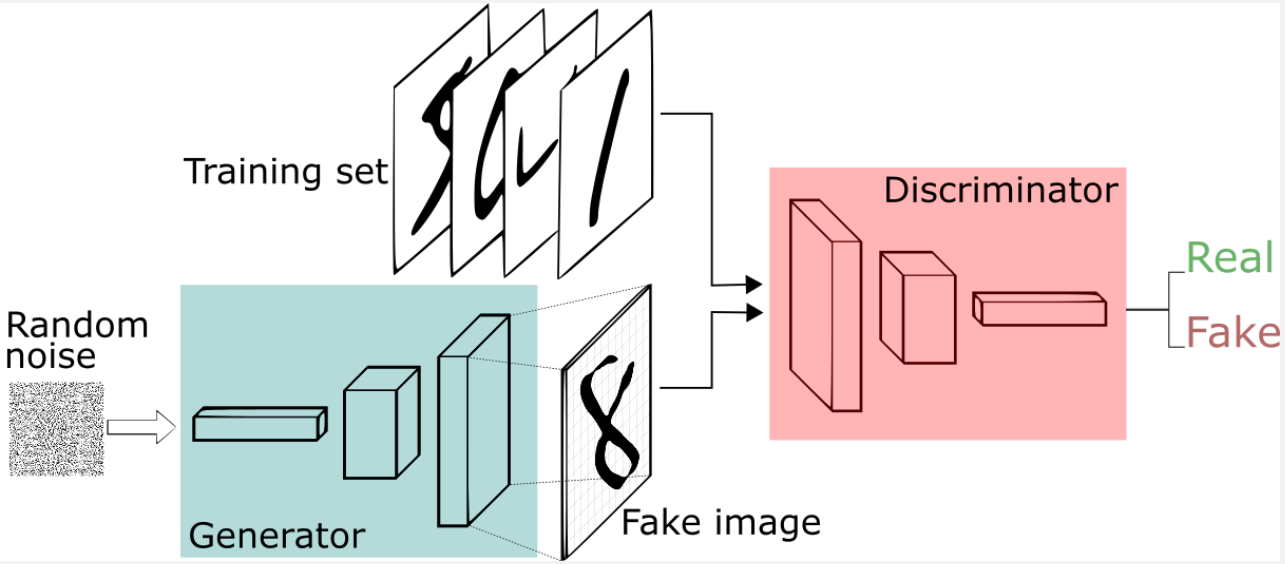
The loss of GANs

$$\arg\min_{\theta_G}\max_{\theta_D}\mathbb{E}_{x\sim\mathcal{P}_{data}}[\log D(x)]+\mathbb{E}_{z\sim\mathcal{P}_{noise}}[\log(1-D(G(z)))] \qquad (1)$$

We denote the generator as G and the discriminator as D with training image $x \sim \mathcal{P}_{data}$ following the distribution of training images and random noise $z \sim \mathcal{P}_{noise}$. The generator is parametrized by θ_G and the discriminator is parametrized by θ_D.



Neural Tangent Kernel(NTK)

According to the paper Neural Tangent Kernel: Convergence and Generalization in Neural Networks(by Jacot, NIPS'18), the Neural Tangent Kernel(NTK) over the training dataset $\mathbf{X}$ corresponding to the architecture of the neural network $f(\cdot;\theta)$ is defined as

$$\mathbf{K} = \nabla_\theta f(\mathbf{X};\theta)^\top \nabla_\theta f(\mathbf{X};\theta) \qquad (2)$$

Let $\mathbf{K}^{n,n} \in \mathbb{R}^{n\times n}$ be the kernel matrix for $\mathbf{X}^n$, i.e.,$\mathbf{K}^{n,n}_{i,j} = k(\mathbf{X}^n_{i,:}, \mathbf{X}^n_{j,:})$. **The mean prediction of the ensemble of the neural network $f(\cdot;\theta)$ after training t steps gradient descent can be approximated by the mean prediction of the NTK-GP with corresponding kernel.** The mean prediction of the NTK-GP over $\mathbf{X}^n$ evolves as

$$(\mathbf{I}^n - e^{-\eta \mathbf{K}^{n,n} t})\mathbf{Y}^n \qquad (3)$$

where $\mathbf{I}^n \in \mathbb{R}^{n\times n}$ is an identity matrix and η is a sufficiently small learning rate.

Our Method: GA-NTK

Let $\mathbf{K}^{2n,2n} \in \mathbb{R}^{2n \times 2n}$ be the kernel matrix for $\mathbf{X}^n \oplus \mathbf{Z}^n$, where the value each element $\mathbf{K}^{2n,2n}_{i,j} = k((\mathbf{X}^n \oplus \mathbf{Z}^n)_{i,:}, (\mathbf{X}^n \oplus \mathbf{Z}^n)_{j,:})$. Let $\lambda = \eta \cdot t$ The discriminator can be written as:

$$D(\mathbf{X}^n, \mathbf{Z}^n; k, \lambda) = \underbrace{(\mathbf{I}^{2n} - e^{-\lambda \mathbf{K}^{2n,2n}})}_{NTK-GP}(\mathbf{1}^n \oplus \mathbf{0}^n) \in \mathbb{R}^{2n} \qquad (4)$$

where $\mathbf{I}^{2n} \in \mathbb{R}^{2n \times 2n}$ is an identity matrix. We formulate the objective of GA-NTK as follows:

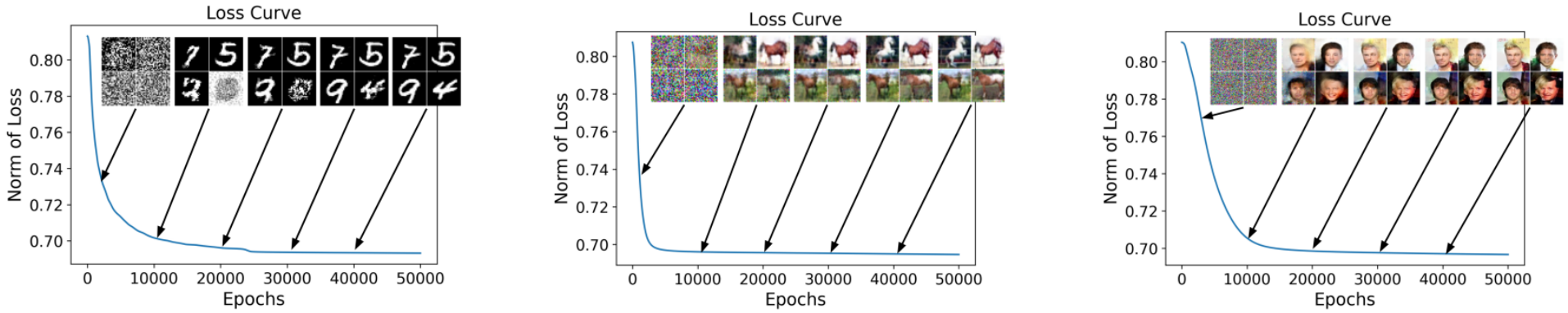
$$\arg\min_{\mathbf{Z}^n} ||\mathbf{1}^{2n} - D(\mathbf{X}^n, \mathbf{Z}^n; k, \lambda)||_2 \qquad (5)$$

where $\mathbf{1}^{2n} \in \mathbb{R}^{2n}$ is a vector of ones. Then, we update the fake image at the last time step

$$\mathbf{Z}^{n+1} = \mathbf{Z}^n + \alpha \nabla_{\mathbf{Z}^n} ||\mathbf{1}^{2n} - D(\mathbf{X}^n, \mathbf{Z}^n; k, \lambda)||_2 \qquad (6)$$

Our Contribution: Solved Failure to Convergence

In our experiment, the image quality is consistent with the loss. As the **loss decreases, the image quality is improved**.

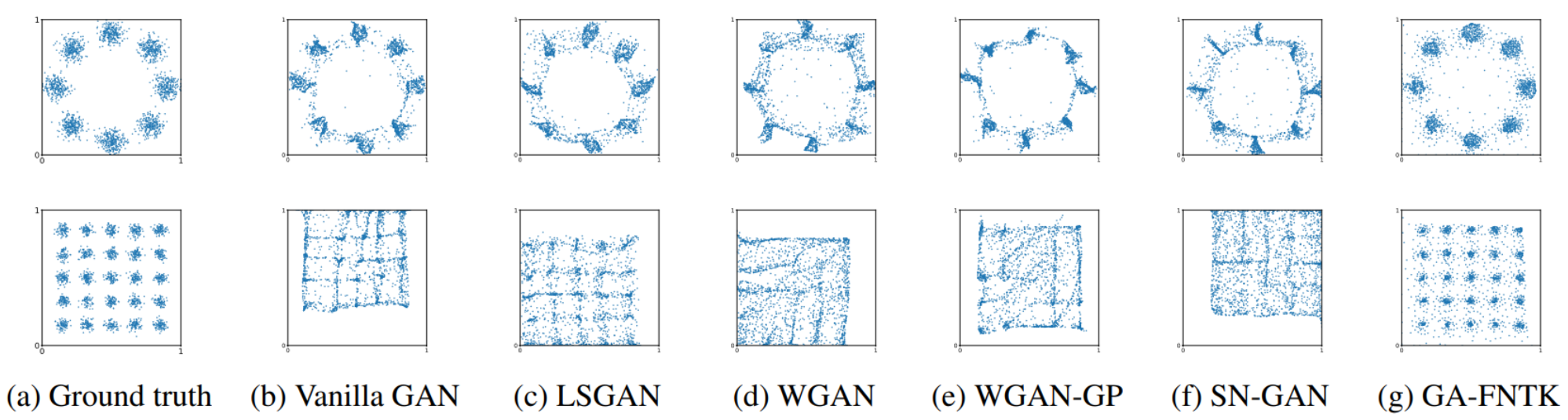


(a) MNIST (b) CIFAR-10 (c) CelebA

Figure 3. Loss curve and image quality at different iterations on different dataset. We can see that image quality issues correlated with loss curve.

Mode Collapse

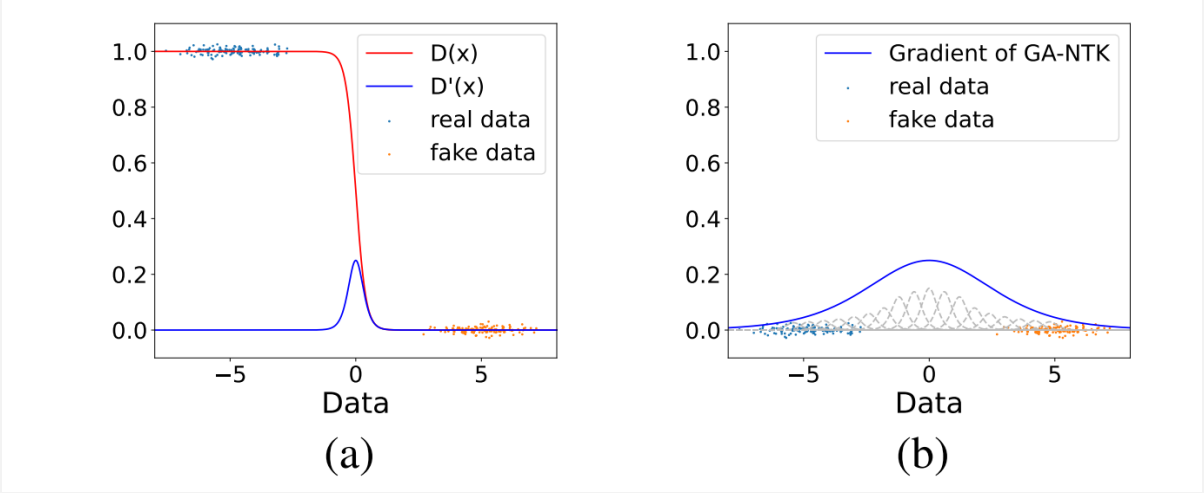
Our method(**GA-FNTK**) **fits to multi-modal distribution perfectly**, rather than other baseline GANs.



Our Contribution: Solved Gradient Vanish

In our experiment, decision boundary of the discriminator and the corresponding gradients for the generator $\mathbf{Z}^n$ in (a) GANs and (b) GA-NTK on 1D toy dataset. The blue points are real data points and the red ones are fake. The discriminator output 1 for the real data points and 0 for the fake ones(fig (a) red line). The **gradients(fig (a) blue line) vanishes while the the data points are far away from the decision boundary**.

The **gray dashed lines indicate the gradients for $\mathbf{Z}^n$ from different element networks of the ensemble discriminator in GA-NTK**.



Compare GA-NTK & GANs Baseline

FID: The **lower**, the **better image quality**

AM-SSIM: The **lower**, the **better creativity**

Our method outperform than other widely-used GANs in small dataset

			GA-NTK	DCGAN	LSGAN	WGAN	WGAN-GP	SN-GAN
MNIST	dataset size = 64	FID	31.10	27.43	69.76	50.69	32.49	57.89
		AM-SSIM	0.49	0.84	0.79	0.77	0.83	0.67
	dataset size = 128	FID	21.14	31.89	38.52	49.28	30.20	38.33
		AM-SSIM	0.52	0.85	0.80	0.74	0.76	0.67
	dataset size = 256	FID	14.96	69.76	35.33	50.33	24.37	29.49
		AM-SSIM	0.54	0.69	0.78	0.72	0.73	0.70
CIFAR-10	dataset size = 64	FID	55.54	312.21	258.41	117.85	49.29	118.16
		AM-SSIM	0.41	0.22	0.25	0.29	0.74	0.28
	dataset size = 128	FID	39.98	229.94	339.27	101.90	68.53	128.65
		AM-SSIM	0.41	0.36	0.10	0.26	0.60	0.21
	dataset size = 256	FID	28.40	181.15	255.19	111.92	85.34	107.29
		AM-SSIM	0.42	0.27	0.22	0.22	0.46	0.20
CelebA	dataset size = 64	FID	30.83	489.82	83.71	122.36	83.71	169.04
		AM-SSIM	0.60	0.02	0.05	0.29	0.56	0.29
	dataset size = 128	FID	33.51	55.01	450.81	125.82	92.73	168.11
		AM-SSIM	0.51	0.03	0.11	0.28	0.54	0.28
	dataset size = 256	FID	63.15	461.95	403.79	108.07	79.36	161.20
		AM-SSIM	0.38	0.04	0.09	0.31	0.39	0.27

We use only **256** training images on CelebA, CIFAR10, MNIST datasets.

(a) GA-CNTK

(b) DCGAN

(c) LSGAN

(d) WGAN

(e) WGAN-GP

(f) SN-GAN

Figure 2. Generated samples from MNIST, CIFAR10, and CelebA. The sizes of the training datasets are 256.