AS-IntroVAE: Adversarial Similarity Distance Makes Robust IntroVAE

Abstract

Recently, introspective models like IntroVAE and S-IntroVAE have excelled in image generation and reconstruction tasks. The principal characteristic of introspective models is the adversarial learning of VAE, where the encoder attempts to distinguish between the real and the fake (i.e., synthesized) images. However, due to the unavailability of an effective metric to evaluate the difference between the real and the fake images, the posterior collapse and the vanishing gradient problem still exist, reducing the fidelity of the synthesized image. In this paper, we propose a new variation of IntroVAE called Adversarial Similarity Distance Variational Autoencoder (AS-IntroVAE). We theoretically analyze the vanishing gradient problem and construct a new Adversarial Similarity Distance (AS-Distance) using the 2-Wasserstein distance and the kernel trick. With weight annealing on AS-Distance and KL-Divergence, the AS-IntroVAE are able to generate stable and high-quality images. The posterior collapse problem is addressed by making per-batch attempts to transform the image so that it better fits the prior distribution in the latent space. Compared with the per-image approach, this strategy fosters more diverse distributions in the latent space, allowing our model to produce images of great diversity. Comprehensive experiments on benchmark datasets demonstrate the effectiveness of AS-IntroVAE on image generation and reconstruction tasks.

Keywords: Image Generation; Variational Autoencoder; Introspective Learning

1. Introduction

In the last decade, two types of deep generative models—Variational Autoencoders (VAEs) (Kingma and Welling (2013)), and Generative Adversarial Networks (GANs) (Goodfellow et al. (2014)) — has gained tremendous popularity in computer vision (CV) applications. Their acceptance is attributed to their success in various CV tasks, such as image generation (Karras et al. (2019); Vahdat and Kautz (2020)), image reconstruction (Gu et al. (2020); Hou et al. (2017)), and image-to-image translation (Zhu et al. (2017); Liu et al. (2017)).

VAEs can produce images with diverse appearances and is easy-to-train. However, the synthesized images of VAEs are often blurry and lack fine details (Larsen et al. (2016)). On the other hand, GANs produce sharper images with more details but often suffer from mode collapse (i.e., lack diversity) and vanishing gradient (Goodfellow (2016)). Many researchers have sought to develop an efficient hybrid model that combines the advantages of VAEs and GANs. Unfortunately, due to the requirement of an extra discriminator, existing hybrid VAE and GAN models (Makhzani et al. (2015); Larsen et al. (2016); Dumoulin et al. (2016); Tolstikhin et al. (2017)) has high computational complexity and heavy memory usage. Even with these delicate architecture designs, those methods still underperform leading GANs (Karras et al. (2017); Brock et al. (2018)) in terms of the quality of the generated images.

Unlike classical hybrid GAN-VAE models, introspective methods (Huang et al. (2018); Daniel and Tamar (2021)) eliminate the need for extra discriminators. Instead, they utilize the decoder as the 'actual' discriminator to distinguish between the fake and the real images and have achieved state-of-the-art results on image generation tasks. Despite their progress, those introspective learning-based methods suffer from the **posterior collapse** problem with insufficiently tuned hyperparameters and the **vanishing gradient** problem, especially during the early-stage model training.

In this paper, we propose Adversarial Similarity Distance Variational Autoencoder (AS-IntroVAE), an introspective VAE that can competently address the posterior collapse and the vanishing gradient problem. Firstly, we present a theoretical analysis and demonstrate that the vanishing gradient problem of introspective models can be addressed by a similarity distance based upon the 2-Wasserstein distance and the kernel trick. We termed this distance as Adversarial Similarity Distance (AS-Distance). The weight annealing strategy applied to the AS-Distance and the KL-Divergence yields highly stable synthesized images with excellent quality. We address the posterior collapse problem by per-batch aligning the image with the prior distribution in the latent space. This strategy allows the proposed AS-IntroVAE to contain diverse distributions in the latent space, thereby promoting the diversity of the synthesized image.

Our contribution are highlighted as follows (1) A new introspective variational autoencoder named Adversarial Similarity Distance Variational Autoencoder (AS-IntroVAE) (2) A new theoretical understanding of the posteriors collapse and the vanishing gradient problem in VAEs. (3) A novel similarity distance named Adversarial Similarity Distance (AS-Distance) for measuring the differences between the real and the synthesized images. (4) Promising results on image generation and image reconstruction tasks with significantly faster convergence speed.

2. Related Work

2.1. Generative Adversarial Network (GAN)

Generative Adversarial Network (GAN) (Goodfellow et al. (2014)) consists of a generator $G(z)$ and a discriminator $D(x)$. The generator tries to confuse the discriminator by generating a synthetic image from the input noise z, whereas the discriminator tries to distinguish that synthetic image from the real image x.

There are two crucial drawbacks with vanilla GANs: mode collapse (insufficient diversity) and vanishing gradient (insufficient stability) (Goodfellow (2016)). To remedy these issues, WGAN (Arjovsky et al. (2017)) replace the commonly used Jensen Shannon divergence with Wasserstein distance to alleviate vanishing gradient and mode collapse. However, the hard weight clipping used by WGAN to satisfy the Lipschitz constraint significantly reduces the network capacity (Gulrajani et al. (2017)). Some better substitutes for weight clipping include gradient penalty of WGAN-GP (Gulrajani et al. (2017)) and spectral normalization (Miyato et al. (2018)) of SN-GAN.

Compared with these GAN approaches, our method also enjoys the advantages of VAE-based methods: more diversity and stability for the synthesized images.

2.2. Variational Autoencoder (VAE)

Variational Autoencoder (VAE) (Kingma and Welling (2013)) consists of an encoder and an decoder. The encoder $q_\phi(z \mid x)$ compress the image x into a latent variable z, whereas the decoder $p_\theta(x \mid z)$ try to reconstruct the image from that latent variable. Besides, variational inference is applied to approximate the posterior.

A common issue during VAE training is posterior collapse (Bowman et al. (2015)). Posterior collapse occurs when the latent variables become uninformative (e.g., weak, noisy) such that the model choose to rely solely on the autoregressive property of the decoder and ignore the latent variables (Subramanian et al. (2018)). Recent approaches mainly address the posterior collapse problems using KL coefficient annealing (Bowman et al. (2015); Fu et al. (2019)), auxiliary cost function (ALIAS PARTH GOYAL et al. (2017)), pooling operations (Long et al. (2019)), variational approximation restraints (Razavi et al. (2019)), or different ELBO learning objectives (Havrylov and Titov (2020)).

Compared with former VAE approaches, our method also shares the strength of GAN-based models: sharp edges and sufficient fine details.

2.3. Integration of VAE and GAN

A major limitation of VAEs is that they tend to generate blurry, photo-unrealistic images (Dosovitskiy and Brox (2016)). One popular approach to alleviate this issue is to integrate GAN's adversarial training directly into VAE to obtain sharp edges and fine details. Specifically, these hybrid models often consists of an encoder-decoder and an extra discriminator. For example, VAE-GAN (Larsen et al. (2016)) and A-VAE (Mescheder et al. (2017)) both utilize a VAE-like encoder-decoder structure and an extra discriminator to constraint the latent space with adversarial learning. ALI (Dumoulin et al. (2016)) and BiGANs (Donahue et al. (2016)) adopt both mapping and inverse mapping with an extra discriminator to determine which mapping result is better.

To save the growing computational cost from the extra discriminators, the recent state-of-the-art image synthesis method IntroVAE (Huang et al. (2018)) proposes to train VAEs in an introspective way such that the model can distinguish between the fake and real images using only the encoder and the decoder. The problem with IntroVAE is that it utilizes a hard margin to regulate the hinge terms (i.e., the KL divergence between the posterior and the prior), which leads to unstable training and difficult hyperparameter selection. To alleviate this issue, S-IntroVAE (Daniel and Tamar (2021)) expresses VAE's evidence lower bound (ELBO) in the form of a smooth exponential loss, thereby replacing the hard margin with a soft threshold function. However, the posterior collapse and the vanishing gradient problem still exist in these introspective methods.

Unlike these methods, our approach has stable training throughout the entire training stage and can generate samples of sufficient diversity without careful hyperparameter tuning.

3. Background

We place our generative model under the variational inference setting (Kingma and Welling (2013)), where we aim to utilize variational inference methods to approximate the in-

tractable maximum-likelihood objective. With this in mind, in this section, we will first revisit vanilla VAE, focusing on its Evidence Lower Bound (ELBO) technique, and then analyze introspective learning-based methods, including IntroVAE and S-IntroVAE.

3.1. Evidence Lower Bound (ELBO)

The learning object of VAE is to maximize the evidence lower bound (ELBO) as below:

$$\log_\theta(x) \geq E_{q_\phi(z|x)} \log p_\theta(x \mid z) - D_{KL}\left(q_\phi(z \mid x) \| p_\theta(z)\right) \tag{1}$$

where x is the input data, z is the latent variable, $q_\phi(z \mid x)$ represents neural network with parameter ϕ, and $p_\theta(x \mid z)$ represents the input data's probability density function with parameter θ. The Kullback-Leibler (KL) divergence term can be expressed as:

$$D_{\mathrm{KL}}(q(z \mid x) \| p(z)) = \mathbb{E}_{q(z)}\left[\log \frac{q(z \mid x)}{p(z)}\right] \tag{2}$$

Reparameterization trick (Kingma and Welling (2013)) is applied to make VAE trainable (i.e., differentiable through backpropagation). Specifically, the reparameterization trick transforms the latent representation z into two latent vectors μ and σ and a random vector ε, thereby excluding randomness from the backpropagation process. The reparameterization trick can be formulated as $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \varepsilon$.

3.2. IntroVAE

Unlike the vanilla VAE, which optimizes upon a single lower bound, IntroVAE incorporates an adversarial learning strategy commonly used by GANs during training. Specifically, the encoder aims to maximize the KL divergence between the fake image and the latent variable and minimize the KL divergence between the actual image and the latent variable. Meanwhile, the decoder aims to confuse the encoder by minimizing the KL divergence between the fake photo and the latent variable. The learning objective (i.e., loss function) of IntroVAE for Encoder and Decoder is:

$$\begin{aligned}
\mathcal{L}_E &= ELBO(x) + \sum_{s=r,g}\left[m - KL\left(q_\phi\left(z|x_s\right)\|p(z)\right]^+ \\
\mathcal{L}_D &= \sum_{s=r,g}\left[KL\left(q_\phi\left(z|x_s\right)\|p(z)\right)\right].
\end{aligned} \tag{3}$$

where x_r is the reconstructed image, x_g is the generated image, and m is the hard threshold for constraining the KL divergence.

3.3. S-IntroVAE

The major limitation of IntroVAE is that it utilizes a hard threshold m to constrain the KL divergence term. S-IntroVAE suggests this design will significantly reduce model capacity and induce vanishing gradient. Instead, S-IntroVAE proposes to utilize the complete ELBO
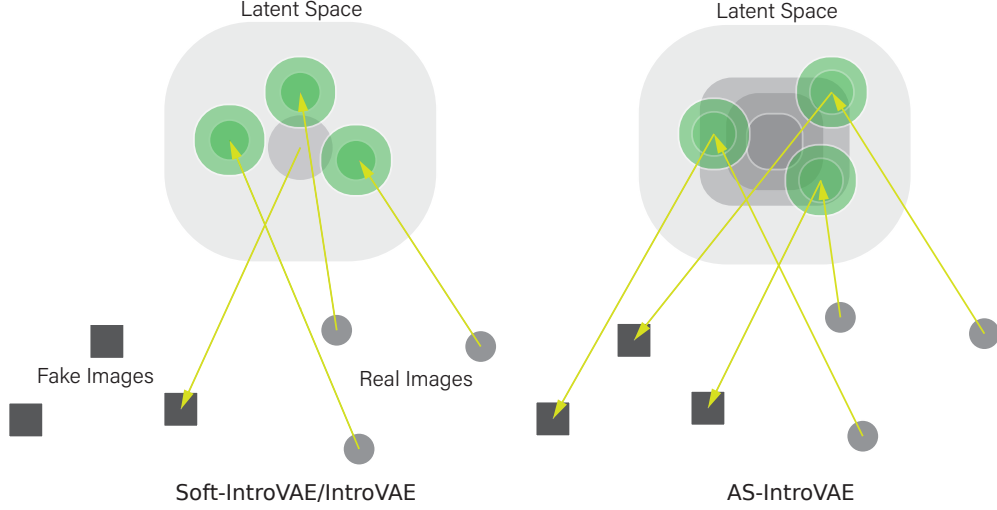
Figure 1: Illustration of how AS-IntroVAE addresses the posterior collapse problem. Both IntroVAE/S-IntroVAE and the proposed AS-IntroVAE project the real images into the latent space. However, IntroVAE/S-IntroVAE force every image to match the prior distribution of the latent space. This enforcement undermines the valuable signal from the real image such that the latent space becomes uninformative for the decoder. In contrast, AS-IntroVAE align the image with the prior distribution in a per-batch manner. Since a batch contains far more variations than a single image, in AS-IntroVAE, the signal becomes strong enough such that the decoder has to leverage the latent space to generate diverse samples.

(instead of just KL) with a soft exponential function (instead of a hard threshold). The learning objective (loss function) of S-IntroVAE is:

$$
\begin{aligned}
\mathcal{L}_E &= ELBO(x) - \frac{1}{\alpha} \sum_{s=r,g} \exp\left(\alpha ELBO\left(x_s\right)\right) \\
\mathcal{L}_D &= ELBO(x) + \gamma \sum_{s=r,g} ELBO\left(x_s\right)
\end{aligned}
\tag{4}
$$

where α, γ are both hyperparameters.

4. Proposed Method

In this section, we will illustrate the proposed AS-IntroVAE, including its strategy for posterior collapse (Fig. 1), its model workflow (Fig. 2), and the theoretical analysis.
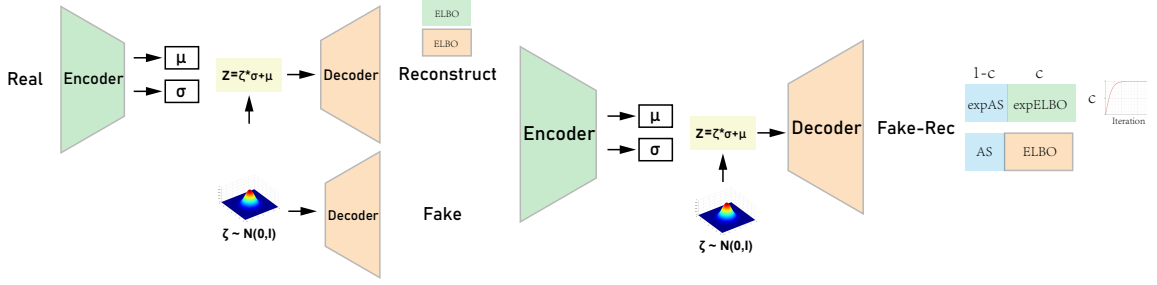
Figure 2: AS-IntroVAE workflow, using image reconstruction task as the example. AS-IntroVAE contains an encoder-decoder architecture. In the first phase, the encoder-decoder receives the real image and produce the reconstructed image. Meanwhile, the decoder will generate fake image from Gaussian noise alone. In the second phase, the **same** encoder-decoder conduct adversarial learning in the latent space for the reconstructed image and the fake image. During adversarial learning, the encoder tries to maximize the adversarial similarity distance between fake image and reconstruction image, whereas the decoder wants to minimize it. After each iteration, the model will calculate the annealing rate c..

4.1. Theoretical Analysis

An effective distance metrics are crucial for generative model like VAEs and GANs. To address the vanishing gradient problems of S-IntroVAE and IntroVAE, we propose a novel similarity distance called Adversarial Similarity Distance (AS-Distance). Inspired by 1-Wasserstein distance, which could provide stable gradients, the AS-Distance is defined as:

$$D\left(p_r, p_g\right) = \mathbb{E}_{x \sim p(z)}[\left(\mathbb{E}_{x \sim p_r}\left[q\left(z|x\right)\right] - \mathbb{E}_{x \sim p_g}\left[q\left(z|x\right)\right]\right)]^2 \tag{5}$$

where p_r is distribution of real data, p_g is distribution of generated data. The encoder and the decoder plays an adversarial game on this distance:

$$arg \min_{Dec} \max_{Enc} D\left(p_r, p_g\right) \tag{6}$$

We use 2-Wasserstein so that we could apply a kernel trick on Equ.5.

$$D(p_r, p_g) = \mathbb{E}_{x \sim p_{r,g}}\left[k\left(x_r^i, x_r^j\right) + k\left(x_g^i, x_g^j\right) - 2k\left(x_r^i, x_g^j\right)\right] \tag{7}$$

where $k\left(x_r^i, x_g^j\right) = \mathbb{E}_{z \sim p(z)}[q(z|x_r^i) \cdot q(z|x_g^i)]$.
Since the latent space is a normal distribution. This kernel k can be deduced as

$$k\left(x_r^i, x_g^j\right) = \frac{-\frac{1}{2}\frac{\left(u_r^i - u_g^j\right)^2}{\lambda_r^i + \lambda_g^j}}{\left(2\pi\right)^{\frac{n}{2}} \cdot \left(\lambda_r^i + \lambda_g^j\right)^{\frac{1}{2}}} \tag{8}$$

where u, λ represent the variational inference on the mean and variance of x, i, j represent the ith, jth pixel in images.

In the maximum mean discrepancy (MMD) method, the distance calculation is conducted *after* the reparameterization. As shown by Wu and Zhuang (2020), this will leads to high variance (i.e., error) for the estimated distance. Instead, we seek to calculate the distance *before* the reparameterization, which can reduce the variance and improve the distance estimation accuracy.

During the experiment, we found that KL term from S-IntroVAE would generate sharp but distort images, whereas our AS term (without KL term) would generate diverse but blur images. If we fix the weight for KL and AS term (e.g. both at 0.5), there will exist two optimal solutions, which induces training stability. Inspired by (Fu et al. (2019)), we decide to gradually increase the weight for KL (from 0 to 1), and decrease the weight for AS (from 1 to 0) during training. In this way, for KL and AS, we can enjoy their advantages and eschew their disadvantages.

Based on the former discussions, we derive the loss function for AS-IntroVAE as:

$$
\begin{aligned}
\mathcal{L}_{E_\phi} = ELBO(x) &- \frac{1}{\alpha} \sum_{s=r,g} \exp(\alpha(\mathbb{E}_{q(z|x_s)}[\log p(x \mid z)] \\
&+ cKL(q_\phi(z|x_s)\|p(z)) + (1-c)D(x_r, x_g))) \\
\mathcal{L}_{D_\theta} = ELBO(x) &+ \gamma \sum_{s=r,g} (\mathbb{E}_{q(z|x_s)}[\log p(x \mid z)] \\
&+ cKL(q_\phi(z|x_s)\|p(z)) + (1-c)D(x_r, x_g))
\end{aligned}
\tag{9}
$$

where $c = min(i * 5/T, 1)$, i is the current iteration and T is total iteration.

Theorem 1 *Introspective Variational Autoencoders (IntroVAEs) have vanishing gradient problems.*

Proof As illustrated in IntroVAEs (IntroVAE and S-IntroVAE), the Nash equilibrium can be attained when $KL\left(q_\phi\left(z|x_r\right)\|q_\phi\left(z|x_g\right)\right) = 0$, where x_r could also represents the real images since the reconstructed images are sampled from real data points. Moreover, with the object $D_{KL}\left(q_\phi(z \mid x)\|p(z)\right) = 0$, we have:

$$
q_\phi\left(z|x_r\right) = q_\phi\left(z|x_g\right) = p(z)
\tag{10}
$$

Replace the term $p(z)$ with $\frac{q_\phi(z|x_r)+q_\phi(z|x_g)}{2}$, the adversarial term for the decoder then becomes:

$$
\begin{aligned}
KL&\left(q_\phi\left(z|x_r\right)\|\frac{q_\phi\left(z|x_r\right)+q_\phi\left(z|x_g\right)}{2}\right) + KL\left(q_\phi\left(z|x_g\right)\|\frac{q_\phi\left(z|x_r\right)+q_\phi\left(z|x_g\right)}{2}\right) \\
&= 2JSD\left(q_\phi\left(z|x_r\right)\|q_\phi\left(z|x_g\right)\right)
\end{aligned}
\tag{11}
$$

∎

Therefore, the gradient of loss for Decoder in IntroVAE becomes:

$$
\nabla\mathcal{L}_D = \nabla 2JSD\left(q_\phi\left(z|x_r\right)\|q_\phi\left(z|x_g\right)\right)
\tag{12}
$$

As shown by (Arjovsky and Bottou (2017)), if P_{x_r} and P_{x_g} are two distributions in two different manifolds that don't align perfectly and don't have full dimension (i.e., the dimension of the latent variable is sparse in the image dimension). With the assumption that P_{x_r} and P_{x_g} are continuous in their manifolds, if there exists a set A with measure 0 in one manifold, then $P(A) = 0$. Consequently, there will be an optimal discriminator with 100% accuracy for classify almost any x in these two manifolds, resulting in $\nabla \mathcal{L}_D = 0$.

For IntroVAE, the above condition (i.e., vanishing gradient problem) occurs since the very beginning of the training process when there is no intersection between the real image distribution and the fake image distribution. The reason why S-IntroVAE alleviate vanishing gradient at the later epoch (See Fig.7) is that the reconstruction loss of S-IntroVAE gradually create a support set between the real and the fake images during training. In comparison, since AS-Distance is based on 2-Wasserstein distance, the proposed AS-IntroVAE provides stable gradient, even there is no intersection between these two distributions.

5. Experiments

In this section, we will explain the implementation details, the comparison on 2D toy datasets, image generation and image reconstruction tasks, and the training stability.

5.1. Implementation Details

We train our model using the Adam (Kingma and Ba (2014)) optimizer with the default setting ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 150 epochs. We implement our framework in Pytorch (Paszke et al. (2019)) with 3 NVIDIA RTX 3090 GPU. Same with (Huang et al. (2018); Daniel and Tamar (2021)), we set a fixed learning rate of 2e-4. It takes around 1 day to converge our model on CelebA-128 dataset and 2 days on CelebA-256 dataset, respectively. The exponential moving average is applied to stabilize training. The encoder and decoder will be updated respectively in each iteration. For the loss function, we set $\alpha = 2$, and $\gamma = 1$. The weight for the real image's KL term and reconstruction term is fixed at 0.5 and 1.0, respectively, whereas the fake image's KL term and reconstruction term is both set at 0.5. For the annealing rate c, we apply a linear function shown in Equ.9. For other hyperparameter settings, we inherit the setting from S-IntroVAE (Daniel and Tamar (2021)).

5.2. 2D Toy Datasets

In this subsection, we evaluate the proposed method's performance on 2D Toy datasets, including Gaussian and Checkerboard (De Cao et al. (2020); Grathwohl et al. (2018)), and compare our approach with baselines including VAE, IntroVAE, and S-IntroVAE using two commonly used metrics, including KL-divergence (KL) and Jensen–Shannon-divergence (JSD). Both KL and JSD measure how far away the model predicted outcome is from the ground truth data distribution. Therefore, a lower score indicates a better result for both KL and JSD.

We design three hyperparameter combinations to assess the robustness of different methods to hyperparameter changes. The hyperparameter includes the weight for real image ELBO, the weight for fake image's KL divergence term, and the weight for fake image's
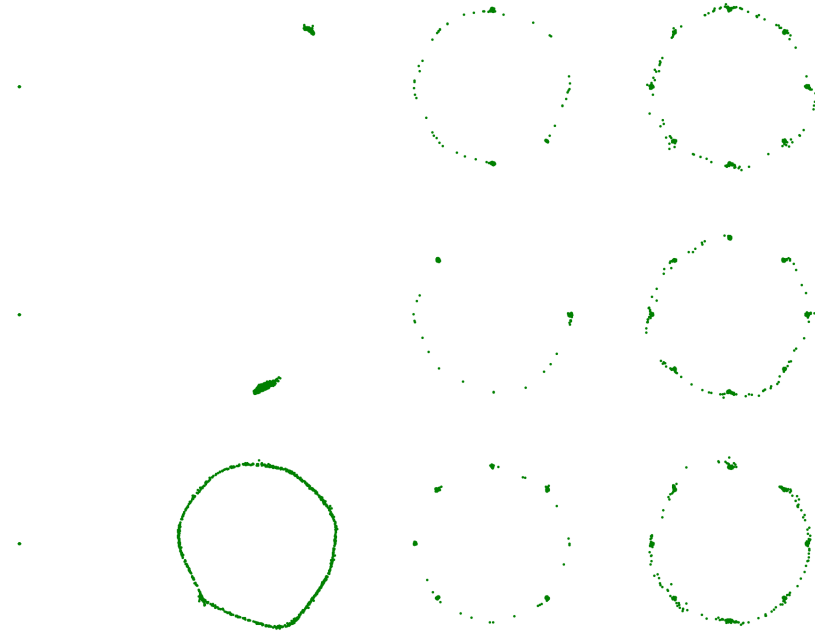
Figure 3: Visual Comparison on 2D Toy Dataset 8 Gaussians. From top to bottom row: results with different hyperparameters. From left to right column: VAE, IntroVAE, S-IntroVAE, Ours. Zoom in to view the detail within each subfigure.

reconstruction term. The value for each combination is as below. *C1: (0.3, 0.1, 0.9) C2: (0.5, 0.1, 0.9) C3: (0.7, 0.2, 0.9).*

Table 5.3 shows the quantitative comparison on 2D Toy Datasets 8 Gaussians. For the 8 Gaussians dataset, we find that our method has the lowest (i.e., best) KL and JSD score under all hyperparameter combinations, outperforming VAE, IntroVAE, and S-IntroVAE by a large margin.

Fig. 3 shows the qualitative comparison of the 8 Gaussians dataset. We notice that VAE has one isolated data point for all three hyperparameter combinations, which means it has severe posterior collapse problems. IntroVAE has a small trace around a specific data point for C1 and C2, indicating nontrivial posterior collapse problems. For C3, IntroVAE produces a ring shape, meaning the generated data is evenly distributed and fails to converge to any designated data point. S-IntroVAE converges to two data points for C1 and C2 and six for C3, which still reflects the posterior collapse problem.

In comparison, our method, under all hyperparameter combinations, successfully converges to all eight data points. Therefore, we can conclude our approach is the only one that avoids the posterior collapse problem in 8 Gaussian datasets experiments. Due to the scope of this paper, the result for the Checkerboard dataset is in the supplementary material.

(a) WGAN-GP   (b) S-IntroVAE   (c) Ours

Figure 4: Image Generation Visual Comparison at CelebA-128 dataset.

5.3. Image Generation

In this subsection, we evaluate the proposed method's performance on image generation tasks, using benchmark datasets including MNIST (LeCun et al. (1998)), CIFAR-10 (Krizhevsky et al. (2009)), CelebA-128, and CelebA-256 (Liu et al. (2015)). The methods for comparison includes WGAN-GP and S-IntroVAE, and the evaluation metric is Frechet Inception Distance (FID). Specifically, we use FID to estimate the distance between the generated dataset's distribution and the source (i.e., training) dataset's distribution. Hence, a lower FID score means a better result.

Table 5.3 shows the quantitative comparison of image generation tasks. For all chosen datasets, the proposed method has the lowest (e.g., best) FID score. Fig. 4 shows the qualitative comparison of image generation at CelebA-128 dataset. We notice both WGAN-GP and S-IntroVAE have apparent facial feature distortion, edging blur, and facial asymmetry. Although S-IntroVAE has less ghost artifact and unnatural texture than WGAN-GP, it has a significant posterior collapse problem: S-IntroVAE's two generated images in the first row are extremely similar. In comparison, the proposed method's generated face is the best in terms of all mentioned aspects.

Fig. 5 shows the qualitative comparison of image generation at CelebA-256 dataset. Compared with the CelebA-128 results in Fig. 4, we find that both WGAN-GP and S-IntroVAE experience less posterior collapse, unnatural texture, and facial asymmetry. However, both WGAN-GP and S-IntroVAE still have significant facial feature distortion. The ghost artifacts, the edging blur, and the over-smoothed hairs from these methods further degrade the perceptual quality. Compared with WGAN-GP and S-IntroVAE, the proposed method is the best in all mentioned aspects. Due to the scope of this paper, the qualitative result of image generation on other datasets will be in the supplementary material.

5.4. Image Reconstruction

In this subsection, we evaluate the proposed method's performance on image reconstruction tasks using benchmark datasets, including MNIST, CIFAR-10, Oxford Building Datasets, CelebA-128, and CelebA-256. The model for comparison is S-IntroVAE, and the evaluation metrics are Peak signal-to-noise ratio (PSNR), Structural Similarity Index (SSIM), and

(a) WGAN-GP　　　　(b) S-IntroVAE　　　　(c) Ours

Figure 5: Image Generation Visual Comparison at CelebA-256 dataset.

		VAE	IntroVAE	S-IntroVAE	Ours
C1	KL	220.2	192.4	50.2	**3.4**
	JSD	110.1	56.0	16.9	**5.6**
C2	KL	220.3	191.1	136.5	**1.3**
	JSD	110.0	68.0	36.6	**4.4**
C3	KL	220.2	64.0	46.2	**2.0**
	JSD	109.8	53.0	9.6	**7.1**

	WGAN-GP	S-IntroVAE	Ours
MNIST	139.02	98.84	**96.16**
CIFAR-10	434.11	275.20	**271.69**
CelebA-128	160.53	140.35	**130.74**
CelebA-256	170.79	143.33	**129.61**

Table 1: 2D Toy Dataset 8 Gaussians Score KL↓/JSD↓ Table

Table 2: Image Generation FID Score↓ Table.

Mean Squared Error (MSE). A higher PSNR, a higher SSIM, and a lower MSE mean better results.

Table 5.4 shows the quantitative comparison of image reconstruction task. For all except CelebA-256, our method has the best PSNR, SSIM, and MSE. For the CelebA-256 dataset, our method has the second-best SSIM but the best for both PSNR and MSE. Fig. 6 shows the qualitative comparison of image reconstruction at CelebA-128 dataset. S-IntroVAE fails to faithfully reconstruct the facial features, facial expressions, and skin textures. The reconstructed image also contains significant edging blur, defects, and artifacts, which significantly distort the perceptual quality.

The proposed method is much closer to the ground truth regarding contrast, exposure, color, edge information, and facial details. In short, our approach surpasses S-IntroVAE by a large margin in image reconstruction with the CelebA-128 dataset.

5.5. Training Stability

Fig. 7 shows the training stability visual comparison at CelebA-128 dataset. We find that IntroVAE fails in image reconstruction and image generation tasks, even if we train the model using its recommended hyperparameters. IntroVAE's reconstructed images at a specific epoch are almost homogeneous: a mixture of blue and green clouds with little semantic information.

We also find that S-IntroVAE has split performances in the early stage (10 & 20 epoch) and the later stage (50 epoch). In the early stage, the reconstructed images contain a loss of blur, defects, and artifacts, whereas the generated images have distorted facial features
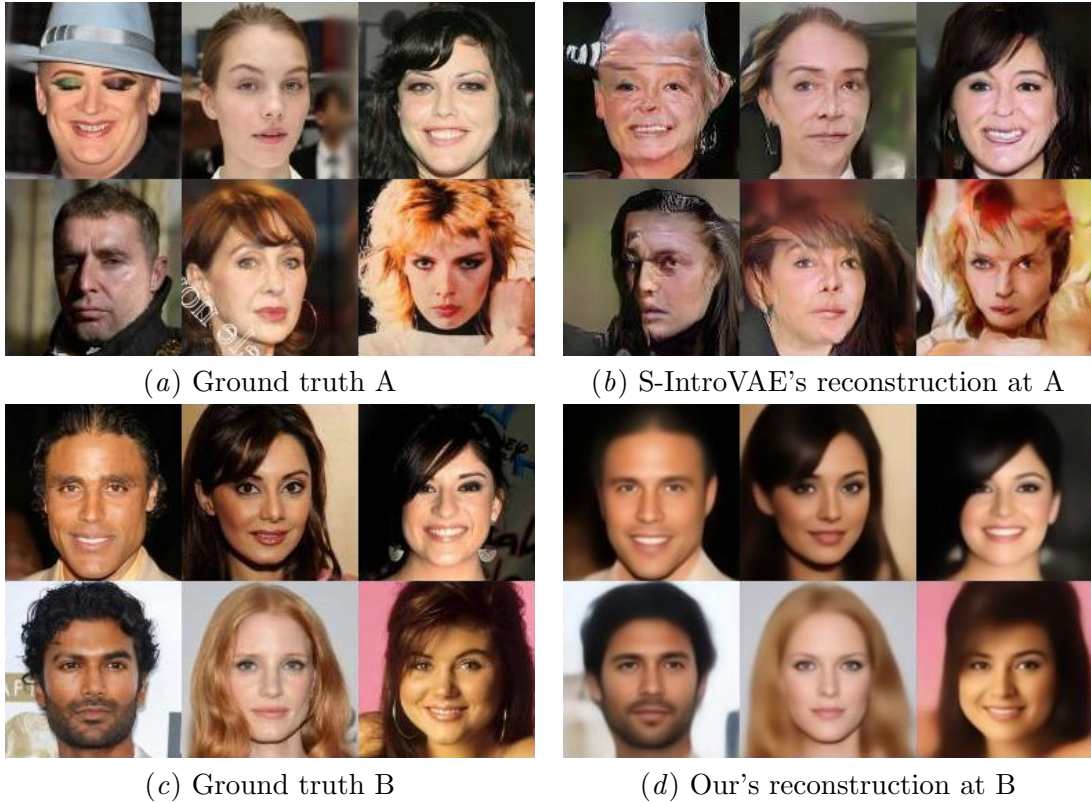
(a) Ground truth A   (b) S-IntroVAE's reconstruction at A

(c) Ground truth B   (d) Our's reconstruction at B

Figure 6: Image Reconstruction Visual Comparison at CelebA-128 dataset.

	PSNR		SSIM		MSE	
	S-IntroVAE	Ours	S-IntroVAE	Ours	S-IntroVAE	Ours
MNIST	20.282	**21.014**	0.885	**0.898**	0.011	**0.009**
CIFAR-10	19.300	**19.445**	0.599	**0.620**	**0.019**	**0.019**
Oxford	15.372	**20.168**	0.348	**0.604**	0.049	**0.013**
CelebA-128	17.818	**22.924**	0.561	**0.801**	0.018	**0.006**
CelebA-256	22.422	**23.156**	**0.790**	0.758	0.007	**0.006**

Table 3: Image Reconstruction PSNR↑/SSIM↑/MSE↓ Score Table

and a significant amount of unnatural artifacts. In the later stage, the quality of both tasks has improved. However, the generated and reconstructed images still contain many defects and edging blur.

In comparison, the proposed method has quick learning convergence in the early stage (10 & 20 epoch) and maintains excellent training stability in the later stage (50 epoch). The reconstructed images are faithfully aligned with the original images, whereas the generated images have superb perceptual quality.

(a) IntroVAE

(b) S-IntroVAE

(c) AS-IntroVAE (Ours)

Figure 7: The training stability visual comparison at CelebA-128 dataset. From left to right panel: 10 epoch, 20 epoch, 50 epoch. For each image grid, the first and the second row are real images, the third and fourth rows are reconstructed images, and the fifth and sixth rows are generated images. Zoom in for a better view.

6. Conclusion

This paper introduces Adversarial Similarity Distance Introspective Variational Autoencoder (AS-IntroVAE), a new introspective approach that can faithfully address the posterior collapse and the vanishing gradient problem. Our theoretical analysis rigorously illustrated the advantages of the proposed Adversarial Similarity Distance (AS-Distance). Our empirical results exhibited compelling quality, diversity, and stability in image generation and construction tasks. In the future, we hope to apply the proposed AS-IntroVAE to high resolution (e.g., 1024×1024) image synthesis. We also hope to extend AS-IntroVAE to reinforcement learning and self-supervised learning tasks.

References

Anirudh Goyal ALIAS PARTH GOYAL, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. Z-forcing: Training stochastic recurrent networks. *Advances in neural information processing systems*, 30, 2017.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Tal Daniel and Aviv Tamar. Soft-introvae: Analyzing and improving the introspective variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4391–4400, 2021.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Block neural autoregressive flow. In *Uncertainty in artificial intelligence*, pages 1263–1273. PMLR, 2020.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Serhii Havrylov and Ivan Titov. Preventing posterior collapse with levenshtein variational autoencoder. *arXiv preprint arXiv:2004.14758*, 2020.

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 1133–1141. IEEE, 2017.

Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. *Advances in neural information processing systems*, 31, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Teng Long, Yanshuai Cao, and Jackie Chi Kit Cheung. Preventing posterior collapse in sequence vaes with pooling. 2019.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400. PMLR, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Ali Razavi, Aäron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*, 2019.

Sandeep Subramanian, Sai Rajeswar Mudumba, Alessandro Sordoni, Adam Trischler, Aaron C Courville, and Chris Pal. Towards text generation with adversarially learned neural outlines. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Fuping Wu and Xiahai Zhuang. Cf distance: a new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):4274–4285, 2020.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.