BourGAN: Generative Networks with Metric Embeddings

Chang Xiao, Peilin Zhong, Changxi Zheng
Columbia University

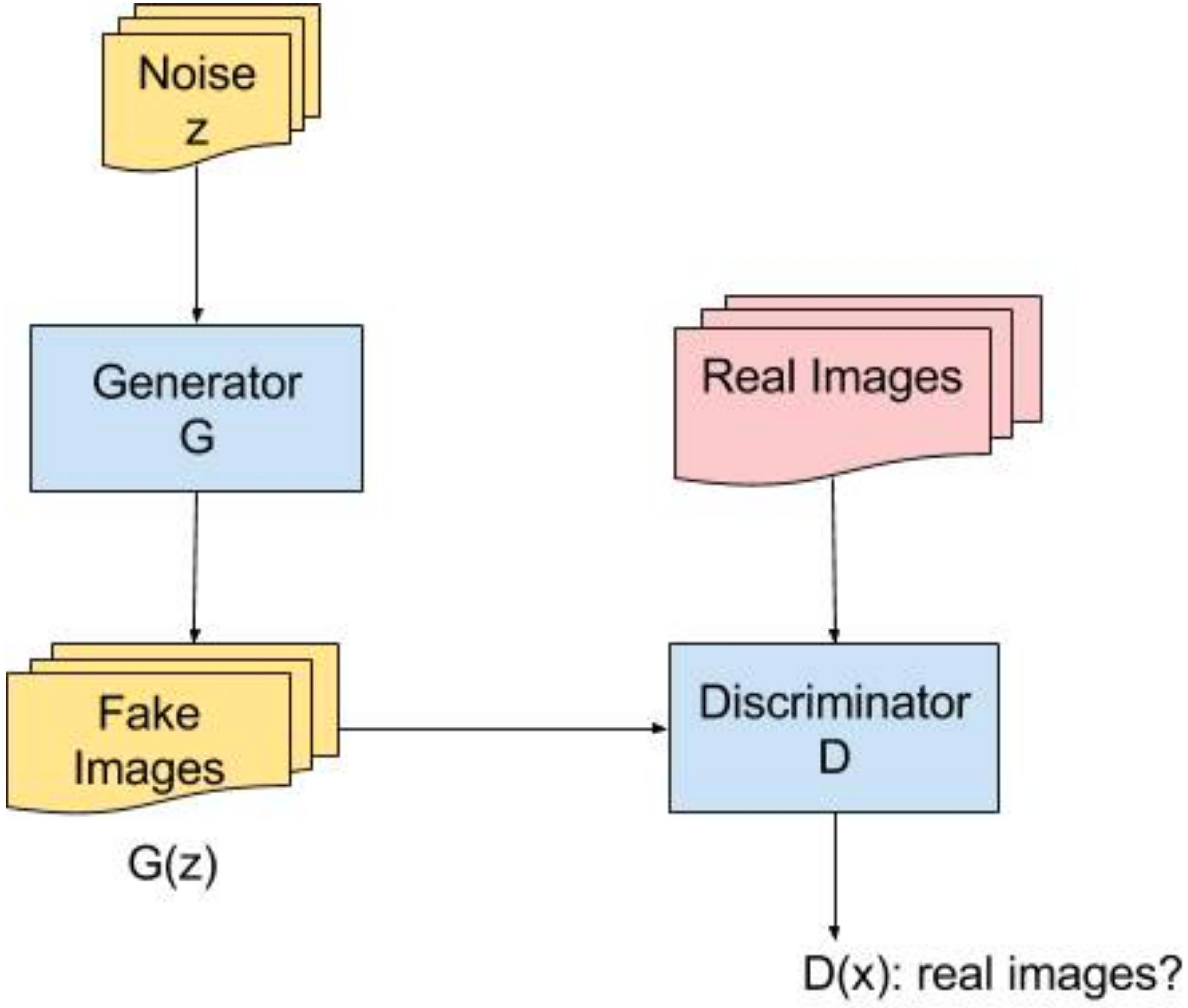
Slides compiled by Hongming Shan

Why it called “BourGAN”

- Jean Bourgain, Belgian mathematician
- Bourgain’s Theorem, explained later

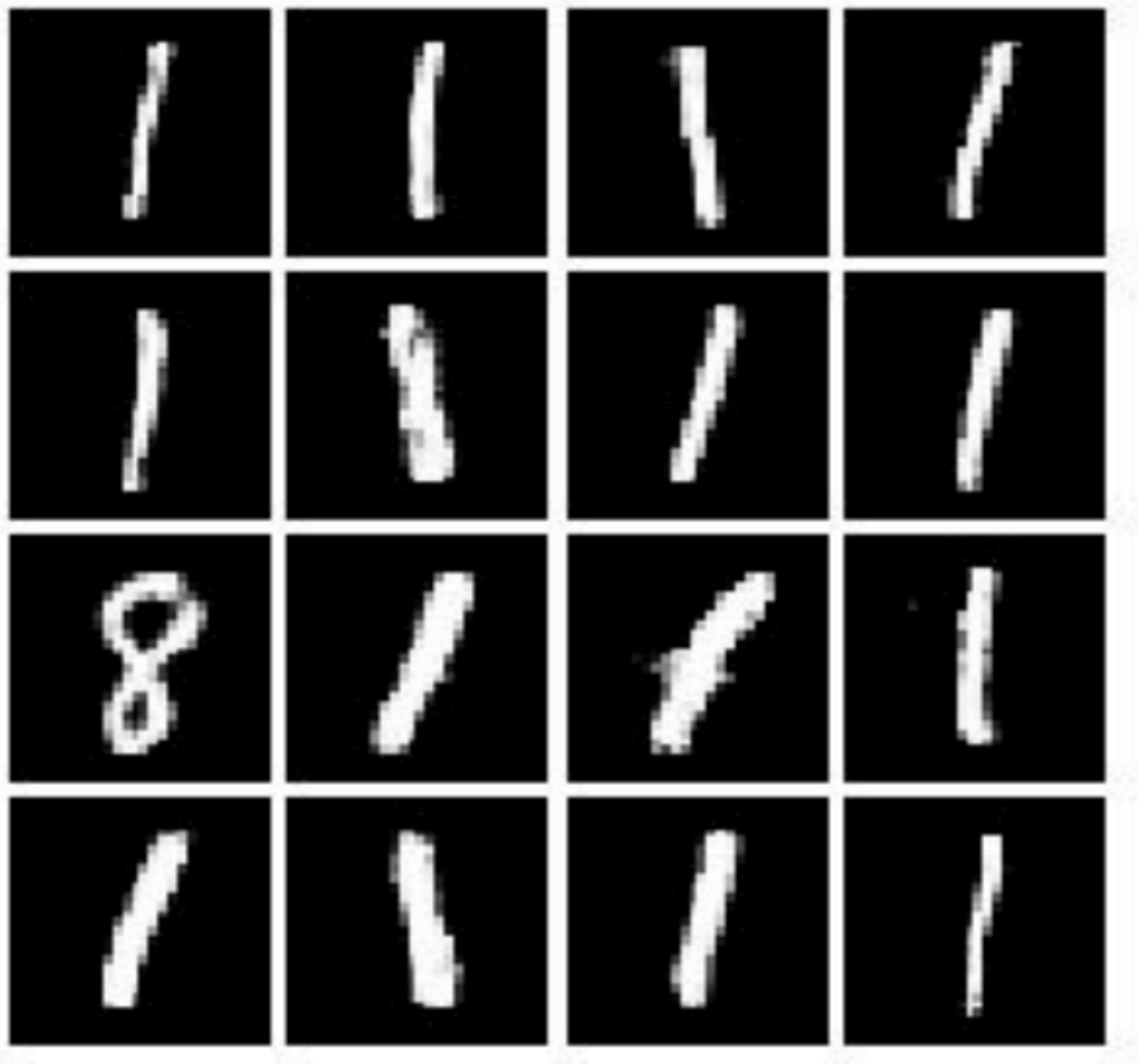
Generative adversarial network

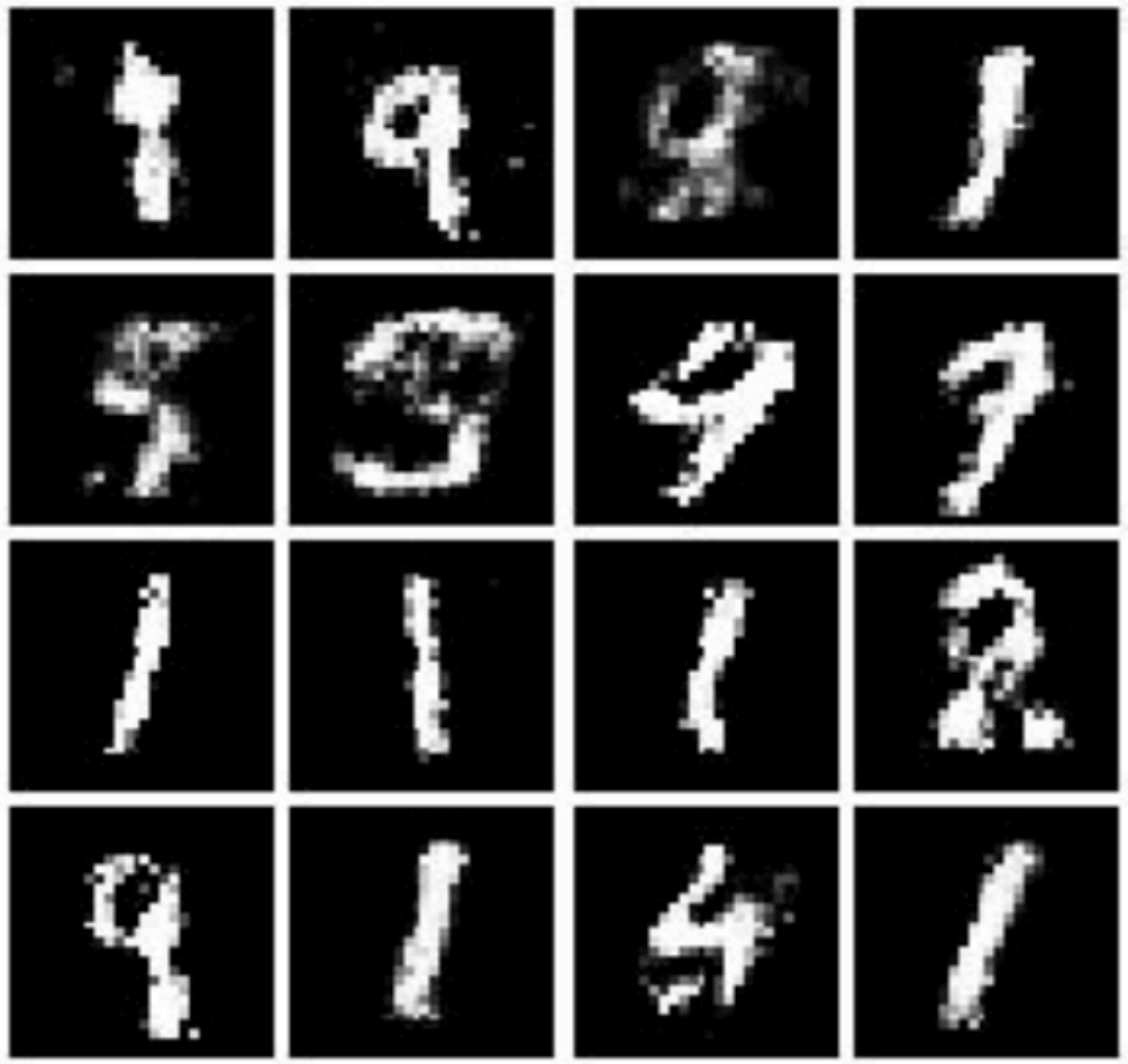
- Generator takes the noise and generates fake images
- Discriminator receives samples from both the generator and real data, and tries to distinguish between two sources.
- The Competition prices the generated samples to be hardly indistinguishable from real data.



Mode Collapse in GAN

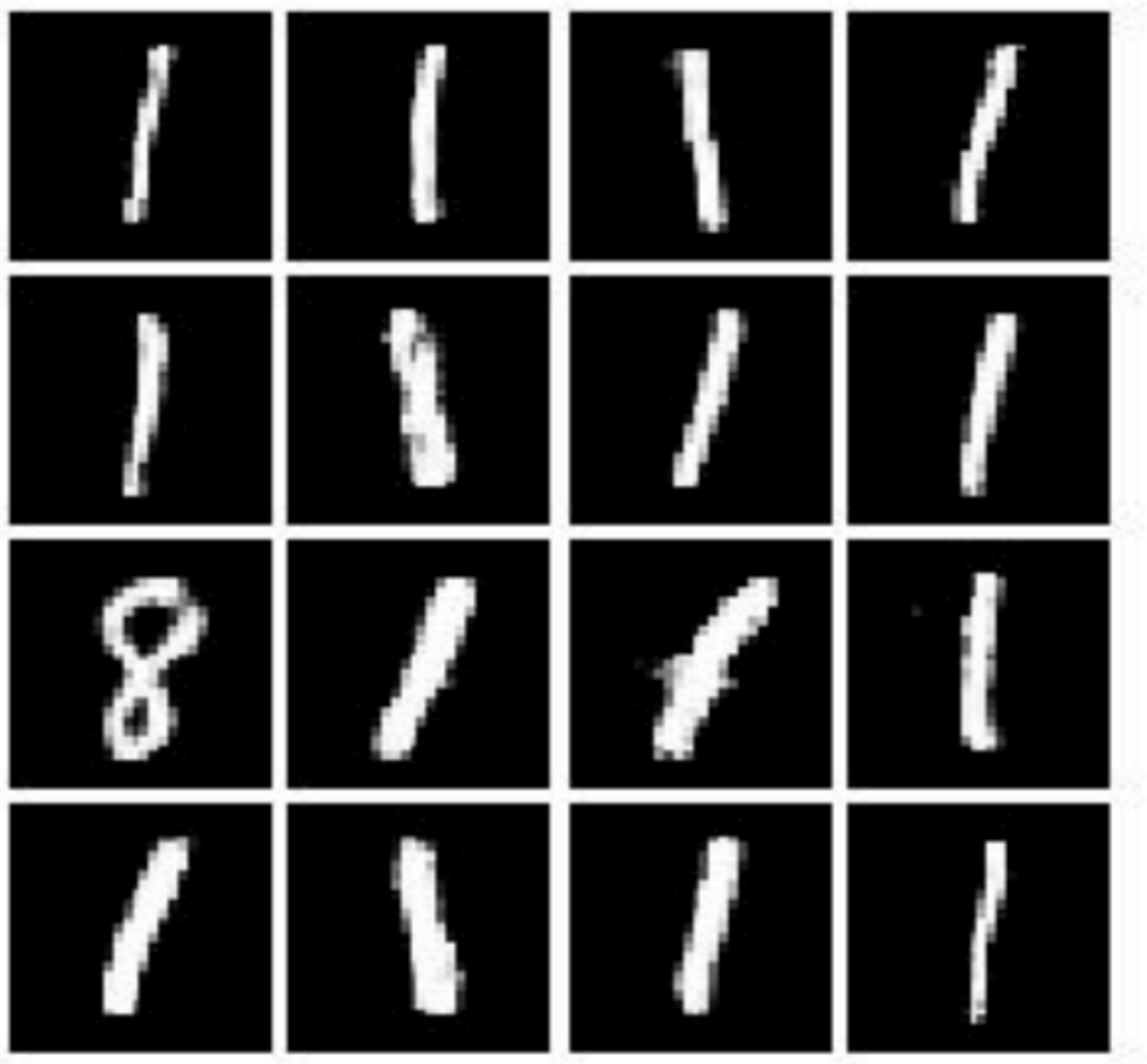
When we train a GAN on MNIST (a digital dataset from 0 to 9)

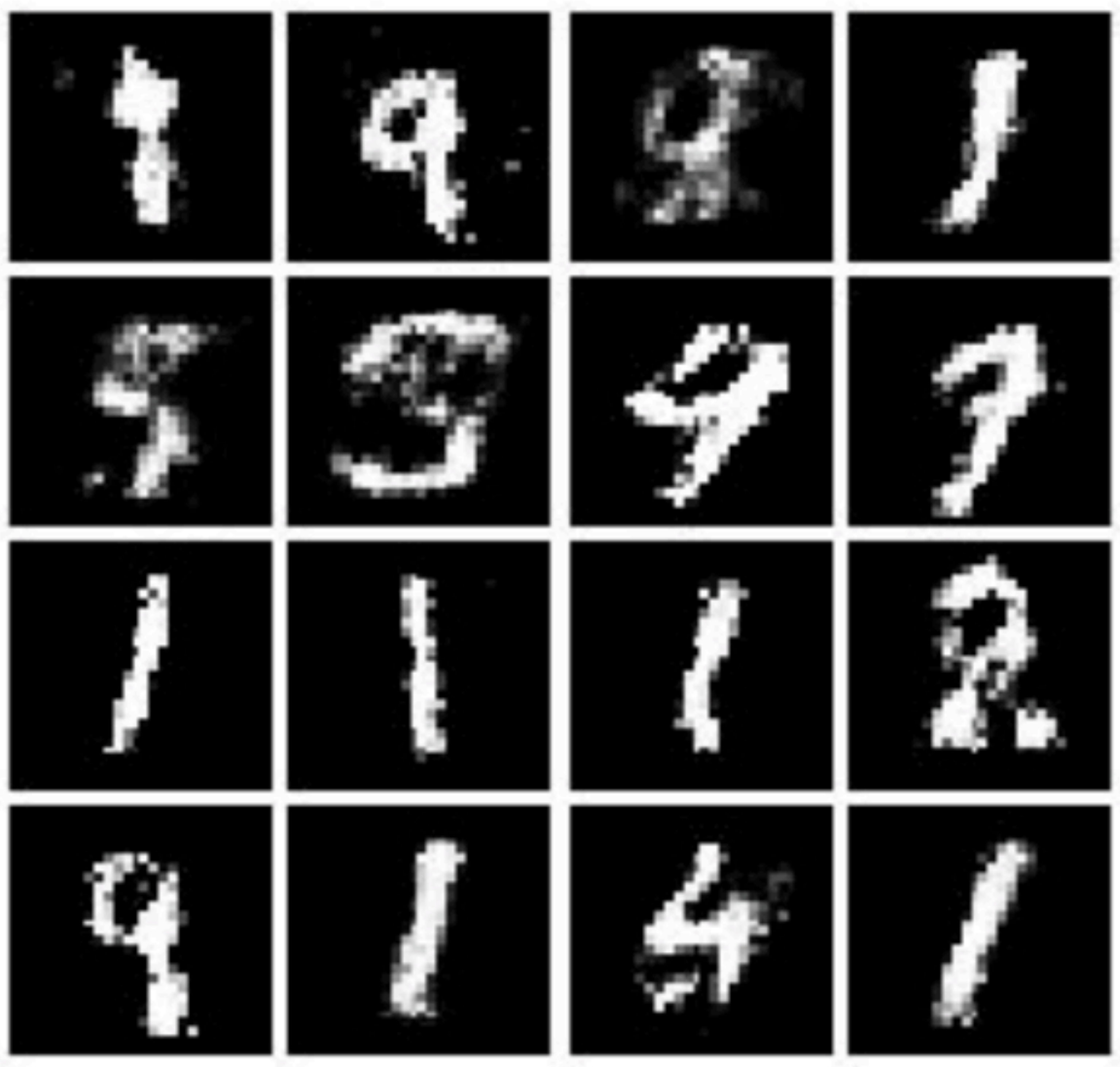




Mode Collapse in GAN

When we train a GAN on MNIST (a digital dataset from 0 to 9)

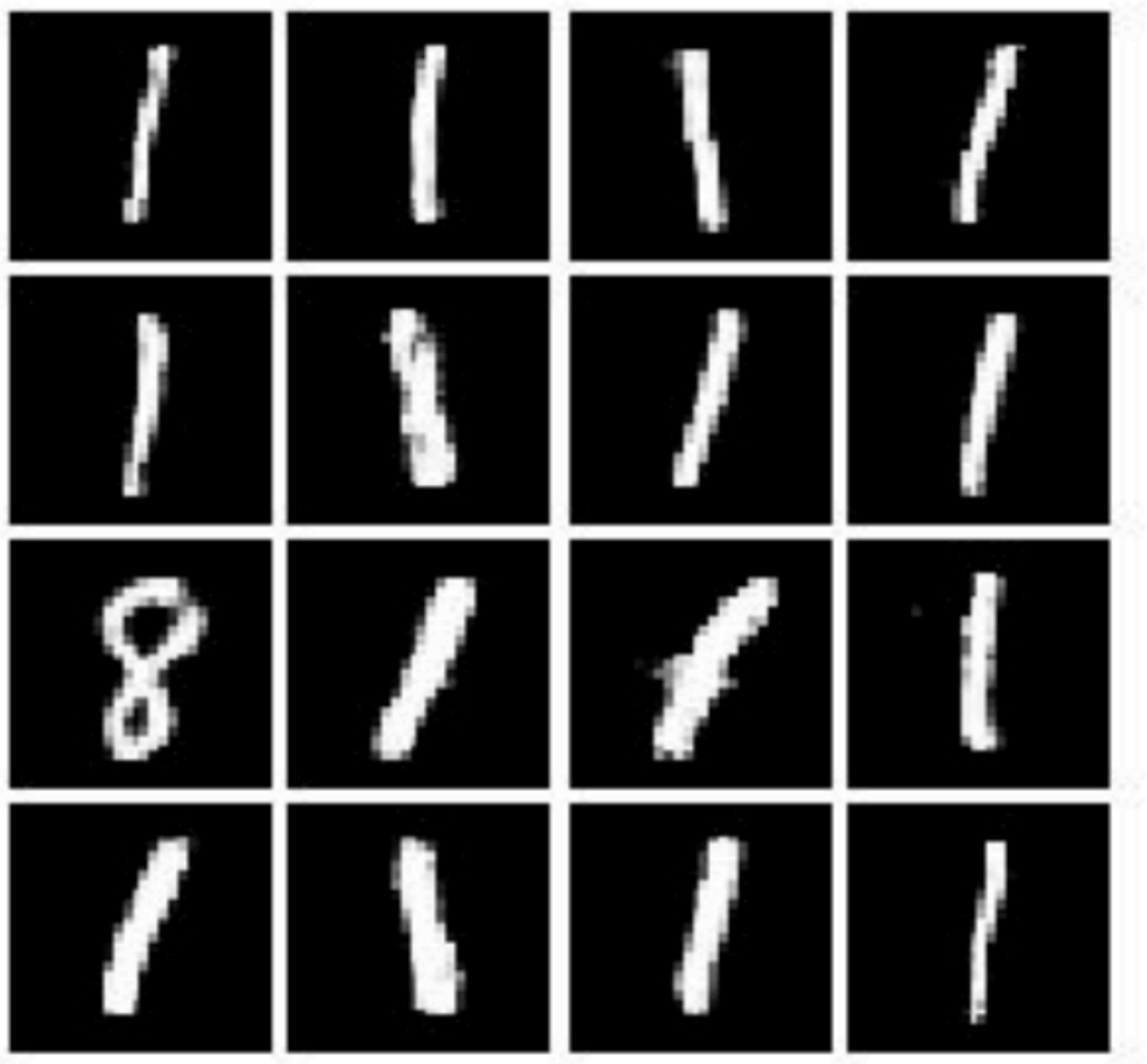




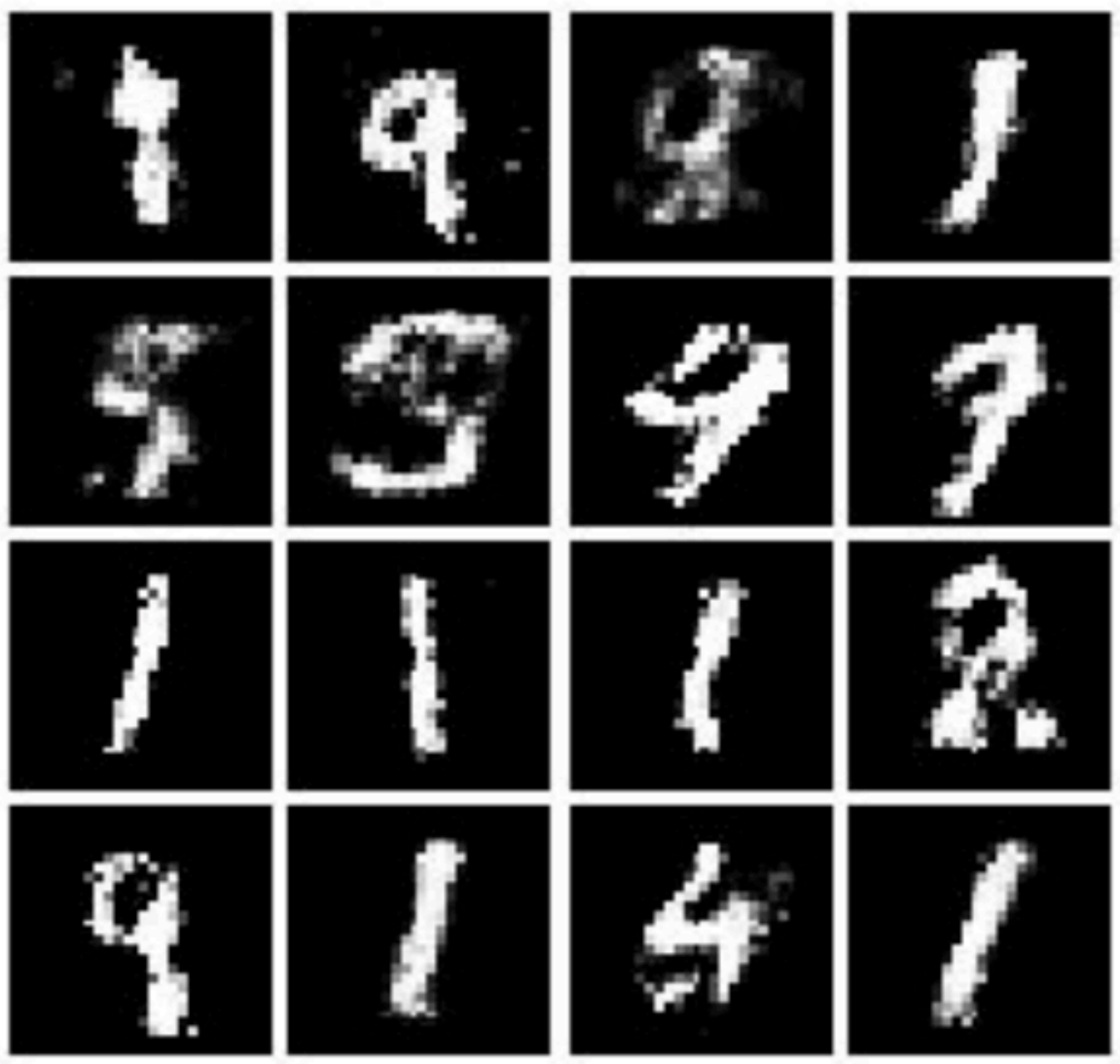
Mode Collapse

Mode Collapse in GAN

When we train a GAN on MNIST (a digital dataset from 0 to 9)

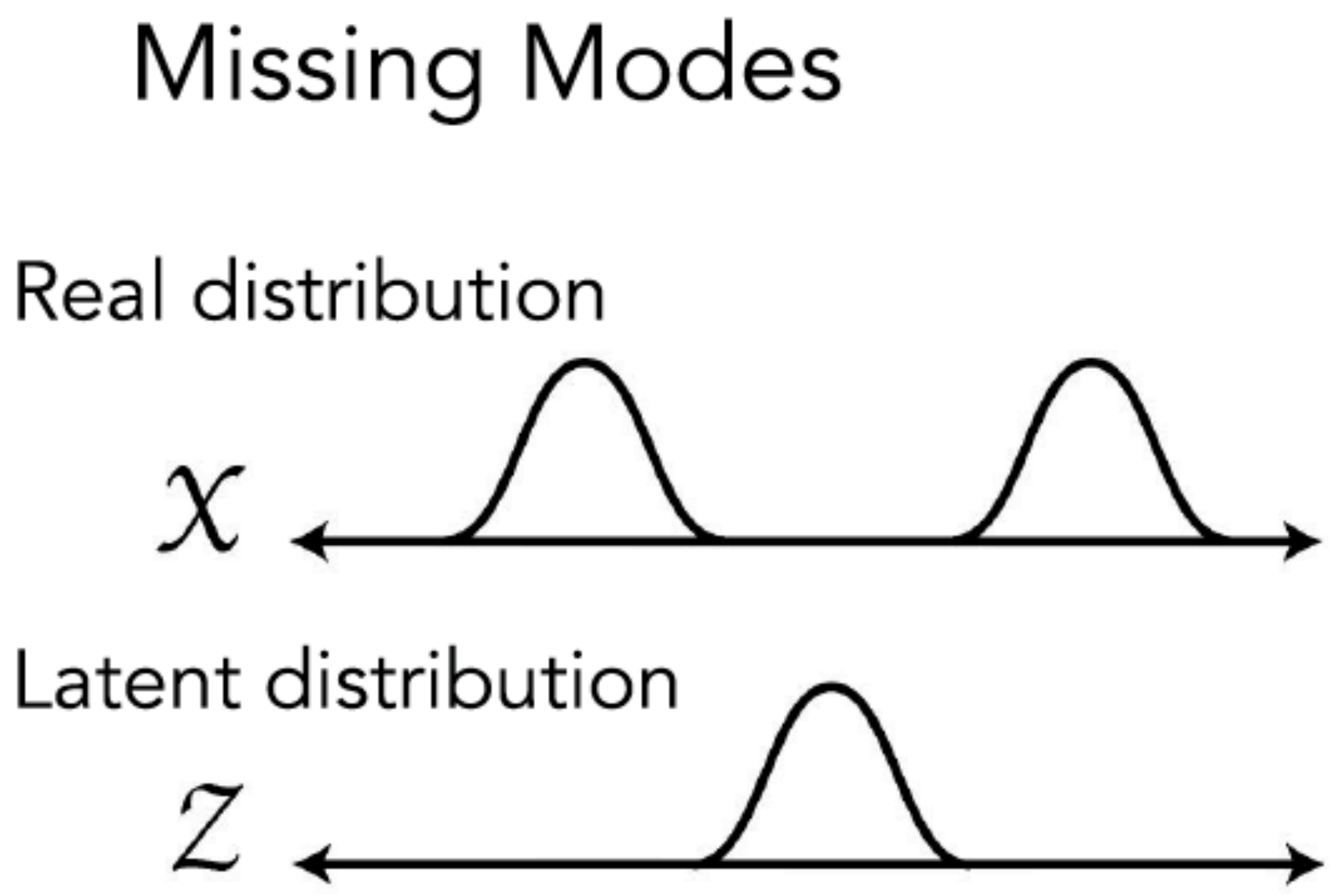


Mode Collapse

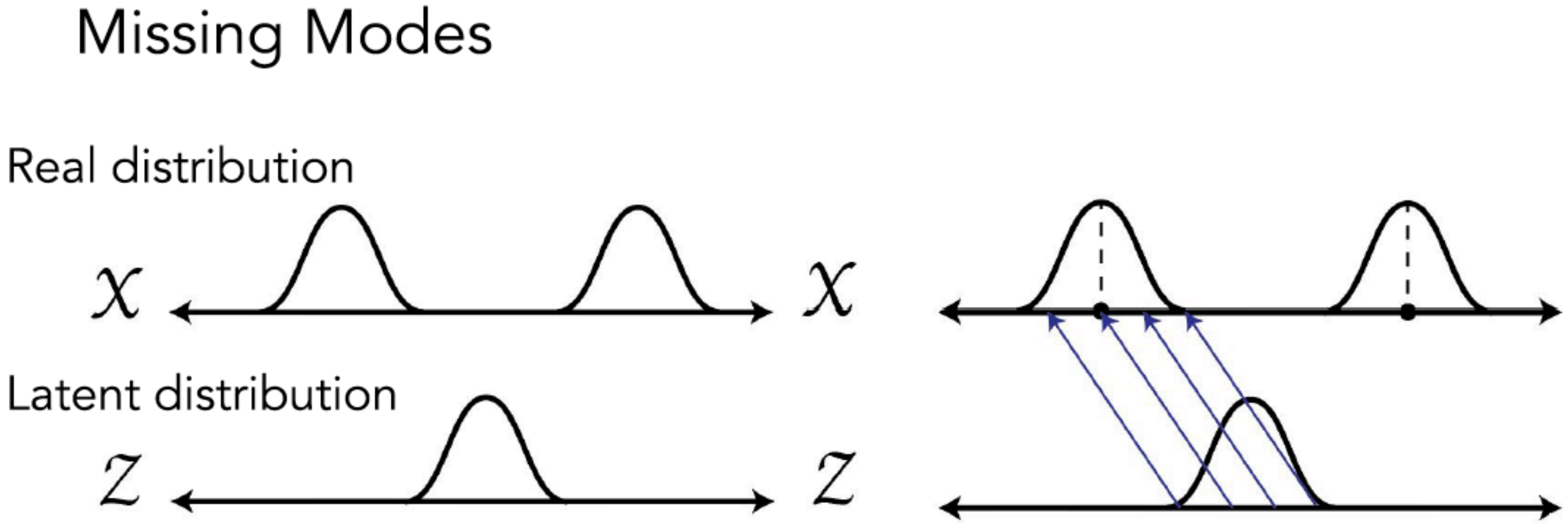


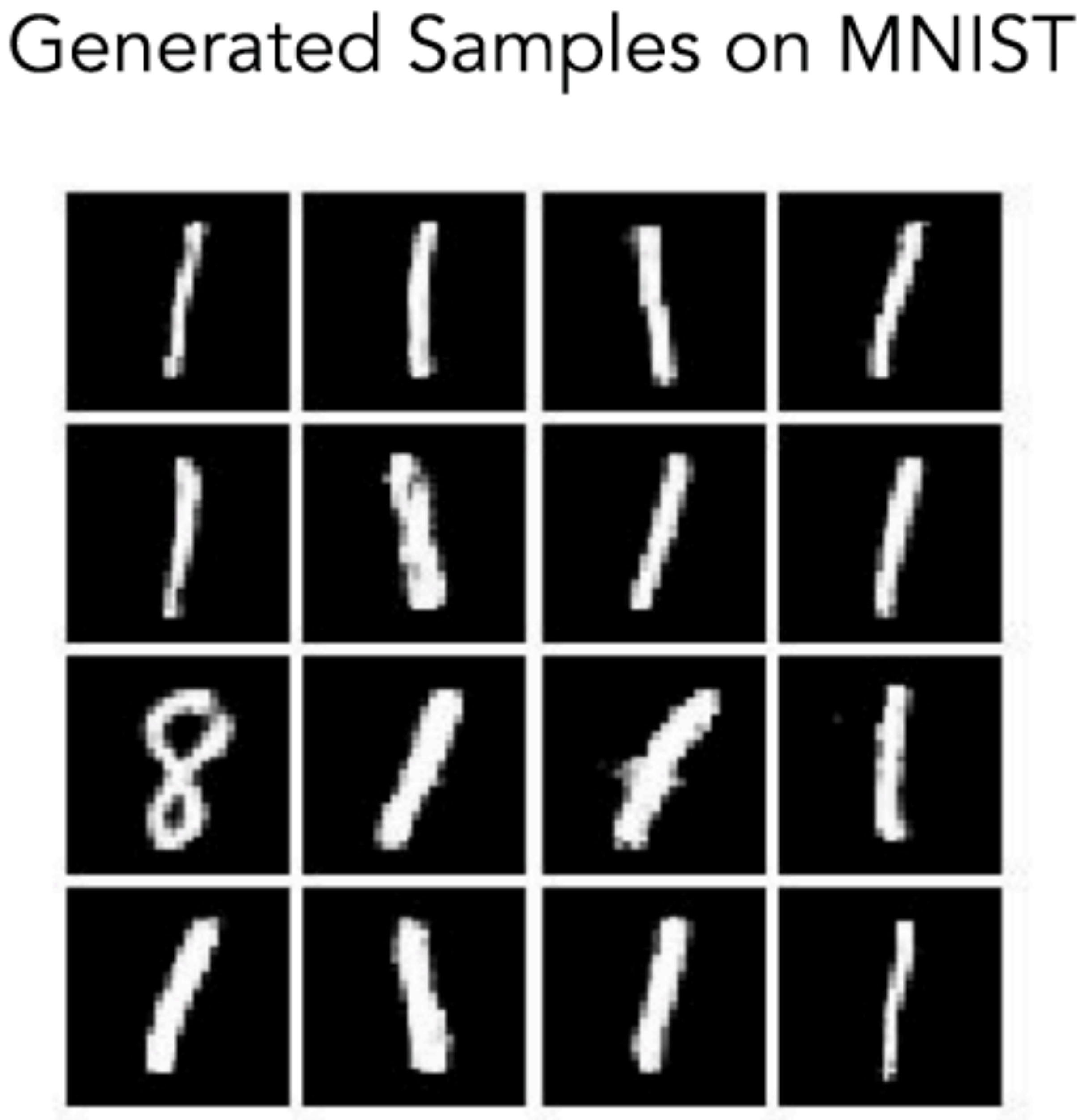
Unwanted data

Why would mode collapse happen?



Why would mode collapse happen?

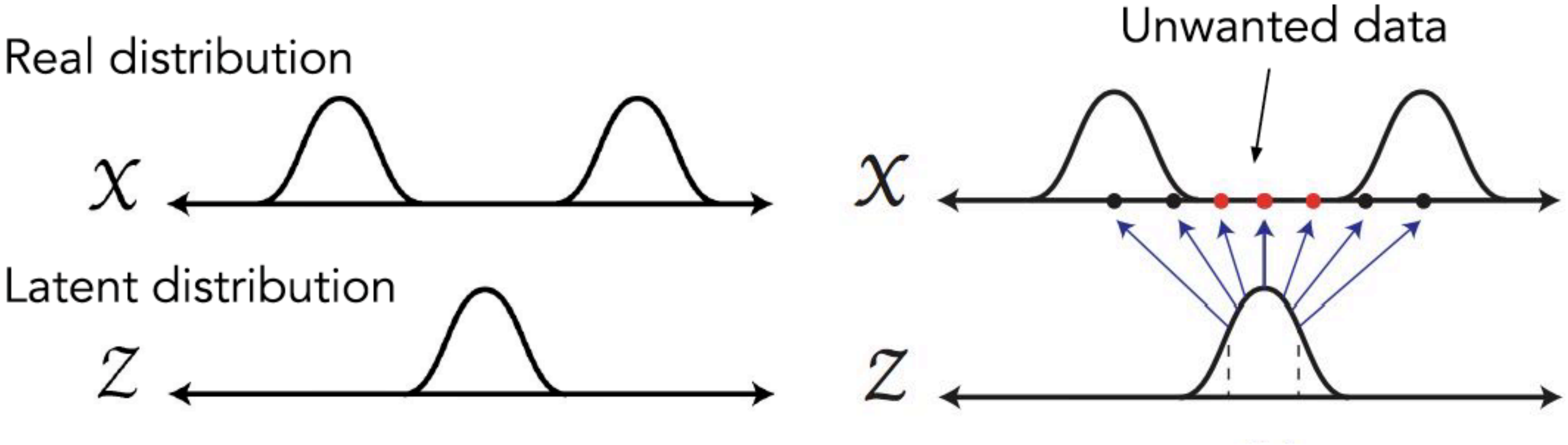




The discriminator can be fooled by generating a subset of data from real distribution.

Why would mode collapse happen?

Unwanted Data



Generated Samples on MNIST

Unwanted data between two modes might be generated.

Large gradients cause the network unstable and hard to train.

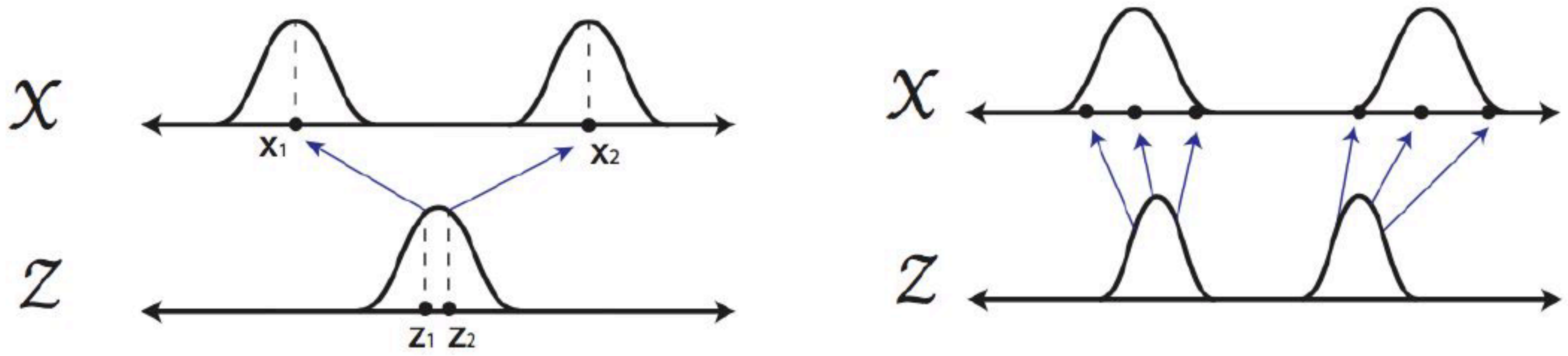
How can we avoid mode collapse and unwanted data?

New latent space

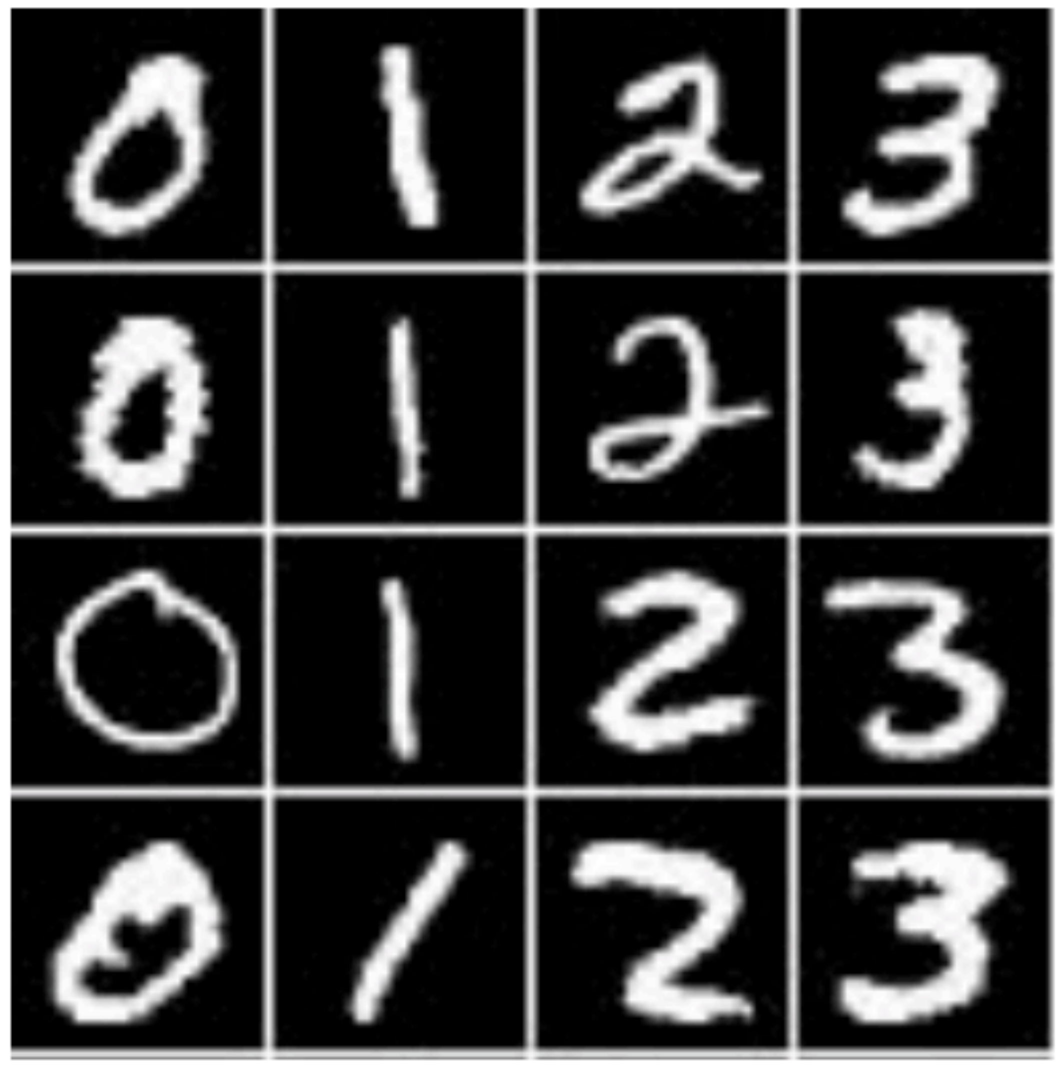
How can we avoid mode collapse and unwanted data?

Solution: Gaussian Mixture model

Sampling from Gaussian mixture model



Generated Samples on MNIST



No missing mode 😀

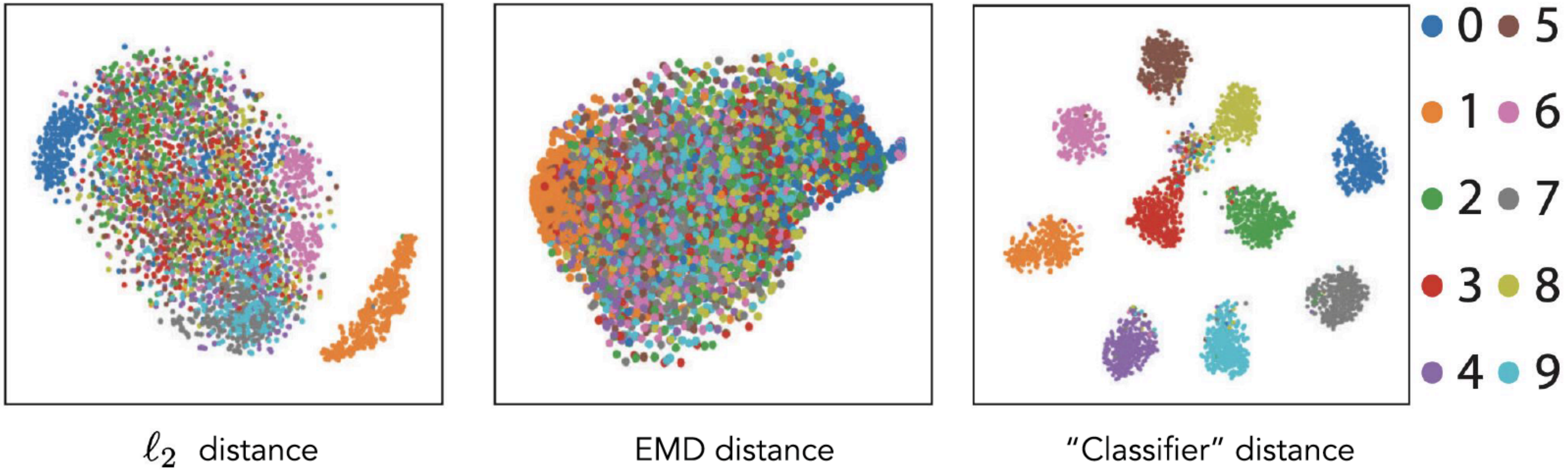
No unwanted data 😀

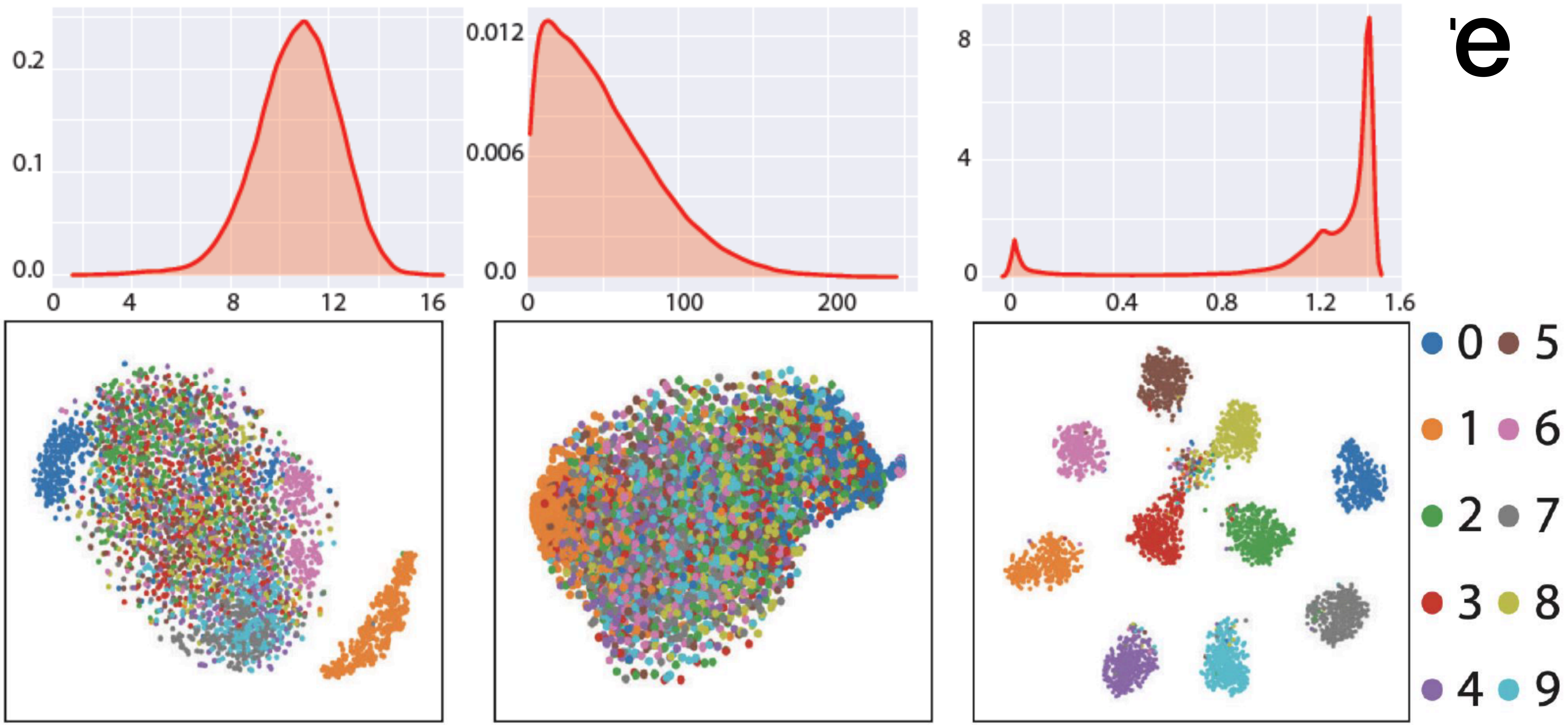
How to construct the Gaussian mixture?

Mode: A Geometric Structure

Modes are defined on a certain distance metric.

t-SNE visualization of MNIST for different metric





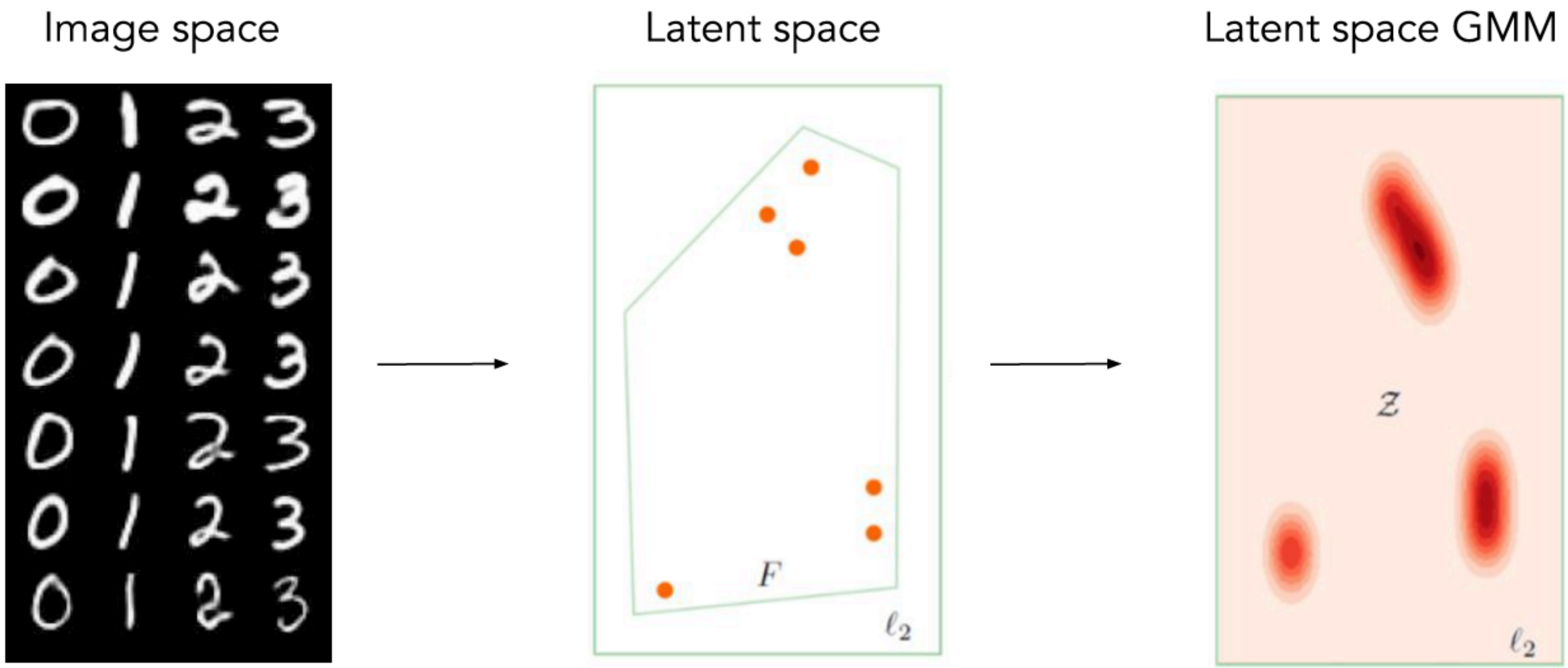
ℓ_2 distance

EMD distance

"Classifier" distance

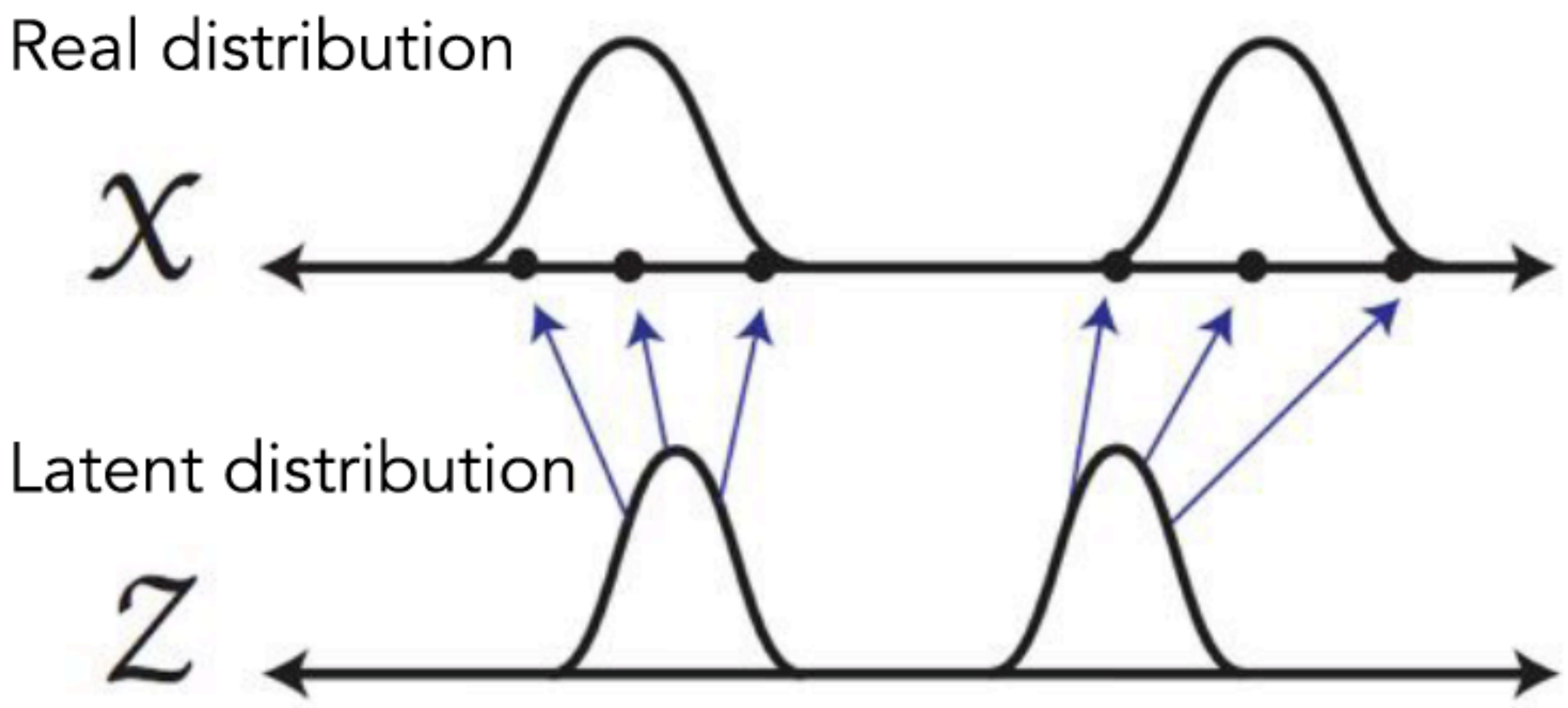
Locating Gaussian Centers

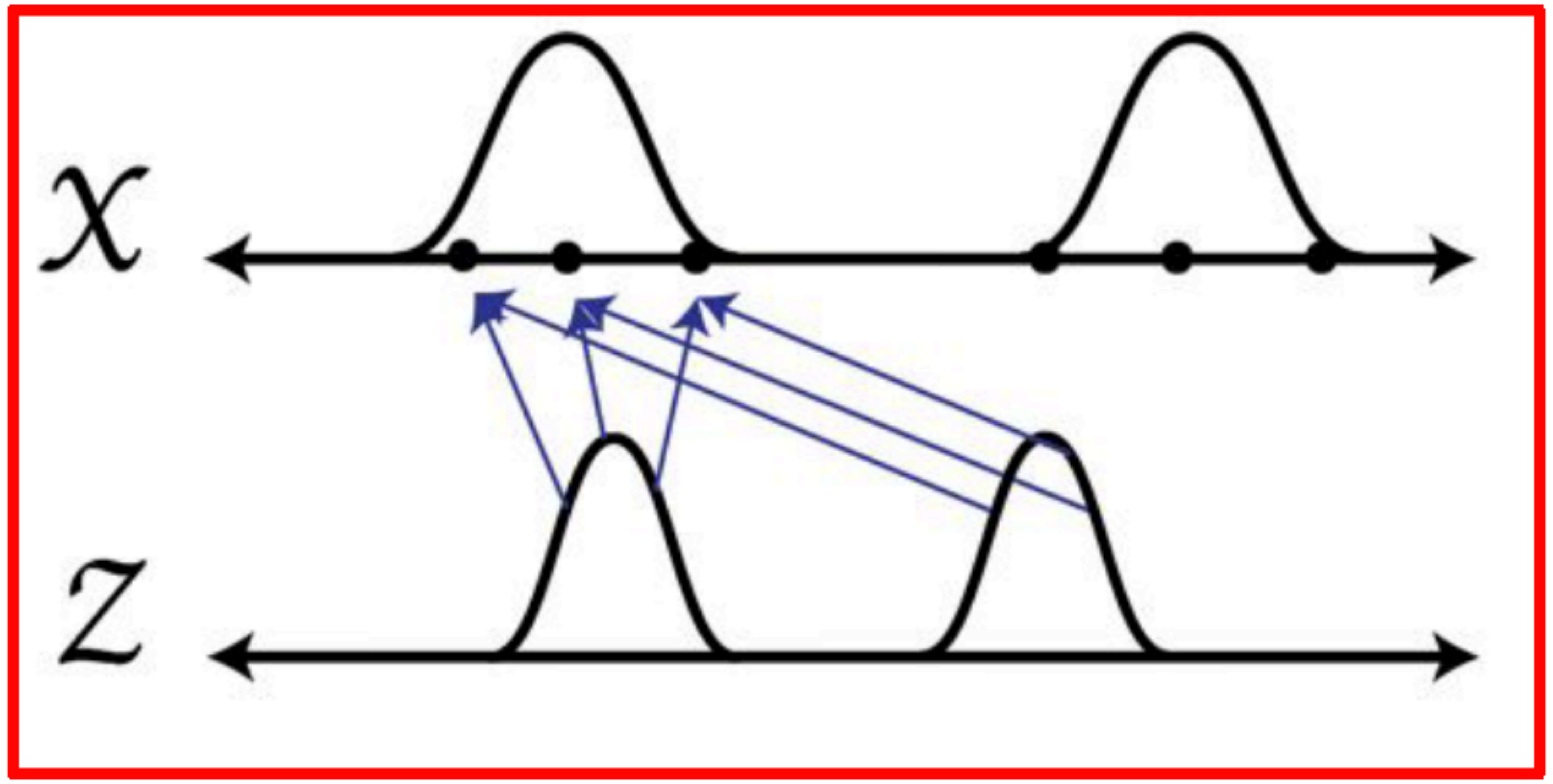
Bourgain's Theorem: an algorithm that embeds data points in an arbitrary metric space into ℓ_2 space with a *bounded* amount of distortion.



Missing Modes

Yet, it is possible to miss certain modes.

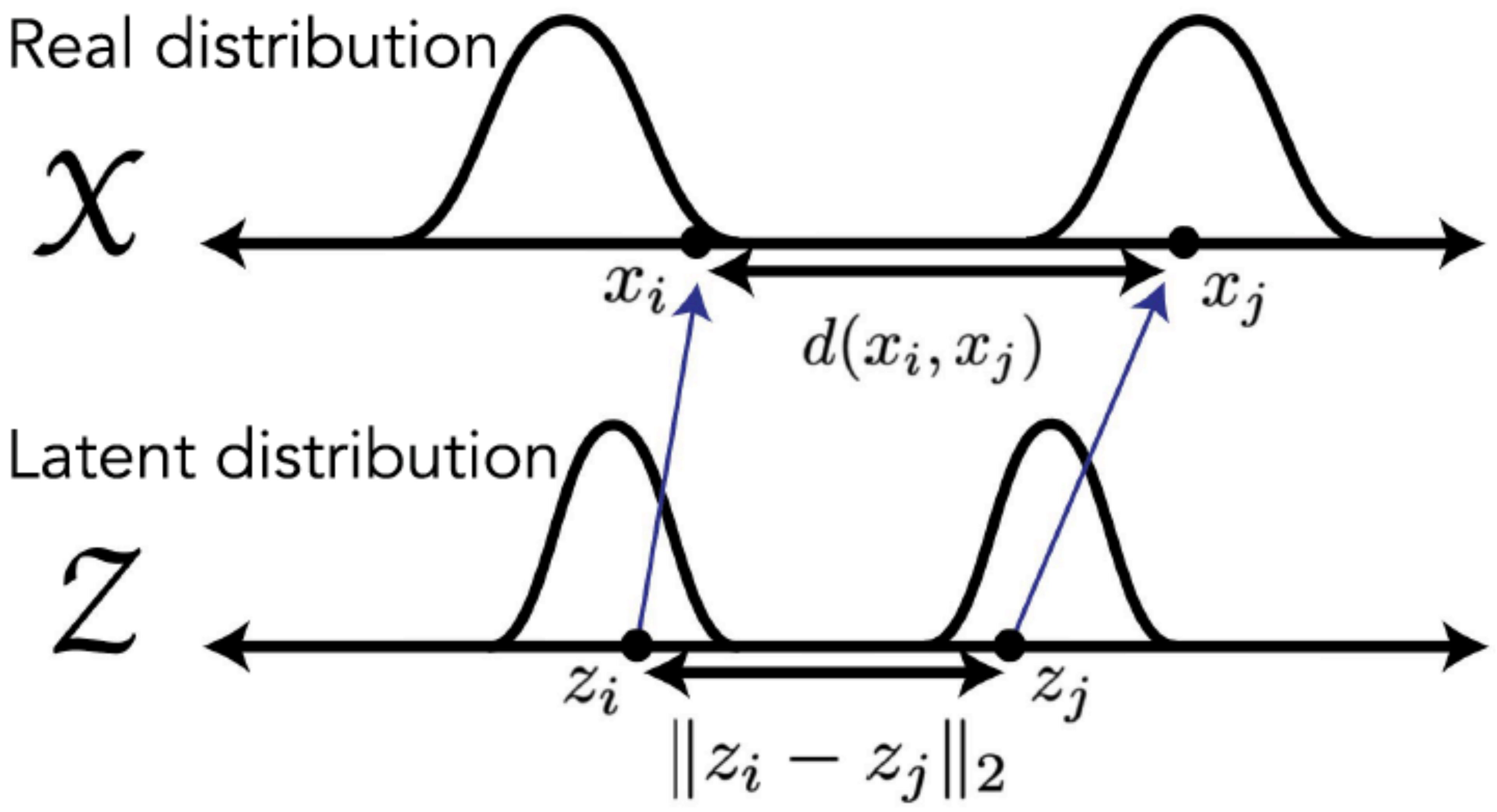


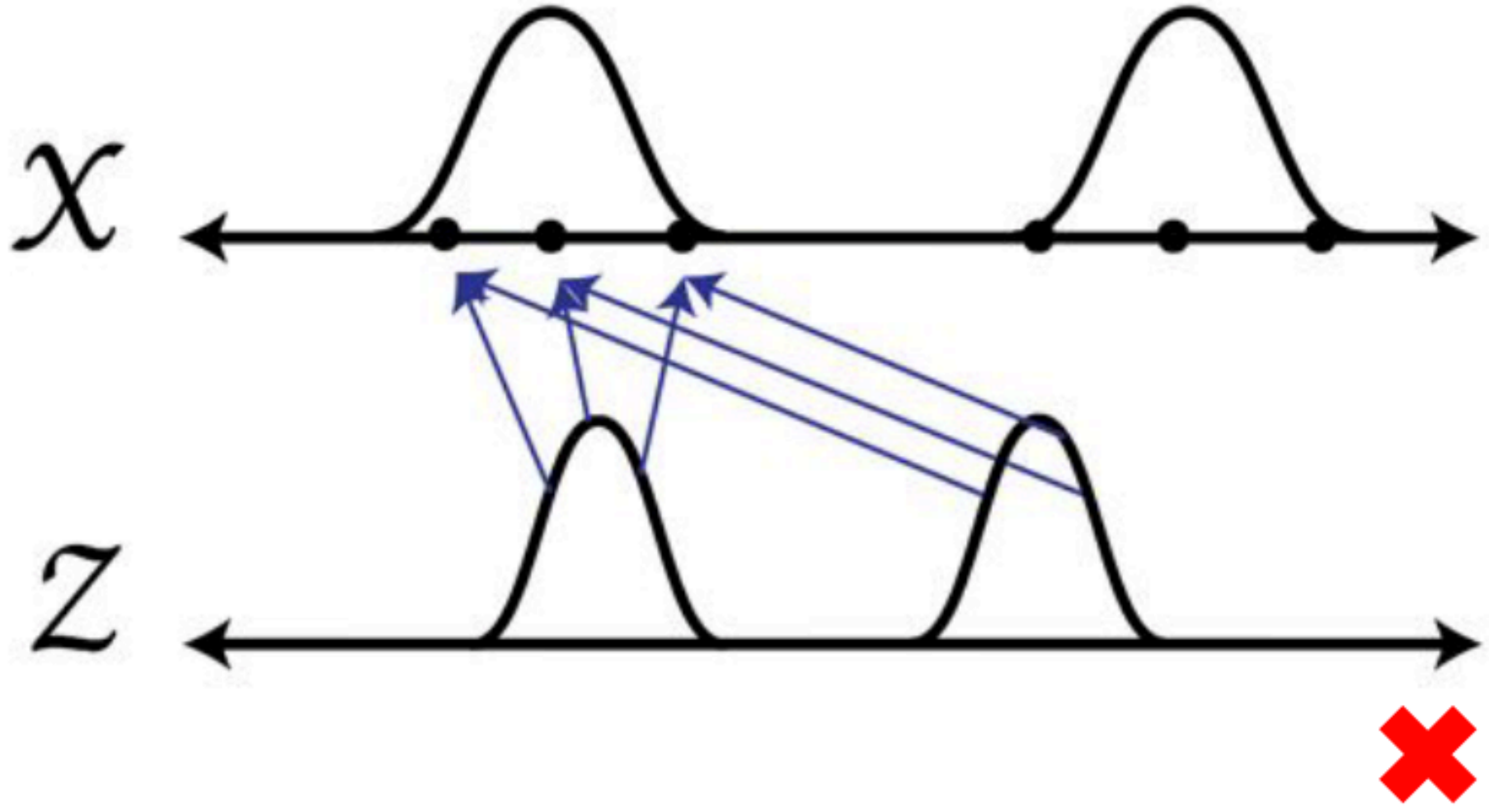


How to solve missing mode?

Solution: Encourage Distance Preservation

Distance constraint: $$\|z_i - z_j\|_2 = d(x_i, x_j)$$





Theoretical Results

Pairwise distance distribution of generated data

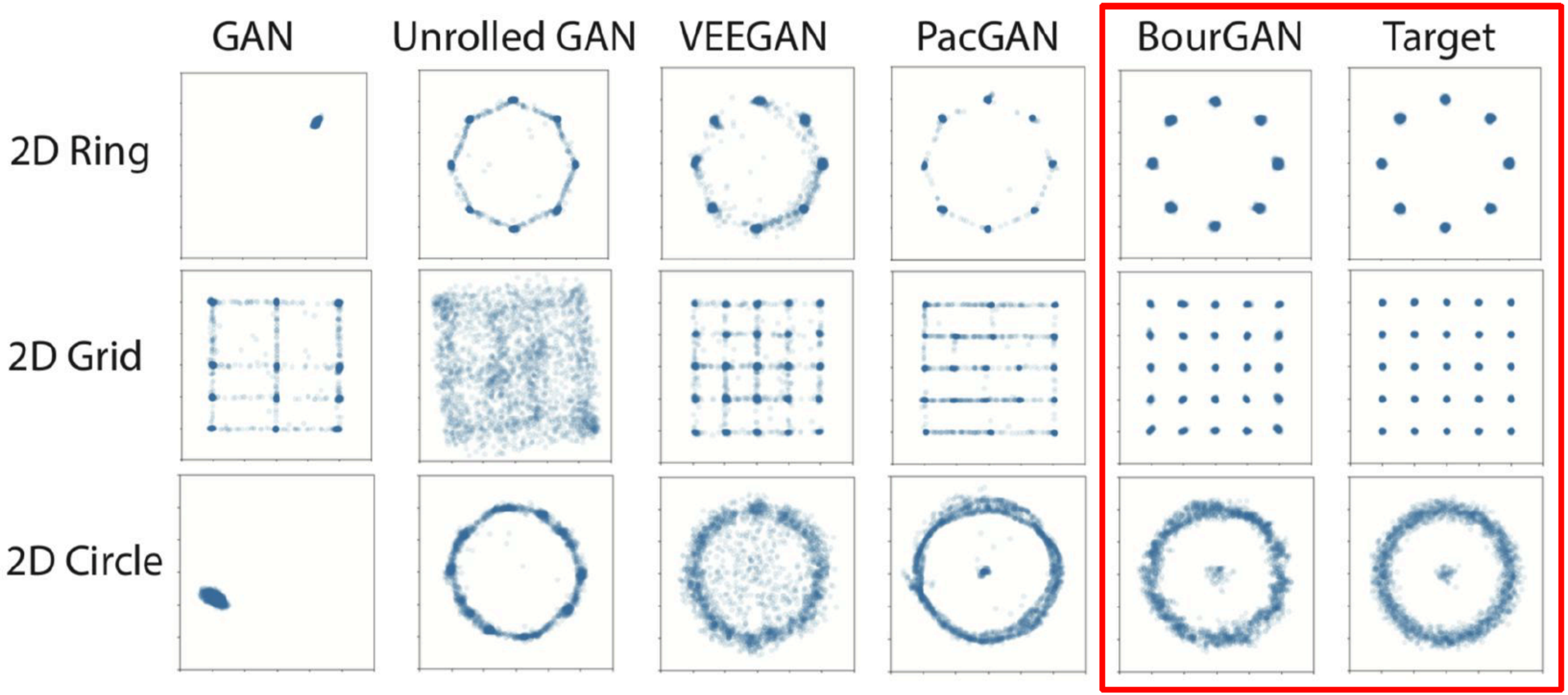
Wasserstein distance

$$W(P, \hat{P}) \leq O(\log \log \log \lambda)$$

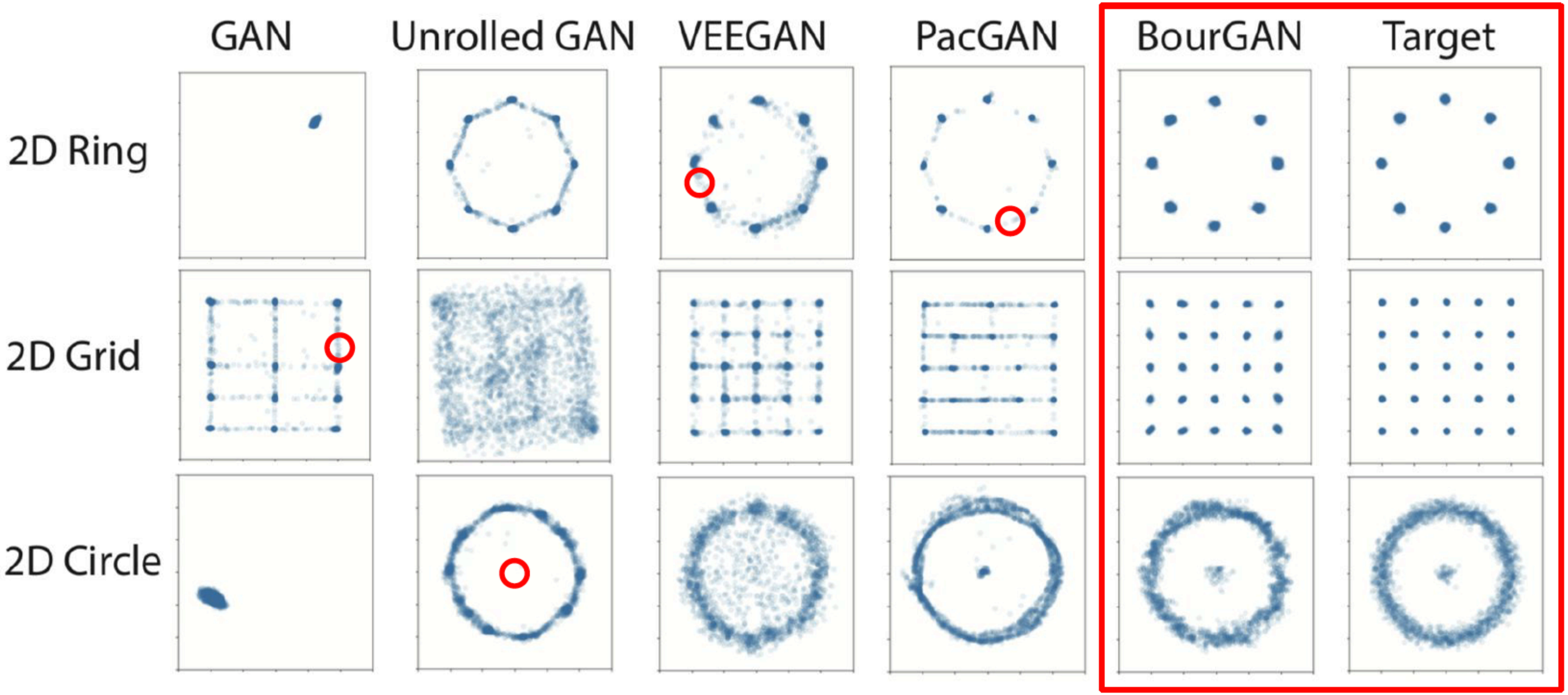
Pairwise distance distribution of real data

A constant related to real distribution

Experimental Results



Experimental Results



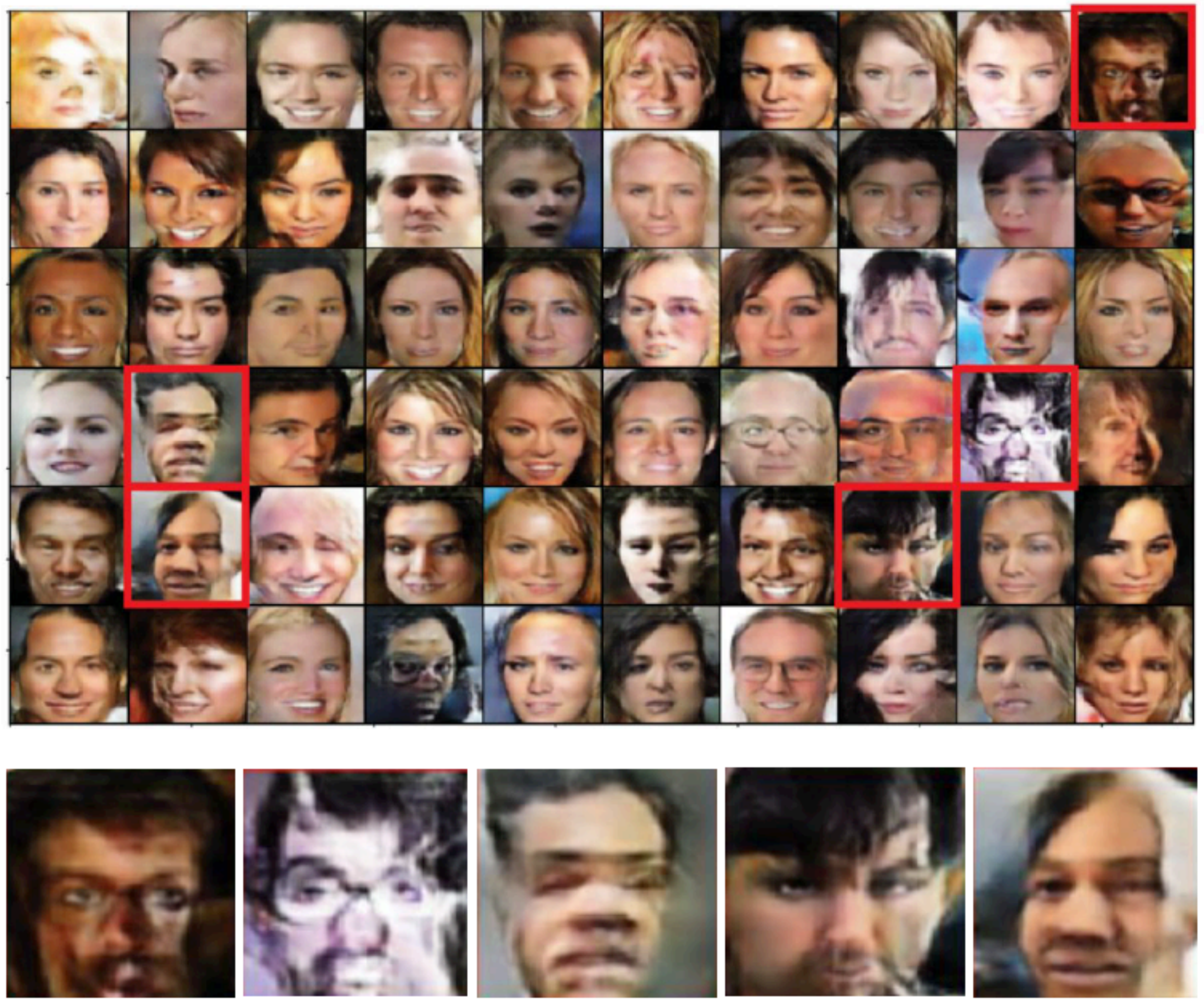
Statistics of Synthetic Experiments

	2D Ring			2D Grid			2D Circle		
	#modes (max 8)	$\mathcal{W}_1$	low quality	#modes (max 25)	$\mathcal{W}_1$	low quality	center captured	$\mathcal{W}_1$	low quality
GAN	1.0	38.60	0.06%	17.7	1.617	17.70%	No	32.59	0.14%
Unrolled	7.6	4.678	12.03%	14.9	2.231	95.11%	No	0.360	0.50%
VEEGAN	8.0	4.904	13.23%	24.4	0.836	22.84%	Yes	0.466	10.72%
PacGAN	7.8	1.412	1.79%	24.3	0.898	20.54%	Yes	0.263	1.38%
BourGAN	8.0	0.687	0.12%	25.0	0.248	4.09%	Yes	0.081	0.35%

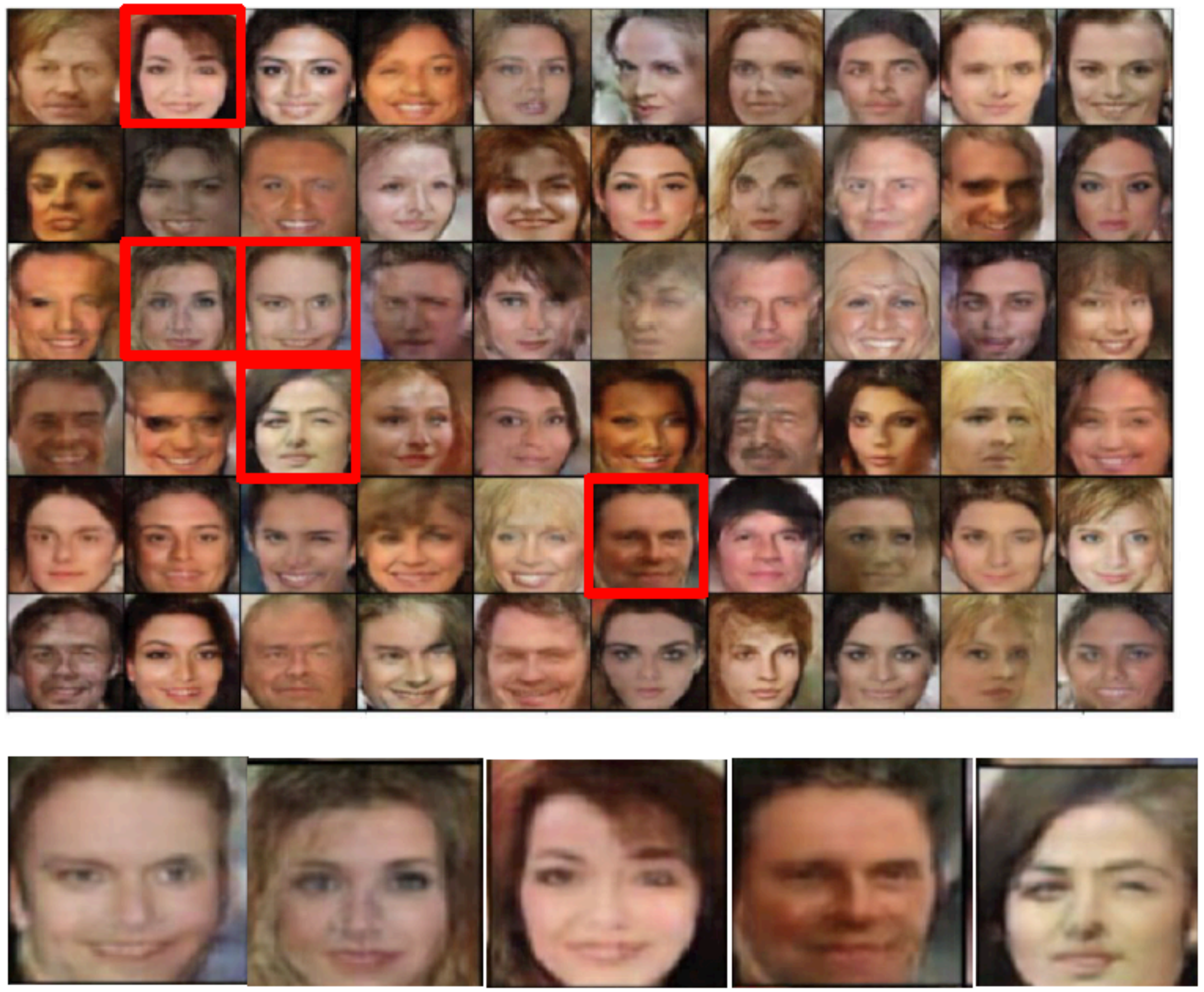
Table 1: Statistics of Experiments on Synthetic Datasets

Experimental Results

DCGAN



BourGAN



More details

Step 1: Subsample of Data Items

- To reduce the computational cost of metric embedding as well as the number of dimension of latent space, randomly sampled a subset of size m ($m << n$) for metric embedding.
- The Wasserstein-1 distance is tightly bounded if m is sufficiently large but much smaller than n.
- m = 4096 in all experiments.

Step 2: (Improved Bourgain's theorem)

- Bourgain's Theorem

Theorem 1 (Bourgain's theorem). *Consider a finite metric space* (Y, d) *with* $m = |Y|$. *There exists a mapping* $g : Y \to \mathbb{R}^k$ *for some* $k = O(\log^2 m)$ *such that* $\forall y, y' \in Y, \mathrm{d}(y, y') \leq \|g(y) - g(y')\|_2 \leq \alpha \cdot \mathrm{d}(y, y')$, *where* α *is a constant satisfying* $\alpha \leq O(\log m)$.

- Improved Bourgain embedding with Johnson-Lindenstrauss Lemma

Corollary 2 (Improved Bourgain embedding). *Consider a finite metric space* (Y, d) *with* $m = |Y|$. *There exist a mapping* $f : Y \to \mathbb{R}^k$ *for some* $k = O(\log m)$ *such that* $\forall y, y' \in Y, \mathrm{d}(y, y') \leq \|f(y) - f(y')\|_2 \leq \alpha \cdot \mathrm{d}(y, y')$, *where* α *is a constant satisfying* $\alpha \leq O(\log m)$.

-

Step 2: (Improved Bourgain's theorem)

- Bourgain's Theorem

Theorem 1 (Bourgain's theorem). *Consider a finite metric space* (Y, d) *with* $m = |Y|$. *There exists a mapping* $g : Y \to \mathbb{R}^k$ *for some* $k = O(\log^2 m)$ *such that* $\forall y, y' \in Y, \mathrm{d}(y, y') \leq \|g(y) - g(y')\|_2 \leq \alpha \cdot \mathrm{d}(y, y')$, *where* α *is a constant satisfying* $\alpha \leq O(\log m)$.

- Improved Bourgain embedding with Johnson-Lindenstrauss Lemma

Corollary 2 (Improved Bourgain embedding). *Consider a finite metric space* (Y, d) *with* $m = |Y|$. *There exist a mapping* $f : Y \to \mathbb{R}^k$ *for some* $k = O(\log m)$ *such that* $\forall y, y' \in Y, \mathrm{d}(y, y') \leq \|f(y) - f(y')\|_2 \leq \alpha \cdot \mathrm{d}(y, y')$, *where* α *is a constant satisfying* $\alpha \leq O(\log m)$.

-

Algorithm 1 Improved Bourgain Embedding

Input: A finite metric space (Y, d).
Output: A mapping $f : Y \to \mathbb{R}^{O(\log |Y|)}$.
//Bourgain Embedding:
Initialization: $m \leftarrow |Y|$, $t \leftarrow O(\log m)$, and $\forall i \in [\lceil \log m \rceil], j \in [t], S_{i,j} \leftarrow \emptyset$.
for $i = 1 \to \lceil \log m \rceil$ **do**
  for $j = 1 \to t$ **do**
    For each $x \in Y$, independently choose x in $S_{i,j}$, i.e. $S_{i,j} = S_{i,j} \cup \{x\}$ with probability 2^{-i}.
  end for
end for
Initialize $g : Y \to \mathbb{R}^{\lceil \log m \rceil \cdot t}$.
for $x \in Y$ **do**
  $\forall i \in [\lceil \log m \rceil], j \in [t]$, set the $((i-1) \cdot t + j)$-th coordinate of $g(x)$ as $\mathrm{d}(x, S_{i,j})$.
end for
//Johnson-Lindenstrauss Dimentionality Reduction:
Let $d = O(\log m)$, and let $G \in \mathbb{R}^{d \times (\lceil \log m \rceil \cdot t)}$ be a random matrix with entries drawn from i.i.d. $\mathcal{N}(0, 1)$.
Let $h : \mathbb{R}^{\lceil \log m \rceil \cdot t} \to \mathbb{R}^d$ satisfy $\forall x \in \mathbb{R}^{\lceil \log m \rceil \cdot t}, h(x) \leftarrow G \cdot x$.
//Rescaling:
Let $\beta = \min_{x,y \in Y : x \neq y} \frac{\|h(g(x)) - h(g(y))\|_2}{\mathrm{d}(x,y)}$.
Initialize $f : Y \to \mathbb{R}^d$. For $x \in Y$, set $f(x) \leftarrow h(g(x))/\beta$.
Return f.

Step 3: Latent-space Gaussian mixture

- Now we construct a distribution using F to draw random vectors in latent space; F is the embedding space.
- A simple choice is the uniform distribution over F, but such a distribution is not continuous over the latent space
- First sample a vector from $\mu \in F$uniformly at random, and draw a vector z from Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$

Remark

Remark. By this definition, the Gaussian mixture consists of m Gaussians (recall $F = \{f(y_i)\}_{i=1}^{m}$). But this does not mean that we construct m "modes" in the latent space. If two Gaussians are close to each other in the latent space, they should be viewed as if they are from the same mode. It is the overall distribution of the m Gaussians that reflects the distribution of modes. In this sense, the number of modes in the latent space is implicitly defined, and the m Gaussians are meant to enable us to sample the modes in the latent space.

Step 4: Loss function

$$L(G, D) = L_{\text{gan}} + \beta L_{\text{dist}},$$

$$\text{where } L_{\text{dist}}(G) = \mathbb{E}_{z_i, z_j \sim \mathcal{Z}} \left[(\log(d(G(z_i), G(z_j))) - \log(\|z_i - z_j\|_2))^2 \right],$$

$$L_{\text{gan}}(G, D) = \mathbb{E}_{x \sim \mathcal{X}} [\log D(x)] + \mathbb{E}_{z \sim \mathcal{Z}} [\log(1 - D(G(z)))] .$$

Summary

- BourGAN, a new GAN variant aiming to address mode collapse in generator network.

- Draw latent space vector using a Gaussian mixture constructed through metric embedding

- Supported by theoretical analysis and experiments, BourGAN enables a well-posed mapping between latent space and multi-modal data distribution.