Recent Topics on Conditional Generative Models

Young-geun Kim

Younggeun.Kim@nyspi.columbia.edu
Department of Psychiatry and Department of Biostatistics
Columbia University

April 18, 2024

What is the Conditional Generative Model

Figure: Examples in video interpolation. Two intermediate frames (X) are generated with two preceding and two following ones (C).

- Generative models are statistical models of the joint distribution of observations (X) and latent factors (Z).
- **Conditional generative models** are statistical models of $\mathbb{P}_{(X,Z)|C}$ where C is additional conditioning data.

Image source: Szeto, Ryan, et al. "A temporally-aware interpolation network for video frame inpainting." *IEEE TPAMI* (2019).

Application: Conditional Image Synthesis

Figure: Images generated by DALL-E-2 which received

> *"There is a clean desk in the middle. Outside the window, a whale shark is swimming in the dark night sky above Manhattan."*

Generated images (X) reflect semantic information in text descriptions (C).

Image source: DALL-E-2.

Application: Medical Imaging Analysis

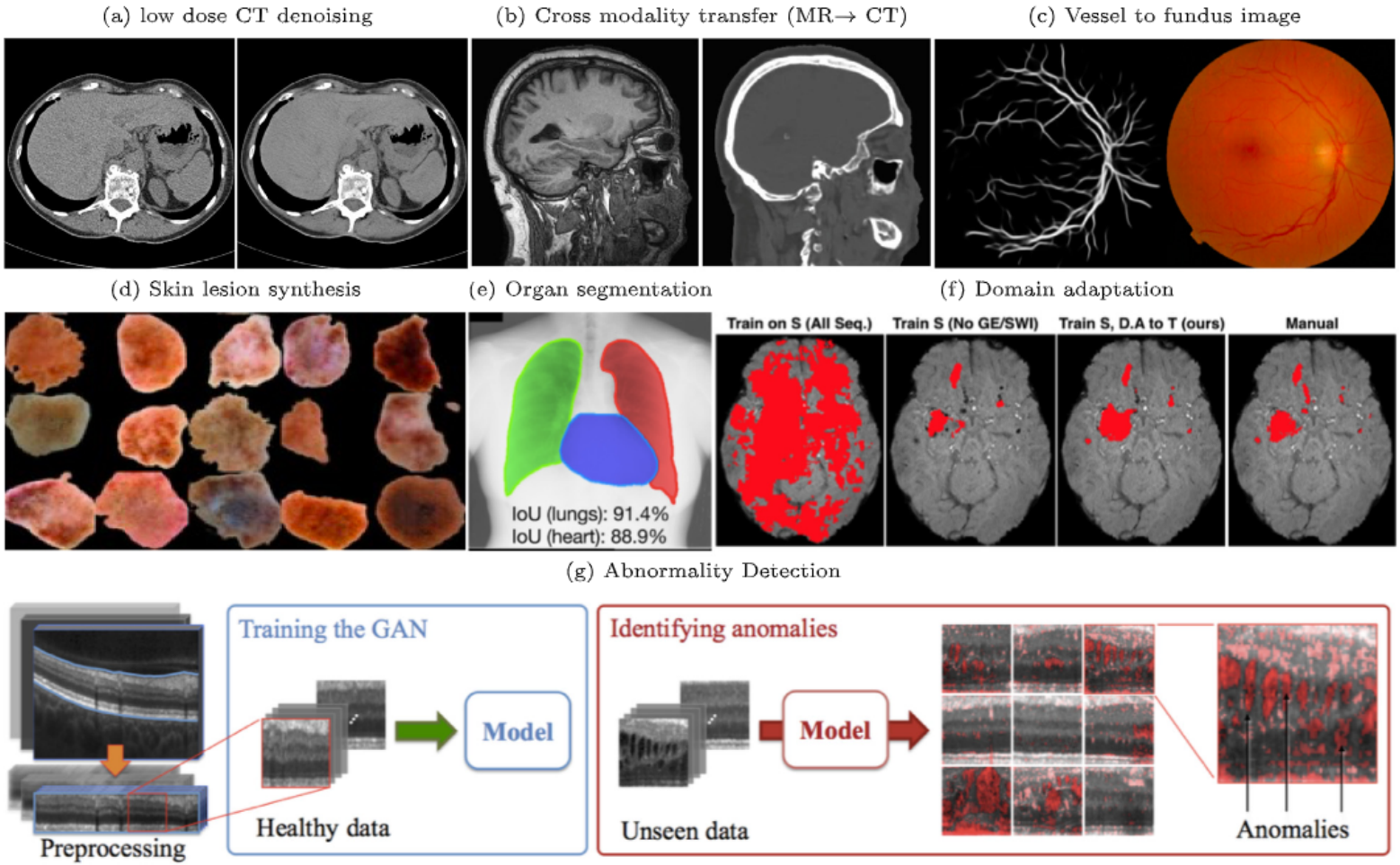


Image source: Yi, Xin, Ekta Walia, and Paul Babyn. "Generative adversarial network in medical imaging: A review." *Medical image analysis* (2019).

Important Topics

- Recently, conditional generative models have shown remarkable performances in
  1. (Density estimation) generating realistic samples following $\mathbb{P}_{X|C}$.
  2. (Representation learning) extracting representations following $\mathbb{P}_{Z|X,C}$.
  3. (Scientific exploratory analysis) understanding high-dimensional genetic/brain imaging data associated with phenotypes.
- In this talk, we focus on (i) statistical distances between conditional distributions and (ii) identifiable representation learning with covariates.

Statistical Distances between Conditional Distributions

Learning Distributions

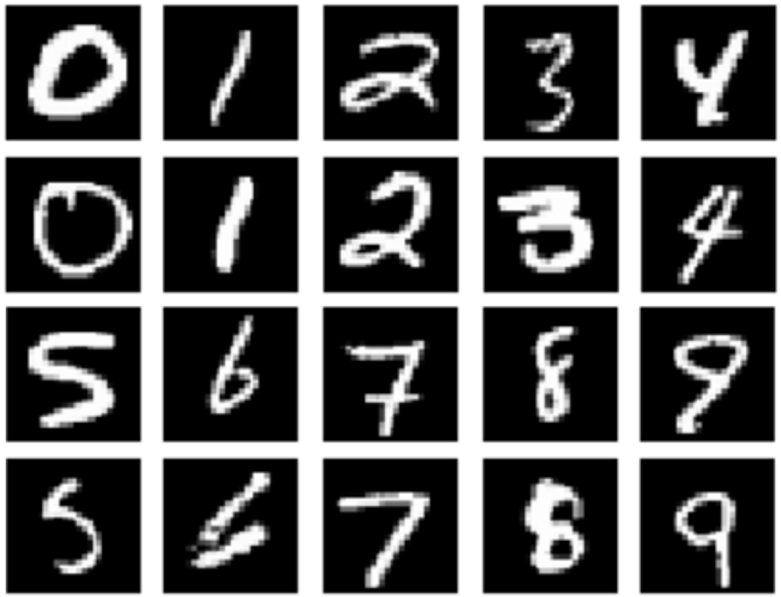
Real

Generated

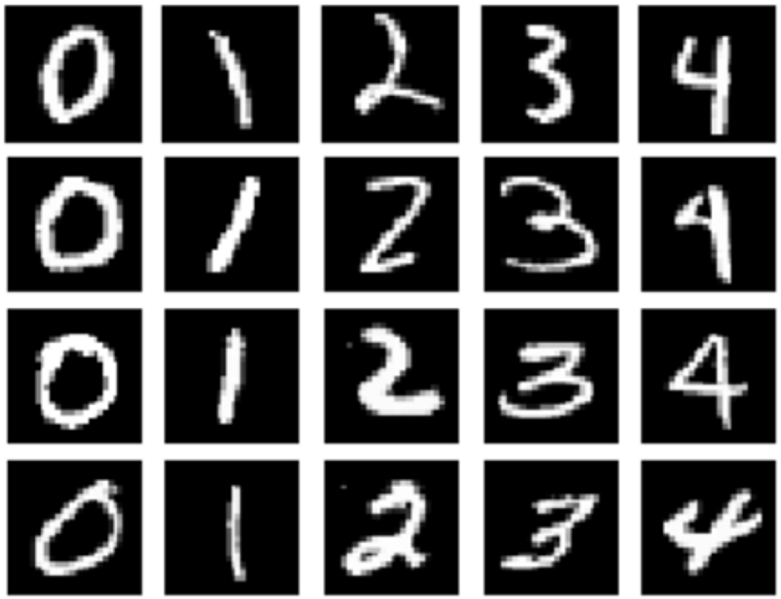
- The generated sample in the right figure seems realistic handwritten digit. Can we say that the generator learned data distributions well?
  $\rightarrow$ The generated data distribution has positive probability for the point where the digit zero image (right) locates. However, the given information is not enough to answer the question.

Image source: MNIST

Learning Distributions



Real



Generated

- When we see collection of generated data, the probability of getting digits from 5 to 9 is zero.
- Thus, we need to compare distributions to learn generative models.

Statistical Distances

- **Statistical distances** are important tools to measure discrepancy between distributions.
- Deep generative models learn $\mathbb{P}_X$ by minimizing

$$\mathcal{D}(\mathbb{P}_X, \mathbb{P}_{G_\theta(Z)})$$

  where G_θ and $\mathcal{D}$ denotes generator networks and statistical distances.
- Popular statistical distances include
  1. f-divergence
  2. Integral probability metrics (IPMs)
  3. Wasserstein distance

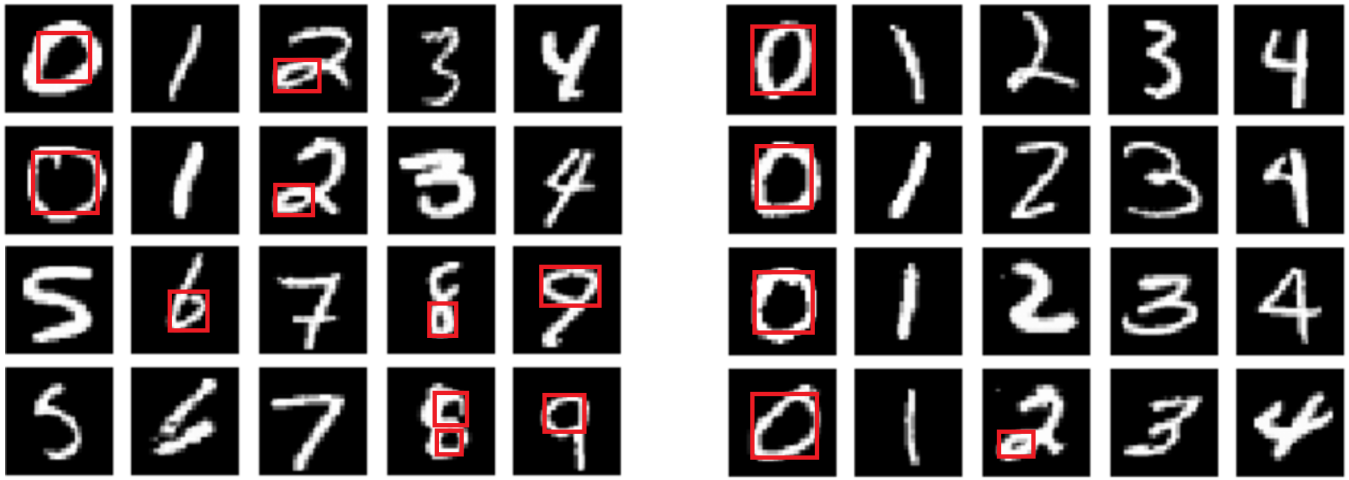
Statistical Distances: f-divergence

- The **f-divergences** are expectations of density ratios mapped by convex functions f satisfying $f(1) = 0$.
- The f-divergence can be expressed as

$$\mathcal{D}_f(\mathbb{P}||\mathbb{Q}) := \int f(p(x)/q(x))q(x)dx$$

  where p and q are density functions of distributions $\mathbb{P}$ and $\mathbb{Q}$.
- Popular generative models using f-divergence include variational autoencoders (VAEs) and generative adversarial networks (GANs).
- For example, $\mathcal{D}_f$ is the KL-divergence when $f(u) = u \log u$ and is the objective of GANs when $f(u) = u \log u - (u+1)\log(u+1) + 2\log 2$.

Statistical Distances: Integral Probability Metrics



Real Generated

- The **integral probability metrics** (IPMs) between distributions are the largest difference between their summary statistics.
- For example, when we use the number of circles in images as summary statistics, the difference is 0.5 (real) − 0.25 (generated) = 0.25.
- We can consider many summary statistics together to precisely compare distributions.

Image source: MNIST

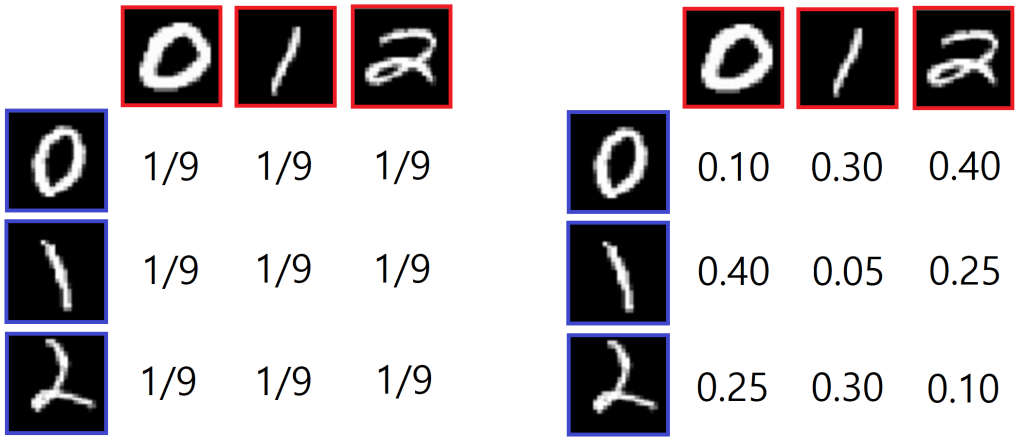
Statistical Distances: Integral Probability Metrics

- The IPMs can be expressed as

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x) d\mathbb{P}(x) - \int_{\mathcal{X}} f(x) d\mathbb{Q}(x) \right|$$

  where $\mathcal{F}$ is a class of real-valued functions.
- Wasserstein GANs are popular methods optimizing 1-Wasserstein distance. The 1-Wasserstein distance is the unique IPM among Wasserstein distances.

Statistical Distances: Wasserstein Distance



	0	1	2
0	1/9	1/9	1/9
1	1/9	1/9	1/9
2	1/9	1/9	1/9

Joint Distribution

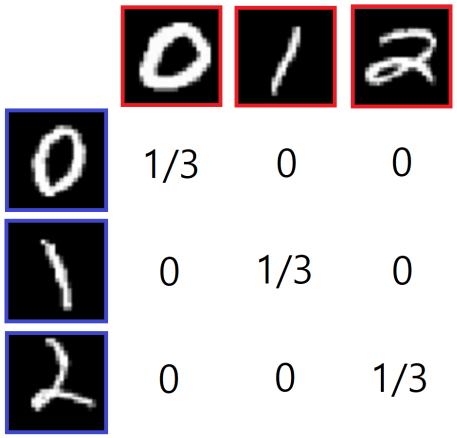
	0	1	2
0	0.10	0.30	0.40
1	0.40	0.05	0.25
2	0.25	0.30	0.10

Transportation Cost

- **Wasserstein distance** is the minimum expected transportation cost between two distributions.
- For example, we can consider the above joint distribution. The transportation cost is

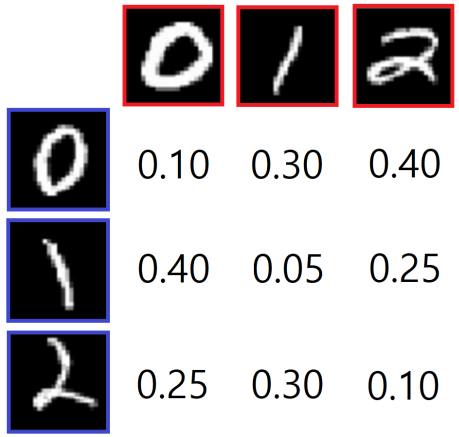
$$
\begin{aligned}
&(1/3)\big(0.10(1/3) + 0.30(1/3) + 0.40(1/3)\big) \\
&+ (1/3)\big(0.40(1/3) + 0.05(1/3) + 0.25(1/3)\big) \\
&+ (1/3)\big(0.25(1/3) + 0.30(1/3) + 0.10(1/3)\big) = 0.24.
\end{aligned}
$$

Statistical Distances: Wasserstein Distance



	0	1	2
0	1/3	0	0
1	0	1/3	0
2	0	0	1/3

Joint Distribution



	0	1	2
0	0.10	0.30	0.40
1	0.40	0.05	0.25
2	0.25	0.30	0.10

Transportation Cost

- We can consider another joint distribution. The transportation cost is

$$
\begin{aligned}
&(1/3)\big(0.10(1) + 0.30(0) + 0.40(0)\big) \\
&+ (1/3)\big(0.40(0) + 0.05(1) + 0.25(0)\big) \\
&+ (1/3)\big(0.25(0) + 0.30(0) + 0.10(1)\big) = 0.08.
\end{aligned}
$$

Statistical Distances: Wasserstein Distance

- The p-**Wasserstein distance** can be expressed as

$$W_p(\mathbb{P}, \mathbb{Q}; d) := \left(\inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X}^2} d^p(x, y) d\pi(x, y) \right)^{1/p}$$

  where $p \in [1, \infty)$ and $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of all joint distributions whose marginals are $\mathbb{P}$ and $\mathbb{Q}$. Closed form in special cases (Gaussian with L)

- Wasserstein autoencoders (WAEs) are generative models using p-Wasserstein distances.
- As an alternative to f-divergences, Wasserstein distances have gained recognition. Wasserstein distances have advantages over f-divergences when data supports are in low-dimensional manifolds.

Advantages of Wasserstein Distance

Example 1

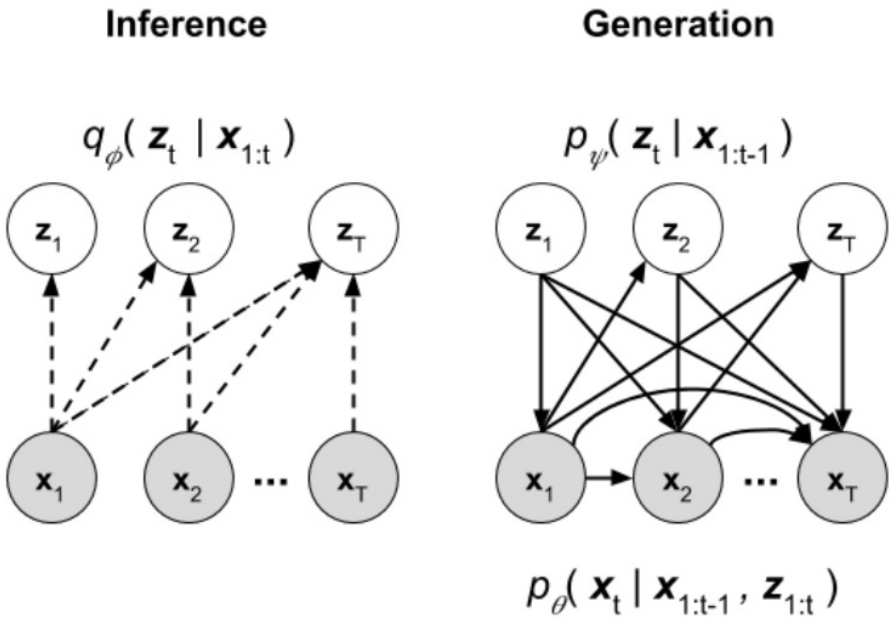
(Example 1 in WGAN paper) Let $Z \sim U[0,1]$, $X = (0, Z)$, and $G_\theta(Z) = (\theta, Z)$. The data supports are lines in the plane.

- Intuitively, $\mathcal{D}(\mathbb{P}_X, \mathbb{P}_{G_\theta(Z)})$ should decrease as θ closes to zero.
- However, when $\theta \neq 0$, JS-divergence is $\log 2$ and KL-divergence is not defined well. In contrast, $W_p(\mathbb{P}_X, \mathbb{P}_{G_\theta(Z)}) = \|\theta\|$.
- Empirically, Wasserstein distance-based methods have produced sharper images than f divergence-based methods in image applications.

Conditional Wasserstein Generator

- What statistical distances should we target to learn conditional generative models?
- I proposed a conditional generative model minimizing Wasserstein distances between conditional distributions.
  - **Kim, Y.-G.**, Lee, K., and Paik, M.C. "Conditional Wasserstein generator." *IEEE TPAMI 2022*.
- This work is (i) an extension of WAE to conditional generation and (ii) a Wasserstein counterpart of f divergences-based conditional generation.

Conditional VAEs in Video Generation



- Stochastic video generation with a learned prior (SVG-LP, Denton and Fergus 2018) is an extension of VAE to conditional generation. The conditioning data C are all the past image frames, $x_{1:t-1}$.
- The prior distribution (without information at the target time step), $\mathbb{P}_{Z|C}(\cdot|\cdot;\psi)$, reflects information in conditioning data.

Image source: Denton, Emily, and Rob Fergus. "Stochastic video generation with a learned prior." *ICML* (2018).

Conditional VAEs in Video Generation

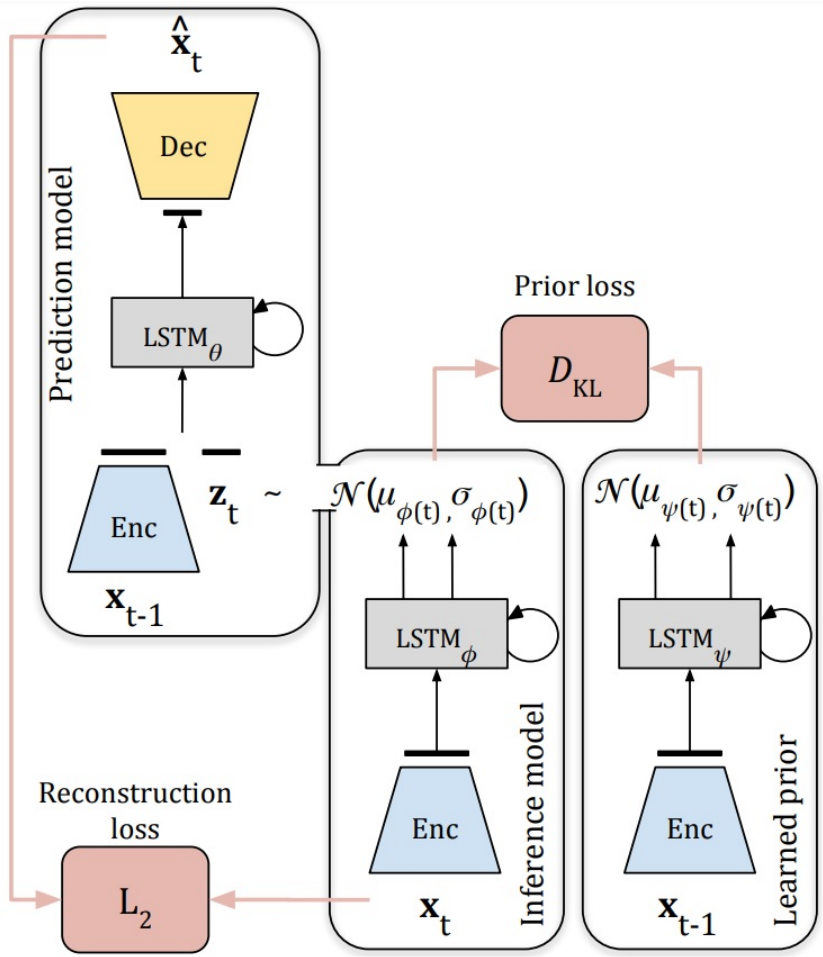


Image source: Denton, Emily, and Rob Fergus. "Stochastic video generation with a learned prior." *ICML* (2018).

Principle Behind Existing Methods

- We derived that SVG-LP minimizes (an upper bound of) **conditional KL-divergence**,

$$\int_{\mathcal{C}} \mathcal{D}_{\mathsf{KL}}(\mathbb{P}_{X|C}(\cdot|c;\theta,\psi)||\mathbb{P}_{X|C}(\cdot|c))d\mathbb{P}_C(c),$$

  where (θ,ψ) is the network parameter.
- A principle behind the existing method is to measure discrepancy between conditional distributions first, and then average them over the distribution of conditioning data.
- We extended this observation to found a conditional generative model framework minimizing

$$\int_{\mathcal{C}} \mathcal{D}(\mathbb{P}_{X|C}(\cdot|c;\theta,\psi),\mathbb{P}_{X|C}(\cdot|c))d\mathbb{P}_C(c),$$

  conditional statistical distances.

Challenges in Minimizing Conditional Statistical Distances

- The conditional statistical distance is intuitively make sense, but measuring $\mathcal{D}(\mathbb{P}_{X|C}(\cdot|c;\theta,\psi)||\mathbb{P}_{X|C}(\cdot|c))$ for all realization c is practically infeasible.
- To overcome this challenge, we derived the relation between conditional statistical distances and

$$\mathcal{D}(\mathbb{P}_{X|C}(\cdot|\cdot;\theta,\psi)\mathbb{P}_C, \mathbb{P}_{X|C}\mathbb{P}_C),$$

distances between joint distributions for three statistical distances: (i) f-divergence, (ii) IPMs, and (iii) Wasserstein distance.

Conditional Statistical Distances and Joint Statistical Distances

Theorem 1

(Kim et al., 2022)

- *(f-divergence) For any convex function $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(1) = 0$,*

$$\int_{\mathcal{C}} \mathcal{D}_f(\mathbb{P}_{X|C}(\cdot|c;\theta,\psi)||\mathbb{P}_{X|C}(\cdot|c))d\mathbb{P}_C(c) = \mathcal{D}_f(\mathbb{P}_{X|C}(\cdot|\cdot;\theta,\psi)\mathbb{P}_C||\mathbb{P}_{X,C}).$$

- *(IPMs) Let $D(\{\mathcal{F}_c\}_{c\in\mathcal{C}}) := \{f|\exists\{\bar{f}_c\}_{c\in\mathcal{C}}$ s.t. $\bar{f}_c \in \mathcal{F}_c$ and $f(\cdot,c) = \bar{f}_c\}$. Then,*

$$\int_{\mathcal{C}} \gamma_{\mathcal{F}_c}(\mathbb{P}_{X|C}(\cdot|c;\theta,\psi),\mathbb{P}_{X|C}(\cdot|c))d\mathbb{P}_C(c) = \gamma_{D(\{\mathcal{F}_c\}_{c\in\mathcal{C}})}(\mathbb{P}_{X|C}(\cdot|\cdot;\theta,\psi)\mathbb{P}_C,\mathbb{P}_{X,C}).$$

- *(Wasserstein distance) For any $p \in [1,\infty)$,*

$$\int_{\mathcal{C}} W_p(\mathbb{P}_{X|C}(\cdot|c;\theta,\psi),\mathbb{P}_{X|C}(\cdot|c);d_{\mathcal{X}})d\mathbb{P}_C(c) \leq W_p(\mathbb{P}_{X|C}(\cdot|\cdot;\theta,\psi)\mathbb{P}_C,\mathbb{P}_{X,C};d_{\mathcal{X}}\oplus d_{\mathcal{C}}).$$

Conditional Wasserstein Generator

- Based on established relations, we proposed a new conditional generative model, *conditional Wasserstein generator*, minimizing

$$\int_{\mathcal{C}} W_p(\mathbb{P}_{X|C}(\cdot|c;\theta,\psi), \mathbb{P}_{X|C}(\cdot|c); d_{\mathcal{X}}) d\mathbb{P}_C(c).$$

- In conditional generative model, no previous work has attempted to apply p-Wasserstein distance in quantifying discrepancy between conditional distributions. Our work filled this gap.
- The proposed method is a Wasserstein counterpart of SVG-LP and an extension of WAEs into conditional generation.

Conditional Wasserstein Generator

Theorem 2

(Kim et al., 2022) Let $G_\theta : \mathcal{C} \times \mathcal{Z} \to \mathcal{X}$ be 1-uniformly Lipschitz continuous in $\mathcal{C}$. Then, for any $p \in [1, \infty)$,

$$W_p(\mathbb{P}_{X|C}(\cdot|\cdot;\theta,\psi)\mathbb{P}_C, \mathbb{P}_{X,C}; d_{\mathcal{X}} \oplus d_{\mathcal{C}})$$
$$= \left(\inf_{\mathbb{Q}_{Z|X,C} \in \mathcal{Q}_\psi} \int_{\mathcal{X}\times\mathcal{C}} \int_{\mathcal{Z}} d^p_{\mathcal{X}}(x, G_\theta(c,z)) d\mathbb{Q}_{Z|X,C}(z|x,c) d\mathbb{P}_{X,C}(x,c) \right)^{1/p},$$

where $\mathcal{Q}_\psi$ is the set of all $\mathbb{Q}_{Z|X,C}$ satisfying $\int_{\mathcal{X}} \mathbb{Q}_{Z|X,C}(\cdot|x,\cdot) d\mathbb{P}_{X|C}(x|\cdot)\mathbb{P}_C = \mathbb{P}_{Z|C}(\cdot|\cdot;\psi)\mathbb{P}_C$.

joint distribution derived by using posterior and using prior should be the same

- We derived a tractable representation of Wasserstein distances for conditional generative models.

Conditional Wasserstein Generator

- Motivated by the previous result, conditional Wasserstein generator solves a constraint optimization problem,

$$\begin{aligned} \underset{G_\theta}{\text{minimize}} \quad & \int_{\mathcal{X}\times\mathcal{C}} \int_{\mathcal{Z}} d_{\mathcal{X}}^{p}(x, G_\theta(c,z)) d\mathbb{Q}_{Z|X,C}(z|x,c) d\mathbb{P}_{X,C}(x,c) \\ \text{subject to} \quad & \int_{\mathcal{X}} \mathbb{Q}_{Z|X,C}(\cdot|x,\cdot) d\mathbb{P}_{X|C}(x|\cdot)\mathbb{P}_C = \mathbb{P}_{Z|C}(\cdot|\cdot;\psi)\mathbb{P}_C \\ & K^{\mathcal{Z}}_{d_{\mathcal{C}},d_{\mathcal{X}}}(G_\theta) \leq 1. \end{aligned}$$

The method of Lagrange multipliers is applied.

Video Generation Results

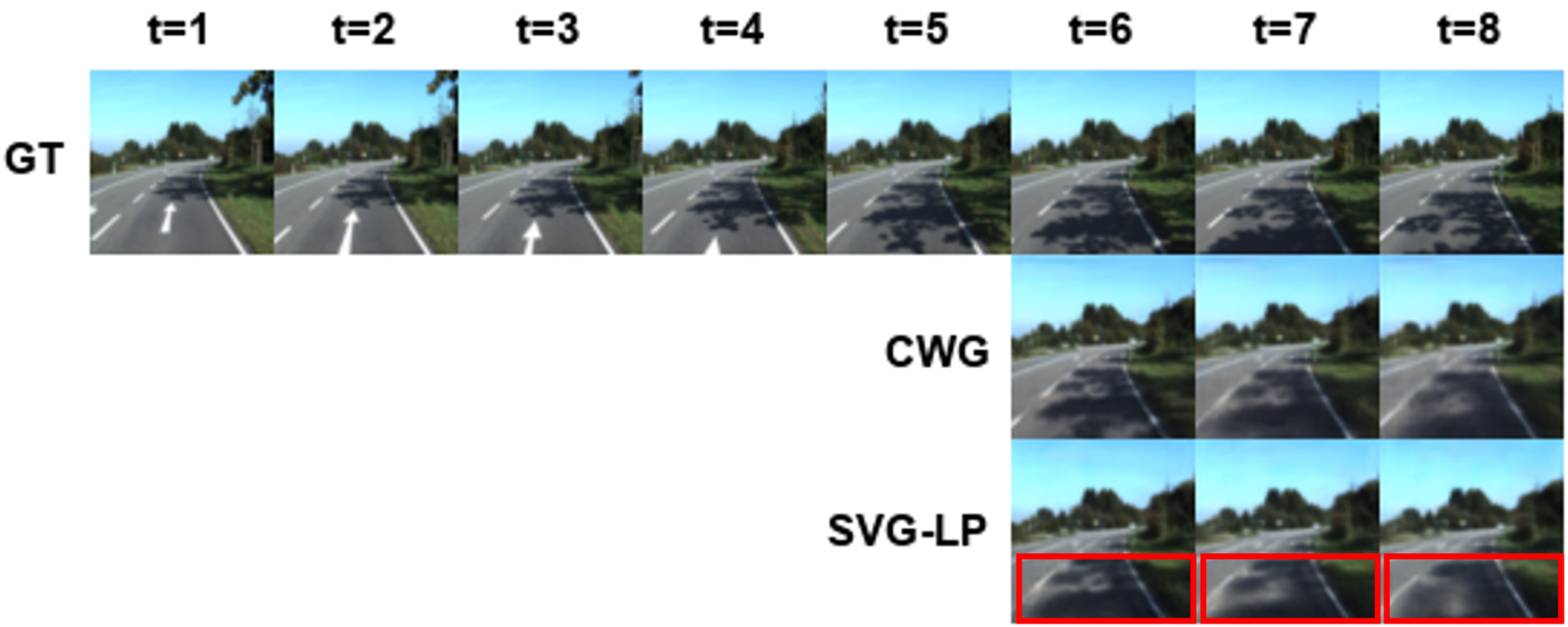
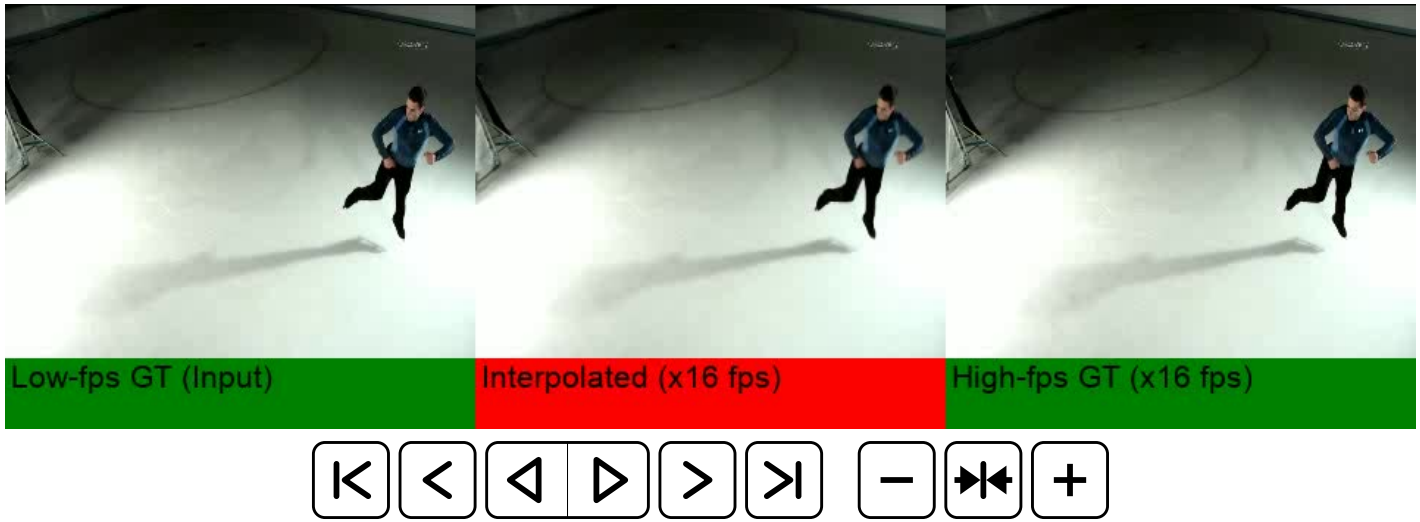


Figure: (KITTI) We used five past frames to predict future frames. In red rectangles, the SVG-LP produces blurry shadows and paints the nearby background black.

Video Generation Results



- We used one past and one future frame to interpolate intermediate fourteen frames.
- Our model smoothly interpolated both the figure skater and the shadow movement on the ice.

Video Generation Results

Table: Means of test scores with standard errors for all **predicted** videos at all time steps.

Dataset	Method	SSIM (↑)	PSNR (↑)	Sharpness (↑)	LPIPS (↓)
KTH	SVG-LP	.885 (.004)	25.784 (.261)	.00257 (.00014)	.128 (.00452)
	CWG	**.890** (.005)	**25.928** (.245)	**.00294** (.00016)	**.119** (.00435)
BAIR	SVG-LP	.798 (.006)	17.275 (.199)	.0514 (.00054)	.107 (.00399)
	CWG	**.802** (.007)	**17.472** (.232)	**.0525** (.00055)	**.096** (.00382)
Towel Pick	SVG-LP	.782 (.002)	20.614 (.064)	**.00806** (.00007)	.141 (.001)
	CWG	**.789** (.002)	**21.180** (.069)	.00804 (.00007)	**.135** (.002)
KITTI	SVG-LP	.531 (.010)	15.802 (.179)	.00535 (.00013)	.266 (.006)
	CWG	**.557** (.009)	**16.386** (.166)	**.00542** (.00013)	**.259** (.006)

Identifiable Representation Learning

Identifiable Representation Learning

- Generative models have extracted features from high-dimensional data. However, their representations have an issue.
- Is learning $p(x)$ sufficient to learn $p(z|x)$? That is, $\int p_{\theta^*}(x, z)dz = p(x)$ implies $p_{\theta^*}(z|x) = p(z|x)$?
- The answer is no. This issue is called by *identifiability issue of generative models*.
- **Conditional generative models** can guarantee the identifiable representation learning.

Formal Definition of Identifiable Representations

- Comon (1994) derived the identifiability of *independent component analysis* (ICA) assuming $X = AZ$.
- When Z consists of independent components,

$$X = \hat{A}\hat{Z} \text{ implies } \hat{Z} = (PD)Z$$

for some permutation matrix P and diagonal matrix D under some mild conditions. That is, ICA is **identifiable up to permutation/re-scaling of latent factors**.

- Note that permutation/re-scaling do not change dependency structures between latent components.

Identifiability Issue of Nonlinear Methods

- The nonlinear ICA assumes $X = f_0(Z)$ where f_0 is an invertible nonlinear function, which is a common form in nonlinear generative models.
- Representations from nonlinear methods are not identifiable (Locatello et al., 2019). There is an invertible function g s.t.

$$X = (f_0 \circ g^{-1})g(Z) \text{ and } \mathbb{P}_Z = \mathbb{P}_{g(Z)}$$

$$\text{while } g(Z) \neq (PD)Z$$

  for all P and D.
- That is, the nonlinear generative model can not distinguish two different representations, $g(Z)$ and Z.

Conditional Generative Models for Identifiable Representation Learning

- Hyvarinen and Morioka (2016) introduced **conditional generative models** to derive identifiable nonlinear ICA.
- Observations (X) were time series, conditioning data (U) were time segments, and ICs (Z) followed exponential family distributions parameterized by time segments.
- Roughly speaking, when the true latent components are clustered well by conditioning data, we can use them as anchors to guarantee identifiability.

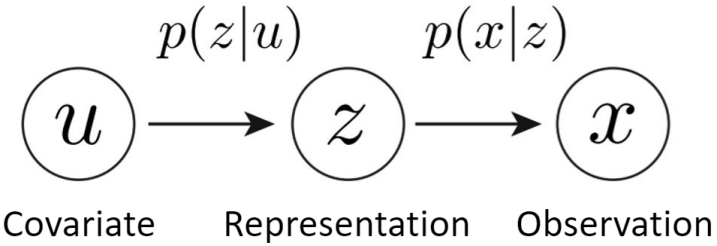
Identifiable Variational Autoencoders

- Khemakhem et al. (2020) extended these results to propose *identifiable variational autoencoders* (iVAEs).
- To explain the background of VAEs, VAEs model $p_\theta(x, z) := p_f(x|z)p(z)$ where f is decoder and $p(z)$ is known prior distribution.
- Since $\log p_\theta(x)$ is intractable, VAEs maximize the evidence lower bound (ELBO),

$$\begin{aligned} &\log p_\theta(x) - \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x)||p_\theta(z|x)) \\ &= \int_{\mathcal{Z}} \log p_\theta(x|z) q_\phi(z|x) dz - \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x)||p(z)), \end{aligned}$$

where q_ϕ is encoder called *posterior*.

Identifiable Variational Autoencoders



- The iVAEs introduce covariates into prior and posterior distributions. They assume the following data generation structure:

$$\begin{cases} Z|U \sim p_{T_0,\lambda_0}(z|u) \\ X = f_0(Z) + \epsilon \end{cases}$$

where p_{T_0,λ_0} is an exponential family distribution with sufficient statistics $T_0(u)$ and natural parameters $\lambda_0(u)$, and ϵ is an observation noise.

- The iVAEs are identifiable, i.e.,

$$p_{\theta_1}(x|u) = p_{\theta_2}(x|u) \text{ implies } T_1(f_1^{-1}(x)) = (PD)\,T_2(f_2^{-1}(x))$$

for some P and D where $\theta = (f, T, \lambda)$.

Objective Function of iVAEs

- The conditional generative model $p_\theta(x, z|u) := p_f(x|z)p_{T,\lambda}(z|u)$ is optimized by maximizing conditional ELBO,

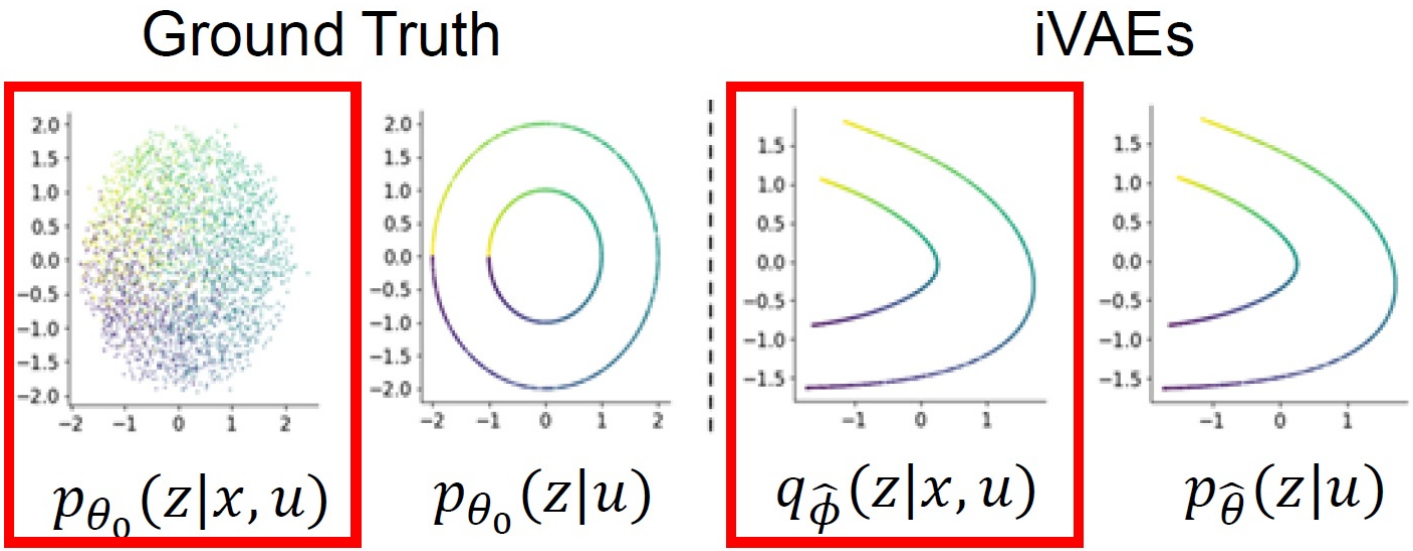
$$\begin{aligned}
&\log p_\theta(x|u) - \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x,u)||p_\theta(z|x,u)) \\
&= \int_{\mathcal{Z}} \log p_f(x|z) q_\phi(z|x,u) dz - \mathcal{D}_{\mathrm{KL}}(q_\phi(z|x,u)||p_{T,\lambda}(z|u)).
\end{aligned}$$

  Here, the modification is on introducing conditioning data U on $p_{T,\lambda}$ and q_ϕ.

Covariate-informed Identifiable Variational Autoencoders

- I derived that iVAEs could have bad local solutions and proposed a new identifiable representation learning method to overcome this limitation.
  - **Kim, Y.-G.**, Liu, Y., and Wei, X. "Covariate-informed Representation Learning to Prevent Posterior Collapse of iVAE." *AISTATS 2023*.

Motivation



- The figure shows example simulation data and results from iVAEs.
- The $p_{\theta_0}(z|x,u)$ has more variability than $p_{\theta_0}(z|u)$.
- However, $q_{\hat{\phi}}(z|x,u)$ reduces to $p_{\hat{\theta}}(z|u)$.
- Consequently, the inference failed.

Posterior Collapse: Bad Local Solutions of iVAEs

Definition 1

(Kim et al., 2023) The $q_\phi(z|x,u)$ is called *collapsed* if $q_\phi(z|x,u) = p_\theta(z|u)$.

- We formulate this issue and coin it by *posterior collapse issue in iVAEs*.
- Under the posterior collapse, the dimension-reduction removes information of observation (X).
- This notion is generally applicable to conditional generative models.

Posterior Collapse: Bad Local Solutions of iVAEs

Theorem 3

(Kim et al., 2023) Let $X|Z \sim N(f(Z), \gamma I)$. Under some conditions, $q_\phi(z|x,u)$ of iVAEs is collapsed for sufficiently large γ.

- We derive that this issue occurs in iVAEs in common cases.
- We show that the main reason is the KL term in ELBO, $\int (\log p_\theta(x|z)) q_\phi(z|x,u) dz - \mathcal{D}_{\mathsf{KL}}(q_\phi(z|x,u)||p_\theta(z|u))$.
- The KL term enforces to match $q_\phi(z|x,u)$ and $p_\theta(z|u)$.

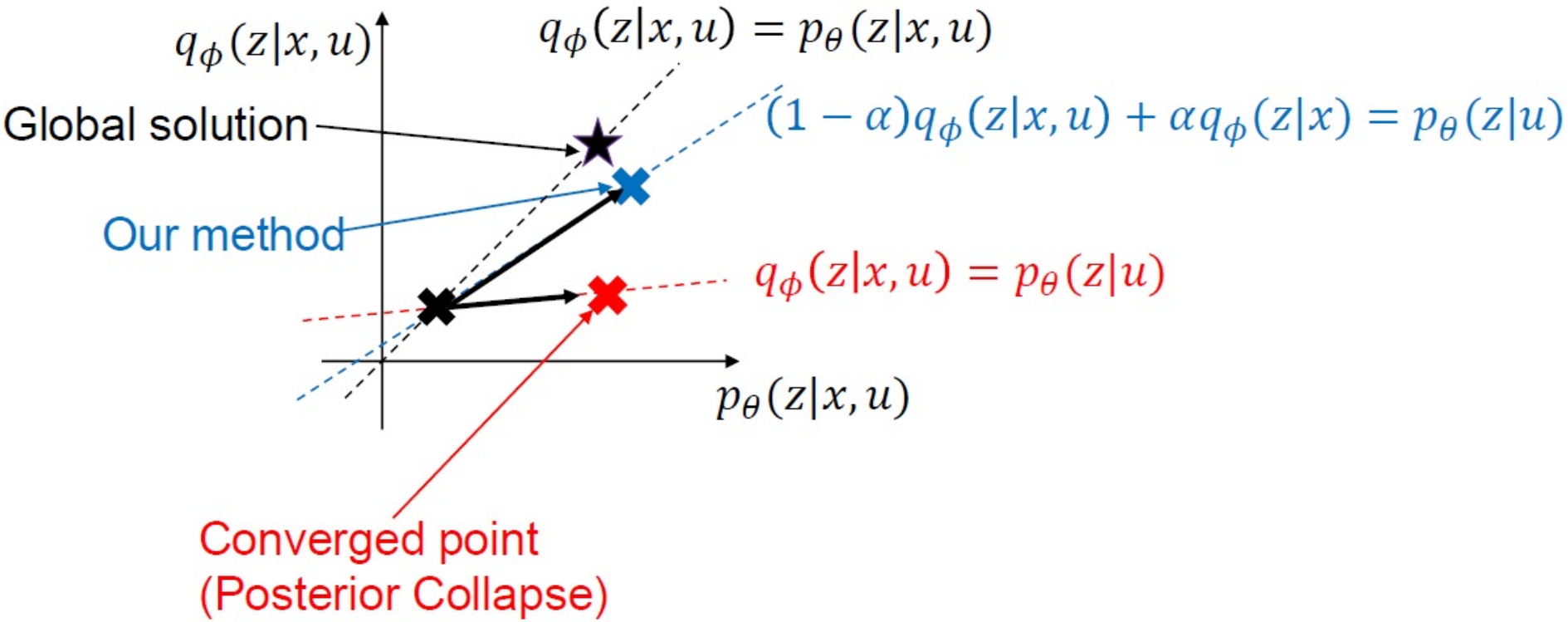
Proposed Method: Covariate-informed iVAEs

Theorem 4

(Kim et al., 2023) If $p_\theta(z|u) \neq p_\theta(z)$, then $q_\phi(z|x) \neq p_\theta(z|u)$.

- We propose a new algorithm, *Covariate-informed iVAEs*.
- Our method introduces $q_\phi(z|x)$ in addition to $q_\phi(z|x, u)$. The $q_\phi(z|x)$ can alleviate the posterior collapse issue.
- When X has information independent of U, $q_\phi(z|x)$ does not reduce to $p_\theta(z|u)$.

Proposed Method: Covariate-informed iVAEs



Proposed Method: Covariate-informed iVAEs

- We propose *covariate-informed posterior distributions*.
- We define $\tilde{q}_\phi(z|x,u;\alpha) := (1-\alpha)q_\phi(z|x,u) + \alpha q_\phi(z|x)$ and use the optimal mixture $\tilde{q}_\phi(z|x,u;\alpha^*(x,u))$ where

$$\alpha^*(x,u) := \underset{\alpha\in(0,1)}{\text{argmax}} \int (\log p_\theta(x|z))\tilde{q}_\phi(z|x,u;\alpha)dz - \mathcal{D}_{\text{KL}}(\tilde{q}_\phi(z|x,u;\alpha)||p_\theta(z|u)).$$

- That is, the $\tilde{q}_\phi(z|x,u;\alpha^*(x,u))$ provides the tightest lower bound of the $\log p_\theta(x|u)$ among all the mixtures.

Proposed Method: Covariate-informed iVAEs

Theorem 5

(Kim et al., 2023) When $p_\theta(z|u) \neq p_\theta(z)$,

$$\alpha^*(x,u)q_\phi(z|x) + (1-\alpha^*(x,u))q_\phi(z|x,u) = p_\theta(z|u)$$

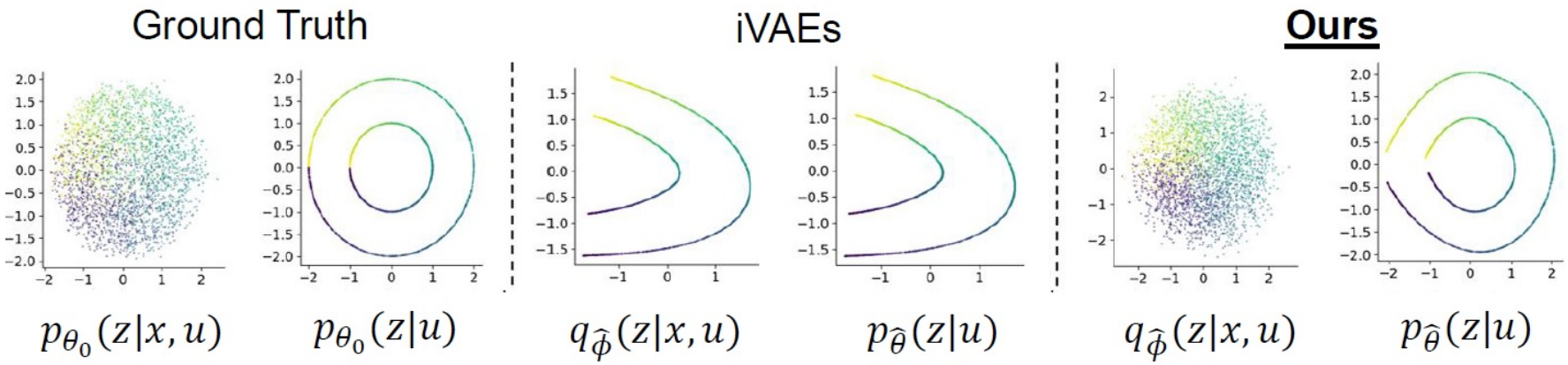
implies $q_\phi(z|x,u) \neq p_\theta(z|u)$.

- We derive that the posterior collapse does not occur under some restrictions on $q_\phi(z|x,u)$.[1]
- Our objective function can be expressed as

$$\begin{aligned}
&\log p_\theta(x|u) - \mathcal{D}_{\mathrm{KL}}(\tilde{q}_\phi(z|x,u;\alpha^*(x,u))||p_\theta(z|x,u)) \\
&= \int (\log p_\theta(x|z))\tilde{q}_\phi(z|x,u;\alpha^*(x,u))dz \\
&\quad - \mathcal{D}_{\mathrm{KL}}(\tilde{q}_\phi(z|x,u;\alpha^*(x,u)), p_\theta(z|u)).
\end{aligned}$$

[1] $q_\phi(z|x,u) \propto q_\phi(z|x)p_\theta(z|u)/p_\theta(z)$ where $p_\theta(z) := \int p_\theta(z|u)p(u)du$.

Simulation Study

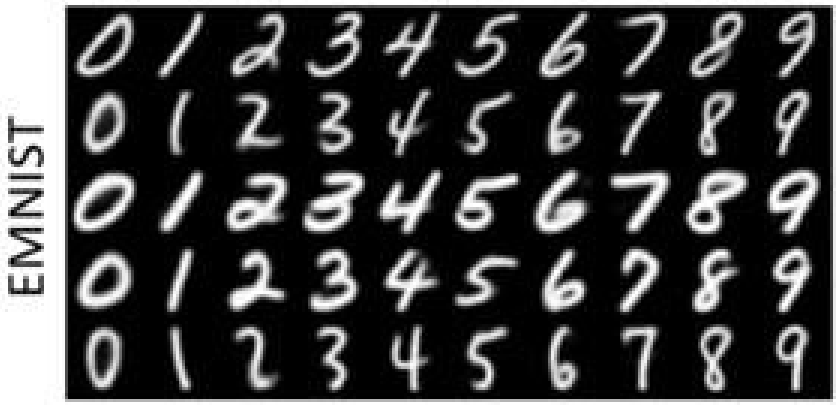


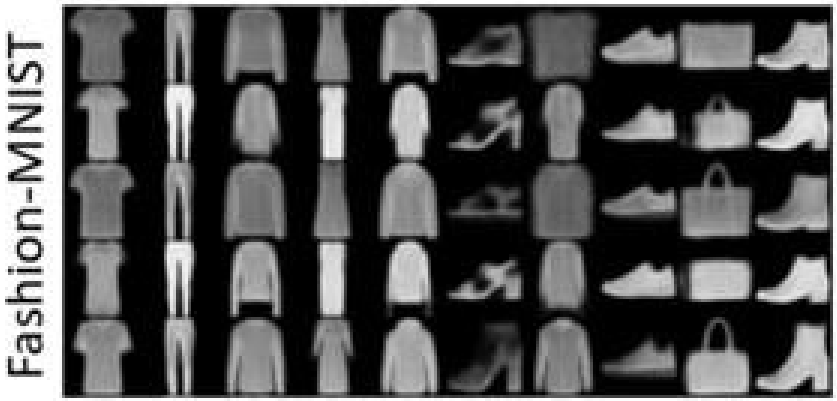
Method	Coefficient of Determination[1] (↑)	Log-likelihood (↑)
iVAEs	.2835 (.0119)	-114.9876 (1.2964)
Covariate-informed iVAE (**ours**)	**.9156** (.0007)	**-102.9267** (0.3198)

[1] Average r^2 between elements in true and estimated representations.

Real Data Analysis: EMNIST and Fashion-MNIST

- We analyzed two benchmark datasets:
  1. **EMNIST**: Handwritten digit images with labels from 0 to 9.
  2. **Fashion-MNIST**: Fashion-item images with labels of ten fashion item classes.
- Representations are expected to be clustered by class labels.
- We use images as X and labels as U.





Real Data Analysis: EMNIST and Fashion-MNIST

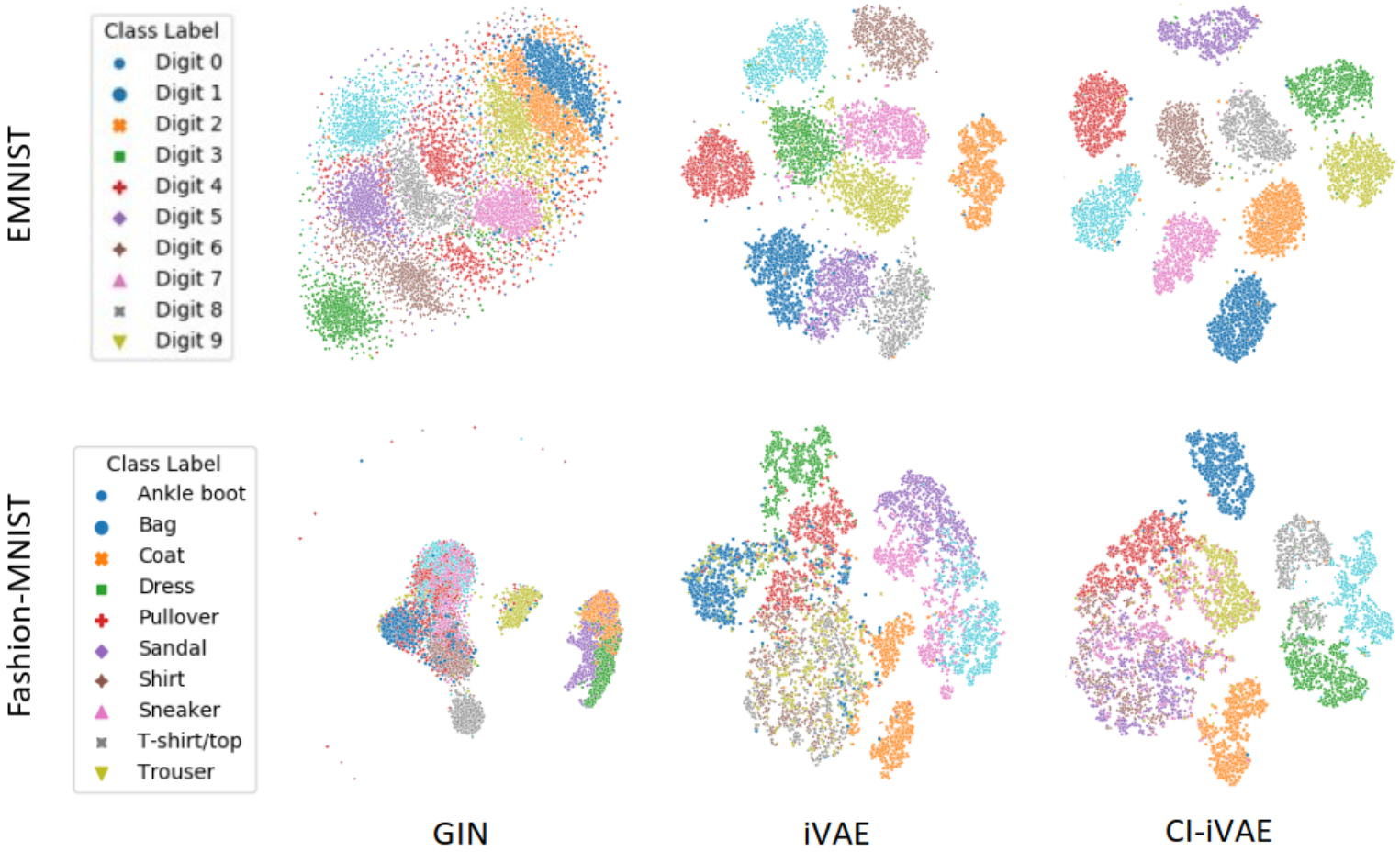


Figure: Visualization of the t-SNE embeddings of representations, $q_\phi(z|x)$, from various methods on EMNIST and Fashion-MNIST datasets.

Real Data Analysis: EMNIST and Fashion-MNIST

Table: Means of evaluation scores with standard errors.

Method	SSW/SST (↓)		5-NN Accuracy (↑)	
	EMNIST	Fashion-MNIST	EMNIST	Fashion-MNIST
GIN	.6130 (.0075)	.8503 (.0026)	.9510 (.0017)	.8340 (.0013)
iVAE	.5486 (.0037)	.6157 (.0046)	.9864 (.0004)	.8086 (.0013)
iVAE with KL-annealing	.6416 (.0031)	.5951 (.0031)	.9885 (.0002)	.8311 (.0009)
iVAE with aggressive posterior	.5838 (.0033)	.6230 (.0032)	.9927 (.0001)	.8424 (.0011)
IDVAE	.5884 (.0071)	.5857 (.0057)	**.9931** (.0001)	.8337 (.0010)
CI-iVAE (ours)	**.4117** (.0032)	**.4926** (.0024)	**.9931** (.0011)	**.8518** (.0006)

Conclusion

- We reviewed recent topics in conditional generative models including (i) conditional statistical distances and (ii) identifiable representation learning with conditioning data.
- With conditional Wasserstein generator, we can train conditional generative models by minimizing the expected Wasserstein distance between conditional distributions.
- With covariate-informed iVAEs, we can get better representations by preventing posterior collapse cases.

References I

Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Kim, Y.-g., Lee, K., and Paik, M. C. (2022). Conditional wasserstein generator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kim, Y.-g., Liu, Y., and Wei, X.-X. (2023). Covariate-informed representation learning to prevent posterior collapse of ivae. In *International Conference on Artificial Intelligence and Statistics*, pages 2641–2660. PMLR.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.