AS-IntroVAE: Adversarial Similarity Distance Makes Robust IntroVAE

Changjie Lu, Shen Zheng, Zirui Wang, Omar Dib, Gaurav Gupta

Wenzhou-Kean University, Carnegie Mellon University, Zhejiang University

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 images. In this paper, we propose a new variation of IntroVAE called Adversarial Similarity Distance Introspective 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.

Background

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.

- The loss function of IntroVAE

$$\begin{aligned}\mathcal{L}_E &= ELBO(x) + \sum_{s=r,g}\left[m - KL\left(q_\phi\left(z|x_s\right)\|p(z)\right]^+\\ \mathcal{L}_D &= \sum_{s=r,g}\left[KL\left(q_\phi\left(z|x_s\right)\|p(z)\right)\right].\end{aligned} \tag{2}$$

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

- The loss function of Soft-IntroVAE

$$\begin{aligned}\mathcal{L}_E &= ELBO(x) - \frac{1}{\alpha}\sum_{s=r,g}\exp\left(\alpha ELBO\left(x_s\right)\right)\\ \mathcal{L}_D &= ELBO(x) + \gamma\sum_{s=r,g} ELBO\left(x_s\right)\end{aligned} \tag{3}$$

  where α, γ are both hyperparameters.

Problems

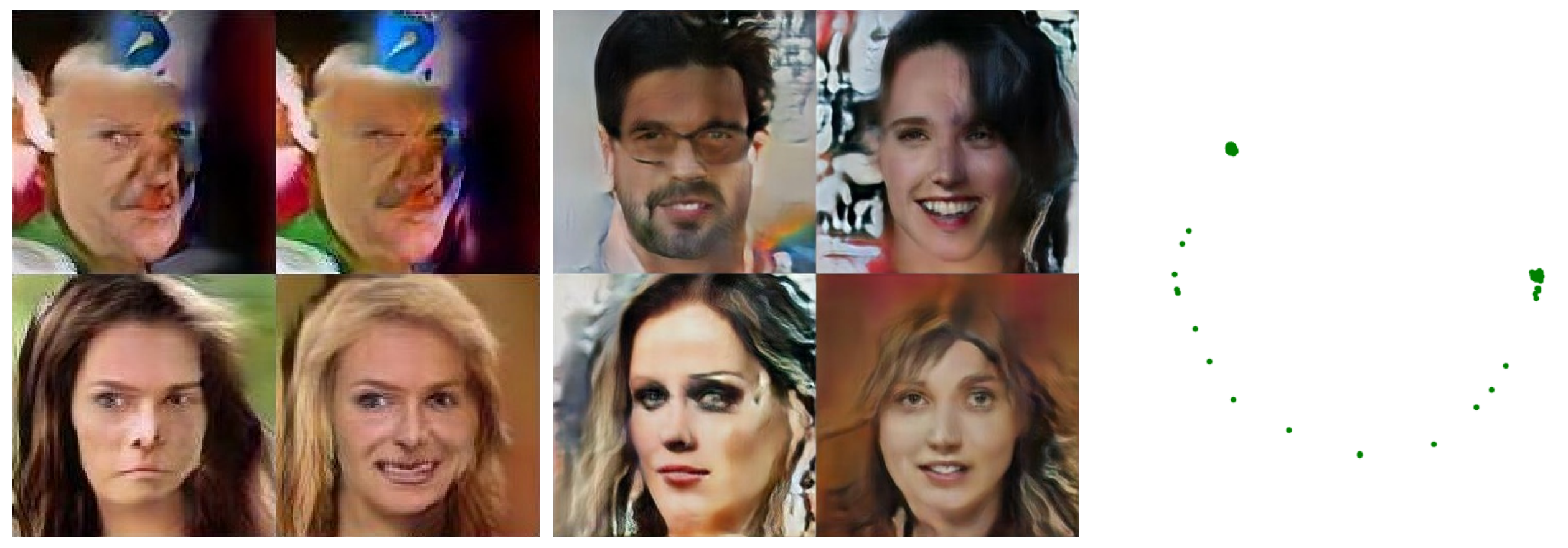


Fig. 1: Image/Numerical generation results of Soft-IntroVAE (from left to right: CelebA128, CelebA256, 8 Gaussian). Drawbacks: vanishing gradient and mode collapse

Proposed Method and Quantitative Results

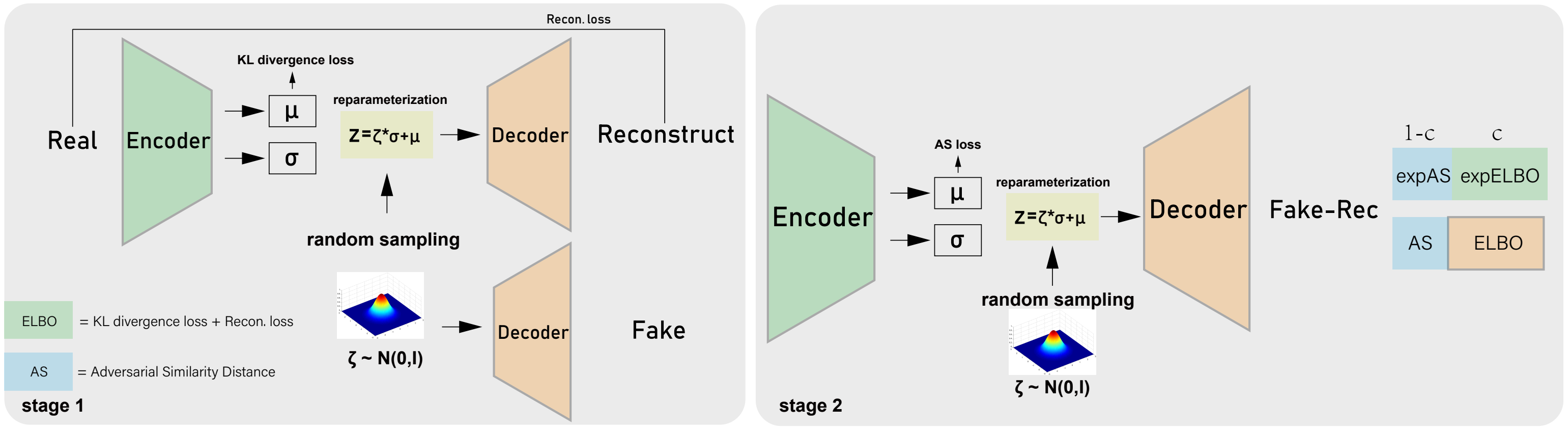


Fig. 2: AS-IntroVAE workflow

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.

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 Wasserstein distance and the distance metrics in unsupervised domain adaptation,

$$D\left(p_r, p_g\right) = \mathbb{E}_{x\sim p(z)}[\left(\mathbb{E}_{x\sim p_r}\left[q\left(z|x\right)\right] - \mathbb{E}_{x\sim p_g}\left[q\left(z|x\right)\right]\right)]^2 \tag{4}$$

where p_r is distribution of real data, p_g is distribution of generated data.

Since the latent space is a normal distribution. This kernel k can be deduced as

$$k\left(x_r^i, x_g^j\right) = \frac{-\frac{1}{2}\frac{\left(u_r^i - u_g^j\right)^2}{\lambda_r^i + \lambda_g^j}}{(2\pi)^{\frac{n}{2}}\cdot\left(\lambda_r^i + \lambda_g^j\right)^{\frac{1}{2}}} \tag{5}$$

where u, λ represent the variational inference on the mean and variance of x, i, j represent the ith, jth pixel in images. we derive the loss function for AS-IntroVAE as:

$$\begin{aligned}\mathcal{L}_{E_\phi} &= ELBO(x) - \frac{1}{\alpha}\sum_{s=r,g}\exp(\alpha(\mathbb{E}_{q(z|x_s)}[\log p(x \mid z)]\\ &\quad + cKL(q_\phi(z|x_s)\|p(z)) + (1-c)D(p_r, p_g)))\\ \mathcal{L}_{D_\theta} &= ELBO(x) + \gamma\sum_{s=r,g}(\mathbb{E}_{q(z|x_s)}[\log p(x \mid z)]\\ &\quad + cKL(q_\phi(z|x_s)\|p(z)) + (1-c)D(p_r, p_g))\end{aligned} \tag{6}$$

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

		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**

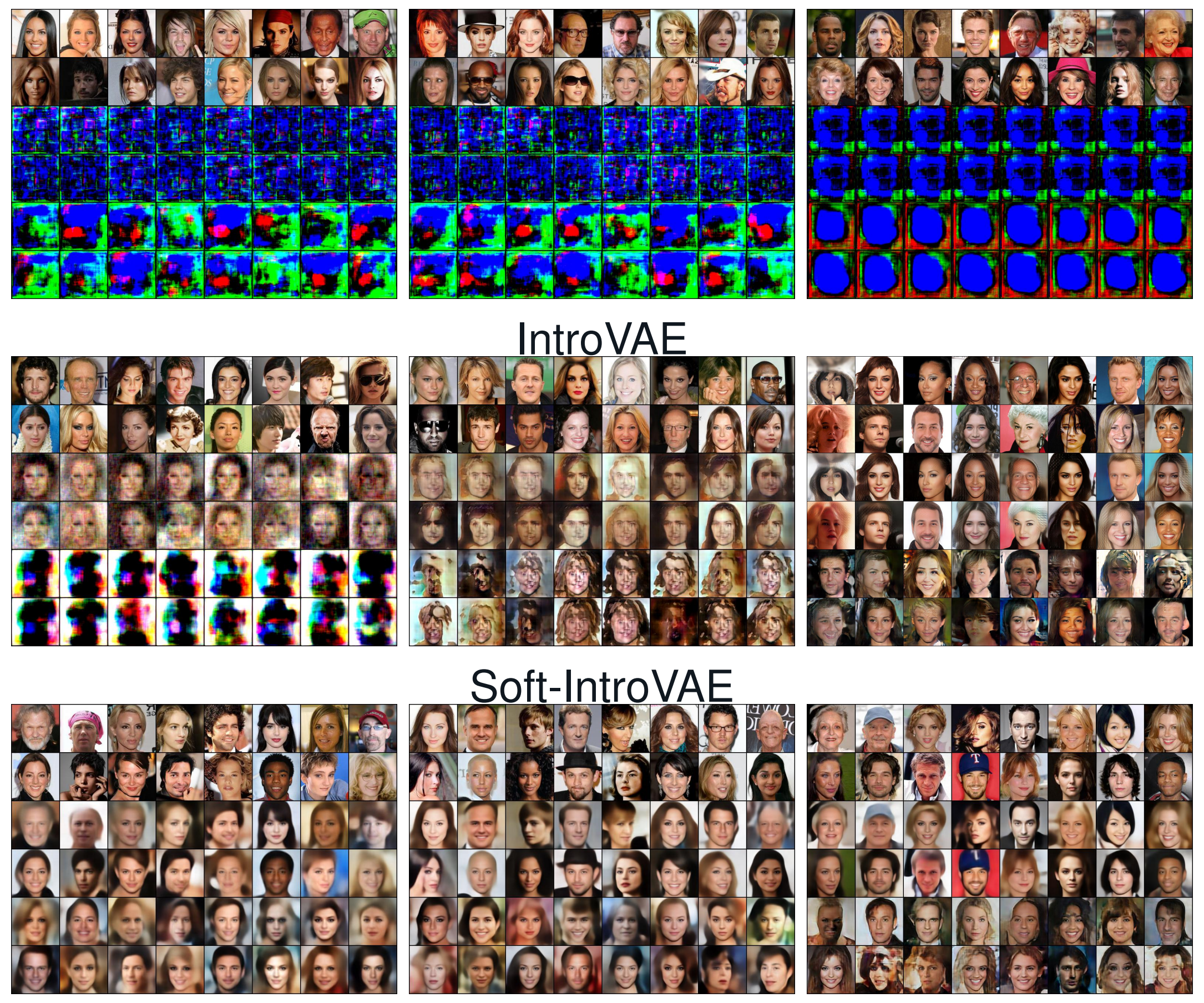
2D Toy Dataset 8 Gaussians Score KL↓/JSD↓ Table.

	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**

Image Generation FID Score↓ Table.

Qualitative comparisons

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.



AS-IntroVAE

The additional visual comparison for image generation tasks. From left to right: WGAN-GP, S-IntroVAE, AS-IntroVAE.

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

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

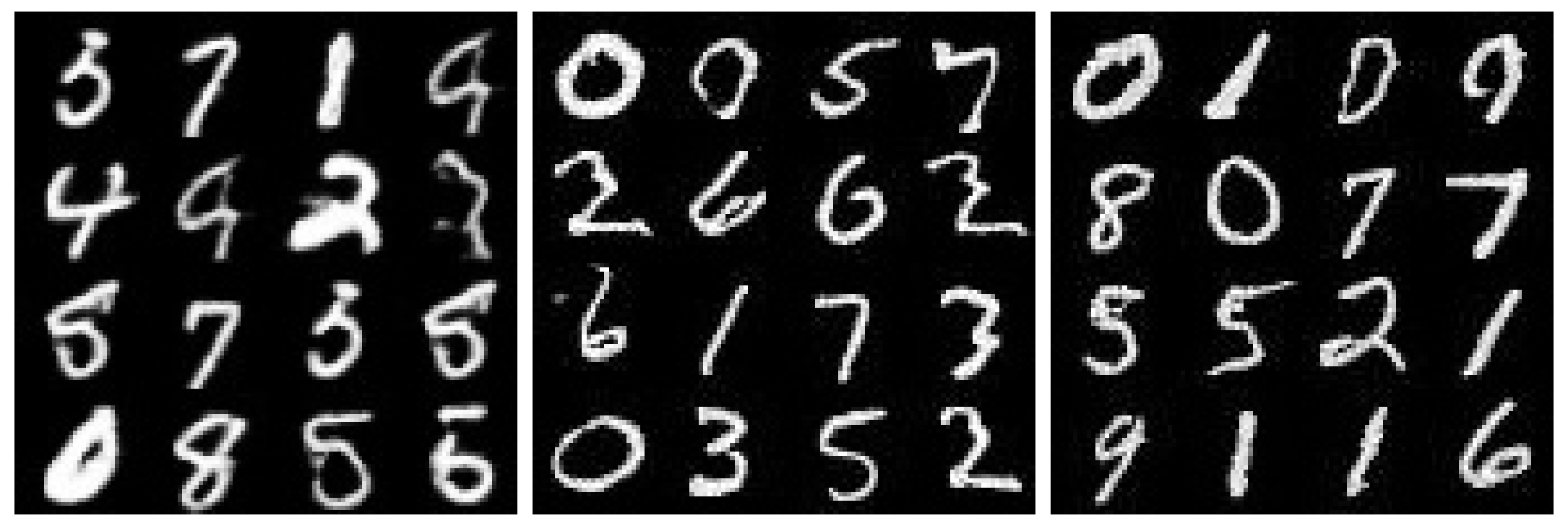


Image generation visual comparisons at MNIST dataset (resolution: 28 × 28).