Enhancing by Decoupling: Exploration of a Unified VAE Framework

Zhiyuan Wu, Xinhui Bian, Dan Li
Research School of Computer Science
Australian National University, Canberra
{u7155144, u7096593, u7076589}@anu.edu.au

Abstract

Variational autoencoders (VAEs) are one of the most popular unconditional generative models with elegant theory and potential capability in unsupervised representation learning. Specifically, they are expected to produce both informative and interpretable features, and approximately optimize over the log likelihood (LL). However, current VAE variants could hardly fulfill these expectations at the same time. We profoundly analyze the pain point and summarize the reconstruction-KL rivalry as well as the model design as the inherent causes. As such, we propose a theoretical enhancement of VAEs to uniformly relieve the drawbacks by introducing an intermdediate feature manifold to decouple the reconstruction and the sampling process. Empirical results demonstrate the superiority of our implementation in terms of producing latent features with the desired property, and hopefully our framework could provide theoretical guidance for subsequent researches in VAE enhancement.

1. Introduction

Unconditional data generation, which is a challenging and significant research area in unsupervised learning, aims to generate data with high fidelity from the underlying distribution of a given dataset. Being essential to many attractive fields in modern machine learning including few-shot learning and representation learning, this area has been developed a lot over recent years, where Variational Autoencoders (VAEs) [20] are proposed as a theoretically-elegant solution with impressive training stability.

In parallel with their generative function, the encoder-decoder structure of VAEs endows them with superior potential capability in terms of unsupervised representation learning. First, while reducing dimensionality, the learnt representation reserves most characteristic information of the input, as the decoder could produce an almost identical piece of data (with minor loss of sharpness) given this latent feature. Second, to fit with the disentangled prior distribution (normal distribution), the feature vector should be an interpretable factorized representation with respect to disentangled data generative factors, each dimension of it corresponding to an independent attribute of the data (e.g. hair color for human faces, brightness for outdoor scenes). This is extremely useful for interpretation-intensive scenarios, such as zero-shot inference and novelty detection, which is simple for humans but remains a pain point for current AI algorithms. Therefore, theoretically VAEs should learn both informative and interpretable representations.

However, as indicated by the literature review in Section 3, the original VAE model [20] is not able to reach this theoretical expectation, and, even worse, it adopts a sub-optimal training objective. Many variants are proposed to solve these issues, but they generally address merely one aspect at the possible cost of behaving more poorly at the others.

Section 4 diagnoses these symptoms of the VAE family, and points out that the intrinsic contradiction between the reconstruction and the sampling process takes the main responsibility. Also, the design choice of the objective limits the potential generative capability of VAEs.

According to the above analysis, in Section 5 we propose our theoretical enhancement of the VAE framework which uniformly relieves the aforementioned drawbacks by introducing an intermdediate feature manifold to decouple the reconstruction and the sampling process, and meanwhile adopts the optimal training objective.

In Section 6, we attempt our idea using the structure of Soft-IntroVAEs [3] and i-ResNets [2], and demonstrate the superiority of our implementation over state-of-the-arts in terms of latent code informativeness and interpretability. Unfortunately we did not outperform the state-of-the-art in generative quality, but please kindly note that we mainly aim to present a theoretical exploration and there's still plenty of room in network architecture optimization.

Our main contribution is three-fold:

- We summarize the main challenges of the VAE family, and find the intrinsic causes.
- We propose a theoretical enhancement of VAEs which

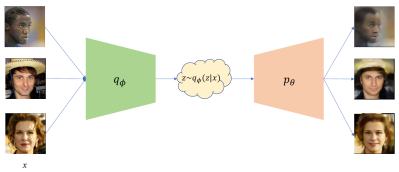


Figure 1: The pipeline of VAEs.

could potentially solve the problems all at once.

- We experimentally attempt our thoughts, and the results are conceptually encouraging.

2. Preliminary

2.1. Generative models

Denoting the generative distribution as p_{model} and the underlying data distribution as p_{data}, the goal of generative models is to minimize the distance between p_{model} and p_{data}. This optimal goal could be mathematically characterized by maximizing the log likelihood (LL),

$$LL = \mathbb{E}_{x\sim p_{\text{data}}} \left[\log p_{\text{model}}(x)\right]. \tag{1}$$

2.2. VAEs

One of the most popular generative models is VAEs, as their theory is elegant and training is stable and fast. The pipeline of VAEs is shown in Fig 1, where its enoder q_ϕ takes a piece of data x as input and produces the posterior distribution $q_\phi(z|x)$, and its decoder p_θ takes the latent feature z as input and produces $p_\theta(x|z)$. Classically, its training objective is to maximize the evidence lower bound (ELBO) of LL [20],

$$\begin{aligned} ELBO = \mathbb{E}_{x\sim p_{\text{data}}}\big[&- KL\left(q_\phi(z|x)\|p_\theta(z)\right) \\ &+ \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right]\big] \leq LL, \end{aligned} \tag{2}$$

where $\mathbb{E}_{x\sim p_{\text{data}}}\big[KL\left(q_\phi(z|x)\|p_\theta(z)\right)\big]$ is referred to as the KL loss while $-\mathbb{E}_{x\sim p_{\text{data}}}\big[\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right]\big]$ being the reconstruction loss. During generation, first sample z from its prior distribution $p_\theta(z)$, and then feed it into the decoder to obtain the generated sample from $p_\theta(x|z)$. A standard choice of the prior distribution $p_\theta(z)$ is the normal distribution, with $q_\phi(z|x)$ and $p_\theta(x|z)$ both being the Gaussian distribution.

2.3. Invertible networks

It is generally impossible to recover input from output for standard networks because of the substantial information loss in the network flow. An invertible network, however, serves as a bijective mapping between its input and output. Many popular solutions of invertible networks are developed in recent years, such as i-ResNet [2], the family of flow models [13, 14, 5, 4]. I-ResNet achieves invertibility by restricting the Lipschitz constant of a regular ResNet [9], while flow models adopt the dimension-splitting and affine transforming strategy.

3. Related work: three main challenges

After the initial VAE framework as in Section 2.2, most subsequent variants aim to improve one of the three targets as follows. To focus on the theoretical limits of VAEs, we exclude methods involving the idea of the Generative Adversarial Nets (GANs) [7], e.g. [3].

3.1. Informativeness of latent features

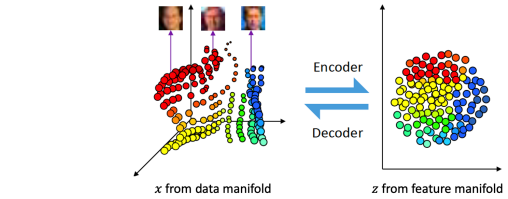
Many papers [18] suggest there is a "posterior collapse" effect, i.e. the KL loss dominants and the posterior distribution $q_\phi(z|x)$ collapses to the prior $p_\theta(z)$. In this case, the encoder tends to produce similar latent codes z ignoring the input x, causing z to be uninformative. [12] attempts to address this issue by using an annealed objective, while [8] lags the decoder behind the coder during training and [21] suggests a structural constraint to the distribution.

3.2. Interpretability of latent features

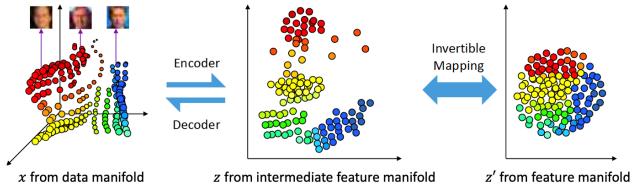
As pointed by [11], another challenge for VAE models is to learn independent latent factors which could help knowledge transfer, zero-shot inference, etc. This typically takes place when the KL loss is too weak such that the posterior distribution $q_\phi(z|x)$ severely deviates from the dimensional-disentangled prior $p_\theta(z)$. [11] relaxes the issue by adding a weighting coefficient β (> 1) to the KL loss, followed by [22] which makes β controllable.

3.3. Sub-optimal training objective

Some criticizers [6] indicate the fact that VAEs' training objective being ELBO (2) instead of LL (1) itself might have side effects on their performance, as even an ideal model reaching the global optimum of ELBO could confront $p_{\text{model}} \neq p_{\text{data}}$. More recently, [19] and [1] get rid of this sub-optimal goal by applying structural changes to the original framework with great success in generative quality. Specifically, [19] forces the latent feature z to be discrete, pairs it with an autoregressive prior $p_\theta(z)$ and applies a semi-2-stage training. Meanwhile, [1] points out that VAEs' training goal suppresses its ability to simultaneously reduce dimensionality and generate high-fidelity data, and implements a 2-stage VAE accordingly to circumvent this intrinsic restriction. This is the most similar work to our approach, as we adopt an analogical 2-stage scheme. The difference comes in that we mainly blame the two competing processes of VAEs and explicitly decouple them.



(a) VAE



(b) Ours

Figure 2: Visualization of VAEs and our model from the manifold mapping perspective. During generation, potential features are first sampled from the normal distribution and then fed to the model from right to left.

4. Diagnosing the symptoms

It is noted that previous works in Section 3 generally focus on one aspect of the proposed challenges, and some of them even borrow from Peter to pay Paul. For instance, papers suggesting a structural limitation [21] to improve the informativeness or an alternative goal exaggerating the influence of KL loss [11, 22] would definitely hurt the overall performance. In this section, we summarize the inner causes of the challenges as 1) the tension between reconstruction and sampling, and 2) the fact that ELBO (2) could never reach the desired optimal goal.

4.1. Reconstruction vs. sampling

To find the potential reason for low interpretability and informativeness, we will inspect VAEs from a manifold mapping perspective, as shown in Fig 2a.

From the figure, the VAE serves as an approximately bijective mapping between the data mainfold and the feature manifold. VAEs generate realistic data by first sampling potential latent representations from the normal distribution and then feeding them to the decoder. To precisely generate samples from the data manifold, VAEs must ensure 1) the effectiveness of potential feature sampling, i.e. the similarity between the feature manifold and the normal distribution, which is charged by the KL loss, and 2) the effectiveness of the decoder, i.e. the reversibility of the encoder-decoder mapping, which is charged by the reconstruction loss. In other words, the sampling and the reconstruction processes should both be effectual.

Since then, the training of VAEs could be intuitively treated as a softly-constrained optimization problem: Maximize the reversibility of the encoder-decoder architecture under a soft structural constraint on the feature manifold, that is keeping it to be a normal distribution. However, as is known to all, even the unconstrained optimization problem itself is not easy to solve, as an autoencoder could never perfectly recover the original data. Intuitions could tell that the additional structural constraint for accurate sampling would to some extent affect the encoder-decoder reversibility for high-quality reconstruction.

Experimentally, a series of trained VAEs with different hyperparameters are fully investigated to validate the above-mentioned intuitions, i.e. the tension between sampling and reconstruction. From Fig 3a, it could be seen that as the weight of the KL loss (soft constraint) increases, the reconstruction loss ascends and the reconstruction quality descends drastically, suggesting that the reconstruction process would favor low weight of the sampling process. In addition, Fig 3b further shows that the similarity between the feature manifold and the normal distribution boosts as the dominance of the reconstruction loss drops, indicating that the sampling process would benefit from low reconstruction pressure. To sum up, qualitative and quantitative results show that there exists a tradeoff between the two processes during the training of VAEs.

Note that the informativeness in Section 3.1 could be measured by the reconstruction loss, as the more information z carries about x, the easier for x to be reconstructed given z. On the other hand, the interpretability could be measured by the KL loss, as the normal distribution is dimensional-disentangled, and higher interpretability, i.e. easier factorization of z, corresponds to lower KL divergence between $q_\phi(z|x)$ and the normal distribution. Hence, it could be summarized that the informativeness and the interpretability of z are opposing each other, caused by the inherent rivalry between reconstruction and sampling.

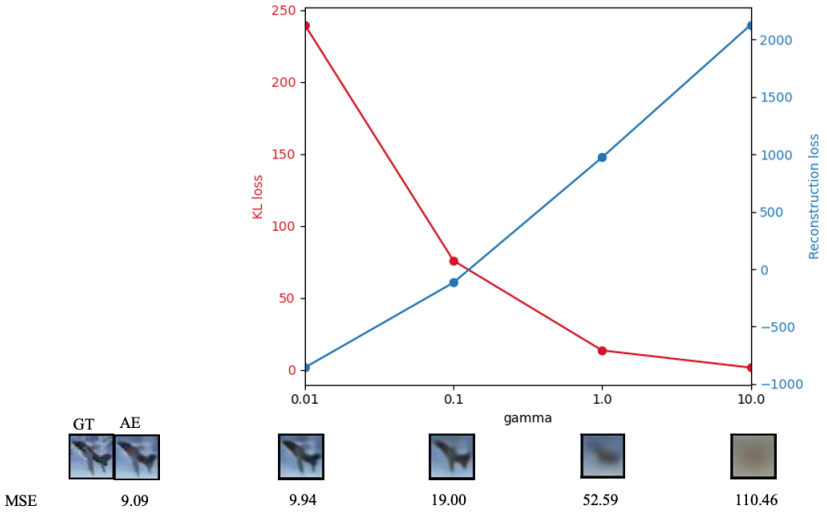
4.2. Sub-optimal objective

As demonstrated by [1], for an optimized VAE, $LL - ELBO = 0$ is possible only if $p_{\text{model}}(x)$ is allowed to be pure Gaussian.

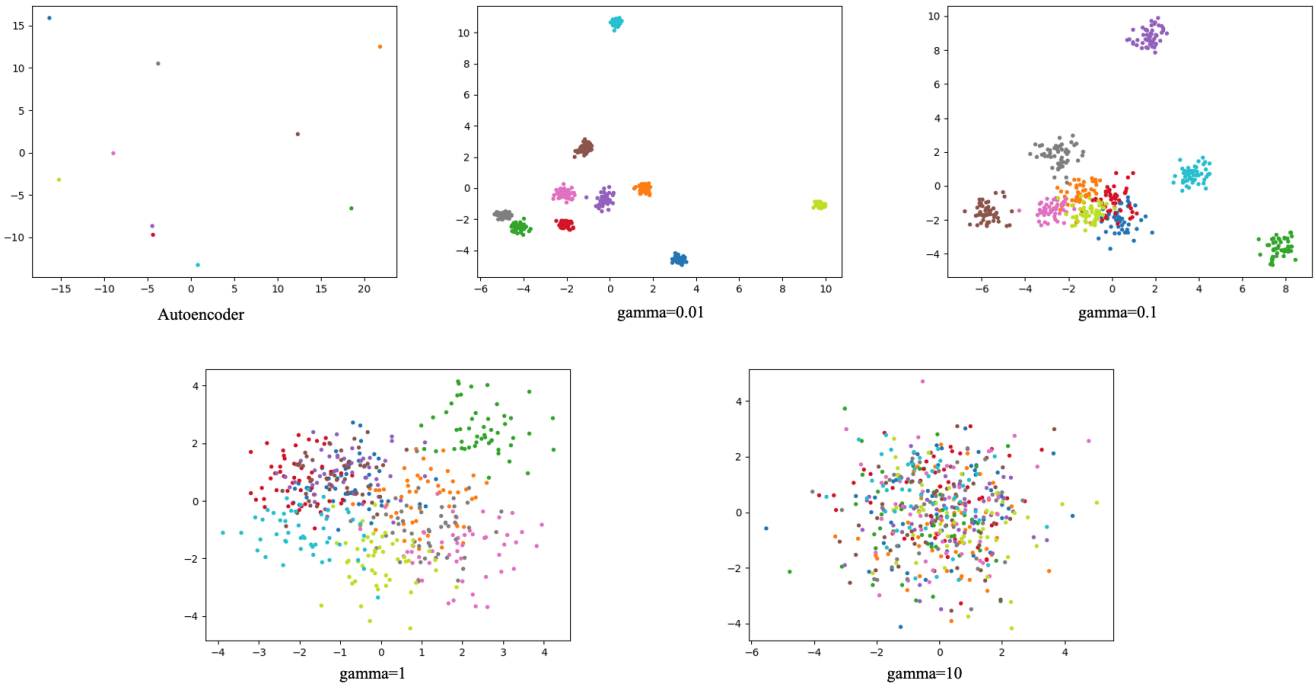
As in real world situations $p_{\text{data}}(x)$ could never be pure Gaussian, this result indicates that VAEs either have a strictly sub-optimal training objective or converge at a sub-optimal solution. Hence, by their design choice VAEs could not achieve the theoretically optimal performance where $p_{\text{model}} = p_{\text{data}}$.

5. Method

These intrinsic issues in the VAE design motivate us to look for a more unified VAE framework which could, at least in theory, solve these problems all at once. We attempt by introducing an intermediate feature manifold between the data manifold and the feature manifold as in Fig 2b,



(a) The two losses of each VAE are plotted at top and the reconstruction quality is exhibited at bottom.



(b) Two-dimensional visualization of $\sum_{i=1}^{10} q_\phi(z|x^{(i)})$, the aggregated posterior distribution of z given 10 samples of data x. This figure roughly shows the shape of the feature manifold.

Figure 3: In ELBO (2), $p_\theta(x|z) = N(\mu_{x|z}, \gamma I)$, and therefore the reconstruction loss is

$$\frac{1}{2\gamma}\mathbb{E}_{x\sim p_{\text{data}}}\left[\mathbb{E}_{z\sim q_\phi(z|x)}\left[\|x-\mu_{x|z}\|_2^2\right]\right] + \text{const},$$

where $\frac{1}{2\gamma}$ could be regarded as the weight of the reconstruction loss. Normally γ is fixed at $\frac{1}{2}$. We train a set of VAEs with different fixed values of γ on the CIFAR-10 dataset [15]. From left to right (and top to bottom), γ increases from 0 (i.e. the autoencoder) to 10, equivalent to increasing the weight of the KL loss or the structural constraint in the softly-constrained optimization analogy.

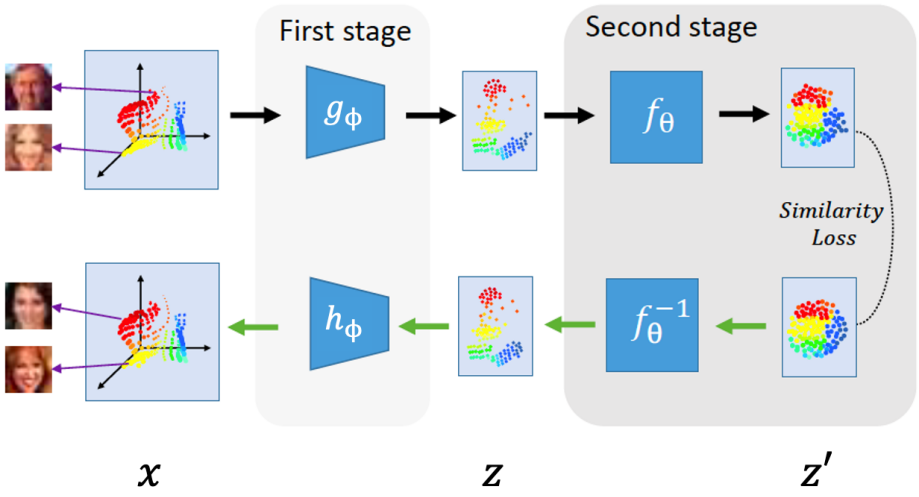


Figure 4: The pipeline of our model. The generation process is indicated by green arrows.

separating the data-feature mapping into two parts. We will argue that this separation could potentially eliminate the undesirable disentanglement between reconstruction and sampling by leaving the former at the first stage and the latter at the second stage, and the probabilistic encoder and decoder are also transformed to deterministic ones for an optimal training objective.

5.1. Model design

As shown in Fig 4, our model adopts a deterministic autoencoder and cascades it with a strictly invertible network as in 2.3, which entails from conventional VAE frameworks the encoder-decoder structure for data reconstruction and the strucural constraint on the feature manifold for potential feature sampling. Our main modification focuses on decoupled training of two stages: the deterministic autoencoder (g_ϕ for the encoder and h_ϕ for the decoder) is optimized using original data x at the first stage, producing respective intermediate features z, and at the second stage, given these features z as the input, we separately train a strictly invertible network f_θ to map them to their dual forms z' in the feature manifold which approximates the normal distribution; during generation, we sample potential features z' from the normal distribution, feeding it to f_θ^{-1}, i.e. the inverse of the invertible mapping f_θ, and again feeding the resulting potential intermediate feature z into the decoder h_ϕ at the first stage.

The training objective of the first stage is to minimize the mean square error (MSE) loss

$$\frac{1}{N}\sum_{i=1}^{N}\left\|x^{(i)}-h_\phi\circ g_\phi\left(x^{(i)}\right)\right\|_2^2, \tag{3}$$

where $x^{(i)}$ denotes the i-th training sample and N is the size of the dataset. For the second stage, the objective becomes the maximization of

$$\frac{1}{N}\sum_{i=1}^{N}\left[\log p_{z'}\left(f_\theta\circ g_\phi\left(x^{(i)}\right)\right)+\log\left|\frac{\partial f_\theta}{\partial z}\right|_{z=g_\phi\left(x^{(i)}\right)}\right], \tag{4}$$

where $p_{z'}$ denotes the normal distribution and $|\partial f_\theta/\partial z|$ denotes the Jacobian determinant. This loss encourages the invertible mapping f_θ to have a normal-distributed codomain. The rationality of this design is further justified in Section 5.3.

5.2. Informative and interpretable features

In this section, we would discuss the advantage of two independent training stages. The first stage performs reconstruction without any structural constraint, which is analogous to the unconstrained optimization problem, while the second stage takes the responsibility of accurate sampling, free of any pressure from the reconstruction process. Note that although both stages take part in the reconstruction process, the strict reversibility of the second stage ensures that the reconstruction quality would be solely determined by the first stage. The discussion in Section 4.1 shows the advantage of this separation, where both processes could potentially achieve their best performances and the learnt features z are both informative and interpretable.

5.3. Optimal objective

Additionally, we would show that our model adopts a theoretically optimal training objective. Under the assumption that the MSE loss (3) makes the autoencoder to fully capture the latent manifold at the first stage, i.e. $h_\phi^{-1}=g_\phi$, our second-stage loss (4) could be mathematically derived from LL (1). First, denoting the generative distribution

Ours	VAE	2-stage VAE
115.22	98.34	86.95

Table 1: Quantitative comparison of different models' generation results. Results are compared by Fréchet inception distance (FID)[10], where lower FID indicates better generation quality.

of the intermediate feature z as p_z, from the generation pipeline in Fig 4 we have

$$\begin{aligned} LL =& \frac{1}{N}\sum_{i=1}^{N}\log p_{\text{model}}\left(x^{(i)}\right) \\ =& \frac{1}{N}\sum_{i=1}^{N}\log p_z\left(h_\phi^{-1}\left(x^{(i)}\right)\right), \end{aligned} \tag{5}$$

omitting constant terms. Next, combine the optimality assumption $h_\phi^{-1} = g_\phi$ and the probability density transformation formula $p_z(z) = p_{z'}(z')\left|\partial z'/\partial z\right|$ into (5), we could obtain the desired objective as (4)

$$\frac{1}{N}\sum_{i=1}^{N}\left[\log p_{z'}\left(f_\theta \circ g_\phi\left(x^{(i)}\right)\right) + \log\left|\frac{\partial f_\theta}{\partial z}\right|_{z=g_\phi\left(x^{(i)}\right)}\right].$$

Hence, the combination of our two losses equivalents to the optimal goal of generative models, LL.

6. Experiment

In this section empirical results of our model are presented to support the claims in Section 5.2 and Section 5.3. We adopt the vanilla VAE [20] as the baseline, and the 2-stage VAE [1] as the state-of-the-art model, since to our best knowledge this is the first-rate VAE model which reserves the representation learning function and focuses on theoretical enhancement.

6.1. Software, datasets and equipment

We use Gitlab/Github for programming collaboration. Python and PyTorch are the main coding languages. Our model is deployed on *CelebAMask-HQ* dataset (2020)[16], which generates 30,000 high-resolution face images from *CelebA*[17]. The training and testing process will be implemented on *MLCV1* Server with 8 Nvidia GeForce RTX 2080Ti GPU, 11GB memory each.

6.2. Informativeness of latent features

As in Section 4.1, we will measure the informativeness of learnt features by comparing the reconstruction quality. Fig 5 shows that our method could recover detailed information, e.g. eye glasses or hats, and also achieve lowest MSE, proving our superiority over the opponents.

6.3. Interpretability of latent features

As in Section 3.2, the interpretability of learnt features could be evaluated by the model's disentangling performance, which is shown in Fig 6. It could be observed that our algorithm is able to automatically discover independent generative factors while other solutions could only exhibit entangled attributes. Hence, our method produces most interpretable features.

6.4. Generative quality

At last, the generation results are qualitatively shown in Fig 7, and quantitatively evaluated in Table 1. The overall performance of our solution in terms of data generation is only better than the baseline (improved visual results and slightly worse FID score), which is somehow understandable as we do not have sufficient time to fully investigate the existing model architectures for optimizing our model design.

7. Conclusion and discussion

First, we review relevant literature and summarize three main challenges of VAEs which could not be addressed simultaneously before our paper. Then we diagnose these symptoms and conclude that the rivalry between the reconstruction loss and the KL loss as well as the design choice of VAEs should take the blame. This conclusion is leveraged to explore a unified theoretical enhancement of the VAE family to solve the issues all at once, where an intermediate feature manifold is introduced to decouple reconstruction and sampling processes. It is shown empirically that the resulting model produces informative and interpretable features, while exhibiting generative quality barely above the baseline (potential room for improvement in model architecture). Although the theoretical expectations are not fully fulfilled, this work serves as a wedge to various future researches and applications, such as exploration of a more expressive invertible network for the second stage to further improve the generative quality, unsupervised representation learning based on generative models, downstream tasks leveraging the produced features.

References

[1] Dai B and Wipf D. Diagnosing and enhancing VAE models. In *Int. Conf. Learn. Represent.*, 2019.

[2] Jens Behrmann and et al. Invertible residual networks. In *Proc. the 36th Int. Conf. Mach. Learn.* PMLR, 2019.

[3] Tal Daniel and Aviv Tamar. Soft-introvae: Analyzing and improving the introspective variational autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4391–4400. Computer Vision Foundation / IEEE, 2021.

[4] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua

Figure 5: Comparison of different models' reconstruction quality.

Figure 6: Qualitative comparison of different algorithms' disentangling performance. In theory, each dimension of an interpretable feature should represent an independent generative factor, e.g. skin color or azimuth. To validate this theoretical expectation, each row sets a generative factor and selects a dimension which mostly corresponds to it. Each sub-figure presents the traversal along the selected dimension while keeping all other dimensions fixed. From the results, our method outperforms others in disentangling performance, e.g. at the last row our model only changes the azimuth of the face as expected, while the other two additionally change the identity, hair style, etc.

Figure 7: Qualitative comparison of different models' generation results.

Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[5] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[6] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[8] Junxian He and et al. Lagging inference networks and posterior collapse in variational autoencoders. In *7th Int. Conf. Learn. Represent., ICLR*, 2019.

[9] Kaiming He and et al. Deep residual learning for image recognition. In *2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. IEEE Computer Society, jun 2016.

[10] 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. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017.

[11] Irina Higgins and et al. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th Int. Conf. Learn. Represent, ICLR*, 2017.

[12] Chin-Wei Huang and et al. Improving explorability in variational inference with annealed variational objectives. In *Adv. Neural Inf. Process. Syst.*, 2018.

[13] Jörn-Henrik Jacobsen, Arnold W. M. Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. *CoRR*, abs/1802.07088, 2018.

[14] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10236–10245, 2018.

[15] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[18] James Lucas and et al. Understanding posterior collapse in generative latent variable models. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop*, 2019.

[19] A. V. D. Oord and et al. Neural discrete representation learning. In *Adv. Neural Inf. Process. Syst.*, volume 30, 2017.

[20] Kingma D P and Welling M. Auto-Encoding Variational Bayes. In *2nd Int. Conf. Learn. Represent., ICLR*, 2014.

[21] Ali Razavi and et al. Preventing posterior collapse with delta-vaes. In *7th Int. Conf. Learn. Represent., ICLR*. OpenReview.net, 2019.

[22] Huajie Shao and et al. Controlvae: Controllable variational autoencoder. *Proc. the 37th Int. Conf. Mach. Learn.*, 2020.