A Geometric Understanding of Deep Learning

Na Lei

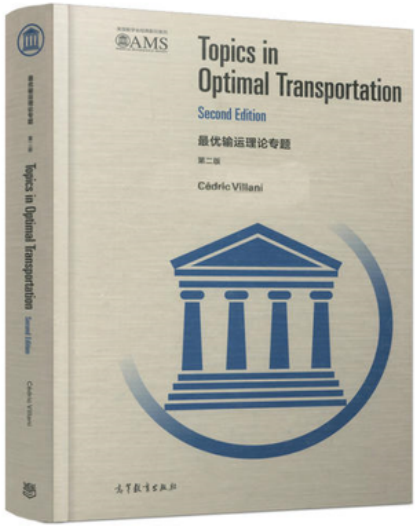
DUT-RU International School of Information Science and Engineering
Dalian University of Technology

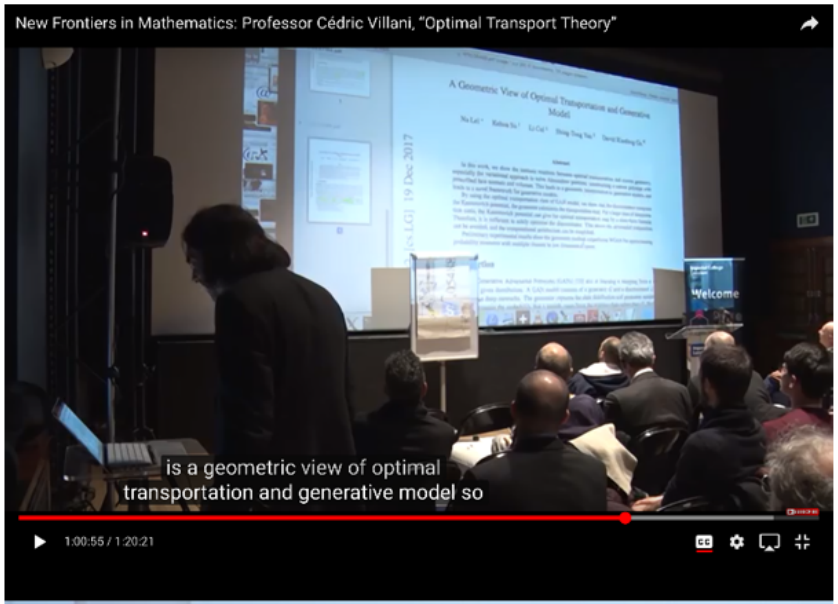
Nov. 6, 2022

References

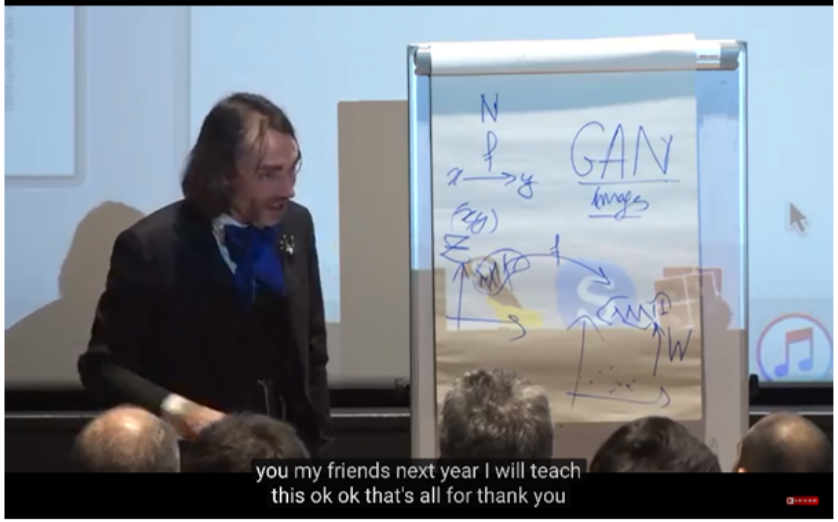
- Na Lei, Kehua Su, Li Cui, Shing-Tung Yau and Xianfeng Gu, “A Geometric View of Optimal Transportation and Generative Model”, Computer Aided Geometric Design, 68(2019), 1-21.
- Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, Xianfeng Gu, A Geometric Understanding of Deep Learning, Engineering, 6(2020), 361-374.
- Dongsheng An, Yang Guo, Na Lei, Zhongxuan Luo, Shing-Tung Yau, Xianfeng Gu, AE-OT: A New Generative Model Based on Extended Semi-Discrete Optimal Transport, ICLR 2020.
- Dongsheng An, Yang Guo, Min Zhang, Xin Qi, Na Lei, Xianfeng Gu, AE-OT-GAN: Training GANs from data specific latent distribution, ECCV 2020.
- Na Lei, Xianfeng Gu, FFT-OT: A Fast Algorithm for Optimal Transportation, ICCV 2021.
- D. An, N. Lei and X. Gu, Approximate Discrete Optimal Transport Plan with Auxiliary Measure Method, ECCV 2022.
- D. An, N. Lei, X. Xu and X. Gu, Efficient Optimal Transport Algorithm by Accelerated Gradient Descent, AAAI, 2022.

International Peer's Opinion





Fields Medalist, Cédric Villani, authoritative expert in optimal transpotation theory, introduced our work in the keynote speach at the New Frontiers in Mathematics International Conference.



International Peer's Opinion

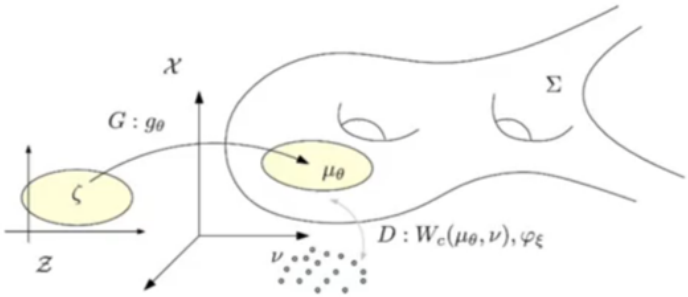
Alessio Figalli, who won the Fields Medal in 2018 for the theory of optimal transmission regularity, introduced our work in a public speech at the International Conference Dynamic, Educations and Applications.

AGH UST DEA 2019
DYNAMICS, EQUATIONS AND APPLICATIONS

ETH zürich

Optimal transport

What else? Wasserstein GAN (Arjovsky et al. 2017, Lei et al. 2019)



Data distribution hypothesis: the distributions of the data live on lower dimensional manifolds (restrictions come from the theme of the contained objects). So real vs fake images are easier to discriminate.

W-GAN: apply optimal transport distance as loss function in the discriminator, to decide whether an image is real or fake.

G acts as a decoding map from the latent space to the sample space. Its goal is to minimize the Wasserstein distance between the real and the fake samples.

D computes the Wasserstein distance with higher and higher accuracy.

Outline

Manifold Distribution Hypothesis

Manifold Assumption

Manifold Assumption[1,2]

Natural high dimensional data concentrates close to a non-linear low-dimensional manifold.

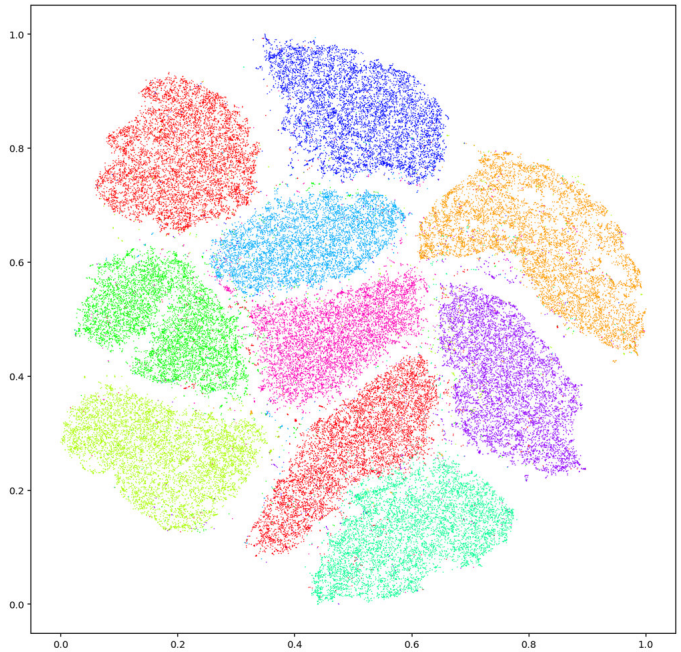
Deep learning method can learn and represent the manifold structure, and transform the probability distributions.

[1] Roweis S T, Saul L K 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323-2326.

[2] Tenenbaum J B, De Silva V, Langford J C 2000. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319-2323.

MNIST tSNE Embedding

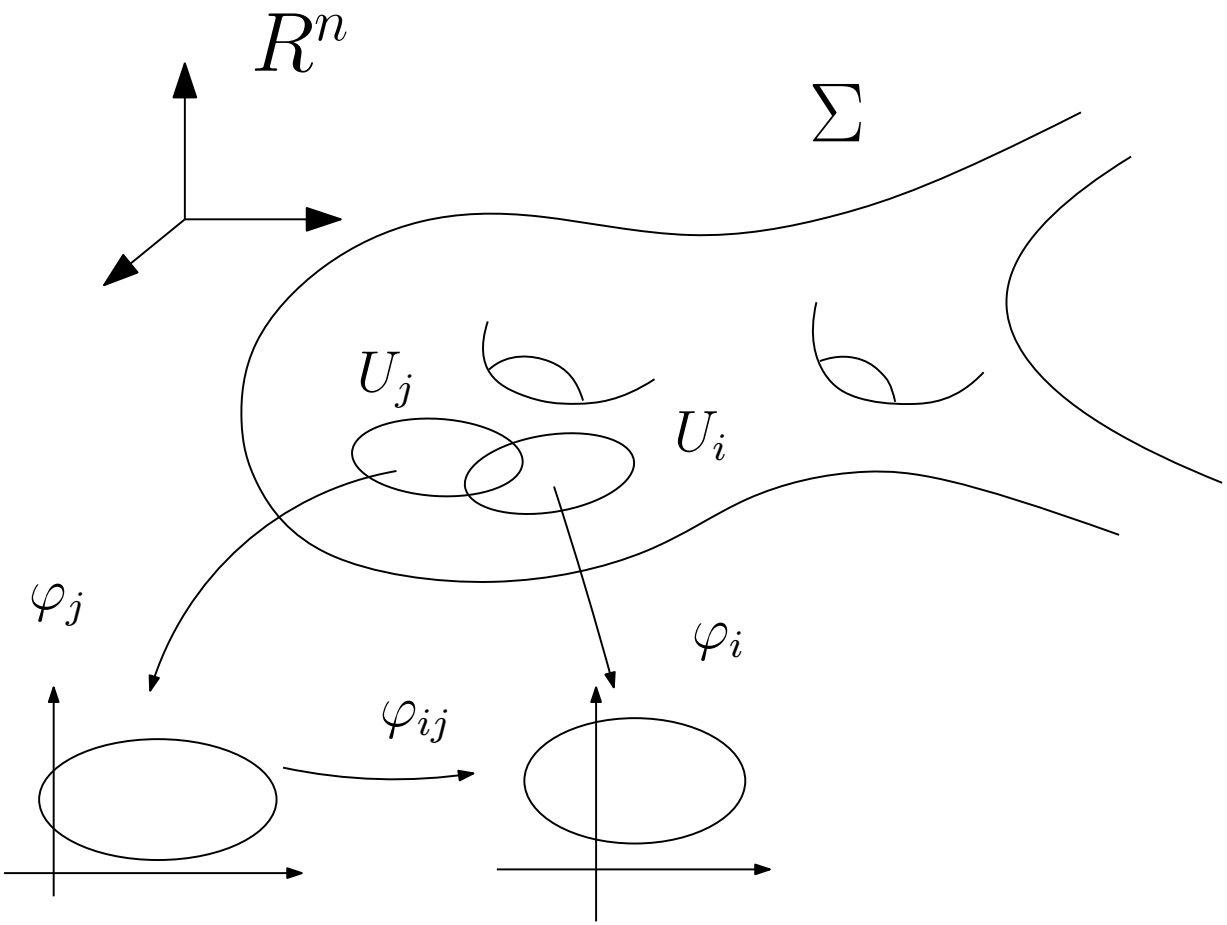
(a) LeCunn's MNIST handwritten digits samples on manifold



(b) Hinton's t-SNE embemdding on latent space

- Each image 28×28 is treated as a point in the 784 dimensional image space;
- The hand-written digits image manifold M is of very low dimension;
- Each digit corresponding to a distribution on M;

General Model



- Ambient Space- data space $\mathbb{R}^n$
- manifold - Support of a distribution μ
- parameter domain - latent space $\mathbb{R}^m$
- coordinates map φ_i- encoding/decoding maps
- φ_{ij} controls the probability measure

Encoding/Decoding

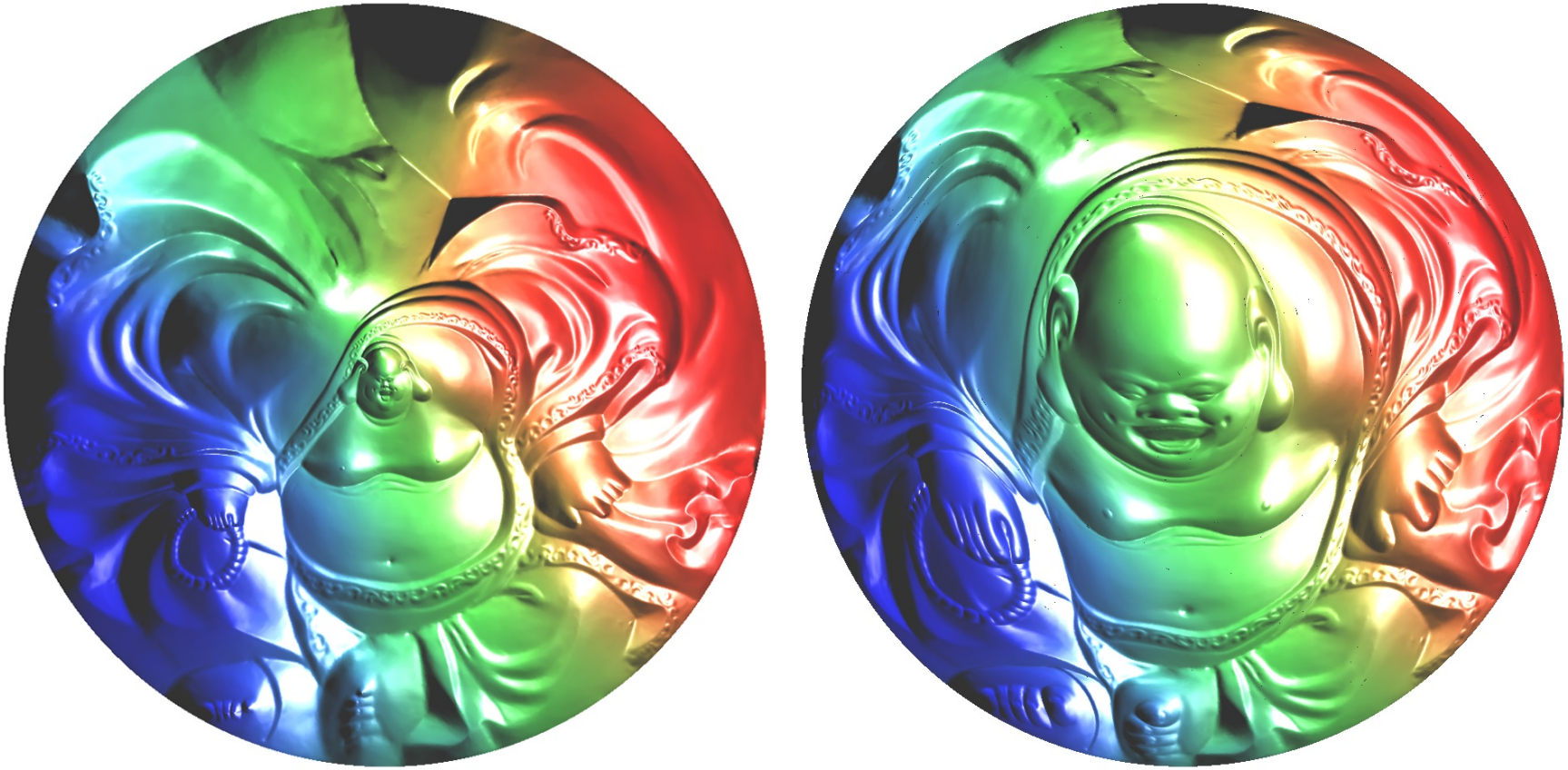
a. Input manifold $M \subset \mathscr{X}$ b. latent representation $D = \varphi_\theta(M) \subset \mathscr{Z}$ c. reconstructed mfld $\tilde{M} = \psi_\theta(D) \subset \mathscr{X}$

Figure: Auto-encoder pipeline.

Example

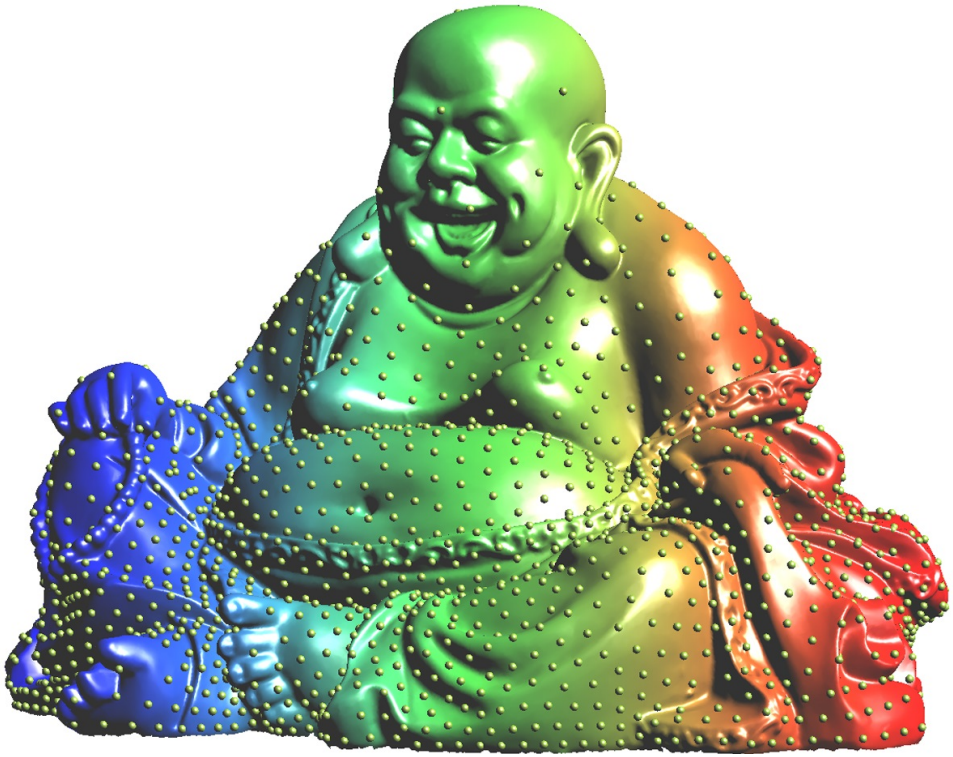
The encoding map is $\varphi_i : \Sigma \to \mathscr{Z}$; the decoding map is $\varphi_i^{-1} : \mathscr{Z} \to \Sigma$.

Example



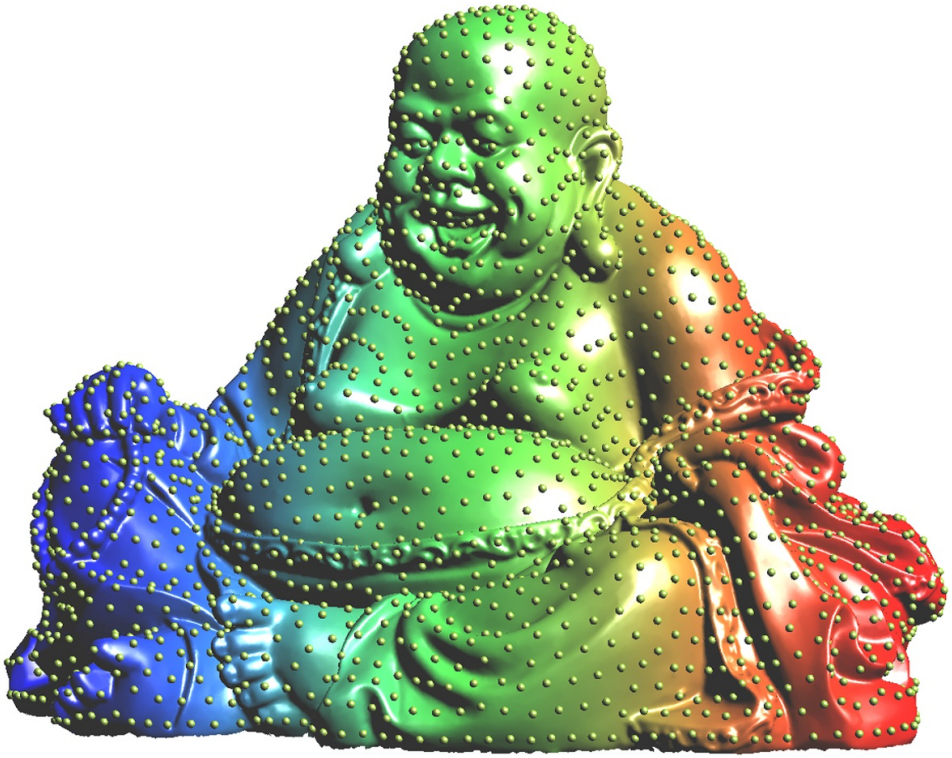
The automorphism of the latent space $\varphi_{ij} : \mathscr{Z} \to \mathscr{Z}$ is the chart transition.

Example



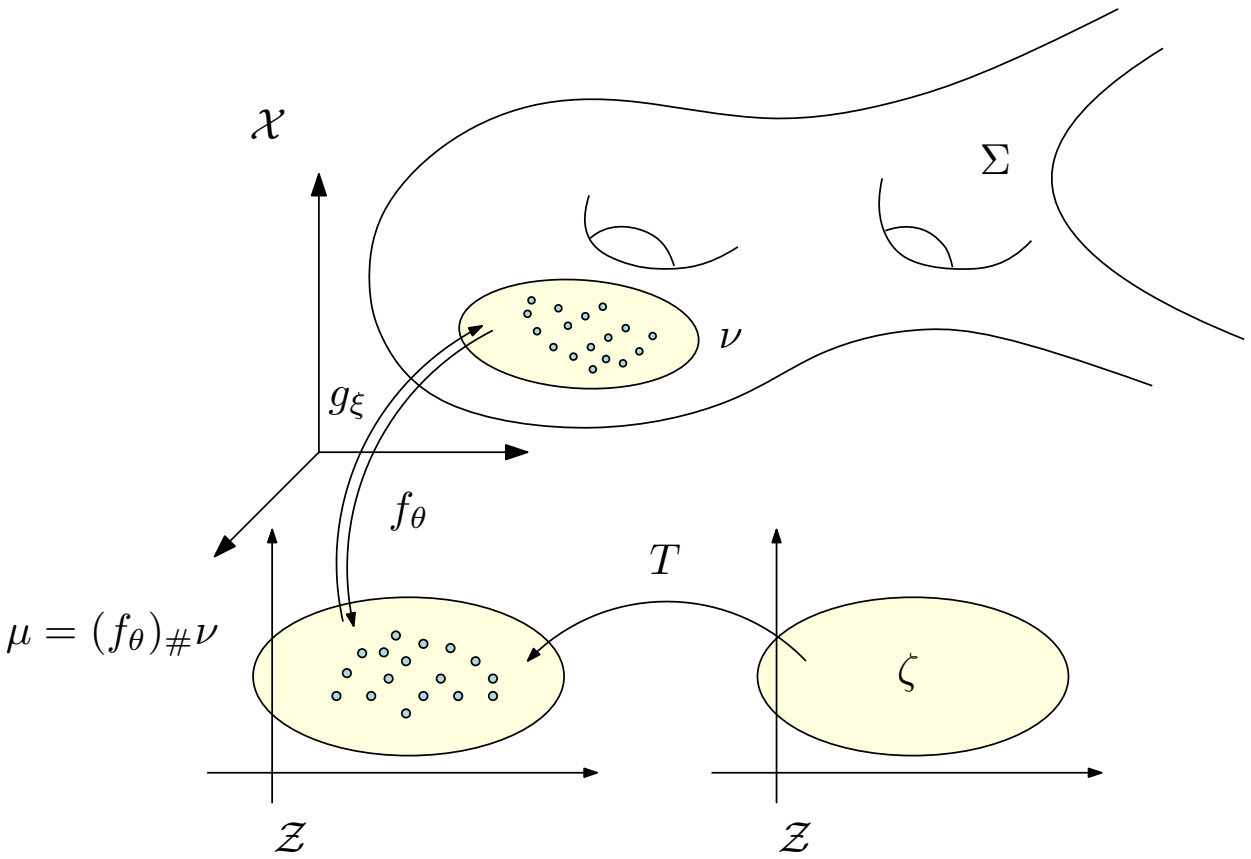
Uniform distribution ζ on the latent space $\mathscr{Z}$, non-uniform distribution on Σ produced by a decoding map.

Example



Uniform distribution ζ on the latent space $\mathscr{Z}$, uniform distribution on Σ produced by another decoding map.

Generative Task



The central tasks for a generative model are

1. Learn the manifold structure from the data (f_θ, g_ξ);
2. Distribution transformation in the latent space (T).

Learn the manifold structure

Non-Adversarial Model: Autoencoder

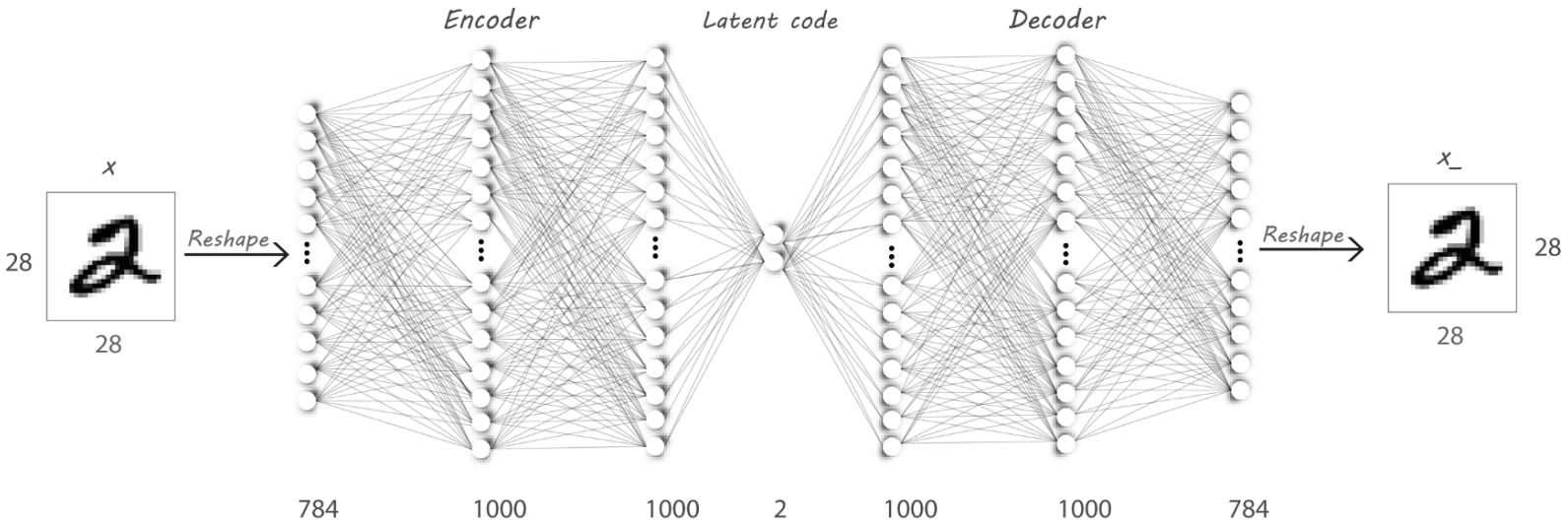


Figure: Auto-encoder architecture.

Ambient space $\mathscr{X}$, latent space $\mathscr{Z}$, encoding map $\varphi_\theta : \mathscr{X} \to \mathscr{Z}$, decoding map $\psi_\theta : \mathscr{Z} \to \mathscr{X}$.

$$(\varphi, \psi) = \text{argmin}_{(\varphi,\psi)} \int_{\mathscr{X}} \mathscr{L}(\mathbf{x}, \psi \circ \varphi(\mathbf{x})) d\mu(\mathbf{x}),$$

ReLU DNN

Definition (ReLU DNN)

For any number of hidden layers $k \in \mathbb{N}$, input and output dimensions $w_0, w_{k+1} \in \mathbb{N}$, a $\mathbb{R}^{w_0} \to \mathbb{R}^{w_{k+1}}$ ReLU DNN is given by specifying a sequence of k natural numbers $w_1, w_2, \ldots, w_k$ representing widths of the hidden layers, a set of k affine transformations $T_i : \mathbb{R}^{w_{i-1}} \to \mathbb{R}^{w_i}$ for $i = 1, \ldots, k$ and a linear transformation $T_{k+1} : \mathbb{R}^{w_k} \to \mathbb{R}^{w_{k+1}}$ corresponding to weights of hidden layers.

The mapping $\varphi_\theta : \mathbb{R}^{w_0} \to \mathbb{R}^{w_{k+1}}$ represented by this ReLU DNN is

$$\varphi = T_{k+1} \circ \sigma \circ T_k \circ \cdots \circ T_2 \circ \sigma \circ T_1, \tag{1}$$

where $\circ$ denotes mapping composition, θ represent all the weight and bias parameters.

Activated Path

Fix the encoding map φ_θ , let the set of all neurons in the network is denoted as $\mathscr{S}$, all the subsets is denoted as $2^{\mathscr{S}}$.

Definition (Activated Path)

Given a point $\mathbf{x} \in \mathscr{X}$, the *activated path* of $\mathbf{x}$ consists all the activated neurons when $\varphi_\theta(\mathbf{x})$ is evaluated, and denoted as $\rho(\mathbf{x})$. Then the activated path defines a set-valued function $\rho : \mathscr{X} \to 2^{\mathscr{S}}$.

Cell Decomposition

Definition (Cell Decomposition)

Fix a encoding map φ_θ represented by a ReLU DNN, two data points $\mathbf{x}_1, \mathbf{x}_2 \in \mathscr{X}$ are *equivalent*, denoted as $\mathbf{x}_1 \sim \mathbf{x}_2$, if they share the same activated path, $\rho(\mathbf{x}_1) = \rho(\mathbf{x}_2)$. Then each equivalence relation partitions the ambient space $\mathscr{X}$ into cells,

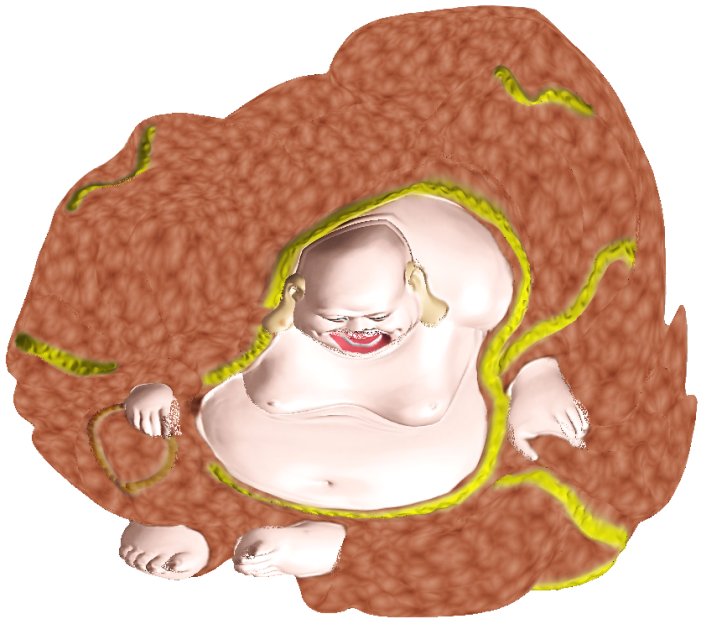
$$\mathscr{D}(\varphi_\theta) : \mathscr{X} = \bigcup_\alpha U_\alpha,$$

each equivalence class corresponds to a cell: $\mathbf{x}_1, \mathbf{x}_2 \in U_\alpha$ if and only if $\mathbf{x}_1 \sim \mathbf{x}_2$. $\mathscr{D}(\varphi_\theta)$ is called the cell decomposition induced by the encoding map φ_θ.

Furthermore, φ_θ maps the cell decomposition in the ambient space $\mathscr{D}(\varphi_\theta)$ to a cell decomposition in the latent space.

Encoding/Decoding

a. Input manifold
$M \subset \mathscr{X}$

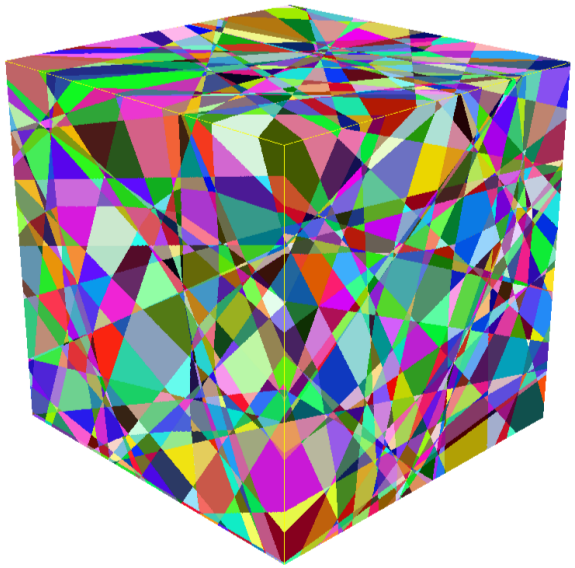


b. latent representation
$D = \varphi_\theta(M)$

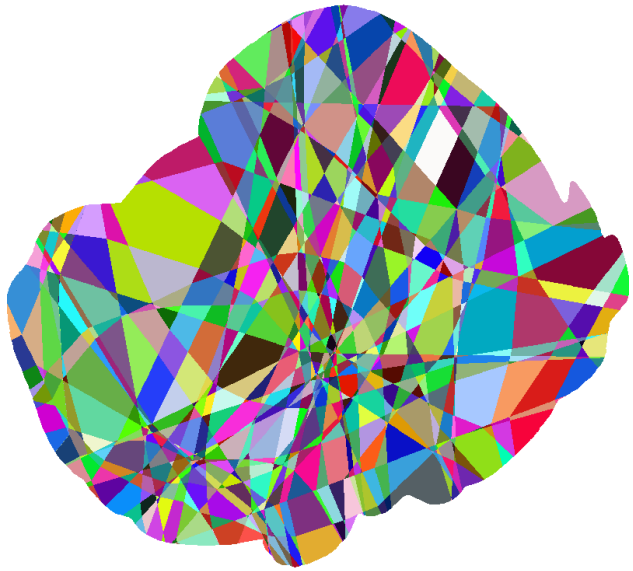
c. reconstructed mfld
$\tilde{M} = \psi_\theta(D)$

Figure: Auto-encoder pipeline.

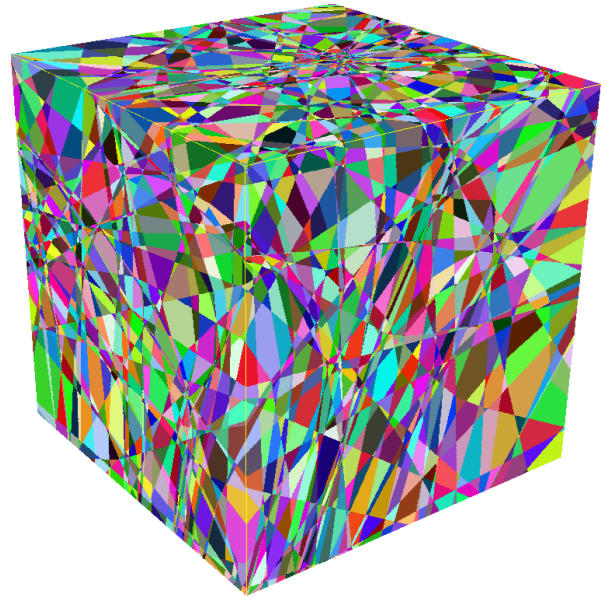
Piecewise Linear Mapping



d. cell decomposition $\mathscr{D}(\varphi_\theta)$



e. latent space cell decomposition



f. cell decomposition $\mathscr{D}(\psi_\theta \circ \varphi_\theta)$

Piecewise linear encoding/decoding maps induce cell decompositions of the ambient space and the latent space.

RL Complexity of a DNN

Definition (Rectified Linear Complexity of a ReLU DNN)

Given a ReLU DNN $N(w_0, \ldots, w_{k+1})$, its rectified linear complexity is the upper bound of the number of pieces of all PL functions φ_θ represented by N,

$$\mathcal{N}(N) := \max_\theta \mathcal{N}(\varphi_\theta).$$

Rectified Linear complexity gives a measurement for the representation capability of a neural network.

RL Complexity Estimate

Lemma

The maximum number of parts one can get when cutting d-dimensional space $\mathbb{R}^d$ with n hyperplanes is denoted as $\mathscr{C}(d,n)$, then

$$\mathscr{C}(d,n) = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{d}. \tag{2}$$

Proof.

Suppose n hyperplanes cut $\mathbb{R}^d$ into $\mathscr{C}(d,n)$ cells, each cell is a convex polyhedron. The $(n+1)$-th hyperplane is π, then the first n hyperplanes intersection π and partition π into $\mathscr{C}(d-1,n)$ cells, each cell on π partitions a polyhedron in $\mathbb{R}^d$ into 2 cells, hence we get the formula

$$\mathscr{C}(d,n+1) = \mathscr{C}(d,n) + \mathscr{C}(d-1,n).$$

It is obvious that $\mathscr{C}(2,1) = 2$, the formula (2) can be easily obtained by induction. □

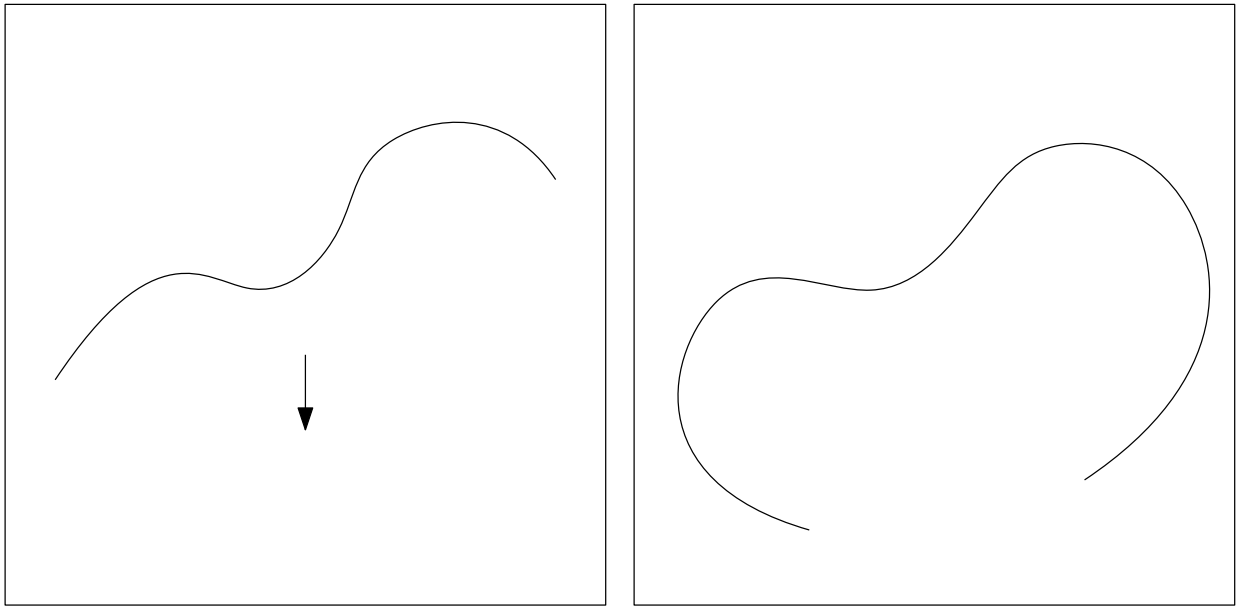
RL Complexity Upper Bound

Theorem (Rectified Linear Complexity of a ReLU DNN)

Given a ReLU DNN $N(w_0, \ldots, w_{k+1})$, representing PL mappings $\varphi_\theta : \mathbb{R}^{w_0} \to \mathbb{R}^{w_{k+1}}$ with k hidden layers of widths $\{w_i\}_{i=1}^k$, then the linear rectified complexity of N has an upper bound,

$$\mathscr{N}(N) \leq \Pi_{i=1}^{k+1} \mathscr{C}(w_{i-1}, w_i). \tag{3}$$

RL Complexity of Manifold



a. linear rectifiable b. non-linear-rectifiable

Definition (Linear Rectifiable Manifold)

Suppose M is a m-dimensional manifold, embedded in $\mathbb{R}^n$, we say M is linear rectifiable, if there exists an affine map $\varphi : \mathbb{R}^n \to \mathbb{R}^m$, such that the restriction of φ on M, $\varphi|_M : M \to \varphi(M) \subset \mathbb{R}^m$, is homeomorphic. φ is called the corresponding rectified linear map of M.

Manifold RL Complexity

Definition (Linear Rectifiable Atlas)

Suppose M is a m-dimensional manifold, embedded in $\mathbb{R}^n$, $\mathscr{A} = \{(U_\alpha, \varphi_\alpha\}$ is an atlas of M. If each chart $(U_\alpha, \varphi_\alpha)$ is linear rectifiable, $\varphi_\alpha : U_\alpha \to \mathbb{R}^m$ is the rectified linear map of U_α, then the atlas is called a linear rectifiable atlas of M.

Definition (Rectified Linear Complexity of a Manifold)

Suppose M is a m-dimensional manifold embedded in $\mathbb{R}^n$, the rectified linear complexity of M is denoted as $\mathscr{N}(\mathbb{R}^n, M)$ and defined as,

$$\mathscr{N}(\mathbb{R}^n, M) := \min\{|\mathscr{A}| \quad |\mathscr{A} \text{ is a linear rectifiable altas of } M\}. \tag{4}$$

Encodable Condition

Definition (Encoding Map)

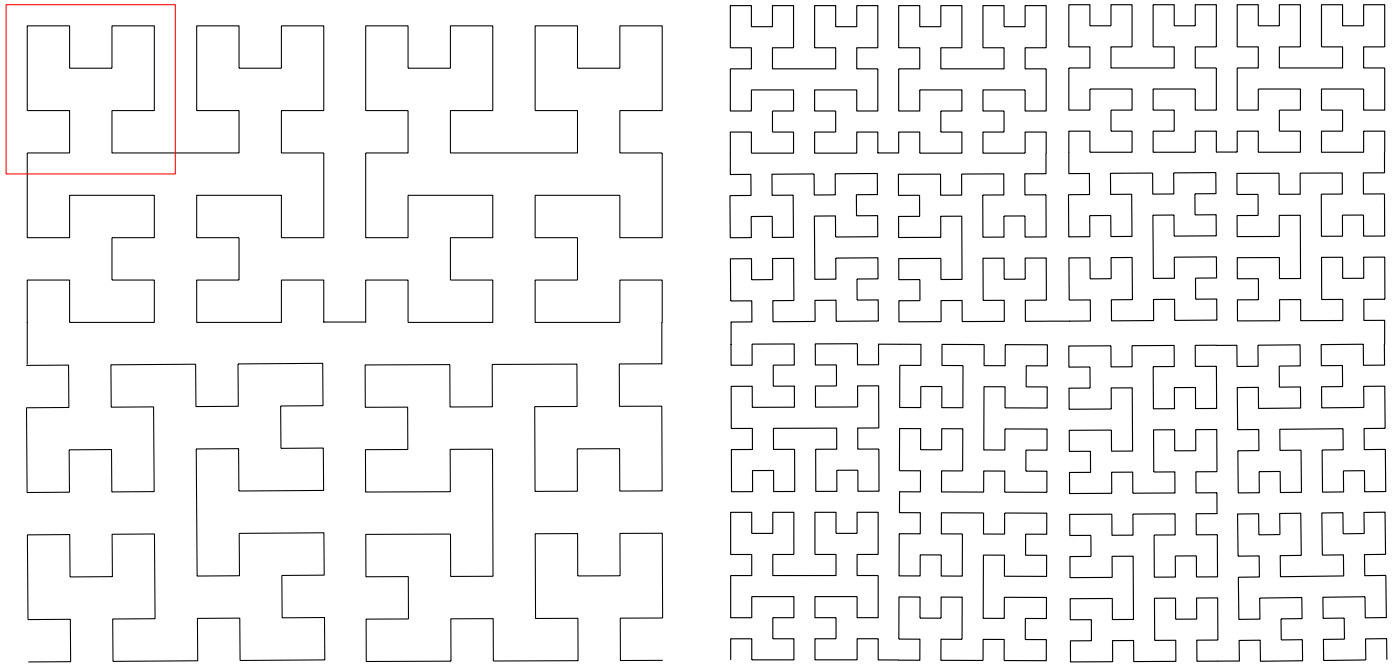
Suppose M is a m-dimensional manifold, embedded in $\mathbb{R}^n$, a continuous mapping $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is called an encoding map of $(\mathbb{R}^n, M)$, if restricted on M, $\varphi|_M : M \to \varphi(M) \subset \mathbb{R}^m$ is homeomorphic.

Theorem (Encodable Condition)

Suppose a ReLU DNN $N(w_0, \ldots, w_{k+1})$ represents a PL mapping $\varphi_\theta : \mathbb{R}^n \to \mathbb{R}^m$, M is a m-dimensional manifold embedded in $\mathbb{R}^n$. If φ_θ is an encoding mapping of $(\mathbb{R}^n, M)$, then the rectified linear complexity of N is no less that the rectified linear complexity of $(\mathbb{R}^n, M)$,

$$\mathcal{N}(\mathbb{R}^n, M) \leq \mathcal{N}(\varphi_\theta) \leq \mathcal{N}(N).$$

Representation Limitation Theorem



C_1 Peano curve C_2 Peano curve

Figure: $\mathcal{N}(\mathbb{R}^2, C_n) \geq 4^{n+1}$

Theorem

Given any ReLU deep neural network $N(w_0, w_1, \ldots, w_k, w_{k+1})$, there is a manifold M embedded in $\mathbb{R}^{w_0}$, such that M can not be encoded by N.

How does DL control the probability distribution?

Concept

Concept

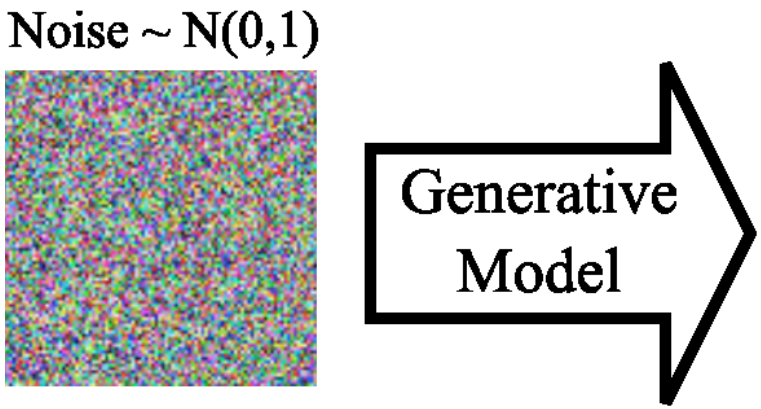
Consider the total space Ω consisting all possible n by n images, a concept is a subset D in the space, $D \subset \Omega$. A concept is represented as a probability measure

$$P(x) = Prob\{x \in D\}.$$

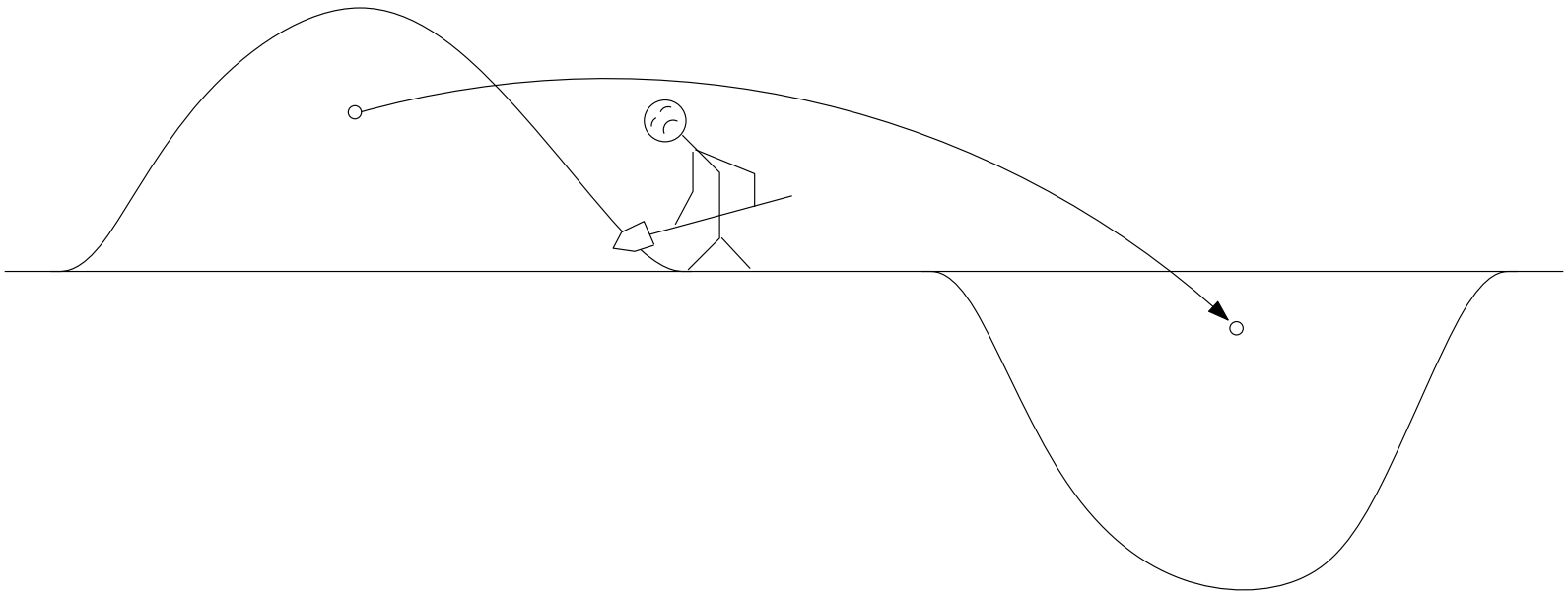
Generative Model

Generative Model

$G : \mathscr{Z} \to \mathscr{X}$ maps a fixed probability distribution ζ to the training data probability distribution ν.



Optimal Transport Problem



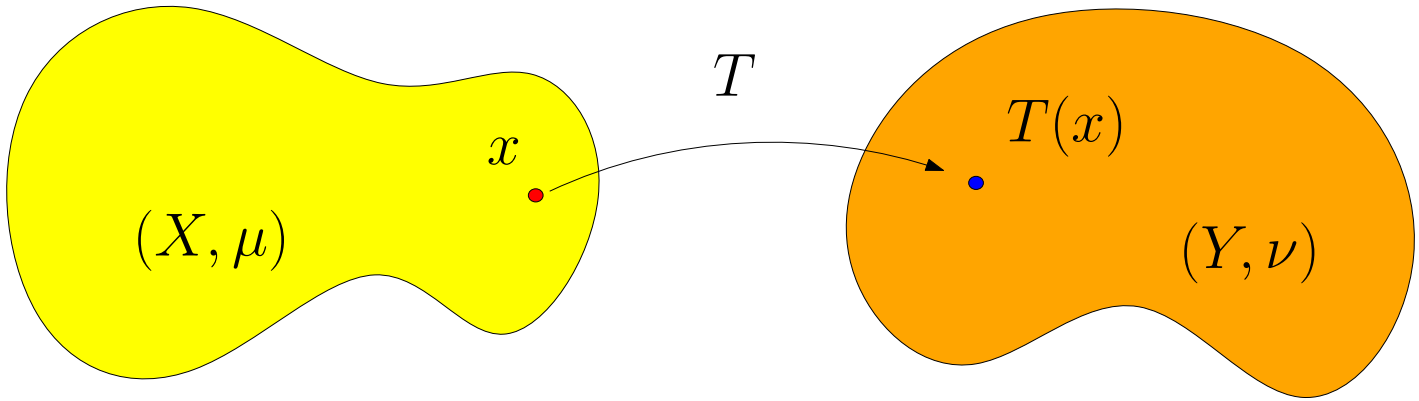
Earth movement cost.

Optimal Transportation

Definition (Transportation Cost)

Given two bounded domains in $\mathbb{R}^n$ with probability measures (X, μ) and (Y, ν), with equal total measure $\mu(X) = \nu(Y)$. Suppose the cost of moving a unit mass from point x to point y is $c(x, y)$, for a transportation map $f : (X, \mu) \to (Y, \nu)$, the total transportation cost is

$$\mathscr{C}(T) = \int_X c(\mathbf{x}, T(\mathbf{x}))d\mu(\mathbf{x}).$$



Optimal Mass Transportation

Definition (Measure-Preserving Mapping)

Given two bounded domains in $\mathbb{R}^n$ with probability measures (X, μ) and (Y, ν), with equal total measure $\mu(X) = \nu(Y)$, a transportation mapping $T : X \to Y$ is measure-preserving, if for any measurable set $B \subset Y$,

$$\int_{T^{-1}(B)} d\mu(x) = \int_B d\nu(y),$$

and denoted as $T_{\#}\mu = \nu$.

Suppose T is a smooth map, then measure-preserving condition is equivalent to the Jacobian equation $\mu(x)dx = \nu(y)dy$

$$det(DT) = \frac{\mu(x)}{\nu \circ T(x)}.$$

Optimal Transport Problem

Problem (Monge)

Find a measure-preserving transportation map $T : (X, \mu) \to (Y, \nu)$ *that minimizes the transportation cost,*

$$(MP) \qquad \min_{T_{\#}\mu=\nu} \mathscr{C}(T) = \min_{T_{\#}\mu=\nu} \int_X c(x, T(x)) d\mu(x).$$

such kind of map is called the optimal transportation map.

Definition (Wasserstein distance)

The transportation cost of the optimal transportation map $T : (X, \mu) \to (Y, \nu)$ is called the Wasserstein distance between μ and ν, denoted as

$$W_c(\mu, \nu) := \min_{T_{\#}\mu=\nu} \mathscr{C}(T).$$

Cost Function $c(x, y)$

The cost of moving a unit mass from point x to point y.

$$Monge(1781) : c(x, y) = |x - y|.$$

This is the natural cost function. Other cost functions include

$$\begin{aligned} c(x, y) &= |x - y|^p, p \neq 0 \\ c(x, y) &= -\log|x - y| \\ c(x, y) &= \sqrt{\varepsilon + |x - y|^2}, \varepsilon > 0 \end{aligned}$$

Any function can be cost function. It can be negative.

Kantorovich Problem

Kantorovich relaxed transportation maps to transportation schemes.

Problem (Kantorovich)

Find an optimal transportation scheme, namely a joint probability measure $\rho \in \mathscr{P}(X \times Y)$, with marginal measures $\rho_{x\#} = \mu$, $\rho_{y\#} = \nu$, that minimizes the transportation cost,

$$(KP) \qquad \min_{\rho} \left\{ \int_{X \times Y} c(x,y) d\rho(x,y) \,\middle|\, \rho_{x\#} = \mu,\ \rho_{y\#} = \nu \right\}.$$

Kantorovich solved this problem by inventing linear programming, and won Nobel's prize in economics in 1975.

Kantorovich Dual Problem

By the duality of linear programming, Kantorovich problem has the dual form:

Problem (Kantorovich Dual)

Find an functions $\varphi : X \to \mathbb{R}$ *and* $\psi : Y \to \mathbb{R}$*, such that*

$$(DP) \quad \max_{\varphi,\psi} \left\{ \int_X \varphi(x) du(x) + \int_Y \psi(y) dv(y), \varphi(x) + \psi(y) \leq c(x,y) \right\}.$$

Kantorovich Dual Problem

Definition (c-transformation)

Given a function $\varphi : X \to \mathbb{R}$, and $c(x, y) : X \times Y \to \mathbb{R}$, its c-transform $\varphi^c : Y \to \mathbb{R}$ is given by

$$\varphi^c(y) := \inf_{x \in X} \{c(x, y) - \varphi(x)\}.$$

Problem (Kantorovich Dual)

The Kantorovich Dual problem can be reformulated as

$$(DP) \qquad \max_{\varphi} \left\{ \int_X \varphi(x) du(x) + \int_Y \varphi^c(y) dv(y) \right\}.$$

φ is called Kantorovich potential.

Brenier's Approach

Theorem (Brenier)

If $\mu, \nu > 0$ and X is convex, and the cost function is quadratic distance,

$$c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}|\mathbf{x} - \mathbf{y}|^2$$

then there exists a convex function $u : X \to \mathbb{R}$ unique upto a constant, such that the unique optimal transportation map is given by the gradient map

$$T : \mathbf{x} \to \nabla u(\mathbf{x}).$$

Problem (Brenier)

Find a convex function $u : X \to \mathbb{R}$, such that

$$(BP) \qquad (\nabla u)_{\#}\mu = \nu,$$

u is called the Brenier potential.

Brenier's Approach

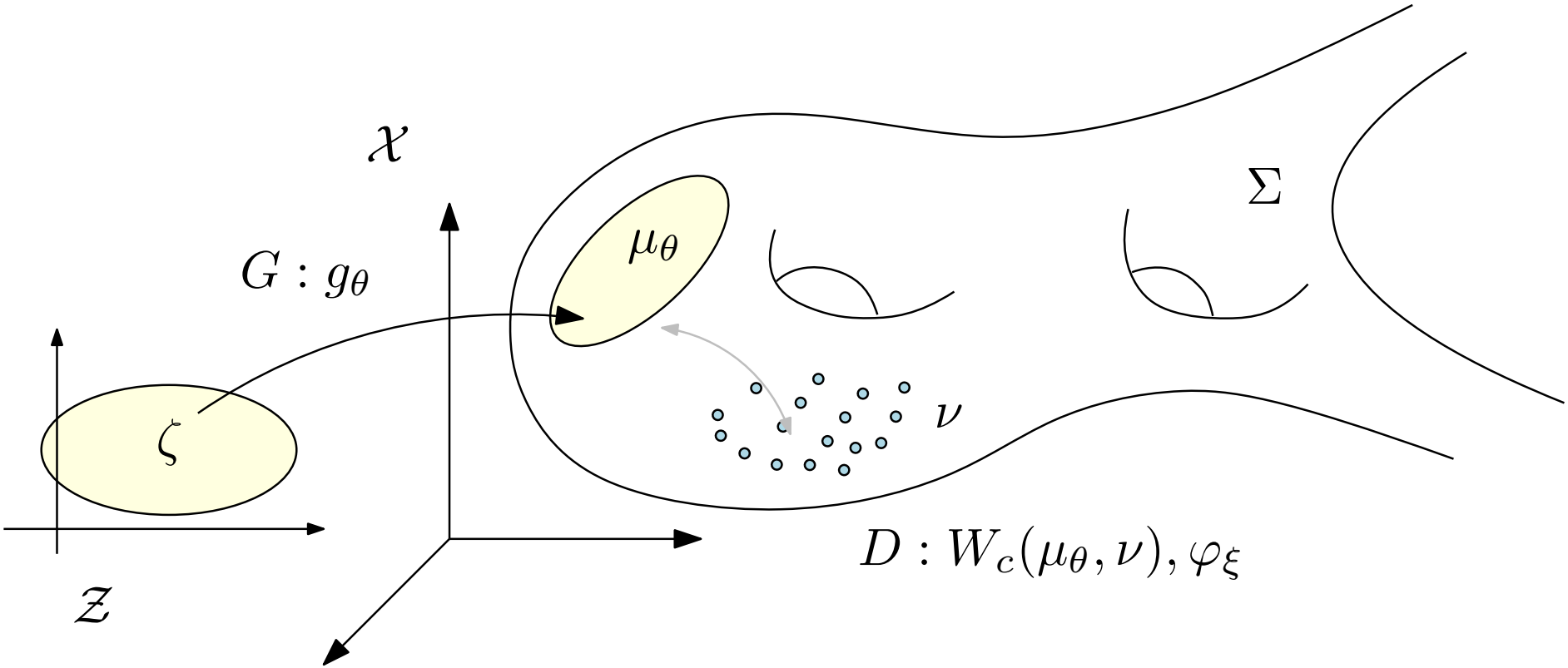
From Jacobian equation, one can get the necessary condition for Brenier potential.

Problem (Brenier)

Find the C^2 Brenier potential $u : X \to \mathbb{R}$ statisfies the Monge-Ampère equation

$$(BP) \qquad det\left(\frac{\partial^2 u}{\partial x_i \partial x_j}\right) = \frac{\mu(\mathbf{x})}{\nu(\nabla u(\mathbf{x}))}.$$

Wasserstein GAN Model



ν-training data distribution; ζ-uniform distribution;
$\mu_\theta = g_{\theta\#}\zeta$-generated distribution; G - generator computes g_θ; D -discriminator, measures the distance between ν and μ_θ, $W_c(\mu_\theta, \nu)$.

OMT view of WGAN

From the optimal transportation point of view, Wasserstein GAN performs the following tasks:

- The discriminator: computes the Wassersteind distance using Kantorovich Dual formula:

$$W_c(\mu_\theta, \nu) = \max_{\varphi_\xi} \int_X \varphi_\xi(x) d\mu_\theta(x) + \int_Y \varphi_\xi^c(y) d\nu(y),$$

  namely computes the Kantorovich potential φ;
- The generator: computes a measure-preserving transportation map $g_\theta : \mathscr{Z} \to \mathscr{X}$, s.t. $g_{\theta\#}\zeta = \mu_\theta = \nu$.
- The WGAN model: min-max optimization

$$\min_\theta \max_\xi \int_X \varphi_\xi \circ g_\theta(z) d\zeta(z) + \int_Y \varphi_\xi^c(y) d\nu(y)$$

OMT view of WGAN

L^1 case

When $c(x,y) = |x-y|$, $\varphi^c = -\varphi$, given φ is 1-Lipsitz, the WGAN model: min-max optimization

$$\min_{\theta}\max_{\xi}\int_X \varphi_\xi \circ g_\theta(z)d\zeta(z) - \int_Y \varphi_\xi(y)d\nu(y).$$

namely

$$\min_{\theta}\max_{\xi}\ \mathbb{E}_{z\sim\zeta}(\varphi_\xi \circ g_\theta(z)) - \mathbb{E}_{y\sim\nu}(\varphi_\xi(y)).$$

with the constraint that φ_ξ is 1-Lipsitz.

OMT view of WGAN

L^2 case

The discriminator computes the Kantorovich potential φ; the generator G computes the optimal transportation map, $T = \nabla u$, where u is the Brenier potential; The Brenier potential equals to

$$u = \frac{1}{2}|x|^2 - \varphi(x).$$

Hence, in theory:

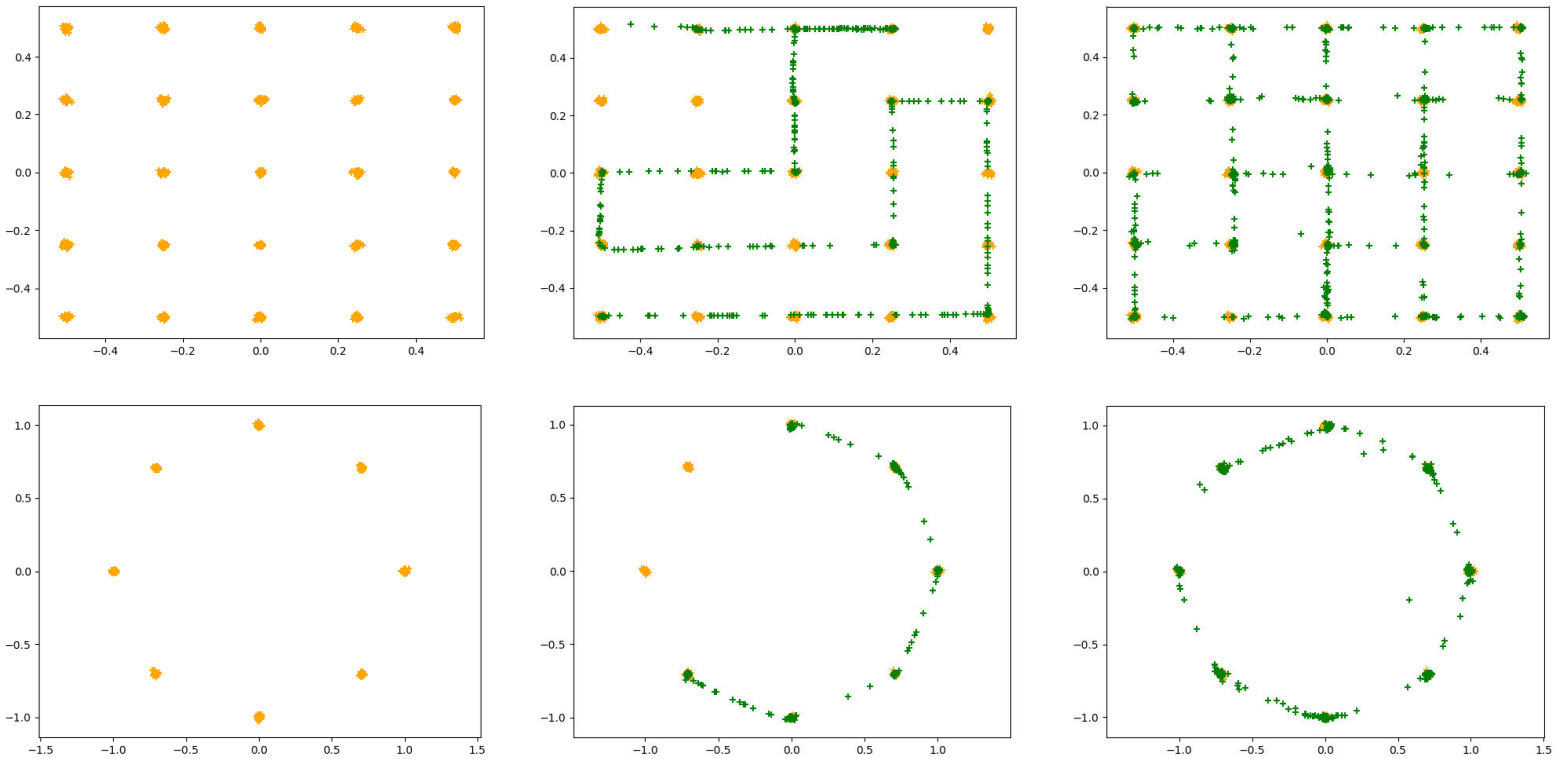
- G can be obtained from the optimal D without training;
- D can be obtained from the optimal G without training;
- The competition between D and G should be replaced by cooperation.

Reason of Mode Collapse

Mode Collapse

- Genereative models are difficult to train and sensitive to hyper-parameters;
- Genereative models suffer from mode collapsing, the generated distributions miss some modes;
- Genereative models may generate unrealistic samples;

Mode Collapse



(a) original (b) GAN (c) pacgan

Singularity Set of OT Maps

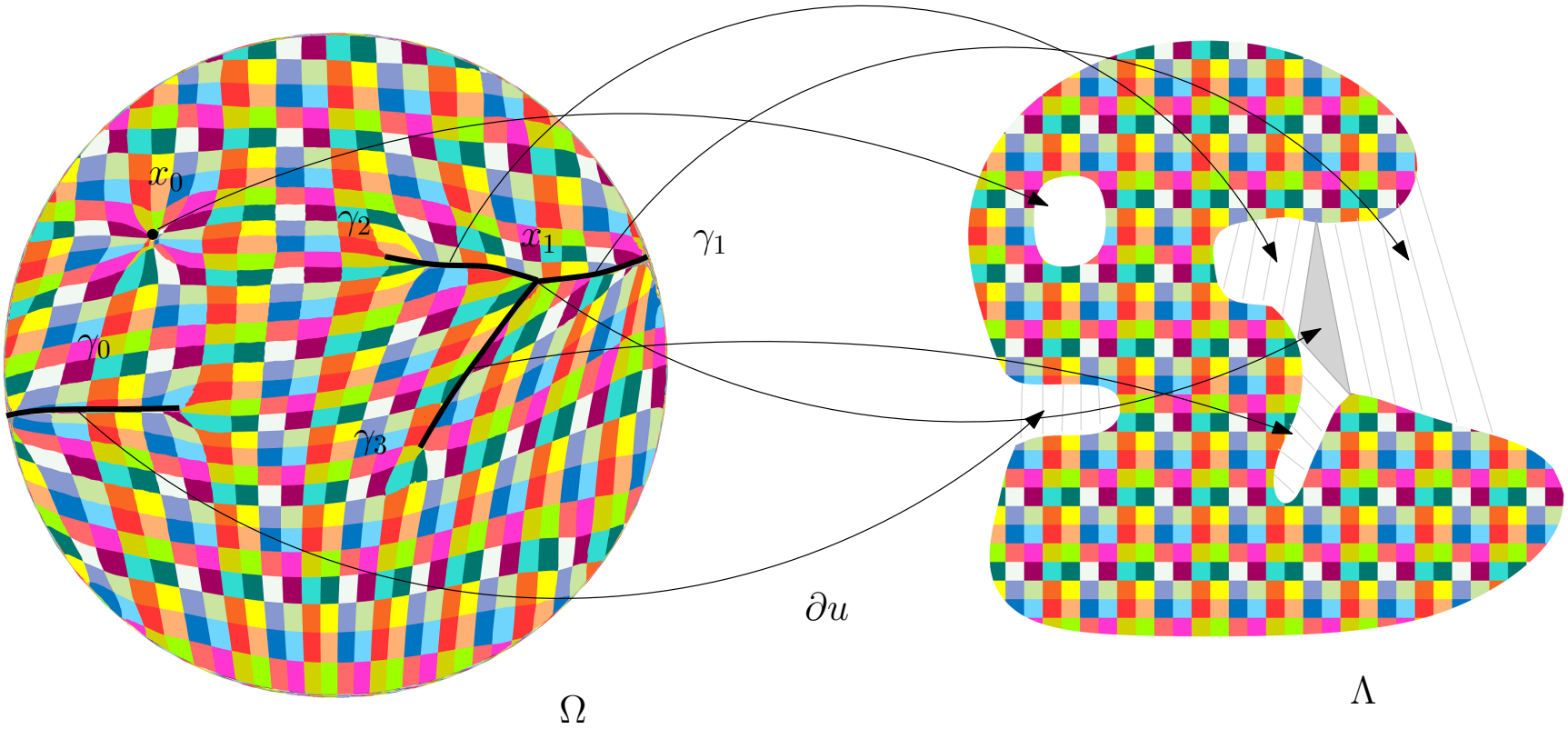


Figure: Singularity structure of an optimal transportation map.

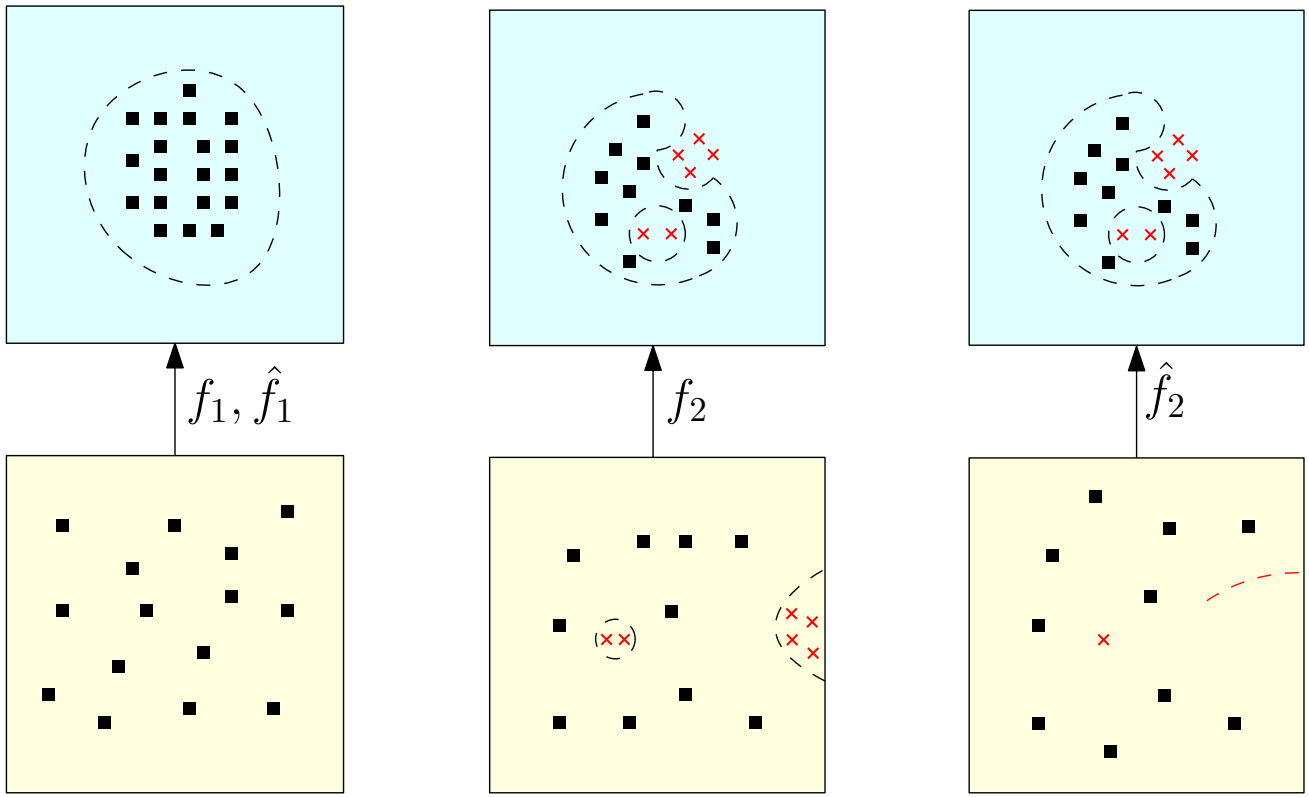
We call Σ_Ω as singular set of the optimal transportation map $\nabla u : \Omega \to \Lambda$.

Mode Collapse

Intrinsic Conflict

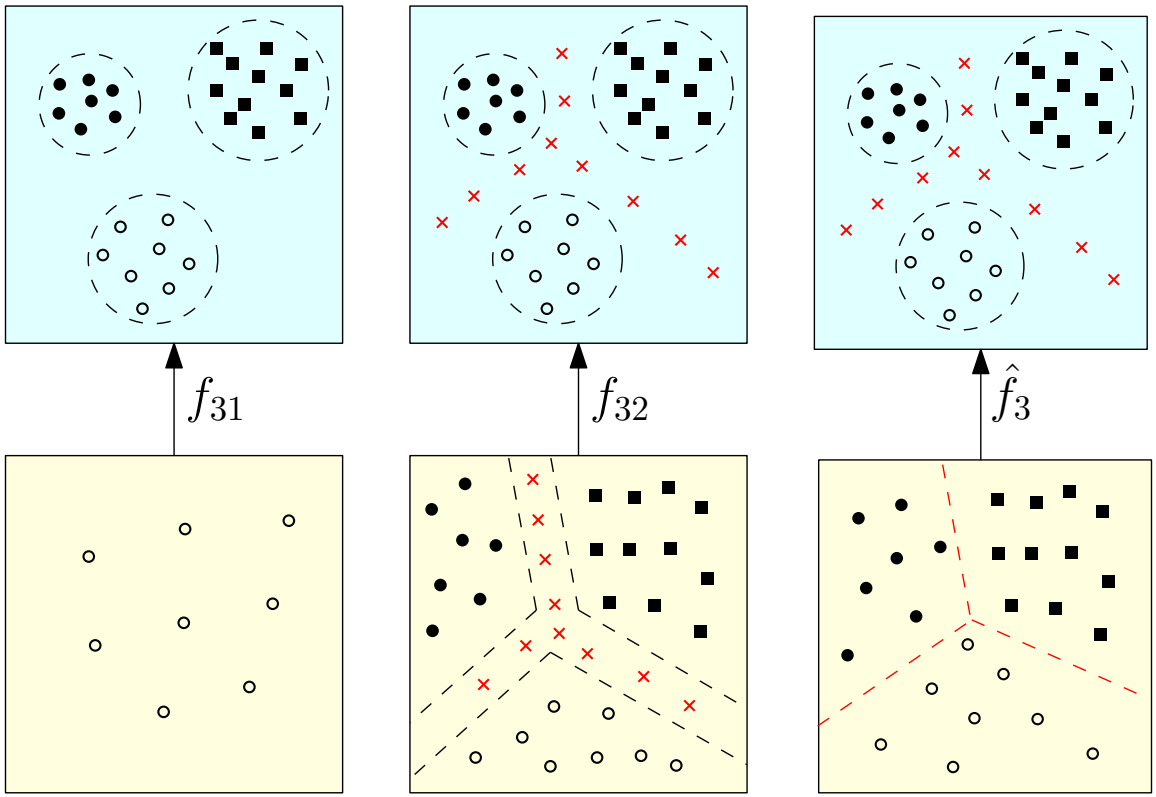
Deep neural networks can only represent continuous mappings, but the transportation maps are discontinuous on singular sets. Namely, the target mappings are outside the functional space of DNNs. This conflict induces mode collapsing.

Discontinuity of Optimal Transportation Map



(a) convex support (b) concave support: single mode

Discontinuity of Optimal Transportation Map



(c) concave support: multi modes

Singularity Set Detection

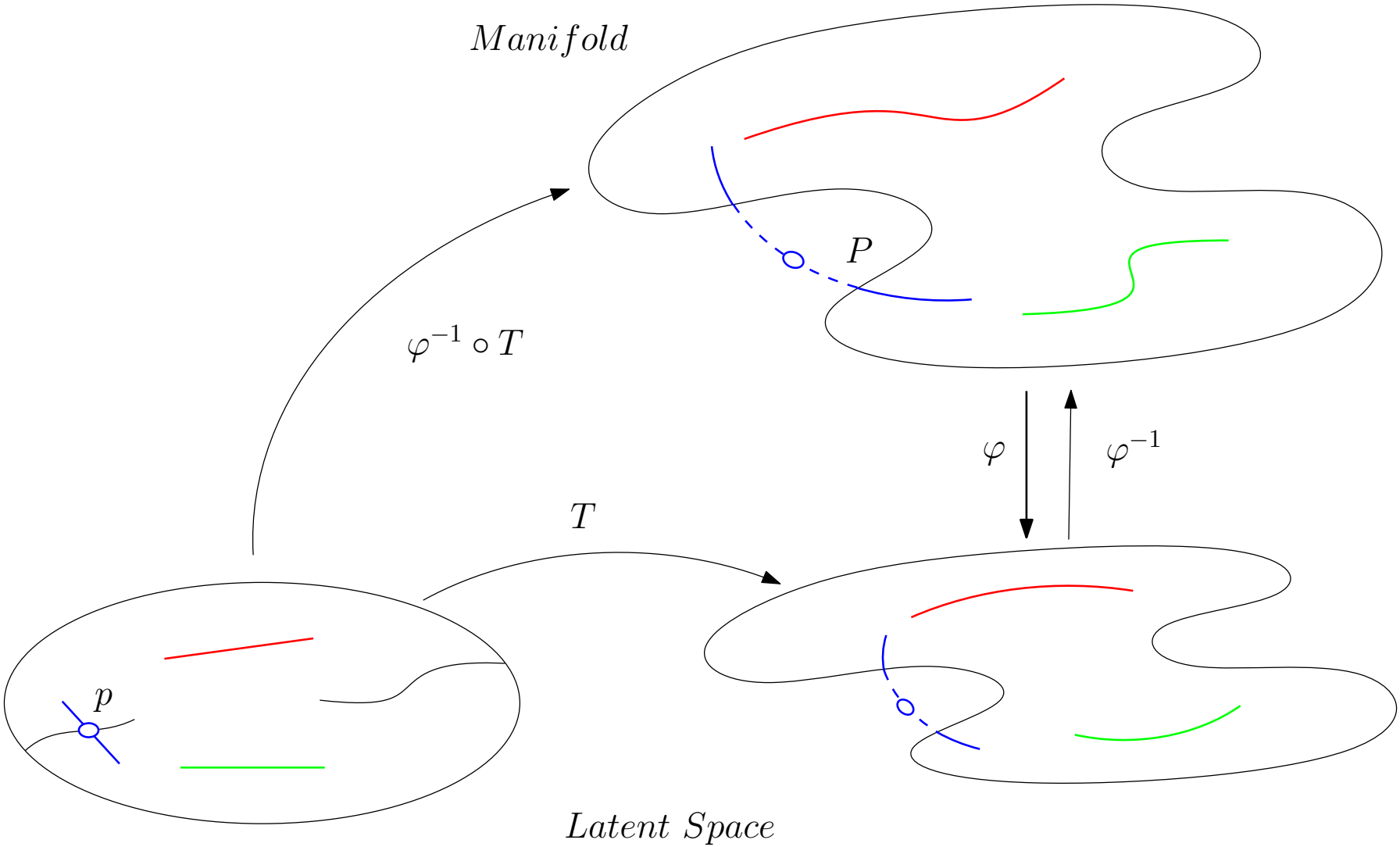
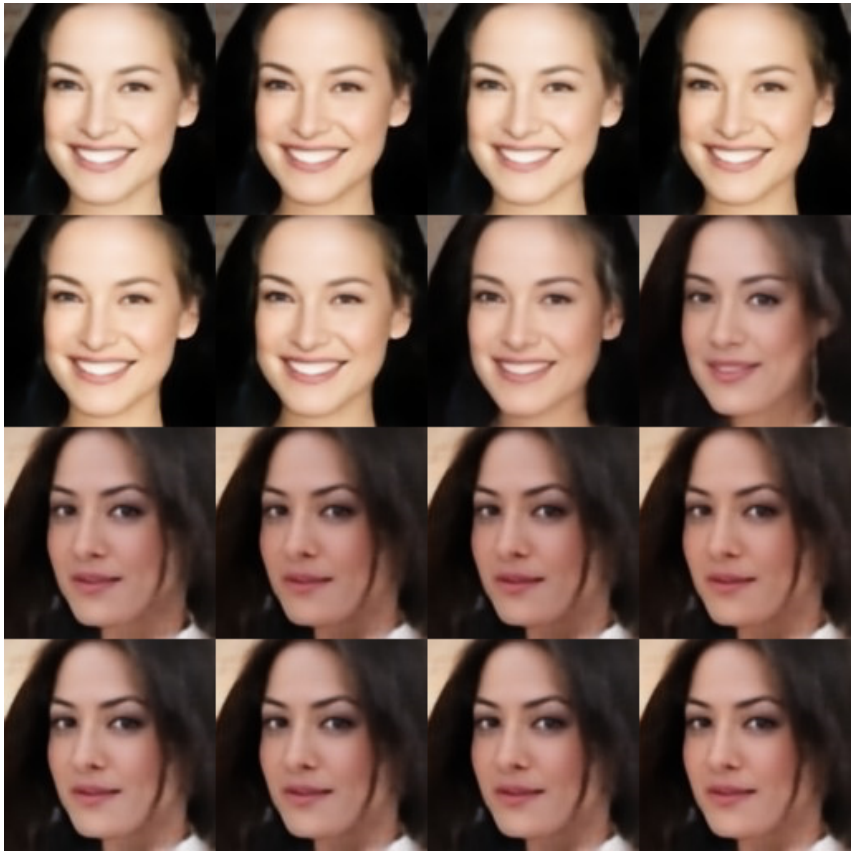


Figure: Singularity set detection.

Mode Collapse



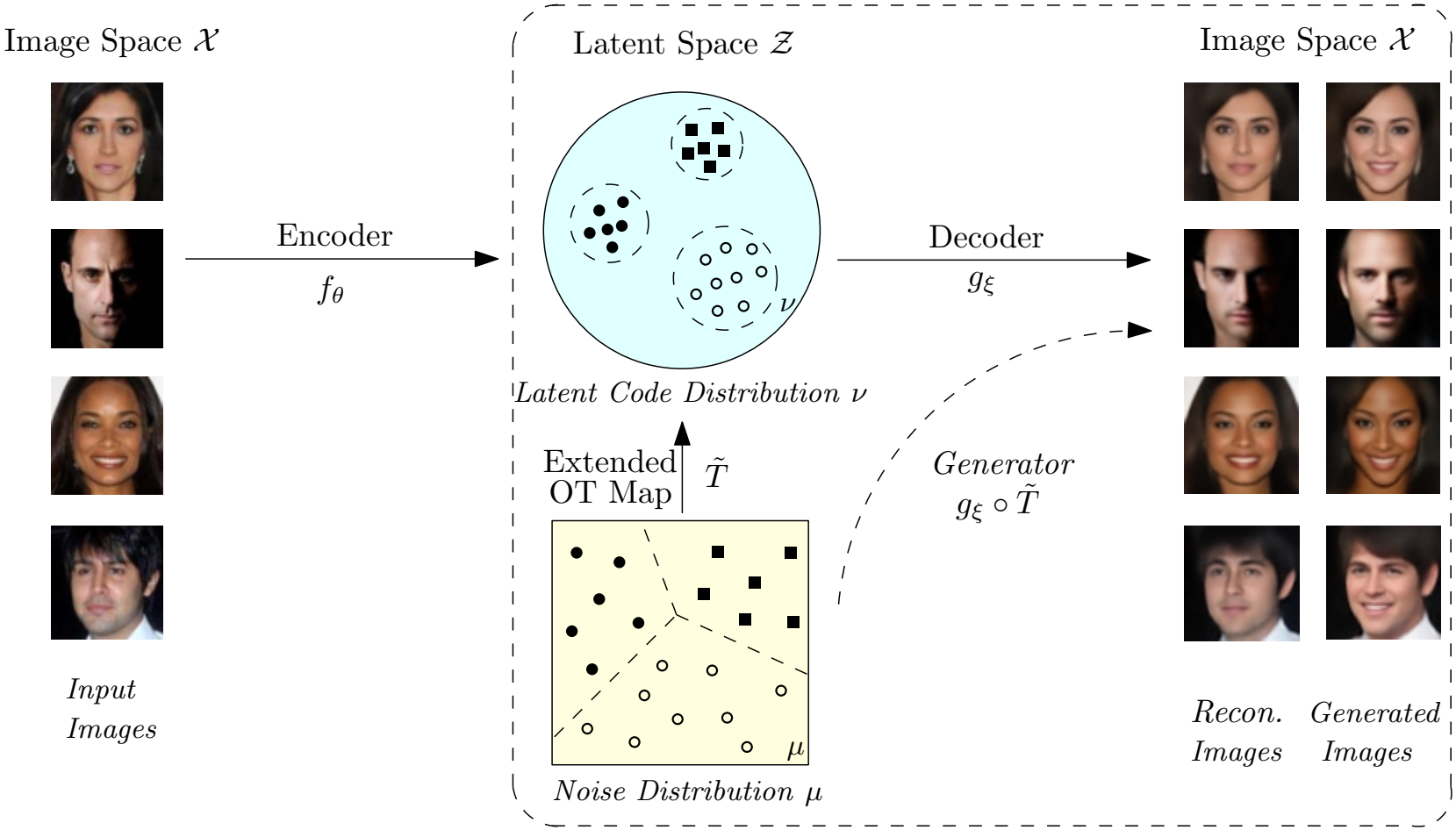
(a) a path inside the manifold

(b) a path through a singularity.

Figure: Facial images generated by an AE-OT model, the image in the center of (b) shows the transportation map is discontinuous.

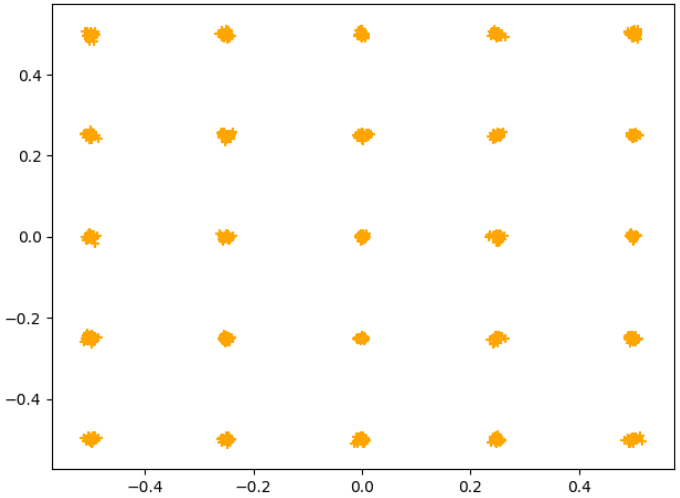
AutoEncoder-Optimal Transportation Framework

AE-OT Model

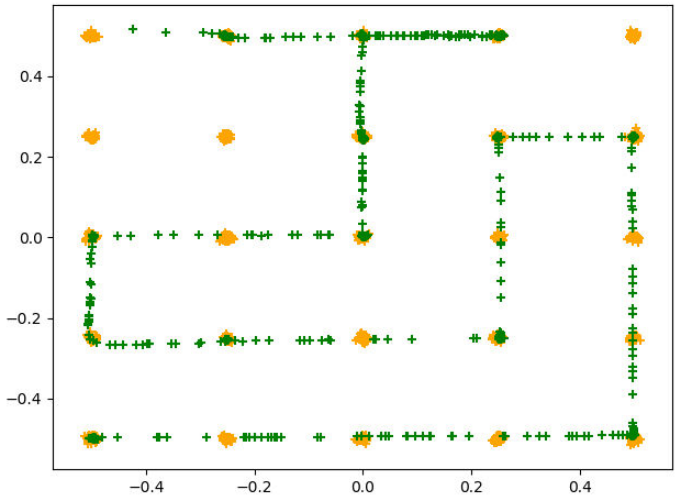


Use autoencoder to realize encoder and decoder, use OT in the latent space to realize probability transformation.

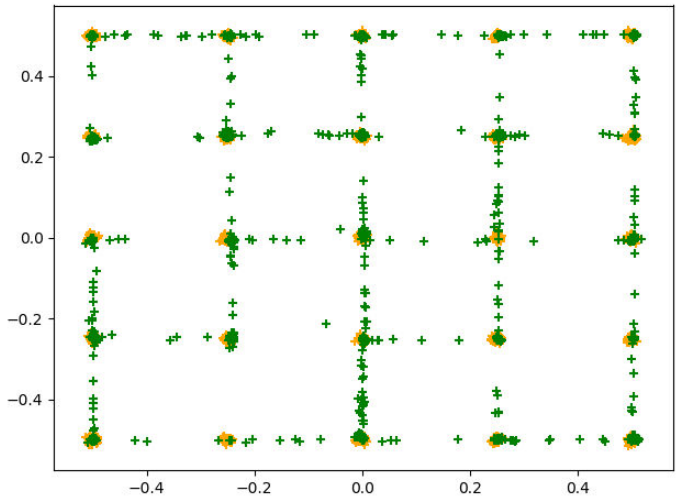
Experiments - Mode Collapse



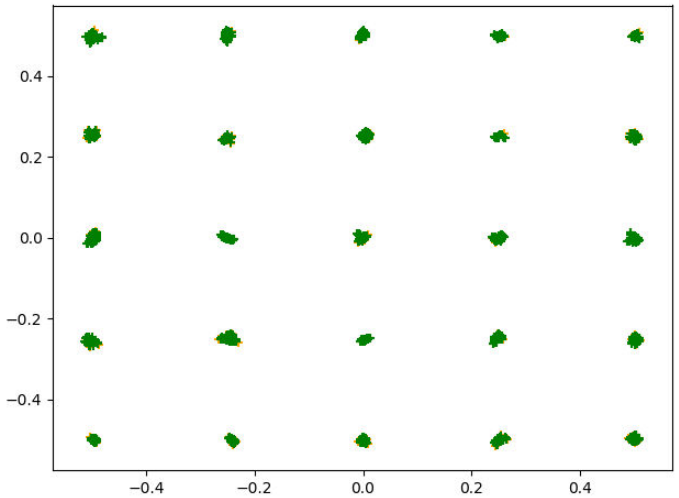
(a) original



(b) GAN



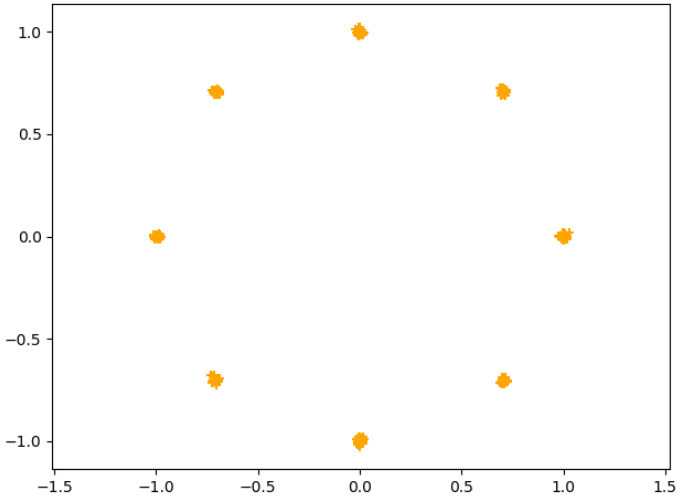
(c) pacgan



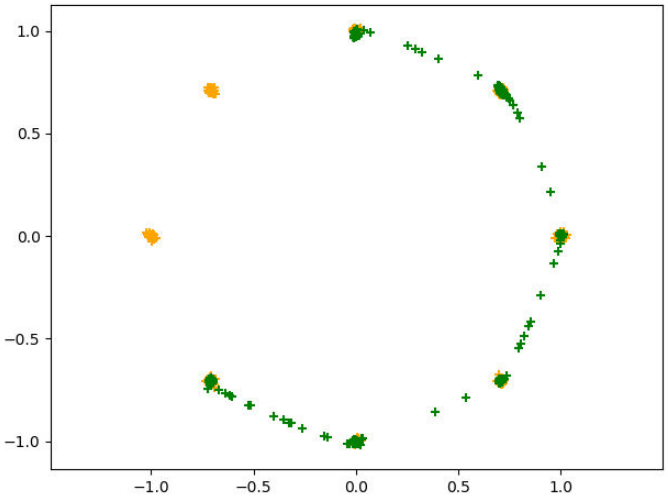
(d) Our model, AE-OT

Figure: Comparison between conventional models with AE-OT

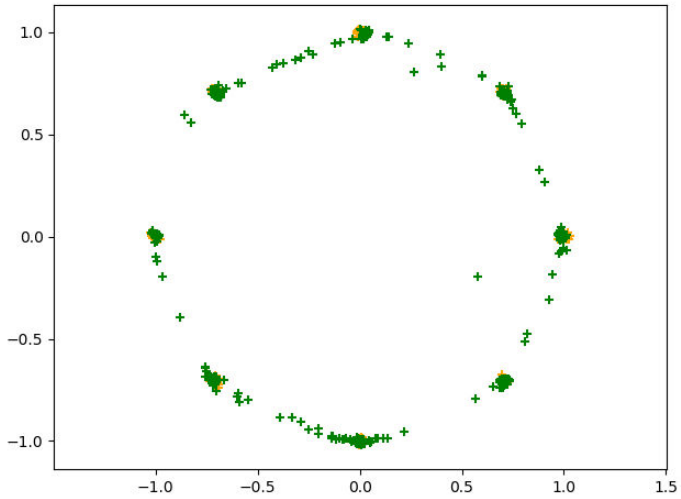
Experiments - Mode Collapse



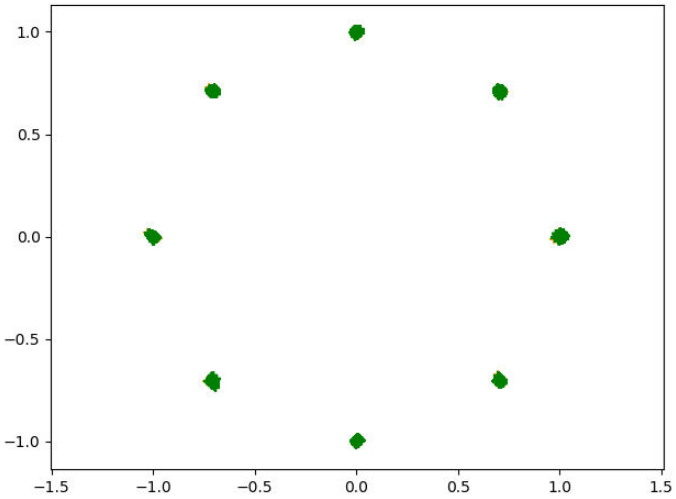
(a) original



(b) GAN



(c) pacgan



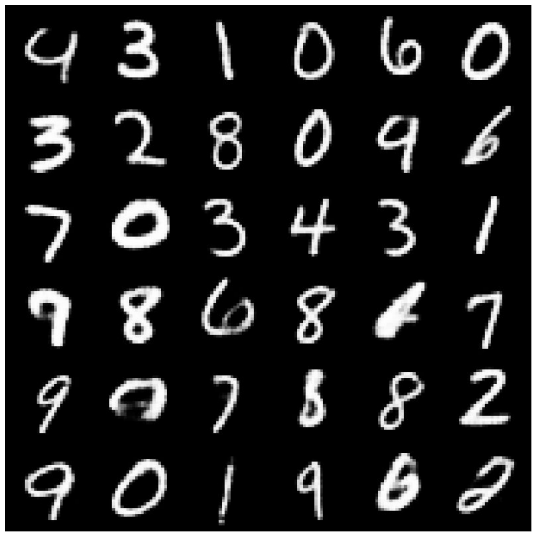
(d) Our model, AE-OT

Figure: Comparison between conventional models with AE-OT

Experiments - Mnist

(a) AE reconstruction

(b) Lucic et al. 2018



(c) Hoshen and Malik 2019

(d) AE-OT

Experiments - Mnist Fashion

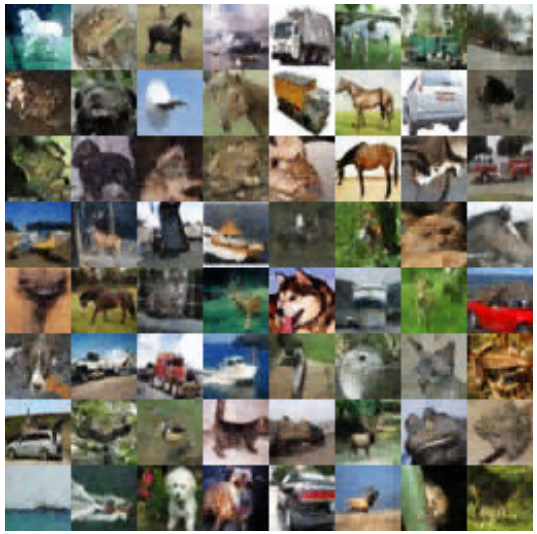
(a) AE reconstruction

(b) Lucic et al. 2018

(c) Hoshen and Malik 2019

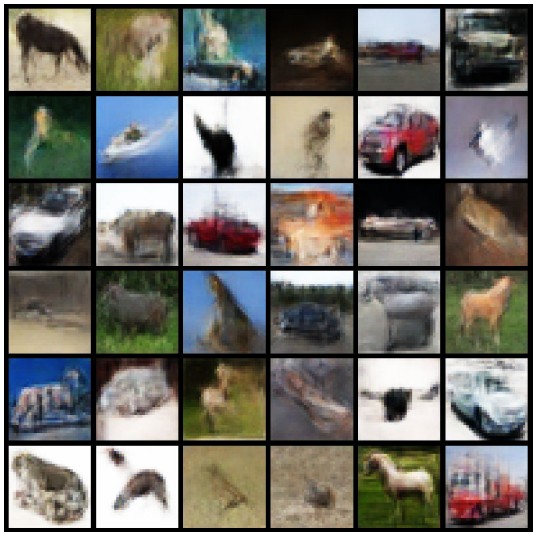
(d) AE-OT

Experiments - Cifar 10

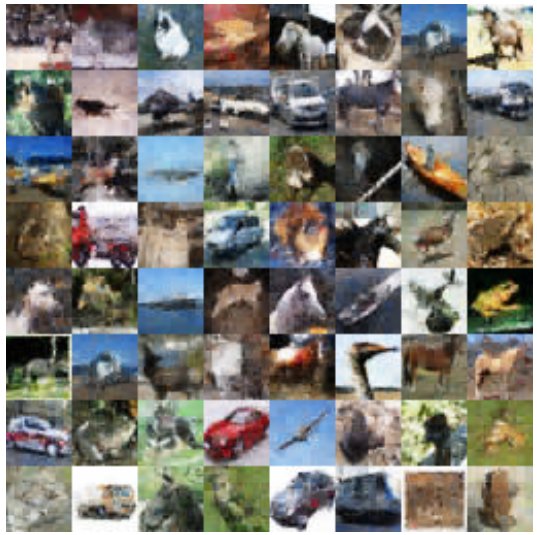


(a) AE reconstruction

(b) Lucic et al. 2018

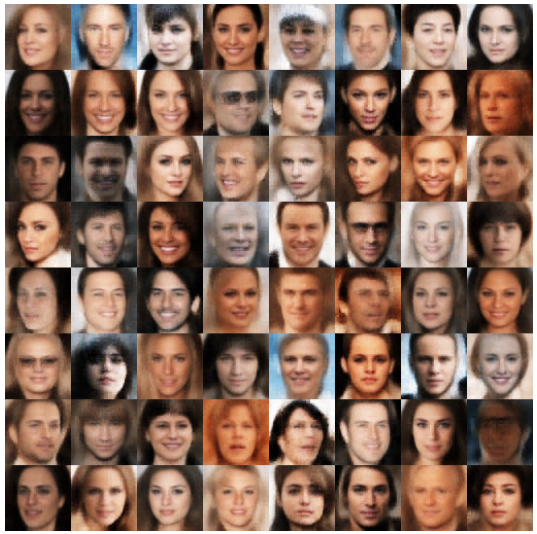


(c) Hoshen and Malik 2019



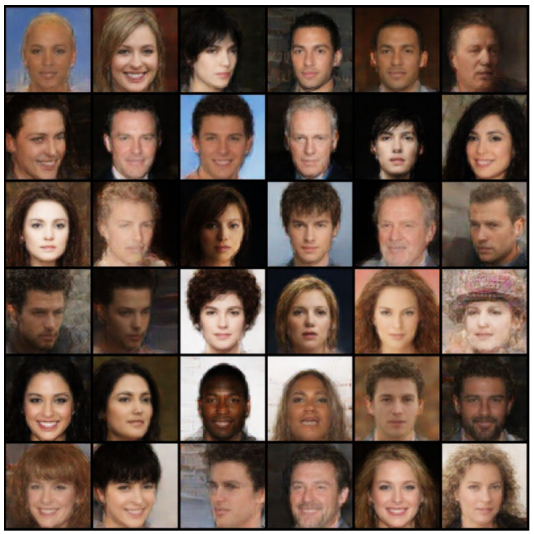
(d) AE-OT

Experiments - CelebA

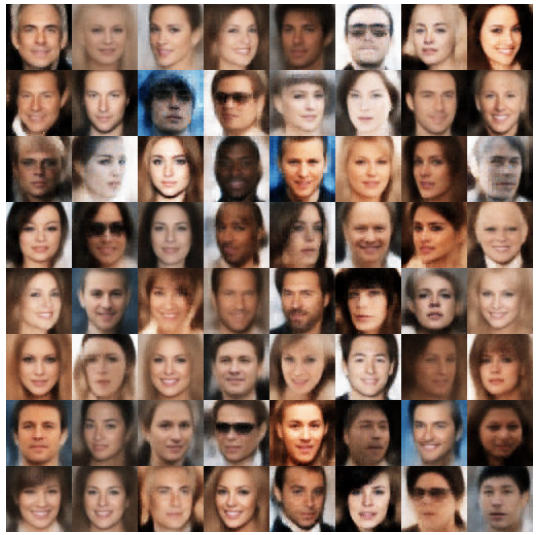


(a) AE reconstruction

(b) Lucic et al. 2018

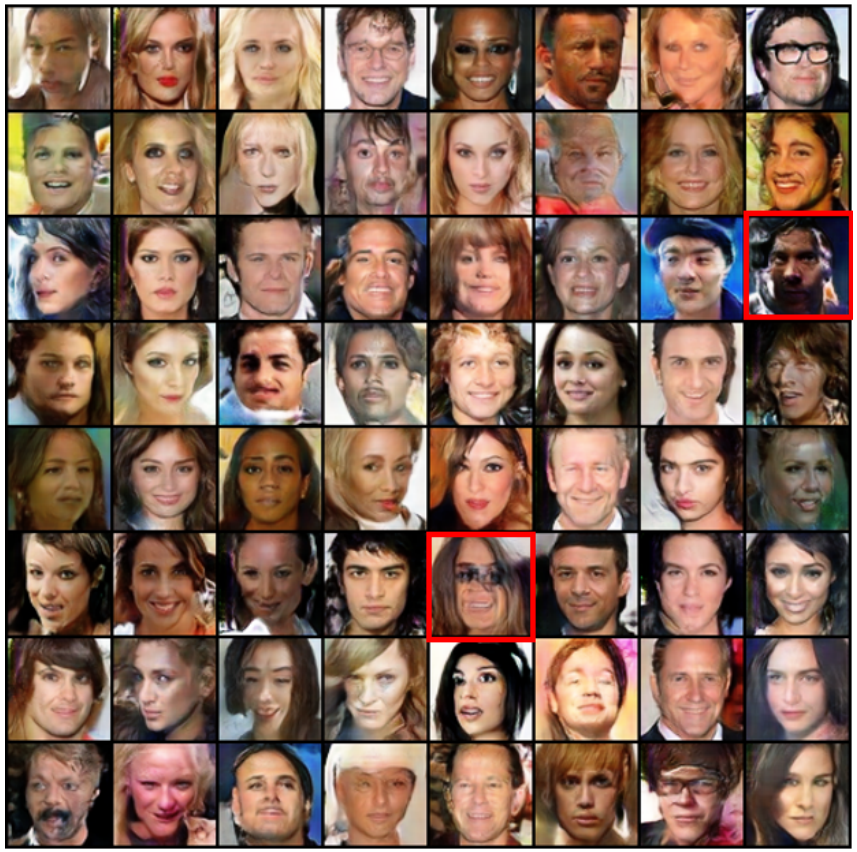


(c) Hoshen and Malik 2019

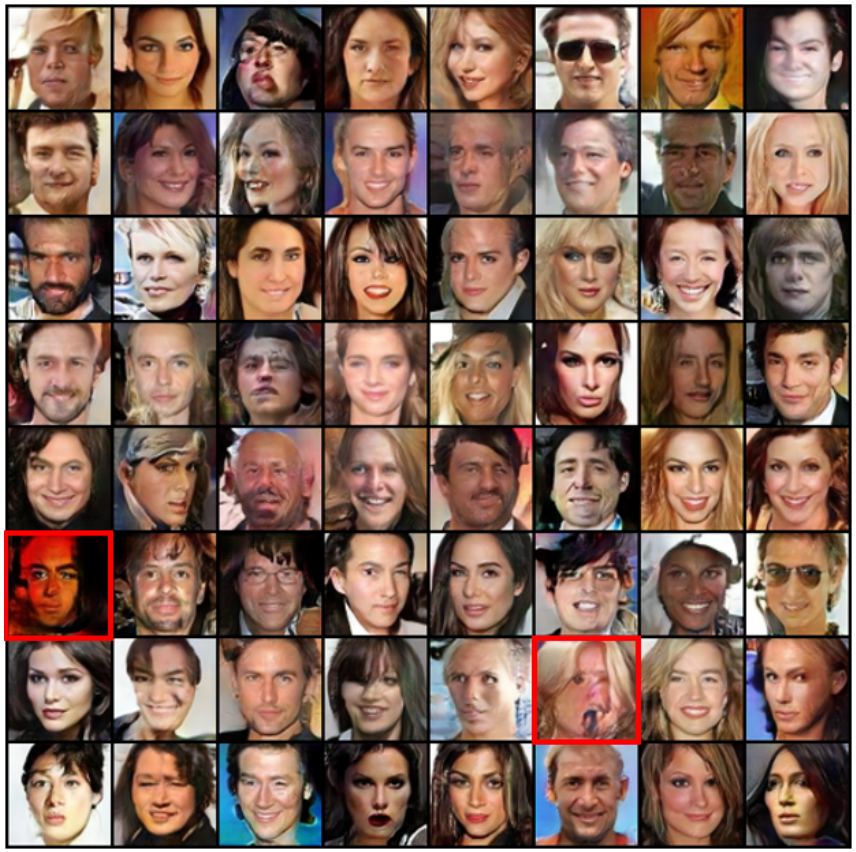


(d) AE-OT

Experiments - CelebA



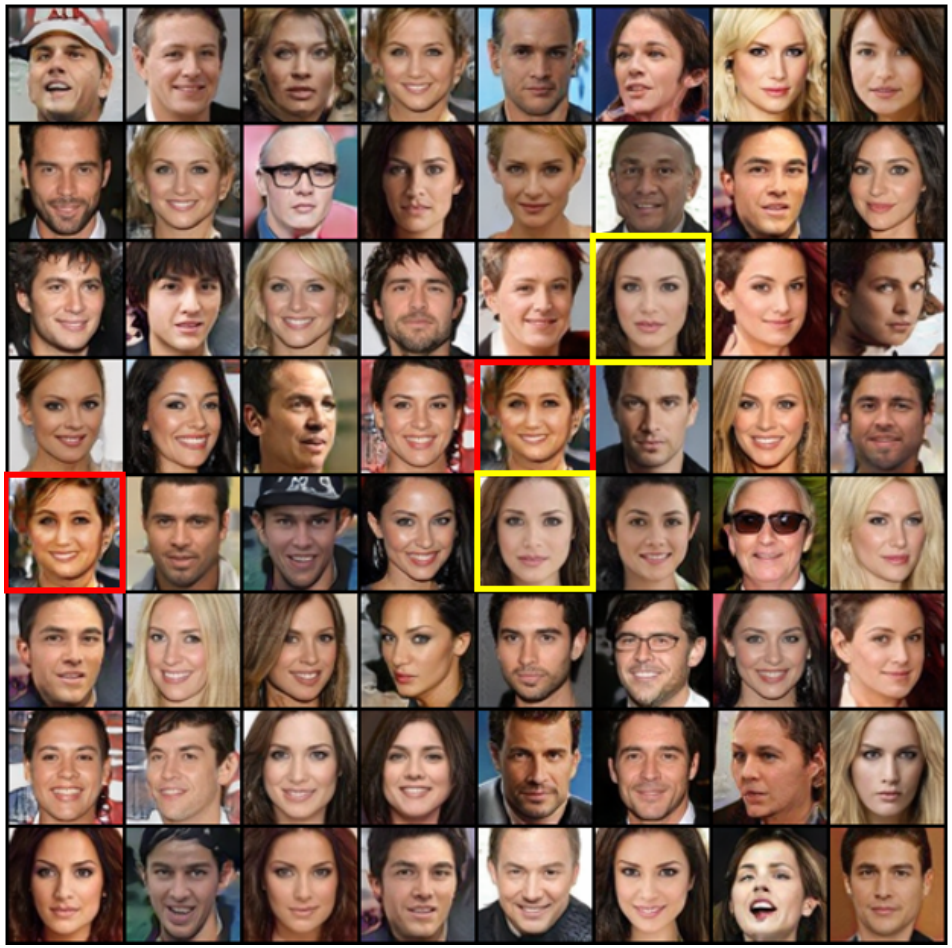
(a) WGAN-GP



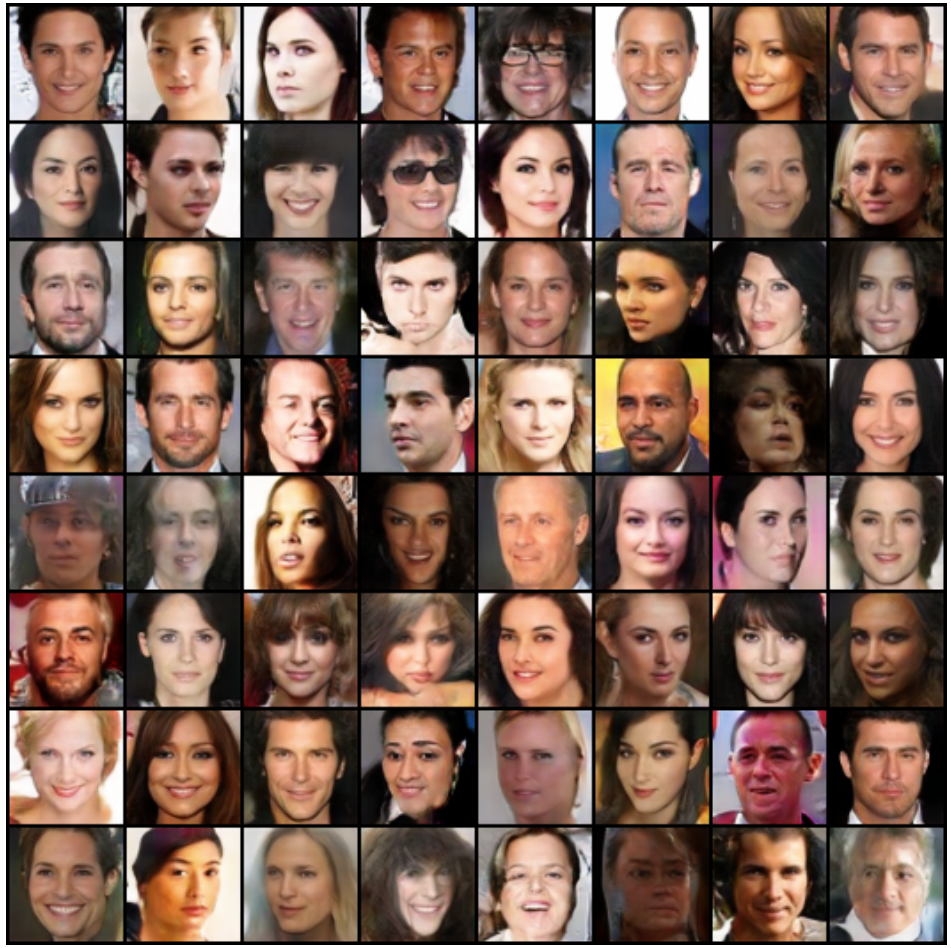
(b) WGAN-div

Figure: Failure cases for WGAN-GP and WGAN-div.

Experiments - CelebA



(c) CRGAN



(d) AE-OT

Figure: Mode collapsing of CRGAN.

Experiments - AE-OT interpolation

Figure: Curves on facial photo manifold.

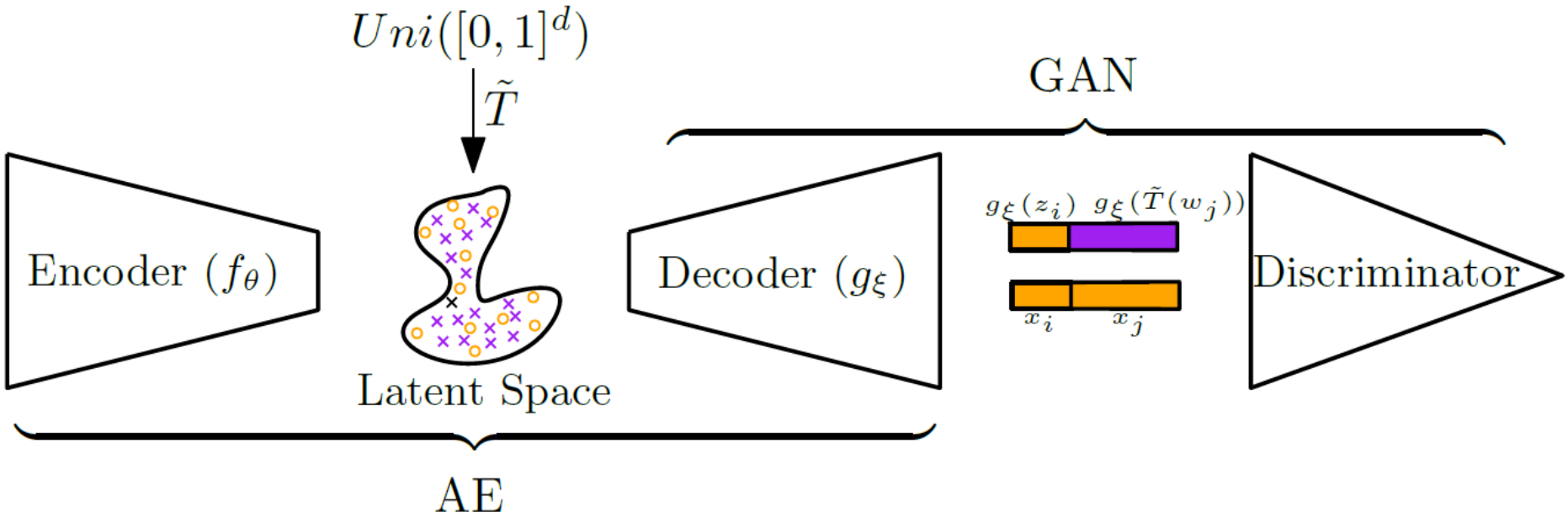
Quantitative Comparison with FID

	Adversarial			
Dataset	NS GAN	LSGAN	WGAN	BEGAN
MNIST	6.8	7.8	6.7	13.1
Fansion	26.5	30.7	21.5	22.9
CIFAR-10	58.5	87.1	55.2	71.4
CelebA	55.0	53.9	41.3	**38.9**

	Non-Adversarial			Reference	
Dataset	VAE	GLO	GLANN	AE Reconstruction	Ours
MNIST	23.8	49.6	8.6	5.5	**6.4**
Fansion	58.7	57.7	13.0	4.7	**10.2**
CIFAR-10	155.7	65.4	46.5	28.2	**38.1**
CelebA	85.7	52.4	46.3	67.5	68.4

AE-OT-GAN Model

AE-OT-GAN Framework

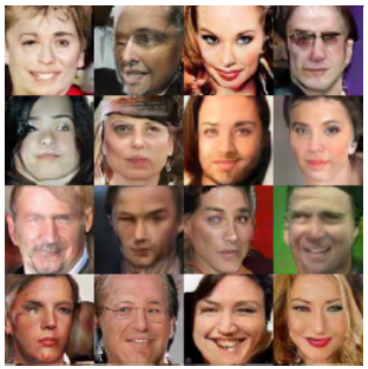


For further improve the quality of the generated images, we proposed AE-OT-GAN model.

AE-OT-GAN Framework

(a) Epoch 0 (b) Epoch 80 (c) Epoch 160 (d) Epoch 240 (e) GT

AE-OT-GAN Framework



CT-GAN [42]

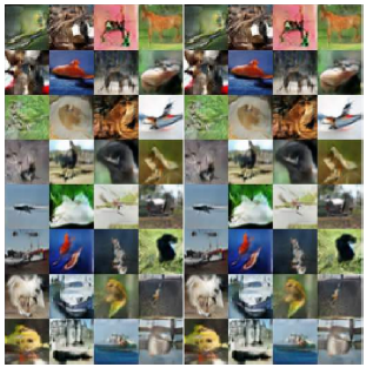
WGAN-GP [12]

WGAN-div [39]

WGAN-QC [24]

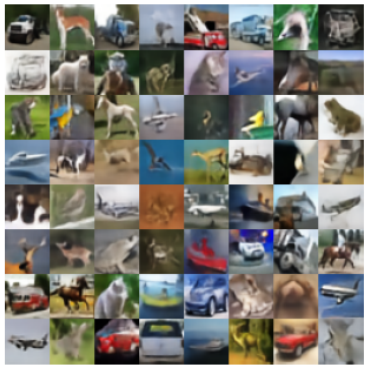
Proposed method

WGAN-GP [12]



SNGAN [27]

WGAN-div [39]



AE-OT [2]

Proposed method

Comparison between the proposed method and the SOTA methods.

AE-OT-GAN Framework

	CIFAR10				CelebA	
	Standard		Resnet		Standard	Resnet
	FID	IS	FID	IS	FID	FID
WGAN-GP [12]	40.2	6.68	19.6	7.86	21.2	18.4
PGGAN [14]	-	-	18.8	8.80	-	16.3
SNGAN [27]	25.5	7.58	21.7	8.22	-	-
WGAN-div [39]	-	-	18.1	-	17.5	15.2
WGAN-QC [24]	-	-	-	-	-	12.9
AE-OT [2]	34.2	6.62	28.5	7.67	24.3	28.6
AE-OT-GAN	25.2	7.62	17.1	8.24	11.2	7.6

AE-OT-GAN Framework

Table 1. Experiments on stacked MNIST.

	Stacked MNIST	
	Modes	KL
DCGAN	99.0	3.40
VEEGAN	150.0	2.95
PacDCGAN4	1000.0 ± 0.00	0.07 ± 0.005
WGAN	314.3 ± 38.54	2.44 ± 0.170
AE-OT	1000.0 ± 0.00	0.03 ± 0.0008
AE-OT-GAN	1000.0 ± 0.0	0.05 ± 0.006

AE-OT-GAN Interpolation

The interpolation of the AE-OT-GAN model. The left column shows the generated images, and the right 5 images are the ones used to generate the left images in the latent space..

Conclusion

This work introduces a geometric understanding of deep learning:

- The intrinsic pattern of natural data can be represented by manifold distribution hypothesis.
- The deep learning system has two major tasks: manifold learning and probability distribution transformation.
- Optimal transportation methods can be used to accomplish the second task.
- The regularity theory of Monge-Ampère equation explains mode collapse.
- The AE-OT framework can avoid mode collapse, and make half the blackbox transparent.

Thanks

Thank you!

For more information, please email to nalei@dlut.edu.cn.