Generating High-Fidelity Images with Disentangled Adversarial VAEs and Structure-Aware Loss

Habibeh Naderi[1], Behrouz Haji Soleimani[1,2], Stan Matwin[1,3]
[1]*Institute for Big Data Analytics, Faculty of Computer Science, Dalhousie University, Halifax, Canada*
[2]*Kinaxis Inc., Ottawa, Canada*
[3]*Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland*
Email(s): habibeh.naderi@dal.ca, bhajisoleimani@kinaxis.com, stan@cs.dal.ca

***Abstract*—While variational autoencoders (VAE) provide the theoretical basis for deep generative models, they often produce "*blurry*" images which is linked to their training objective. In this paper, we propose the "Sharpened Adversarial Variational Auto-Encoder" (AVAE-S) which uses an adversarial training mechanism to fine-tune the learned latent code vector of the VAE with a specialized objective function. The loss function is designed to uncover global structure as well as the local and high frequency features in VAE and leading to the smaller variance in the aggregated posterior and hence, reducing the blurriness of their generated samples. AVAE-S leverages the learned representations to the meaningful latent features by enforcing feature consistency between the model distribution and the target distribution leading to the sharpened output with better perceptual quality. Then, AVAE-S starts training a GAN network, which generator has been collapsed on the VAE's decoder, upon that learned latent code vector. Moreover, we augment the standard VAE's evidence lower bound objective function with other element-wise similarity measures. Our experiments show that AVAE-S achieves the state-of-the-art sample quality in the common MNIST and CelebA datasets. AVAE-S shares many of the good properties of the VAE (stable training, encoder-decoder architecture, nice latent manifold structure) while generating more realistic images, as measured by the sharpness score.**

***Index Terms*—variational autoencoders, adversarial training, information bottleneck, constrained optimization**

I. Introduction

Learning effective representations without the supervision that can capture all the variability in the true data distribution remains a key challenge in machine learning. Generative models adopt probabilistic approaches to learn the low-dimensional manifold that data is assumed to live on and generate new data by sampling from that latent space. By learning a generative model of the data with the appropriate hierarchical structure of latent variables, it is hoped that the model will somehow identify and disentangle the underlying causal sources of variations in the data. In particular, variational autoencoders (VAEs) [14], [23], [25] constitute a theoretically well-founded probabilistic approach to model high-dimensional distributions. VAEs use a prediction network to predict the posterior distribution over the latent variables while encouraging it to follow a fixed prior distribution. Even though VAEs provide an elegant way to learn low-dimensional code vector via performing variational inference, they tend to generate blurry samples. This has been attributed to (1) the restrictiveness of the Gaussian encoder/decoder assumption [5], (2) the use of relatively simple distributions for the prior in the hope that the interactions between high level features are disentangled, and can be well approximated with a Gaussian or uniform distribution [27], or (3) the over-regularization induced by the KL divergence term in the VAE objective function [26]. In addition to the inappropriate choice of the inference distribution, the original VAE objective function has tendency to generate blurry samples by admitting the trivial solutions that decouple the latent space from input data [4], [30]. This makes the latent code completely non-informative leading to the "posterior collapse" phenomenon where the latents are ignored when they are paired with powerful decoders [28]. Moreover, the learned aggregated posterior rarely matches the assumed latent prior in practice leading to fuzziness in the generated samples [1], [5], [15].

Generative adversarial networks (GANs) [9] are another frameworks of choice for generative modeling, which use an adversarial training procedure and generate more impressive images in terms of visual fidelity. However, GANs are harder to train in comparison with VAEs as they come without the encoder and suffer from the inherent saddle point optimization problem, known as "mode collapse", when the resulting model fails to cover all the variability in the input data distribution. There has been a lot of effort in investigating various configurations of GANs as well as combinations of the VAEs and GANs. For instance, adversarial autoencoder (AAE) [20] imposed an arbitrary prior on the latent representation of the autoencoder using GANs framework. Similarly, Wasserstein autoencoder (WAE) [26] proposed the more general regularization technique for VAE by training a discriminator over its latent space to penalize the discrepancy between the encoded distribution and the target distribution. These methods mostly offer different regularization approaches to VAE as a substitute of KL divergence term in the standard VAE objective. AVAE-S is different from these techniques as they perform variational inference through direct regularization of the matching distance between the aggregated posterior of the latent code vector and the prior distribution. However, AVAE-S modifies the smooth latent space implicitly by adding an adversarial term to the VAE loss to not only minimizes the discrepancy between the input image and the reconstructed image, but also compensates the over-regularization effect of the KL divergence.

Since VAEs provide the theoretically well-established basis

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

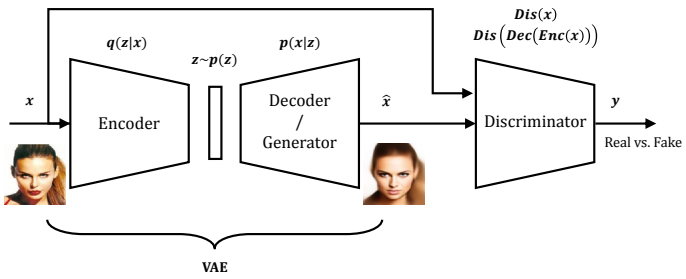


Fig. 1: Adversarial Variational Autoencoder model architecture

for generative modelings with more stable training than GANs and more efficient sampling mechanisms than autoregressive models [11], [16], [21], they become the most promising framework for image generation [12], [19]. Many research studies have been developed to devise new encoder-decoder architectures [13], [15] along with deriving different formulation for the original VAEs' objective function to resolve the blurriness issue of their generated samples. Regularized autoencoder (RAE) [8], substitute the stochastic VAE frameworks with simple deterministic autoencoder with explicit regularization term to enforce the smooth latent space. They relax the constraint on the VAE's posterior to conform with the given prior and replace it with post inference density estimation to generate less blurry samples.

In this work, we propose "adversarial variational autoencoder" (AVAE-S) to tackle VAEs' shortcomings in generating sharp samples while preserving the KL divergence term in its objective. In fact, we employ an adversarial mechanism to fine-tune the pre-trained VAE framework. Hence, we combine the best of VAEs as a method with stable training and nice latent manifold structure, and GAN as a high quality generative model in a unified framework. It is worth mentioning that the pre-trained VAE enables us to start GAN's generator training on a meaningful code vector instead of random vector. We augment the element-wise VAEs objective with more powerful similarity measures that can model the properties of human visual perception such as structural similarity index metric (ssim) [24] to learn shape and edge structure more explicitly. Then, we combine this reinforced VAE loss with the higher-level feature-wise similarity metric expressed in the GAN's discriminator. AVAE-S leverages the learned representations to the meaningful latent features by uncovering the global structure of the data distribution as well as the local and high frequency sources of variations leading to smaller variance on the aggregated posterior which in turn decreases the blurriness of samples. Figure 1 illustrates the high-level architecture of our proposed adversarial VAE model.

Although AVAE-S shares the similar architecture with VAE/GAN model as proposed in [17], the two models first, have different objective functions, and second, have been trained differently. AVAE-S generates sharper images and allows sample generation via interpolations while exploring through the latent space. In AVAE-S, the latent space captures all the informative modes of the target data distribution since the reconstruction loss, generation loss, and discriminator's error, all are back-propagated to the VAE's encoder leading to the smoother latent space where the similar data points are mapped to similar latent code vectors, and small variations in latent space lead to reconstructions by decoder that vary only slightly. However, in VAE/GAN the reconstruction loss does not back-propagate to the encoder makes the latent space somehow independent from the input data. Furthermore, in AVAE-S, we first pre-train the VAE and then start to train the whole VAE combined GAN architecture. Thus, GAN is trained over the learned latent code vector while in VAE/GAN model, they train GAN firstly over random code vector. Additionally, in AVAE-S, the discriminator has to discriminate between the actual input image (real class) and its reconstruction which is the output of the decoder (fake class) whereas, in VAE/GAN model, the discriminator also has been fed with the third input which is the generated samples.

In summary, our contributions are as follows:

- We propose AVAE-S framework as an unsupervised generative model with an adversarial training mechanism to fine-tune the learned latent code vector of the VAE with a specialized objective function.
- We enhance the standard VAEs objective with both element-wise and feature-wise similarity measures that greatly improve sample quality for the VAEs. This customized objective includes (1) an extensive reconstruction loss which captures discrepancies in terms of pixel-wise information, structural dissimilarity based on SSIM metric, edge information calculated using Sobel kernel, and texture information extracted by Gabor wavelet, (2) KL divergence loss between posterior and prior distributions (3) total correlation (TC) KL loss [29] to enforce statistical independence between latent dimensions, and (4) classification error of the discriminator.
- Our interpolations quality demonstrate that AVAE-S learns the smooth, disentangled latent embedding for the true data distribution which successfully represents its factors of variation.
- Our empirical evaluations achieve the state-of-the-art sharpness scores for reconstruction and random samples generation on common image datasets of MNIST and CelebA.

II. Proposed Method

The key reason why VAEs generate blurry samples is ingrained in their limited discriminatory power to learn different representations for different input data. Mapping different data points to the same spot in the latent space, makes the posterior too complex to be approximated with the simple prior. In consequence, VAEs are less faithful to the true data distribution while performing inference in reconstruction and sample generation and their reconstruction error is connected to the mutual information between the latent variables and the data variables [2] (Equation (9)). Furthermore, to learn a

disentangled representation, we need to minimize the intra-latent mutual information between the latent variables [3]. GANs, on the other hand, can measure the similarity between different samples, effectively. These observations motivate us to combine the VAEs and GAN in an unified architecture to provide VAEs with more discriminative power to improve its reconstruction quality. In this section, we describe our adversarial variational autoencoder framework and discuss how it resolves the blurriness issue associated with VAEs' generated samples.

VAE basically is a pair of encoder-decoder where the encoder network encodes a data sample x to a latent vector z and the decoder network maps the latent vector back to the input space:

$$Enc_\phi(x) = z \sim q_\phi(z|x) \ , \quad Dec_\theta(z) = x \sim p_\theta(x|z) \tag{1}$$

where $q_\phi(z|x)$ is a posterior distribution, $p_\theta(x|z)$ is a likelihood, and ϕ,θ are the parameters of VAE's encoder and decoder, respectively. VAE estimates the true data distribution $p_{data}(x)$ as the infinite mixture model $p_\theta(x) = \int p_\theta(x|z)p(z)dz$. Computing this marginal log-likelihood $\log p_\theta(x)$ is generally intractable. Hence, we follow a variational approach, maximizing the evidence lower bound (ELBO) for a sample x:

$$\log p_\theta(x) \geq \text{ELBO}(\phi, \theta, x) = \mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] - D_{KL}(q_\phi(z|x)\|p(z)) \tag{2}$$

where D_{KL} is the Kullback-Leibler divergence which acts as a regularizer during training. Maximizing the ELBO in (2) over data $\mathcal{X}$ w.r.t the model parameters ϕ and θ corresponds to minimizing the loss:

$$\arg\min_{\phi,\theta} \mathbb{E}_{x\sim p_{data}} \mathcal{L}_{\text{ELBO}} = \mathbb{E}_{x\sim p_{data}} \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} \tag{3}$$

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

$$\begin{aligned} \mathcal{L}_{\text{VAE}} &= \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} \\ \mathcal{L}_{\text{REC}} &= \neg\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] \\ \mathcal{L}_{\text{KL}} &= D_{KL}(q_\phi(z|x)\|p(z)) \end{aligned} \tag{4}$$

Hence, the VAE loss, $\mathcal{L}_{\text{VAE}}$, is equal to the sum of the reconstruction loss, $\mathcal{L}_{\text{REC}}$, and the KL divergence term, $\mathcal{L}_{\text{KL}}$. More specifically, VAE aims to reduce the discrepancy between the input sample x and its reconstruction $\tilde{x} = Dec_\theta(Enc_\phi(x))$ which in traditional VAE has been calculated element-wise using the L2 loss [6]. VAE simultaneously encourages the posterior $q_\phi(z|x)$ to match the prior $p(z)$ which is typically assumed to be a standard Gaussian prior $p(z) \sim \mathcal{N}(0, I)$.

We adapt two strategies to tackle the blurriness problem in VAEs' generated samples: (1) incorporating the feature-wise and more comprehensive element-wise similarity measures into our VAE's reconstruction loss in addition to the dimension-wise total correlation (TC) KL divergence term as a measure of dependence between latent variables, and (2) fine-tuning the VAE model with an adversarial discriminator.

A. Sharpness-Focused Specialized $\mathcal{L}_{REC}$ for VAE

Since the quality of the generated samples of VAEs is directly reflected into their reconstruction error and commonly associated with VAEs element-wise similarity measure, we add more comprehensive perceptual similarity metrics to our VAE reconstruction loss function to enforce it to generate more realistic samples. The ultimate reconstruction loss will be the combination of the following sharpness enhancement metrics:

$$\begin{aligned} \mathcal{L}_{\text{REC}} = \lambda_1\mathcal{L}_{\text{MAE}} + \lambda_2\mathcal{L}_{\text{MSE}} \\ + \lambda_3\mathcal{L}_{\text{1-SSIM}} + \lambda_4\mathcal{L}_{\text{Sobel}} + \lambda_5\mathcal{L}_{\text{Gabor}} \end{aligned} \tag{5}$$

where λ_i for $i = 1, ..., 5$ are the corresponding weights to these similarity metrics and have been calculated empirically in a way that all of the losses stay in the same scale, so that the criterion with the smaller value has not been ignored. The L_1 and L_2 losses are sensitive to the pixel-wise discrepancies between the input sample x and its reconstruction $\tilde{x} = Dec_\theta(Enc_\phi(x))$. The $\mathcal{L}_{\text{Sobel}}$ captures the edge information discrepancies as follows:

$$\mathcal{L}_{\text{Sobel}} = \|E_{Sobel}(x^{(i)}) - E_{Sobel}(\tilde{x}^{(i)})\|_2 \tag{6}$$

The $\mathcal{L}_{\text{Gabor}}$ provides the difference between the texture information of the input sample x and its reconstruction $\tilde{x}$.

$$\mathcal{L}_{\text{Gabor}} = \|T_{Gabor}(x^{(i)}) - T_{Gabor}(\tilde{x}^{(i)})\|_2 \tag{7}$$

The $\mathcal{L}_{\text{1-SSIM}}$ measures the structural dissimilarity (DSSIM) index between the input sample x and its reconstruction $\tilde{x}$. SSIM index characterizes as a metric to provide the human perceptual quality of the images and is defined based on comparing every window in input image x with its corresponding window in the reconstructed image $\tilde{x}$ regarding luminance (l), contrast (c), and structure (s).

$$\text{SSIM}(x, y) = \big[l(x, y)^\alpha.c(x, y)^\beta.s(x, y)^\gamma\big] \tag{8}$$

where α, β, and γ are the associated weights to these three comparative measures.

As we discussed earlier in this section, the reconstruction error in VAEs is also connected to the mutual information between observed and latent variables [2], [18]. Since the mutual information is characterized as the KL divergence between the joint probability distribution and the product of the marginals (i.e., $I(X; Z) = D_{KL}(P_{XZ}\|P_X \bigotimes P_Z)$), the reconstruction error, $\mathcal{L}_{\text{REC}}$, is bounded by (9) [2].

$$\mathcal{L}_{\text{REC}} \leq D_{KL}(q(x, z)\|p(z, x)) - I_q(x, z) + H_q(z) \tag{9}$$

where the two joint distributions of $q(x, z) = q(z|x)q(x)$ and $p(z, x) = p(x|z)p(z)$ are induced by the encoder and decoder models parameterized by q_ϕ and p_θ , respectively. H is the Shannon entropy and the $I_q(x, z)$ is the mutual information between the latent variables and the data variables. The disentanglement is not directly related to blurriness. However, since $I_q(x, z) = H_q(z) - H_q(z|x)$, the equation (9) can be rewritten as $\mathcal{L}_{\text{REC}} \leq D_{KL}(q(x, z)\|p(z, x)) + H_q(z|x)$. Therefore, minimizing the conditional entropy of the input samples given the latent representations leads to minimizing

the upper bound of the reconstruction error (similar to ELBO that implicitly maximizes the lower bound of the likelihood). Hence, if the joint distributions are matched, $H_q(z)$ tends to $H_p(z)$, which is fixed as long as the prior, $p(z)$, is itself fixed. Subsequently, maximizing the mutual information minimizes the expected reconstruction error. In other words, the disentanglement implicitly helps in reducing the reconstruction error and it can decrease blurriness as well.

Similarly, to enforce the model to find disentangled and statistically independent factors of variation in the data distribution, we consider another KL term among the latent variables as defined in (10). This means that each latent dimension is sensitive to changes in one factor of variation such as pose (azimuth and elevation), lighting condition, and attributes of the face such as skin tone, gender, face width, etc. and relatively invariant to changes in other.

$$\mathcal{L}_{\text{KL-TC}} = D_{KL}(q(z)\|\prod_j q(z_j)) \tag{10}$$

B. Adversarial Training of VAEs

We first pre-train the VAE model and then, combine it with an adversarial discriminator which is trained to discriminate between the original input images x as real class samples and their reconstructed images $\tilde{x} = Dec_\theta(Enc_\phi(x))$ as fake class samples. The discriminator fine-tunes the learned latent code vector of the VAE and leverages the learned representations to the meaningful latent features. More specifically, in our AVAE-S model, we have a dual-purposed decoder network which responsible for the input images reconstruction and the random samples generation of realistic images. Hence, AVAE-S objective function can be defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{AVAE}} &= \mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{Dis}} \\ \mathcal{L}_{\text{VAE}} &= \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{KL-TC}} \\ \mathcal{L}_{\text{Dis}} &= \log\left(Dis(x)\right) + \log\left(1 - Dis(Dec(Enc(x)))\right) \end{aligned} \tag{11}$$

where the first term in (11), $\mathcal{L}_{\text{VAE}}$, equals to sum of the VAE's reconstruction loss, $\mathcal{L}_{\text{REC}}$, KL divergence loss between the posterior and prior, $\mathcal{L}_{\text{KL}}$, and the total correlation KL loss between latent variables, $\mathcal{L}_{\text{KL-TC}}$. The $\mathcal{L}_{\text{Dis}}$ shows the discriminator's binary cross entropy loss for a sample x which calculated based on the GANs objective.

GAN consist of two neural networks: a generator network, $Gen(z)$, that maps the latent samples z from the prior $p(z)$ to the data space and a discriminator network, $Dis(x)$, that classifies real versus fake inputs. In our AVAE-S framework, the decoder acts as a generator as well, $Gen(z) = Dec(z) = Dec(Enc(x))$, and tries to makes the reconstructed image similar to the input image as much as possible to confuse the discriminator into believing that the reconstructed images come from the actual data distribution. The objective of GAN is to find the binary classifier that gives the best possible discrimination between true and generated (reconstructed images in our model) data while simultaneously encouraging the $Gen(z)$ to fit the true data distribution. Hence, this problem is usually formulated as a min/max optimization:

$$\begin{aligned} \min_{Gen}\max_{Dis} \mathbb{E}_{x\sim p_{data}}\big[\log Dis(x)\big] \\ +\mathbb{E}_{z\sim p(z)}\big[\log\left(1 - Dis(Gen(z))\right)\big] \end{aligned} \tag{12}$$

Algorithm 1 Sharpened Adversarial Variational Autoencoder

Input: $\mathbf{X}, d, k_{max} = 50, l = 3, Iter = 250$
Output: Encoder, Decoder, Discriminator
$\theta_{VAE}, \theta_{Dis} \leftarrow$ initialize network parameters
for $i = 1$ **to** $epoch\_pretrain$ **do**
  $X \leftarrow$ random mini-batch from dataset
  $Z \leftarrow Encode(X)$
  $\tilde{X} \leftarrow Decode(Z)$
  $\mathcal{L}_{\text{KL}} \leftarrow D_{KL}(q_\phi(Z|X)\|p(Z))$
  $\mathcal{L}_{\text{KL-TC}} = D_{KL}(q(z)\|\prod_j q(z_j))$
  $\mathcal{L}_{\text{MAE}} \leftarrow \|X - \tilde{X}\|_1, \quad \mathcal{L}_{\text{MSE}} \leftarrow \|X - \tilde{X}\|_2$
  $\mathcal{L}_{\text{1-SSIM}} \leftarrow 1 - SSIM(X, \tilde{X})$
  $\mathcal{L}_{\text{Sobel}} \leftarrow \|E_{Sobel}(X) - E_{Sobel}(\tilde{X}\|_2$
  $\mathcal{L}_{\text{Gabor}} \leftarrow \|T_{Gabor}(X) - T_{Gabor}(\tilde{X})\|_2$
  $\mathcal{L}_{\text{REC}} \leftarrow \lambda_1\mathcal{L}_{\text{MAE}} + \lambda_2\mathcal{L}_{\text{MSE}} + \lambda_3\mathcal{L}_{\text{1-SSIM}} + \lambda_4\mathcal{L}_{\text{Sobel}} + \lambda_5\mathcal{L}_{\text{Gabor}}$
  $\mathcal{L}_{\text{VAE}} \leftarrow \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{KL-TC}}$
  $\theta_{VAE} \leftarrow \theta_{VAE} - \nabla_{\theta_{VAE}}\mathcal{L}_{\text{VAE}}$
end for
for $i = 1$ **to** $epoch\_adversarial$ **do**
  $X \leftarrow$ random mini-batch from dataset
  $Z \leftarrow Encode(X)$
  $\tilde{X} \leftarrow Decode(Z)$
  $\mathcal{L}_{\text{REC}} \leftarrow \lambda_1\mathcal{L}_{\text{MAE}} + \lambda_2\mathcal{L}_{\text{MSE}} + \lambda_3\mathcal{L}_{\text{SSIM}} + \lambda_4\mathcal{L}_{\text{Sobel}} + \lambda_5\mathcal{L}_{\text{Gabor}}$
  $\mathcal{L}_{\text{VAE}} \leftarrow \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{KL-TC}}$
  if $Accuracy_{Dis} < 0.7$ **then**
    $\mathcal{L}_{\text{Dis}} \leftarrow \log\left(Dis(X)\right) + \log\left(1 - Dis(\tilde{X})\right)$
    $\theta_{Dis} \leftarrow \theta_{Dis} - \nabla_{\theta_{Dis}}\mathcal{L}_{\text{Dis}}$
  end if
  if $Accuracy_{Dis} > 0.5$ **then**
    $\mathcal{L}_{\text{AVAE}} \leftarrow \mathcal{L}_{\text{VAE}} + \log Dis(\tilde{X})$
    $\theta_{VAE} \leftarrow \theta_{VAE} - \nabla_{\theta_{VAE}}\mathcal{L}_{\text{AVAE}}$
  end if
end for

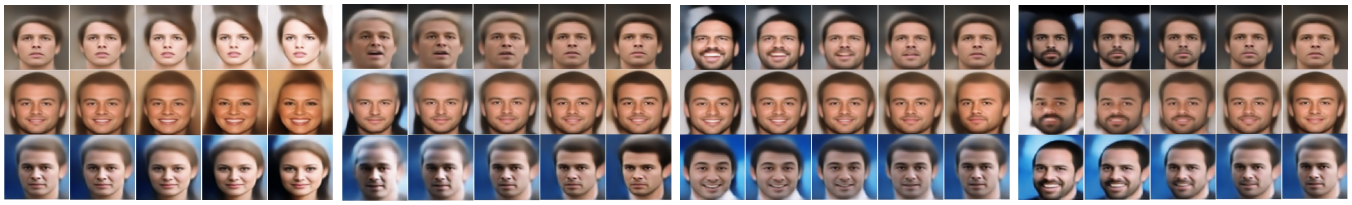
III. Experiments

In this section, we present the results of training various generative models on CelebA dataset. Measuring the quality of generative models has always been challenging as current evaluation methods are problematic for larger natural images and cannot quantify the visual appeal. For instance, the traditional log likelihood or mean squared error measures do not capture visual fidelity. In this work, we use images of size 96x96 and mostly focus on qualitative assessments of the models. In this section we investigate the performance of different generative models:

- **VAE:** This is the vanilla Variational Autoencoder (VAE) that has Mean Squared Error (MSE) reconstruction loss and a Gaussian prior Kullback-Leibler loss.
- **AVAE:** Adversarial fine-tuning of pre-trained VAE. In this case, we pre-train the vanilla VAE for 100 epochs and use adversarial training to fine-tune it in order to produce sharper and more realistic images.

TABLE I: Variational Autoencoder (VAE) architecture details.

Encoder	Decoder
24 conv. 3×3, stride 1, bnorm, ReLU	48 fully-connected, bnorm, tanh
24 conv. 4×4, stride 2, bnorm, ReLU	6×6×192 fully-connected, bnorm, tanh
48 conv. 3×3, stride 1, bnorm, ReLU	192 conv. 3×3, bnorm, ReLU, NN_upsample
48 conv. 4×4, stride 2, bnorm, ReLU	192 conv. 3×3, bnorm, ReLU
96 conv. 3×3, stride 1, bnorm, ReLU	96 conv. 3×3, bnorm, ReLU, NN_upsample
96 conv. 4×4, stride 2, bnorm, ReLU	96 conv. 3×3, bnorm, ReLU
192 conv. 3×3, stride 1, bnorm, ReLU	48 conv. 3×3, bnorm, ReLU, NN_upsample
192 conv. 4×4, stride 2, bnorm, ReLU	48 conv. 3×3, bnorm, ReLU
48 fully-connected, bnorm, tanh	24 conv. 3×3, bnorm, ReLU, NN_upsample
32 latent dim, z_mean, z_sigma	24 conv. 3×3, bnorm, ReLU
32 sampling layer, z_sample	3 conv_transpose 1×1, sigmoid



(a) Gender (b) Age (c) Smile (d) Beard

Fig. 2: Traversing in four different dimensions of the latent space of AVAE-S demonstrating the disentanglement in the learned representation. Each of the dimensions are learning unique attributes of the input distribution.

TABLE II: FID (smaller is better) and sharpness (larger is better) scores for samples of various models for CelebA (best model in bold).

	WAE	RAE	RAE-SN	Ours			
				VAE	AVAE	VAE-S	AVAE-S
Sharpness	0.006	-	-	0.0269 ± 0.0091	0.0315 ± 0.0105	0.0389 ± 0.0107	**0.0452 ± 0.0125**
FID	53.67	48.20	44.74	45.4467	44.3741	40.3900	**38.4775**

TABLE III: Quantitative evaluation scores for various generative models.

	Ground Truth	VAE	AVAE	VAE-S	AVAE-S
Sharpness (Laplace)	0.112015	0.026969	0.031597	0.038958	**0.045204**
Cosine similarity	1	0.978226	0.978675	0.979577	**0.980018**
Mean Squared Error	0	0.010207	0.009804	0.009412	**0.009189**
FID - first max pooling (d=64)	-	0.6911	0.6370	0.5650	**0.5224**
FID - second max pooling (d=192)	-	11.2447	10.1863	8.9121	**8.2571**
FID - pre-aux classifier (d=768)	-	1.0433	1.0433	**1.0367**	1.0521
FID - final average pooling (d=2048)	-	45.4467	44.3741	40.3900	**38.4775**

- **VAE-S:** Variational Autoencoder with sharpness-focused specialized loss function. This model uses the compound loss function defined in Section II that incorporates edges, texture and structural similarity in addition to the color.
- **AVAE-S:** Adversarial fine-tuning of pre-trained VAE-S. In this case, we pre-train the VAE-S model for 100 epochs and use adversarial fine-tuning in order to produce even crisper and more realistic images.

All models share the same architectures for encoder, decoder and discriminator. The architectures of encoder and decoder are explained in Table I. The discriminator is very similar to the encoder with the difference that it uses max pooling for down-sampling instead of stride 2 convolution.

One of the common ways that many researchers use for up-sampling the image in the decoder is to use transposed convolution with stride 2. However, this often generates checkerboard artifacts in the output image. For this reason, we have used nearest neighbor up-sampling in the decoder and based on our experiments it works better than transposed convolution. Similarly in the encoder, the most common way to down-sample is to use pooling layers such as max pooling and we have indeed tried it. However, based on our empirical findings the extra convolution with stride 2 extracts better features and passes a richer representation forward.

Figure 3 illustrates the qualitative evaluation of sample quality for different generative models. As we can see from the figure, sharpness-focused loss functions in VAE-S and AVAE-S lead to crisper images and better visual quality

(a) Reconstruction

(b) Generation

(c) Interpolation

Fig. 3: Qualitative evaluation of sample quality for different generative models on CelebA dataset (a) Reconstruction of randomly selected images (top row represents the ground truth images), (b) Random sampling from the latent and random generation, and (c) Linear interpolation between randomly selected images.

overall. Moreover, the adversarial fine-tuning improves the generation and reconstruction quality. Our proposed models provide overall sharper samples and reconstructions while interpolating smoothly in the latent space.

We also evaluated the models quantitatively using the sharpness score, mean squared error, and Cosine similarity. For calculating the sharpness score, we convert every image to grey-scale and convolve it with the Laplace filter $\begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, which acts as an edge detector. We then compute the variance of the resulting activations and consider it as the sharpness score for the images [26]. Table III shows the average evaluation metrics of different models on the reconstructions of the validation set. As we can see from the table, our specialized loss function for VAE improves the overall sharpness and reconstruction quality of generated images. Additionally, the adversarial fine-tuning takes it a step further to make crisp and realistic images. In terms of sharpness, our AVAE-S model outperforms state-of-the-art VAE models such as RAE [7], WAE [26], and hybrid VAE/GAN [17] networks.

To demonstrate the quantitative superiority of our proposed AVAE-S model in addition to its qualitative outstanding results over other VAE-based generative algorithms, we also used the ubiquitous Fréchet Inception Distance (FID) [10] to compare different models. Even though the obtained result are not consistent and highly sensitive to the depth of the selected activations layer in the Inception network, Table II shows that

AVAE-S has improved the VAE's FID scores significantly in all cases. None of the SOTA VAE-based generative models such as RAE, WAE, and hybrid VAE/GAN have disclosed the details of their reported FID scores calculation specifically from which activation layer they obtained those FID scores. Therefore, we calculated the FID scores correspond to four common activation layers of Inception network. Moreover, since the Frechet distance is biased to the sample size, it is important to use the same sample size to compute the FID score when comparing two generative models. Furthermore, the source codes for these algorithms are not available which makes it hard for us to reproduce their exact FID scores when implementing their algorithms. For the above reasons, we think it will not be a fair comparison to rely on Frechet distance as a criterion for generative models evaluation unless we are aware of these unknowns in the FID score calculation. We have found Laplacian filter as a more tangible metric for sharpness evaluation which better matches to the human perception of the image quality. This quality enhancement can be clearly observed in less blurry generated samples with our AVAE-S model which is also confirmed by Laplacian sharpness score.

We chose the baseline models based on the availability of their source codes, availability of results on FID and sharpness scores, as well as similarity of datasets used. The recent trend on the evaluation of generative models is to use FID score and the recent SOTA methods such as WAE and RAE papers evaluated their results on FID and both claimed to better than the VAE/GAN model. Therefore, we did not include the VAE/GAN base model. It is also not feasible to compare VAE-based generative models to pure GAN models since we cannot compute FID for GAN models as there is no notion of reconstruction. We also did not include VQ-VAE-2 [22] and similar models since they are trained on the HQ celeb dataset and the visual quality on high resolution (1024x1024) data are not comparable with that of standard Celeb-A dataset.

To learn the disentangled latent space, we trained our AVAE-S model with additional term of the total correlation (TC) KL loss. This minimizes the intra-latent mutual information between the latent variables and the interpolation results of the latent space confirms the independence of its dimensions. As shown in Figure 2, even though we have not pass any labels with the input images when training our model, it inherently decouples the sources of variations (e.g. gender, age, hair color, smile, etc.) to a great extent.

Regarding the hyper-parameter tuning of our model, we understand that exhaustive hyper-parameter tuning is not fair for comparisons and is also not practical and portable to other domains. Therefore, we did not tune the hyper-parameters exhaustively, and the only thing we did was to make all loss components to be in the same scale. This way, all components contribute almost equally in the final loss and don't get outweighed by others. For instance, if average L1 loss of pixels in 0-255 range was 30, then their corresponding L2 loss was about 900, so we determined the λ_{l1} as 30 to make them in the same scale. We probed the loss components for 1 epoch and determined the weights and used those weights in all the experiments thereafter.

IV. Conclusion

The main reason for blurriness of outputs of VAEs lies with the loss function being optimized. The pixel-wise reconstruction error measure simply cannot capture the high-level structure and global shapes and objects in the images. In this work, we combined simple yet effective methods that have been used for decades in computer vision with the recent advancements in deep neural networks to address some of the drawbacks of VAE-based generative models. This compound loss function combines element-wise and feature-wise error measures and captures local (pixel-wise) and global (structure-based) characteristics of images. Optimizing this constrained objective forces VAE to not only learn the color information in the pixels, but also to pay attention to the texture, edge information as well as objects and structural similarity of images.

Additionally, we propose new training strategies and optimizing the latent representation through disentanglement. Using an adversarial training framework, we fine-tune the VAEs to create even sharper images by improving the latent representation. Our adversarial loss takes into account both the reconstruction error and the classification error of the discriminator. We showed that the adversarial fine-tuning procedure can improve the latent and consequently the sampling in VAEs. The main advantage of our proposed architecture over GANs is that we do not start from a random latent space and we can simply pre-train the latent code vectors in the VAE part. This pre-training of latent stabilizes the adversarial training part and avoids mode-collapse and other challenges in training GANs. This way our model combines the benefits of both VAEs and GANs. It has stable training, encoder-decoder architecture, smooth latent representation, sharp image generation, without suffering from limitations and challenges of training GANs.

Moreover, by reducing the mutual information between latent variables, we learned a disentangled representations of data which improved the inference and reconstructions in adversarially-learned VAE. Our experimental results confirm that these incremental enhancements altogether helped in achieving more realistic outputs with better visual quality. In particular, result on CelebA dataset shows remarkable improvements in terms of sharpness of generated or reconstructed images compared to all existing forms of VAEs.

References

[1] Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. *arXiv preprint arXiv:1810.11428*, 2018.

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 531–540, 2018.

[3] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[4] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

[5] Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

[6] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, pages 658–666, 2016.

[7] Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *CoRR*, abs/1903.12436, 2019.

[8] Partha Ghosh, Mehdi SM Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436*, 2019.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[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 *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[11] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P. Xing. On unifying deep generative models. In *International Conference on Learning Representations*, 2018.

[12] Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. In *Advances in Neural Information Processing Systems*, pages 52–63, 2018.

[13] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364*, 2018.

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

[16] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 29–37, 2011.

[17] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[18] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, pages 5495–5503, 2017.

[19] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 700–709, 2018.

[20] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *ICLR (2016)*, 2015.

[21] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2391–2400. JMLR. org, 2017.

[22] 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*, pages 14837–14847, 2019.

[23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[24] Karl Ridgeway, Jake Snell, Brett Roads, Richard S Zemel, and Michael C Mozer. Learning to generate images with perceptual similarity metrics. *arXiv preprint arXiv:1511.06409*, 2015.

[25] Mihaela Rosca, Balaji Lakshminarayanan, and Shakir Mohamed. Distribution matching in variational inference. *arXiv preprint arXiv:1802.06847*, 2018.

[26] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

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

[28] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

[29] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

[30] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.