AS-IntroVAE: Adversarial Similarity Distance Makes Robust IntroVAE

Changjie Lu[1], *ShenZheng*[1,3], *ZiruiWang*[2], *OmarDib*[1], *GauravGupta*[1]

Wenzhou Kean University, Zhejiang University, Carnegie Mellon University

SSPF Presentation

Table of Contents

Image Generation

Figure 1: Image generation results from our model: AS-IntroVAE

Classical Methods

1. Generative Adversarial Networks(GAN)[Goo+14]

$$\min_G \max_D V(D,G) = E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))]. \tag{1}$$

where D is discriminator, G is generator, x is the datasets, z is the latent variable.

2. Variational AutoEncoder(VAE)[KW13]

$$\begin{aligned} \log p(x) &\geq \mathbb{E}_{q(z\mid x)}[\log p(x,z) - \log q(z\mid x)] \\ &:= ELBO \\ &= \underbrace{\mathbb{E}_{q(z\mid x)}[\log p(x\mid z)]}_{\text{Reconstruct term } L_{\text{Rec}}} - \underbrace{D_{KL}(q(z\mid x)\|p(z))}_{\text{KL term } L_{KL}} \end{aligned} \tag{2}$$

where KL is KL Divergence, p is decoder, q is encoder.

Related Works

To tackle with the drawbacks of VAE(Posterior collapse[Bow+15], vague visual quality[DB16]) and GAN(mode collpase, vanishing gradient[Goo16]), here are some related works.

1. GAN: WGAN[ACB17],WGAN-GP[Gul+17],SN-GAN[Miy+18]
2. VAE: VAE-GAN[Lar+16],AAE[MNG17a],ALI[Dum+16],BiGAN[DKD16]

Limitation: Quality not good enough, Need extra networks.

Introspective VAE(Intro-VAE)[Hua+18]

Combine VAE(statistical analysis) and GAN(adversarial learning) together.

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

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

Soft-IntroVAE[DT21]

The hard threshold makes training stability sensitive to the hyper parameter, S-IntroVAE introduces a soft expression.

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

where α, γ are both hyperparameters.

Limitation and Solution

Those introspective learning-based methods suffer from the **posterior collapse** problem and the **vanishing gradient** problem.

Contribution:

1. A new introspective variational autoencoder named Adversarial Similarity Distance Introspective 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

Table of Contents

Theoretical Analysis

Inspired by 1-Wasserstein distance, which could provide stable gradients, the AS-Distance is defined as:

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

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:

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

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

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

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

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.

Figure 2: S-IntroVAE performance at CelebA-128, when the weight for KL divergence and AS-Distance are both 0.5. The upper/middle/bottom two rows refer to real/reconstructed/generated images.

Inspired by ([Fu+19]), 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.

We derive the loss function for AS-IntroVAE as:

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

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

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

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 ([AB17]), 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). Consequently, there will be an optimal discriminator with 100% accuracy for classify almost any x in these two manifolds, resulting in $\nabla \mathcal{L}_D = 0$.

Table of Contents

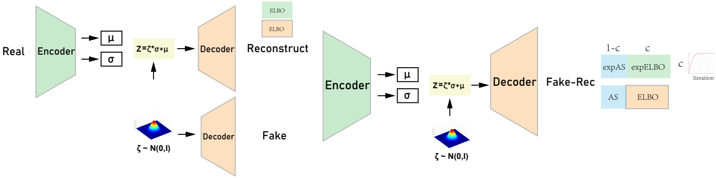

Figure 3: AS-IntroVAE workflow. In the first phase, the encoder-decoder receives the real image and produce the reconstructed image. In the second phase, the **same** encoder-decoder conduct adversarial learning in the latent space for the reconstructed image and the fake image.

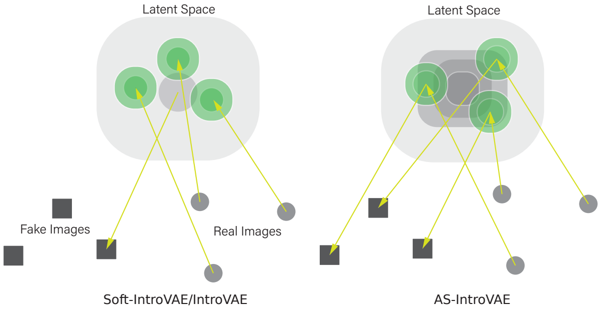

Figure 4: Illustration of how AS-IntroVAE addresses the posterior collapse problem. Both IntroVAE/S-IntroVAE and the proposed AS-IntroVAE project the real images into the latent space. However, IntroVAE/S-IntroVAE force every image to match the prior distribution of the latent space. AS-IntroVAE align the image with the prior distribution in a per-batch manner.

Results

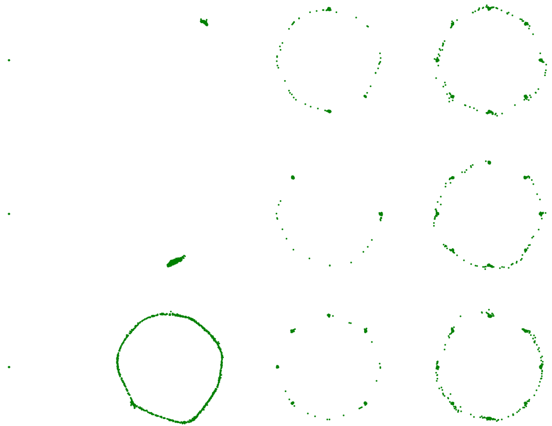

Figure 5: Visual Comparison on 2D Toy Dataset 8 Gaussians. From top to bottom row: results with different hyperparameters. From left to right column: VAE, IntroVAE, S-IntroVAE, Ours.

Figure 6: Image Generation Visual Comparison at CelebA-128 dataset. From left to Right: WGAN-GP, S-IntroVAE, Ours

Figure 7: Image Generation Visual Comparison at CelebA-256 dataset. From

Figure 8: Image Reconstruction Visual Comparison at CelebA-128 dataset.

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

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

	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 2: Image Generation FID Score↓ Table.

	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

Figure 9: The training stability visual comparison at CelebA-128 dataset. From left to right panel: 10 epoch, 20 epoch, 50 epoch.

Figure 10: Image generation visual comparisons at CelebA-128 dataset (resolution: 128×128).

Figure 11: Image generation visual comparisons at CelebA-256 dataset (resolution: 256×256).

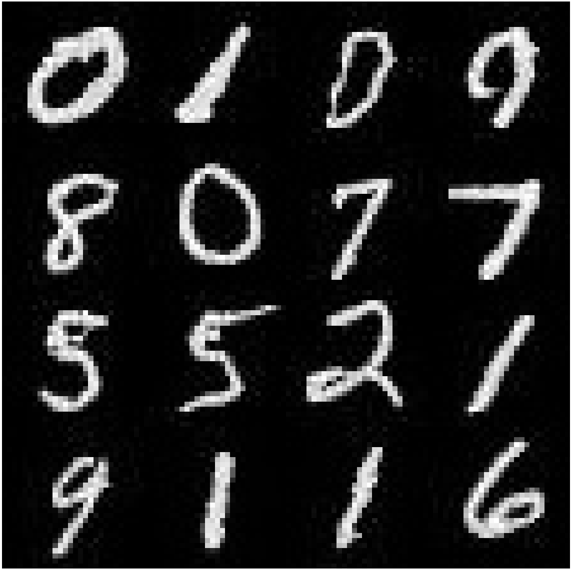

Figure 12: Image generation visual comparisons at MNIST dataset (resolution: 28×28).

Table of Contents

1. 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.
2. Our theoretical analysis rigorously illustrated the advantages of the proposed Adversarial Similarity Distance (AS-Distance).
3. Our empirical results exhibited compelling quality, diversity, and stability in image generation and construction tasks.
4. 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 I

[AB17] Martin Arjovsky and Léon Bottou. "Towards principled methods for training generative adversarial networks". In: *arXiv preprint arXiv:1701.04862* (2017).

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International conference on machine learning*. PMLR. 2017, pp. 214–223.

[ALI+17] Anirudh Goyal ALIAS PARTH GOYAL et al. "Z-forcing: Training stochastic recurrent networks". In: *Advances in neural information processing systems* 30 (2017).

[BDS18] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis". In: *arXiv preprint arXiv:1809.11096* (2018).

[Bow+15] Samuel R Bowman et al. "Generating sentences from a continuous space". In: *arXiv preprint arXiv:1511.06349* (2015).

References II

[CGJ19] Lei Cai, Hongyang Gao, and Shuiwang Ji. "Multi-stage variational auto-encoders for coarse-to-fine image generation". In: *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM. 2019, pp. 630–638.

[DAT20] Nicola De Cao, Wilker Aziz, and Ivan Titov. "Block neural autoregressive flow". In: *Uncertainty in artificial intelligence*. PMLR. 2020, pp. 1263–1273.

[DB16] Alexey Dosovitskiy and Thomas Brox. "Generating images with perceptual similarity metrics based on deep networks". In: *Advances in neural information processing systems* 29 (2016).

[DKD16] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. "Adversarial feature learning". In: *arXiv preprint arXiv:1605.09782* (2016).

References III

[DT21] 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*. 2021, pp. 4391–4400.

[Dum+16] Vincent Dumoulin et al. "Adversarially learned inference". In: *arXiv preprint arXiv:1606.00704* (2016).

[Fu+19] Hao Fu et al. "Cyclical annealing schedule: A simple approach to mitigating kl vanishing". In: *arXiv preprint arXiv:1903.10145* (2019).

[Goo+14] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[Goo16] Ian Goodfellow. "Nips 2016 tutorial: Generative adversarial networks". In: *arXiv preprint arXiv:1701.00160* (2016).

References IV

[Gra+18] Will Grathwohl et al. "Ffjord: Free-form continuous dynamics for scalable reversible generative models". In: *arXiv preprint arXiv:1810.01367* (2018).

[GSZ20] 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*. 2020, pp. 3012–3021.

[Gul+17] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *Advances in neural information processing systems* 30 (2017).

[Hou+17] Xianxu Hou et al. "Deep feature consistent variational autoencoder". In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2017, pp. 1133–1141.

References V

[HT20] Serhii Havrylov and Ivan Titov. "Preventing posterior collapse with levenshtein variational autoencoder". In: *arXiv preprint arXiv:2004.14758* (2020).

[Hua+18] Huaibo Huang et al. "Introvae: Introspective variational autoencoders for photographic image synthesis". In: *Advances in neural information processing systems* 31 (2018).

[Kar+17] Tero Karras et al. "Progressive growing of gans for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196* (2017).

[Kar+20] Tero Karras et al. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8110–8119.

References VI

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[KLA19] 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*. 2019, pp. 4401–4410.

[KW13] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[Lai+17] Wei-Sheng Lai et al. "Deep laplacian pyramid networks for fast and accurate super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 624–632.

References VII

[Lar+16] Anders Boesen Lindbo Larsen et al. "Autoencoding beyond pixels using a learned similarity metric". In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566.

[LBK17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks". In: *Advances in neural information processing systems* 30 (2017).

[LCC19] Teng Long, Yanshuai Cao, and Jackie Chi Kit Cheung. "Preventing posterior collapse in sequence vaes with pooling". In: (2019).

[LeC+98] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[Liu+15] Ziwei Liu et al. "Deep learning face attributes in the wild". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 3730–3738.

References VIII

[Mak+15] Alireza Makhzani et al. "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644* (2015).

[MGN18] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. "Which training methods for GANs do actually converge?" In: *International conference on machine learning*. PMLR. 2018, pp. 3481–3490.

[Miy+18] Takeru Miyato et al. "Spectral normalization for generative adversarial networks". In: *arXiv preprint arXiv:1802.05957* (2018).

[MNG17a] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2391–2400.

References IX

[MNG17b] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. "The numerics of gans". In: *Advances in neural information processing systems* 30 (2017).

[PAD20] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. "Adversarial latent autoencoders". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 14104–14113.

[Pas+19] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[Raz+19] Ali Razavi et al. "Preventing posterior collapse with delta-vaes". In: *arXiv preprint arXiv:1901.03416* (2019).

[SC21] Divya Saxena and Jiannong Cao. "Generative adversarial networks (GANs) challenges, solutions, and future directions". In: *ACM Computing Surveys (CSUR)* 54.3 (2021), pp. 1–42.

References X

[Sha+20] Huajie Shao et al. "Controlvae: Controllable variational autoencoder". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8655–8664.

[Søn+16] Casper Kaae Sønderby et al. "Amortised map inference for image super-resolution". In: *arXiv preprint arXiv:1610.04490* (2016).

[Sta+21] Jan Stanczuk et al. "Wasserstein GANs work because they fail (to approximate the Wasserstein distance)". In: *arXiv preprint arXiv:2103.01678* (2021).

[Sub+18] Sandeep Subramanian et al. "Towards text generation with adversarially learned neural outlines". In: *Advances in Neural Information Processing Systems* 31 (2018).

[Tol+17] Ilya Tolstikhin et al. "Wasserstein auto-encoders". In: *arXiv preprint arXiv:1711.01558* (2017).

References XI

[VK20] Arash Vahdat and Jan Kautz. "NVAE: A deep hierarchical variational autoencoder". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19667–19679.

[WZ20] Fuping Wu and Xiahai Zhuang. "CF distance: a new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation". In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4274–4285.

[Zha+17] Han Zhang et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5907–5915.

[Zha+18] Han Zhang et al. "Stackgan++: Realistic image synthesis with stacked generative adversarial networks". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 1947–1962.

References XII

[Zhu+17] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[ZML16] Junbo Zhao, Michael Mathieu, and Yann LeCun. "Energy-based generative adversarial network". In: *arXiv preprint arXiv:1609.03126* (2016).

[ZSE17] Shengjia Zhao, Jiaming Song, and Stefano Ermon. "Towards deeper understanding of variational autoencoding models". In: *arXiv preprint arXiv:1702.08658* (2017).

[ZXY18] Zizhao Zhang, Yuanpu Xie, and Lin Yang. "Photographic text-to-image synthesis with a hierarchically-nested adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6199–6208.