mathbf{v}_{i-1}} \right|,$$

where $\left| \frac{\partial f_i(\mathbf{v}_{i-1})}{\partial \mathbf{v}_{i-1}} \right|$ is the Jacobian-determinant of the ith transformation.

Thus, using the transformed prior we end up with the following training objective function:

$$\mathcal{L}(\theta, \phi, \lambda) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \Big[\log p_\theta(\mathbf{x}|\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}) + \log p_V(\mathbf{v}_0) + \sum_{i=1}^{L} \log \left| \frac{\partial f_i^{-1}(\mathbf{v}_i)}{\partial \mathbf{v}_i} \right| \Big].$$

In this work, we utilize RealNVP (Dinh et al., 2016) as the prior, however, any other flow-based model could be used (Kingma & Dhariwal, 2018). For the experiments and ablation study that shows the impact of the bijective prior on VAEs, please refer to the appendix (Section A.2).

3 Method

3.1 Motivation

The idea of self-supervised learning is about utilizing original unlabeled data to create additional context information. It could be achieved in multiple manners, e.g., by adding noise to data (Vincent et al., 2008) or masking data during training (Zhang et al., 2017). Self-supervised learning could be also seen as turning an unsupervised model into a supervised one by, e.g., treating prediction of next pixels as a classification task (Hénaff et al., 2019; Oord et al., 2018). These are only a few examples of a quickly growing research line (Liu et al., 2020).

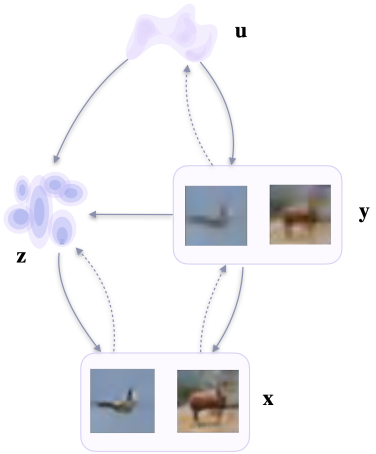

Figure 2: The proposed approach.

Here, we propose to use non-trainable transformations to obtain information about image data. Our main hypothesis is that since working with highly-quality images is difficult, we could alleviate this problem by additionally considering partial information about them. Thus, fitting a model to images of lower quality, and then enhancing them to match the target distribution seems to be overall an easier task (Chang et al., 2004; Gatopoulos et al., 2020). By incorporating compressed transformations (i.e., the self-supervised representations) that still contain global information, with the premise that it would be easier to approximate, the process of modelling a high-dimensional complex density breaks down into simpler tasks. In this way, the expressivity of the model will grow and gradually result into richer, better generations. A positive effect of the proposed framework is that the model allows us to integrate prior knowledge through the image transformations, without losing its unconditional generative functionality. Overall, we end up with a two-level VAE with three latent variables, where one is a data transformation that can be obtained in a self-supervised fashion. In Figure 2 a schematic representation of the proposed approach with downscaling is presented.

[1]That is, invertible and differentiable transformations.

A number of exemplary image transformations are presented in Figure 7. We can see that with these transformations, even though we discard a lot of information, the global structure is preserved. As a result, in practice the model should have the ability to extract a general concept of the data at the first stage, and add local information afterwards. In this work, we focus on downscaling (Figure 7.b, c & d in the appendix) and edge detection or *sketching* (Fig. 7.i in the appendix).

3.2 Model formulation

Let us introduce additional variables $\mathbf{y} \in \mathbb{R}^{\mathrm{C}}$ that result from self-supervised transformations of $\mathbf{x}$, where $C \leq D$. Further, let $\mathbf{u} \in \mathbb{R}^N$ and $\mathbf{z} \in \mathbb{R}^M$ be stochastic latent variables that interact with the above observed variables in a manner that is presented in Figure 2. The dependencies of the considered model yield the following objective function (see the appendix A.4 for the full derivation):

$$\begin{aligned}\mathcal{L}(\mathbf{x}) = &\mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})\, q(\mathbf{y}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z}) - \mathbb{E}_{q(\mathbf{u}|\mathbf{y})} D_{\mathrm{KL}}(q(\mathbf{y}|\mathbf{x})||p_\theta(\mathbf{y}|\mathbf{u})) + \\ &- \mathbb{E}_{q(\mathbf{u}|\mathbf{y})q(\mathbf{y}|\mathbf{x})} D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x},\mathbf{y})||p(\mathbf{z}|\mathbf{y},\mathbf{u})) - \mathbb{E}_{q(\mathbf{y}|\mathbf{x})} D_{\mathrm{KL}}(q(\mathbf{u}|\mathbf{y})||p(\mathbf{u})),\end{aligned} \tag{1}$$

where $D_{\mathrm{KL}}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

In addition, to simplify the training objective function and allow it to scale with high-dimensional data, we propose to consider the distribution $q(\mathbf{y}|\mathbf{x})$ as a *deterministic* and *discrete* density, i.e., a degenerate distribution (the Dirac's delta), that gives $\mathbb{E}_{q(\mathbf{y}|\mathbf{x})}\Big[\log q(\mathbf{y}|\mathbf{x})\Big] = 0$. This property is important, because it does not introduce any complexity and allows us to exploit prior knowledge (i.e., image transformations) of the generative process, giving emphasis on specific desired characteristics of the processed data. Next, we simplify the distribution $q(\mathbf{z}|\mathbf{y},\mathbf{x})$ to $q(\mathbf{z}|\mathbf{x})$ since we assume that $\mathbf{y}$ does not introduce extra information about $\mathbf{x}$. Thus, with these two properties in mind, the lower bound of the marginal likelihood of $\mathbf{x}$ is the following:

$$\begin{aligned}\mathcal{L}(\mathbf{x}) = &\underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})\, q(\mathbf{y}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})}_{\mathrm{RE}_x} + \underbrace{\mathbb{E}_{q(\mathbf{u}|\mathbf{y})q(\mathbf{y}|\mathbf{x})} \log p_\theta(\mathbf{y}|\mathbf{u})}_{\mathrm{RE}_y} + \\ &- \underbrace{\mathbb{E}_{q(\mathbf{u}|\mathbf{y})q(\mathbf{y}|\mathbf{x})} D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{y},\mathbf{u}))}_{\mathrm{KL}_z} - \underbrace{\mathbb{E}_{q(\mathbf{y}|\mathbf{x})} D_{\mathrm{KL}}(q(\mathbf{u}|\mathbf{y})||p(\mathbf{u}))}_{\mathrm{KL}_u}.\end{aligned} \tag{2}$$

Further, in order to fully define our model, we propose to choose the following distributions:

$$\begin{aligned}
p(\mathbf{v}) &= \mathcal{N}(\mathbf{v}|\mathbf{0},\mathbf{1}) \\
p_\lambda(\mathbf{u}) &= p(\mathbf{v}) \prod_{i=1}^{F} \left| \det \frac{\partial f_i(\mathbf{v}_{i-1})}{\partial \mathbf{v}_{i-1}} \right|^{-1} & q(\mathbf{y}|\mathbf{x}) &= \delta(\mathbf{y} = d(\mathbf{x})) \\
p_{\theta_1}(\mathbf{y}|\mathbf{u}) &= \sum_{i=1}^{I} \pi_i^{(\mathbf{u})} \mathrm{Dlogistic}\Big(\mu_i^{(\mathbf{u})}, s_i^{(\mathbf{u})}\Big) & q_{\phi_1}(\mathbf{u}|\mathbf{y}) &= \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_{\phi_1}(\mathbf{y}), \mathrm{diag}(\boldsymbol{\sigma}_{\phi_1}(\mathbf{y}))) \\
p_{\theta_2}(\mathbf{z}|\mathbf{y},\mathbf{u}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\theta_2}(\mathbf{y},\mathbf{u}), \mathrm{diag}(\boldsymbol{\sigma}_{\theta_2}(\mathbf{y},\mathbf{u}))) & q_{\phi_2}(\mathbf{z}|\mathbf{x}) &= \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi_2}(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}_{\phi_2}(\mathbf{x}))) \\
p_{\theta_3}(\mathbf{x}|\mathbf{z},\mathbf{y}) &= \sum_{i=1}^{I} \pi_i^{(\mathbf{z},\mathbf{y})} \mathrm{Dlogistic}\Big(\mu_i^{(\mathbf{z},\mathbf{y})}, s_i^{(\mathbf{z},\mathbf{y})}\Big)
\end{aligned}$$

where $\mathrm{Dlogistic}$ is defined as the discretized logistic distribution (Salimans et al., 2017), $\delta(\cdot)$ is the Dirac's delta, and $d(\mathbf{x})$ denotes the deterministic transformation that returns discrete values. In order to highlight the self-supervised part in our model, we refer to it as the *self-supervised Variational Auto-Encoder* (or selfVAE for short).

It is noteworthy that considering RE_y and KL_u in Eq. 2 alone is equivalent to a standard VAE objective, but for the transformed data $\mathbf{y}$. Intuitively, the premise for selfVAE is that the latents $\mathbf{u}$ will capture the global structure of the input data and the latents $\mathbf{z}$ will encode the missing information between $\mathbf{y}$ and $\mathbf{x}$, guiding the model to discover the distribution of the target observations.

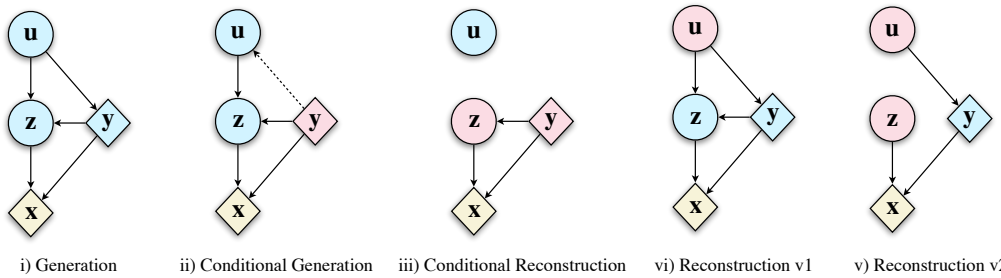

Figure 3: Operations of the self-supervised Variational Auto-Encoder. The blue and pink nodes represent the sampled and inferred latent codes, respectively.

3.3 Generation and Reconstruction in selfVAE

As generative models, VAEs can be used to synthesize novel content through the following process:

$$\mathbf{z} \sim p(\mathbf{z}) \rightarrow \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}),$$

but also to reconstruct a data sample $\mathbf{x}$ by using the following scheme:

$$\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}^*) \rightarrow \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}).$$

Interestingly, our approach allows to utilize more operations regarding data generation and reconstruction, which are illustrated in Figure 3. First, analogously to VAEs, the selfVAE allows to generate data by applying the following hierarchical sampling process (*generation*; Fig 3:i):

$$\mathbf{u} \sim p(\mathbf{u}) \rightarrow \mathbf{y} \sim p(\mathbf{y}|\mathbf{u}) \rightarrow \mathbf{z} \sim p(\mathbf{z}|\mathbf{u}, \mathbf{y}) \rightarrow \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{y}).$$

Given that we can sample or infer the latent codes, we can use a variety of options on how to reconstruct or generate synthetic data $\mathbf{x}$. In particular, in *conditional generation* and *conditional reconstruction* (Fig 3:ii and iii) we use the ground-truth $\mathbf{y}$ (that is the $q(\mathbf{y}|\mathbf{x})$, a deterministic and discrete transformation) and sampled or inferred $\mathbf{z}$. The generative process for the former would be:

$$\mathbf{z} \sim q(\mathbf{z}|\mathbf{x}^*) \rightarrow \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{y}),$$

and for the latter:

$$\mathbf{u} \sim q(\mathbf{u}|\mathbf{y}) \rightarrow \mathbf{z} \sim p(\mathbf{z}|\mathbf{y}, \mathbf{u}), \rightarrow \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{y}).$$

Notably, we can take advantage of the first process to tackle self-supervised or even supervised tasks, which relates to the chosen representation of $\mathbf{y}$. For example, if $\mathbf{y}$ is a downscaling transformation of the input image, it can be used in a manner similar to the super-resolution Gatopoulos et al. (2020).

Alternatively, given the flexibility of the model, we can instead sample (or generate) $\mathbf{y}$ and choose to sample or infer $\mathbf{z}$. In this way, we can reconstruct in two ways, namely, the *reconstruction 1* (Figure 3:iv):

$$\mathbf{y}^* \sim q(\mathbf{y}^*|\mathbf{x}^*) \rightarrow \mathbf{u} \sim q(\mathbf{u}|\mathbf{y}^*) \rightarrow \mathbf{y} \sim p(\mathbf{y}|\mathbf{u}) \rightarrow \mathbf{z} \sim p(\mathbf{z}|\mathbf{y}, \mathbf{u}), \rightarrow \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{y}),$$

and the *reconstruction 2* (Figure 3:v):

$$\Big(\mathbf{y}^* \sim q(\mathbf{y}^*|\mathbf{x}^*) \rightarrow \mathbf{u} \sim q(\mathbf{u}|\mathbf{y}^*) \rightarrow \mathbf{y} \sim p(\mathbf{y}|\mathbf{u})\Big) \text{ and } \mathbf{z} \sim q(\mathbf{z}|\mathbf{x}^*) \rightarrow \mathbf{x} \sim p(\mathbf{x}|\mathbf{z}, \mathbf{y}).$$

As we will see in the experiments, each option creates a different ratio of the reconstruction quality against the memory that we need to allocate to send information. Note, however, that every inferred variable needs to be sent, thus, more sampling corresponds to lower memory requirements.

3.4 HIERARCHICAL SELF-SUPERVISED VAE

The proposed approach can be further extended and generalized by introducing multiple transformations, in the way that it is illustrated in Figure 4. By incorporating a single (or multiple) self-supervised representation(s) of the data, the process of modelling a high-dimensional complex density breaks down into K simpler modeling tasks. Thus, we obtain a K-level VAE architecture, where the overall expressivity of the model grows even further and gradually results into generations of higher quality. Some transformations cannot be applied multiple times (e.g., edge detection), however, others naturally could be used sequentially, e.g., downscaling.

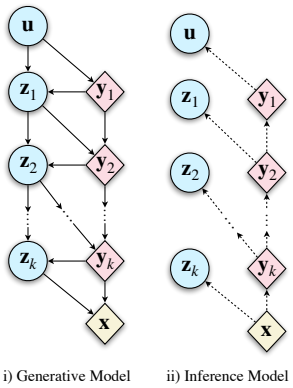

Figure 4: Hierarchical selfVAE.

Formally, the general case of Eq. (2) for K self-supervised data transformations, where we define $\mathbf{z}_0 = \mathbf{u}$ and $\mathbf{y}_{K+1} = \mathbf{x}$, can be expressed as follows:

$$\mathcal{L}(\mathbf{x}) = \sum_{k=1}^{K} \mathbb{E}_{q(\mathbf{z}_k|\mathbf{y}_{k+1})q(\mathbf{y}_k|\mathbf{y}_{k+1})} \log p_\theta(\mathbf{y}_{k+1}|\mathbf{y}_k, \mathbf{z}_k) + \mathbb{E}_{q(\mathbf{u}|\mathbf{y}_1)q(\mathbf{y}_1|\mathbf{y}_2)} \log p_\theta(\mathbf{y}_1|\mathbf{u}) +$$
$$- \mathbb{E}_{q(\mathbf{y}_1|\mathbf{y}_2)} D_{\mathrm{KL}}(q(\mathbf{u}|\mathbf{y}_1)||p(\mathbf{u})) - \sum_{k=2}^{K} \mathbb{E}_{q(\mathbf{z}_{k-1}|\mathbf{y}_k)q(\mathbf{y}_k|\mathbf{y}_{k+1})} D_{\mathrm{KL}}(q(\mathbf{z}_k|\mathbf{y}_{k+1})||p(\mathbf{z}_k|\mathbf{y}_k, \mathbf{z}_{k-1})).$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate the proposed model on datasets of natural images: CIFAR-10, Imagenette64 and CelebA. We design a neural network architectures for the encoder and the decoder that consists building blocks composed of DenseNets (Huang et al., 2016) in order to encourage features reuse and help to preserve visual information, channel-wise attention (Zhang et al., 2018) to favor the most informative channel features, and ELUs (Clevert et al., 2015) as activation functions. The dimensionality of all the latent variables were kept at $8 \times 8 \times 16 = 1024$ and all models were trained using AdaMax optimizer (Kingma & Ba, 2014) with data-depended initialisation (Salimans & Kingma, 2016). Regarding the selfVAEs, in CIFAR-10 we used an architecture with a single downscaled transformation (selfVAE-downscale), while on the remaining two datasets (CelebA and Imagenette64) we used a hierarchical 3-leveled selfVAE with downscaling, and a selfVAE with sketching. All models were employed with the bijective prior (RealNVP) comparable in terms of the number of parameters (the range of the weights of all models was from 32M to 42M). For more information about the neural network architecture and the datasets that where used, please refer to the appendix sections A.3 and A.5, respectively. As evaluation metrics, we approximate the negative log-likelihood (*nll*) using 512 IW-samples (Burda et al., 2015) and express the scores in bits per dimension (*bpd*). Additionally, for CIFAR-10, we use the *Fréchet Inception Distance* (FID) (Heusel et al., 2017).

4.2 QUANTITATIVE RESULTS

We present the results of the experiments on the three benchmark datasets in Table 1. First we notice that on CIFAR-10 our implementation of the VAE is still lacking behind other generative models in terms of *bpd*, however, it is better or comparable in terms of FID. The selfVAE-downscale achieves worse *bpd* than the VAE. A possible explanation may lie in the small image size (32px $\times$ 32px), as the benefits of breaking down the learning process in two or more steps are not obvious given the small target dimensional space. Nevertheless, the selfVAE-downscale achieves significantly better FID scores than the other generative models. This result could follow from the fact that downscaling allows to maintain context information about the original image and, as a result, a general coherence is of higher quality.

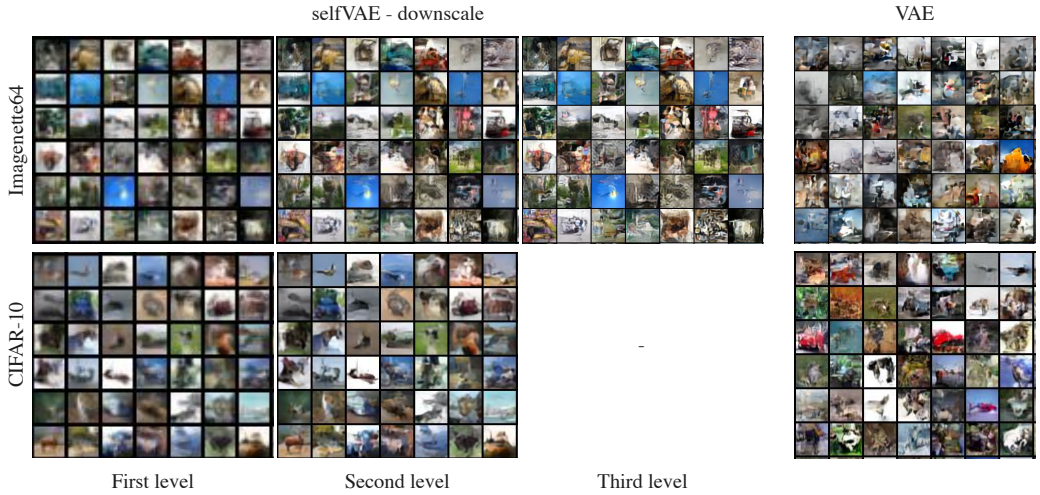

Figure 5: Uncoditional generations on Imagenette64 and CIFAR-10.

Interestingly, on the two other datasets, a three-level selfVAE-downscale achieves significantly better *bpd* scores than the VAE with the bijective prior. This indicates that the benefit of employing a multi-leveled self-supervised framework against the VAE is more obvious when we move to higher-dimensional data, where the plain model fails to scale efficiently. Unlike before, the RE is now better, but the biggest difference lies in the KL term. Apparently, it seems that the hierarchical structure of self-supervised random variables allows to encode the missing information more efficiently in $\mathbf{z}_k$, in contrast to the vanilla VAE, where all information about images must be coded in $\mathbf{z}$. This result is promising and indicates that the proposed approach of using information in data in a self-supervised manner is of great potential for generative modelling.

4.3 Qualitative results

We present generations on CIFAR-10 and Imagenette64 in Figure 5 and on CelebA in Figure 1. In addition, we provide reconstructions on CIFAR-10 and CelebA in Figure 6. We first notice that the generations from selfVAE seem to be more coherent, in contrast with these from VAE that produces overall more contextless and distorted generations. This result seems to be in line with the FID scores. Especially for CelebA, we observe impressive synthesis quality, great sampling diversity and coherent

Table 1: Quantitative comparison on test sets from CIFAR-10, CelebA, and Imagenette64. *Measured on training set.

Dataset	**Model**	*bpd*	RE	KL	FID
CIFAR-10	PixelCNN (van den Oord et al., 2016)	3.14	-	-	65.93
	GLOW Kingma & Dhariwal (2018)	3.35	-	-	65.93
	ResidualFlow (Chen et al., 2019)	3.28	-	-	46.37
	BIVA (Maaløe et al., 2019)	3.08	-	-	-
	NVAE Vahdat & Kautz (2020)	**2.91**	-	-	-
	WGAN-GP (Gulrajani et al., 2017)	-	-	-	36.40
	VAE (ours)	3.51	5540	1966	41.36 (37.25*)
	selfVAE-downscale	3.65	6348	1438	**34.71** (**29.95***)
CelebA	VAE (ours)	3.12	24096	2502	-
	selfVAE-sketch	3.24	23797	3816	-
	selfVAE-downscale-3lvl	**2.88**	23336	1214	-
Imagenette64	VAE (ours)	3.85	30809	1961	-
	selfVAE-downscale-3lvl	**3.70**	30394	1171	-

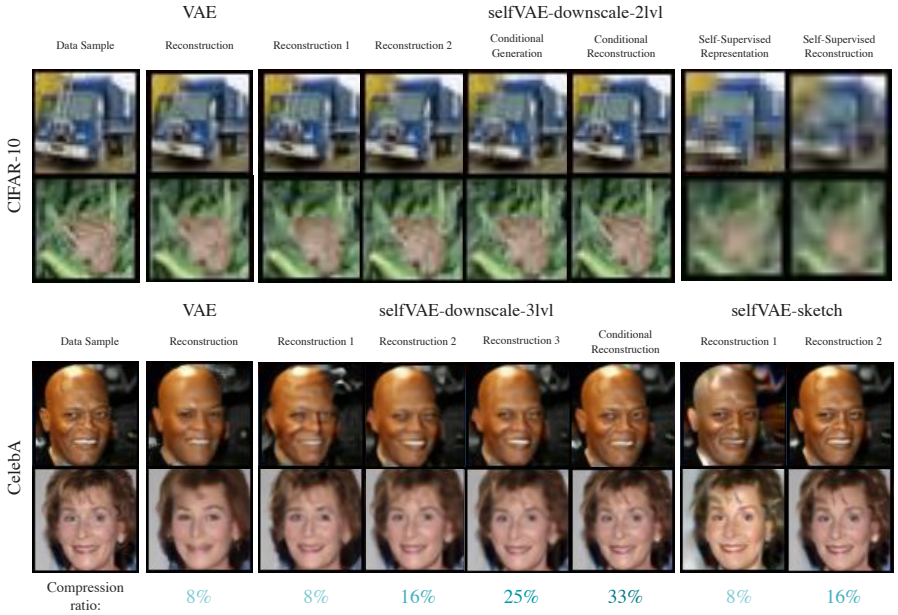

Figure 6: Comparison on image reconstructions with different amount of sent information.

generations (Figure 1). On the Imagenette64 dataset, we can also observe crisper generations for our method compared to the VAE (Figure 5). Furthermore, the hierarchical selfVAE seems to be of a great potential for compression purposes. In contrast to the VAE, which is restricted to using a single way of reconstructing an image, the selfVAE allows four various options with different quality/memory ratios (see Figure 6). Impressively, with the selfVAE-sketch we can retrieve the image with high accuracy by using only 16% of the original data, as it manages to encode all the texture of the image to $\mathbf{z}$ (see Figure 12 in the appendix). This shows the advantage of choosing prior knowledge into the learning process. Lastly, the selfVAE-downscale can perform super-resolution out-of-the-box successfully that further highlights the importance of encoding details in $\mathbf{z}$. The latents $\mathbf{z}$ learns to add extra information, which defines the end result, and, when it sampled, we can alter details of an image like facial expressions (see Figure 13.ii in the appendix).

5 Conclusion

We proposed a novel class of generative models, called self-supervised Variational Auto-Encoder (selfVAE), where we extend the plain VAE architecture with self-supervised representations. The application of transforming data in a self-supervised manner allows to integrate prior knowledge without loosing its unconditional generative functionality. We showed that taking deterministic and discrete transformations results in a simplified ELBO which incorporates self-supervised representations. In the experimental studies, we indicated that the proposed approach results in coherent generations of high visual quality. Moreover, the results seem to confirm that hierarchical architectures perform better and allow to obtain both better *bpd* scores and better generations and reconstructions. In this paper, we considered two classes of image transformations, namely, *downscaling* and edge detection (*sketching*). However, there is a vast of possible other transformations (see Figure 7), and we leave investigating them for future work. Moreover, we find the proposed approach interesting for the compression task. A similar approach with a multi-scale auto-encoder for image compression was proposed, e.g, by Mentzer et al. (2019) or Razavi et al. (2019). However, we still use a probabilistic framework and indicate that various non-trainable image transformations (not only multiple scales) could be of great potential.

Acknowledgments

Anonymized for the double-blind review.

REFERENCES

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv*, 2015.

Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pp. I–I. IEEE, 2004.

Ricky T. Q. Chen, Jens Behrmann, David Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *arXiv*, 2019.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv*, 2016.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv*, 2015.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv*, 2016.

Eduardo S. L. Gastal and Manuel M. Oliveira. Domain transform for edge-aware image and video processing. *ACM TOG*, 30(4):69:1–69:12, 2011. Proceedings of SIGGRAPH 2011.

Ioannis Gatopoulos, Maarten Stol, and Jakub M. Tomczak. Super-resolution variational auto-encoders. *arXiv.2006.05218*, 2020.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. Pixelvae: A latent variable model for natural images. *arXiv*, 2016.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv*, 2017.

Amirhossein Habibian, Ties van Rozendaal, Jakub M Tomczak, and Taco S Cohen. Video compression with rate-distortion autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7033–7042, 2019.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv*, 2017.

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

Emiel Hoogeboom, Victor Garcia Satorras, Jakub M Tomczak, and Max Welling. The convolution exponential and generalized sylvester flows. *arXiv preprint arXiv:2006.01910*, 2020.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *arXiv*, 2016.

Michael I. Jordan, Zoubin Ghahramani, and et al. An introduction to variational methods for graphical models. In *Machine Learning*, pp. 183–233. MIT Press, 1999.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013.

Frantzeska Lavda, Magda Gregorová, and Alexandros Kalousis. Data-dependent conditional priors for unsupervised learning of multimodal data. *Entropy*, 22(8):888, 2020.

Shuyu Lin and Ronald Clark. Ladder: Latent data distribution modelling with a generative prior. *arXiv preprint arXiv:2009.00088*, 2020.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *arXiv preprint arXiv:2006.08218*, 2020.

Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. *arXiv*, 2016.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. *arXiv*, 2019.

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Practical full resolution learned lossless image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10629–10638, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, pp. 14866–14876, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv*, 2014.

Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *arXiv*, 2018.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv*, 2016.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv*, 2017.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *arXiv*, 2016.

Jakub M Tomczak and Max Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv*, 2020.

Rianne van den Berg, Leonard Hasenclever, Jakub M. Tomczak, and Max Welling. Sylvester normalizing flows for variational inference. *arXiv*, 2018.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv*, 2016.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. *arXiv*, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from deep generative models. In *International Conference on Machine Learning*, pp. 4091–4099, 2017.