Disentangling and Learning Robust Representations with Natural Clustering

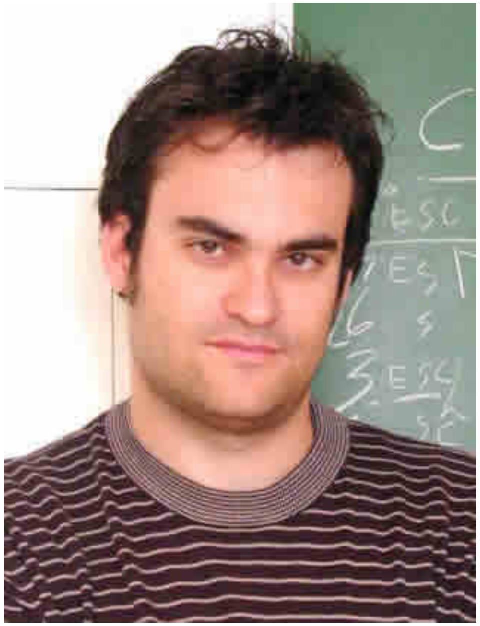
ICMLA 2019

Javier Antorán, Antonio Miguel

About Us

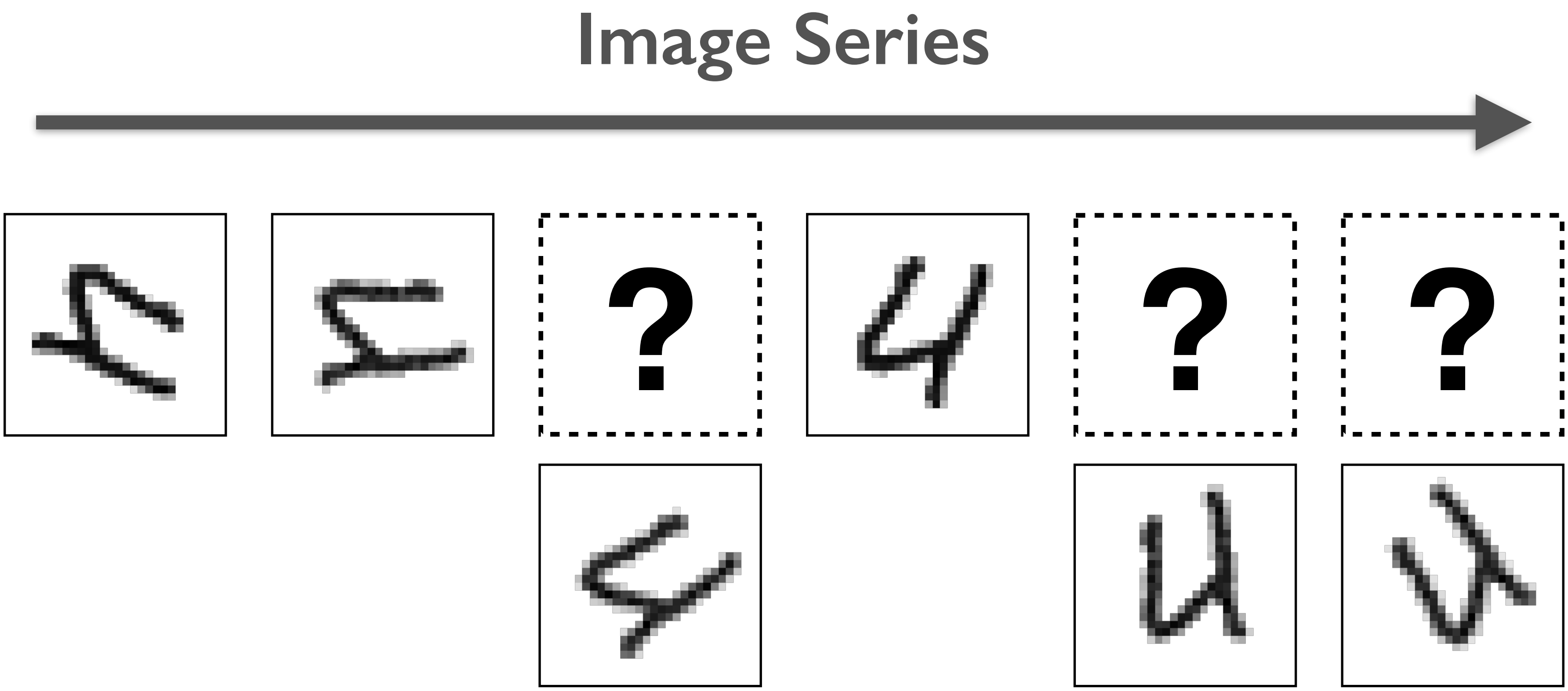
- **Javier Antorán** ja666@cam.ac.uk
- **Antonio Miguel** amiguel@unizar.es

(Now at University of Cambridge)



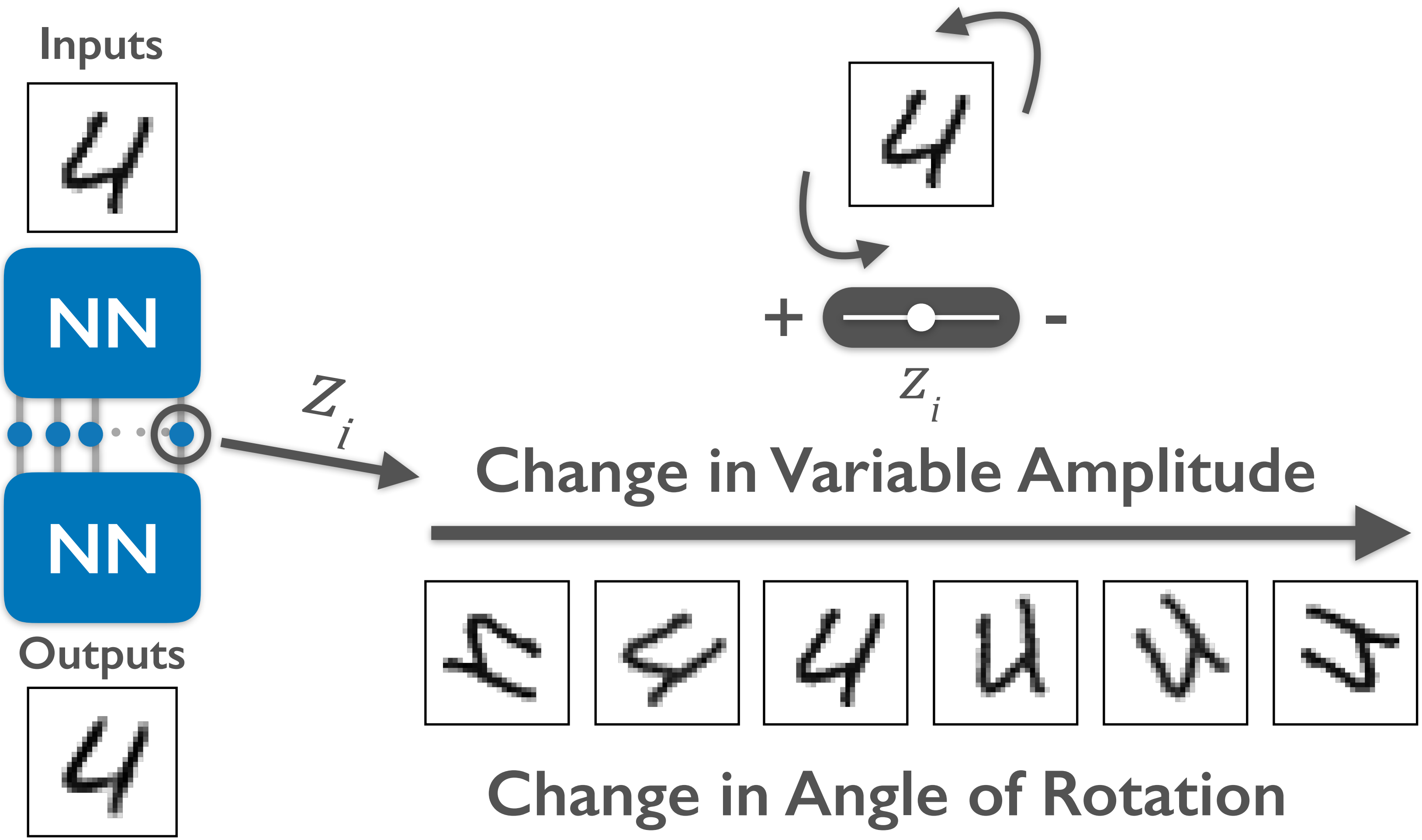
VIVOLAB

Motivation: Disentangling High Level Concepts



- Can learn images individually or single digit and concept of rotation

Desired Behaviour

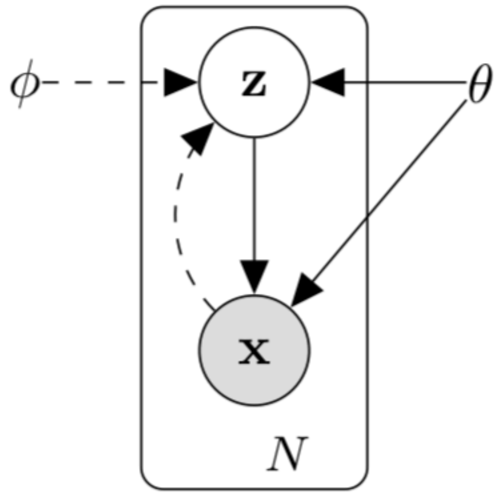


More Motivation: Applications

- High level analysis of complex data:
  - Single cell RNA sequencing
  - Pharmaceutical drug molecules
- High level editing of complex data:
  - Image / Audio manipulation
- Feature extraction for interpretable decision making

Variational Autoencoders

(Inference model params)



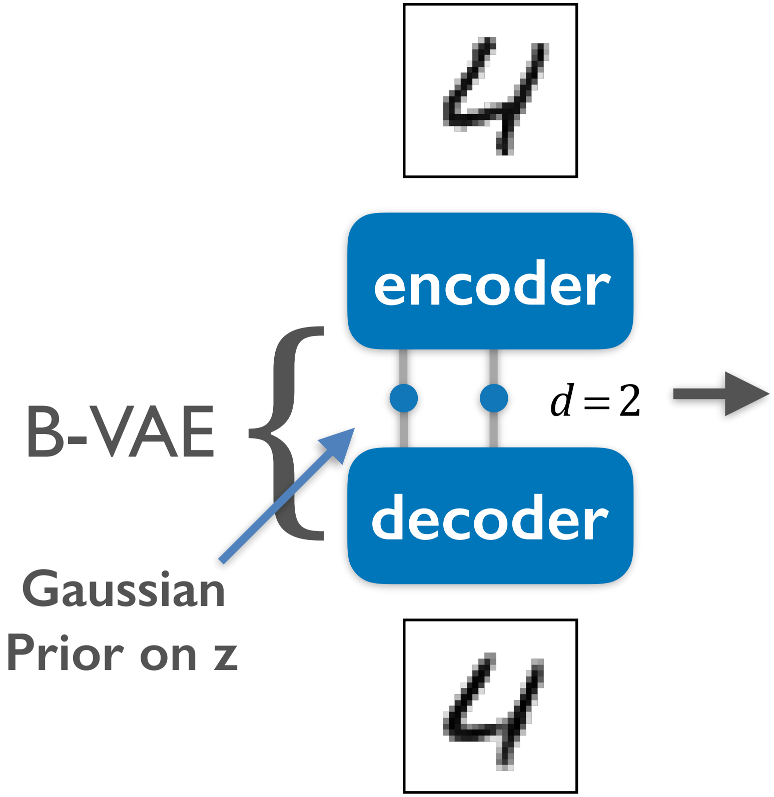
(Generative model params)

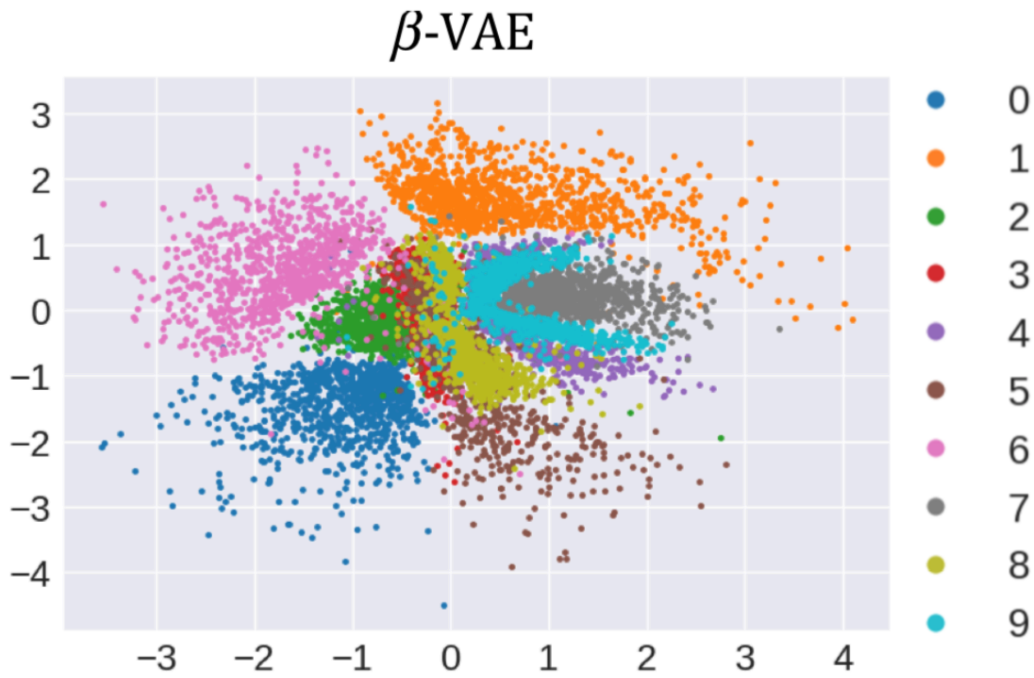
Inference

Generation

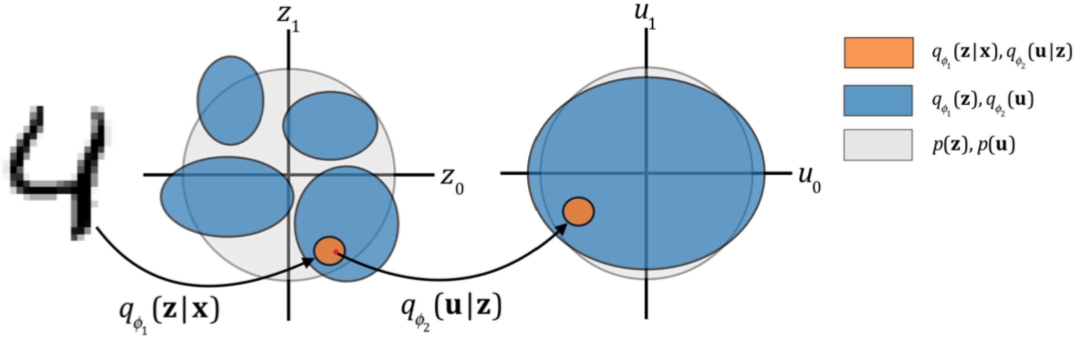
$$\mathcal{L}(\mathbf{x}) = -\underbrace{D_{KL}(q(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))}_{\text{Regularization cost}} + \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruction cost}} \qquad (1)$$

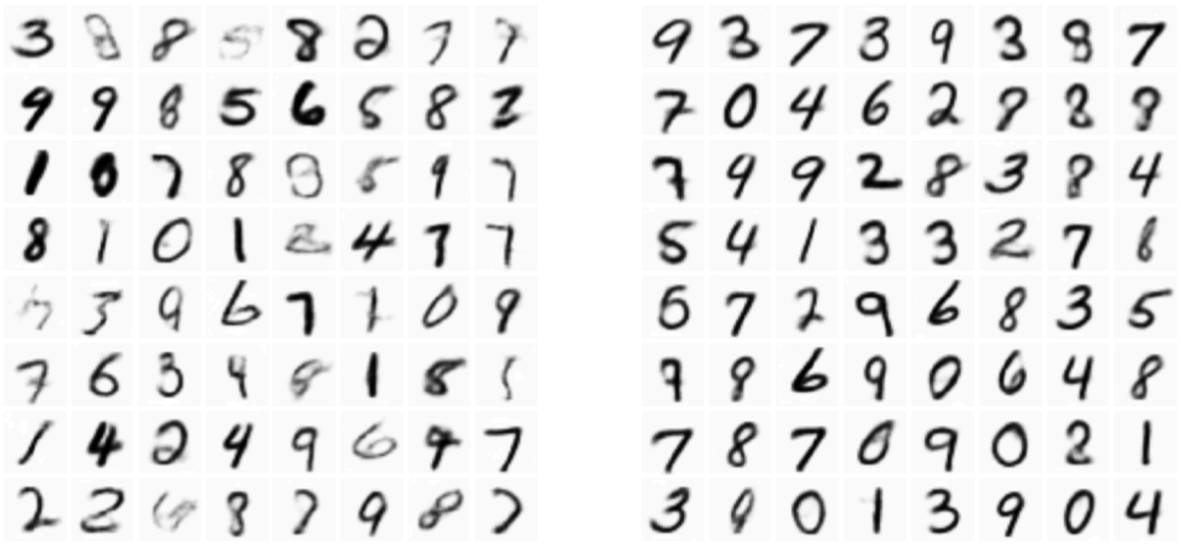
Multimodality in VAE Latent Space





More VAE Underfitting: Ancestral Sampling

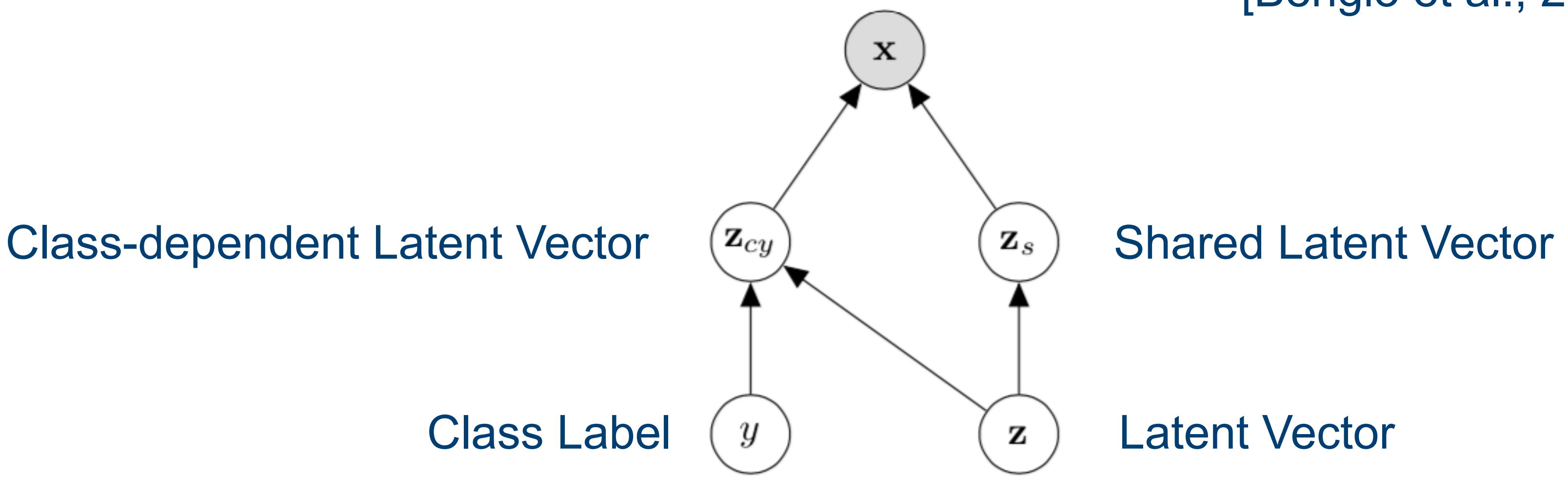




Natural Clustering as an Inductive Bias

- **Natural clustering:** “different values of categorical variables such as object classes are associated with separate manifolds.”
- “(…) the local variations on the manifold tend to preserve the value of a category, and a linear interpolation between examples of different classes in general involves going through a low density region.”

[Bengio et al., 2012]



A Lower Bound on the Joint Likelihood

$$\log p(\mathbf{x}, y) \geq \mathcal{L}(\mathbf{x}, y) = \mathbb{E}_{q(\mathbf{z},\boldsymbol{\pi}|\mathbf{x},y)}[-\log q(\mathbf{z}, \boldsymbol{\pi}|\mathbf{x}, y) + \log p(\mathbf{x}, y, \mathbf{z}, \boldsymbol{\pi})] \quad (2)$$

Where $\boldsymbol{\pi}$ is a probability distribution over categorical outcomes

Rearrange + Lower Bound

$$\mathcal{L}_{obj}(\mathbf{x}, y) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|y, \mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z})) + \log(q_\phi(y|\mathbf{x})) \quad (6)$$

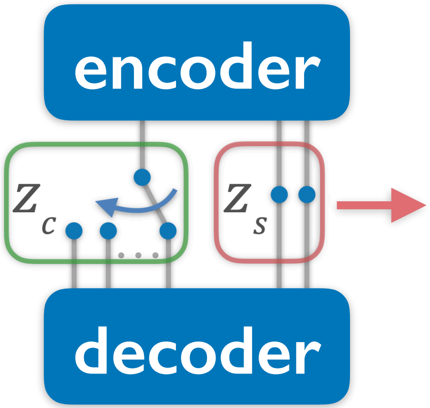
$$\mathbf{z} = [\mathbf{z}_c^\intercal, \mathbf{z}_s^\intercal]^\intercal$$

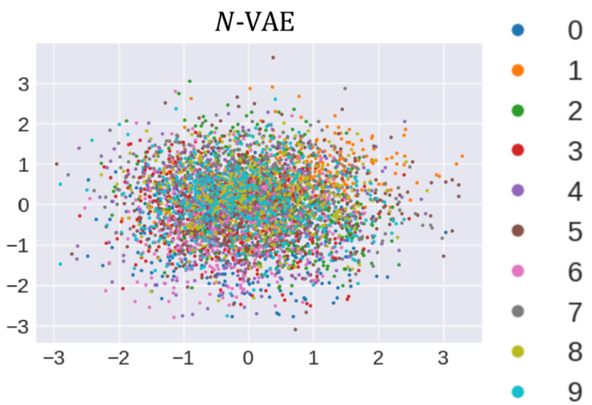
$$z_{cy} = vec(\mathbf{c}_y \odot [\mathbf{1}_L, [\mathbf{z}_{c1}, \mathbf{z}_{c2}, \ldots, \mathbf{z}_{cL}]^\intercal])$$

\+ Reweigh

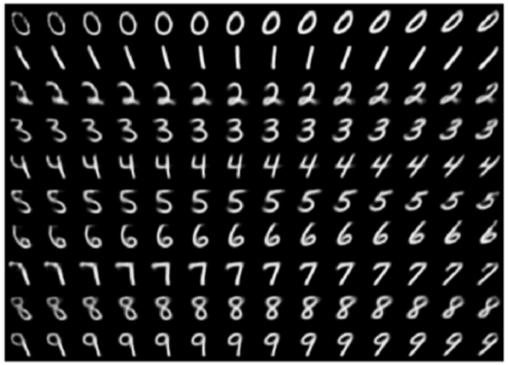
$$\mathcal{L}_{\beta_c} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|y, \mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}_s|\mathbf{x}) \,\|\, p(\mathbf{z})) - \beta_c D_{KL}(q_\phi(\mathbf{z}_c|\mathbf{x}) \,\|\, p(\mathbf{z})) + \log(q_\phi(y|\mathbf{x})) \quad (7)$$

Shared Latent Space: MNIST





Azimuth

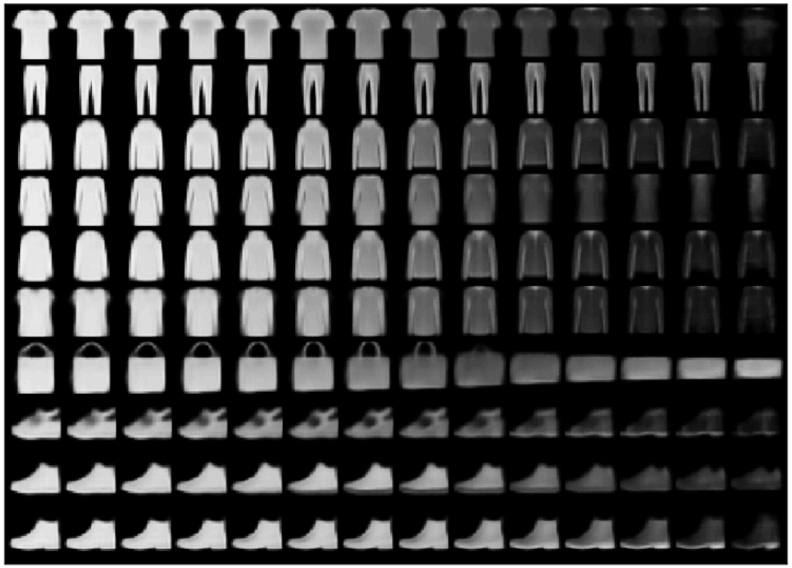


Stroke thickness

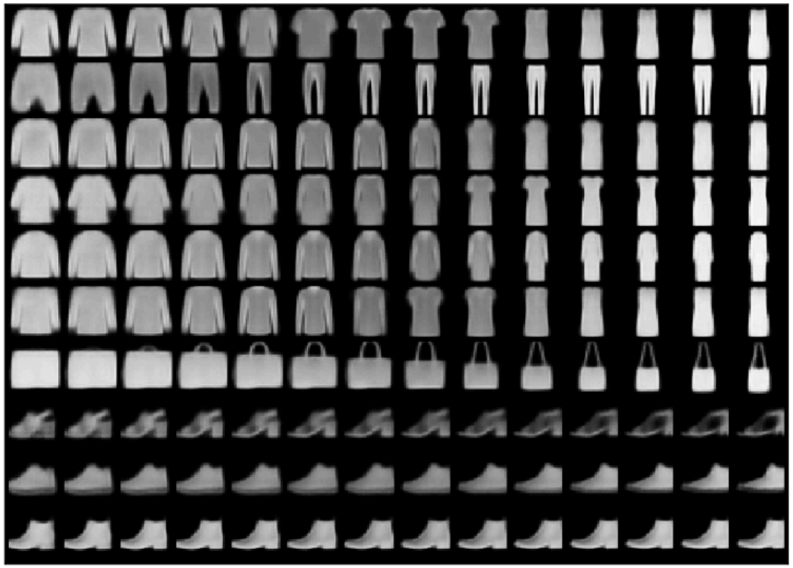
Width

Shared Latent Space: FMNIST

Color Intensity



Width

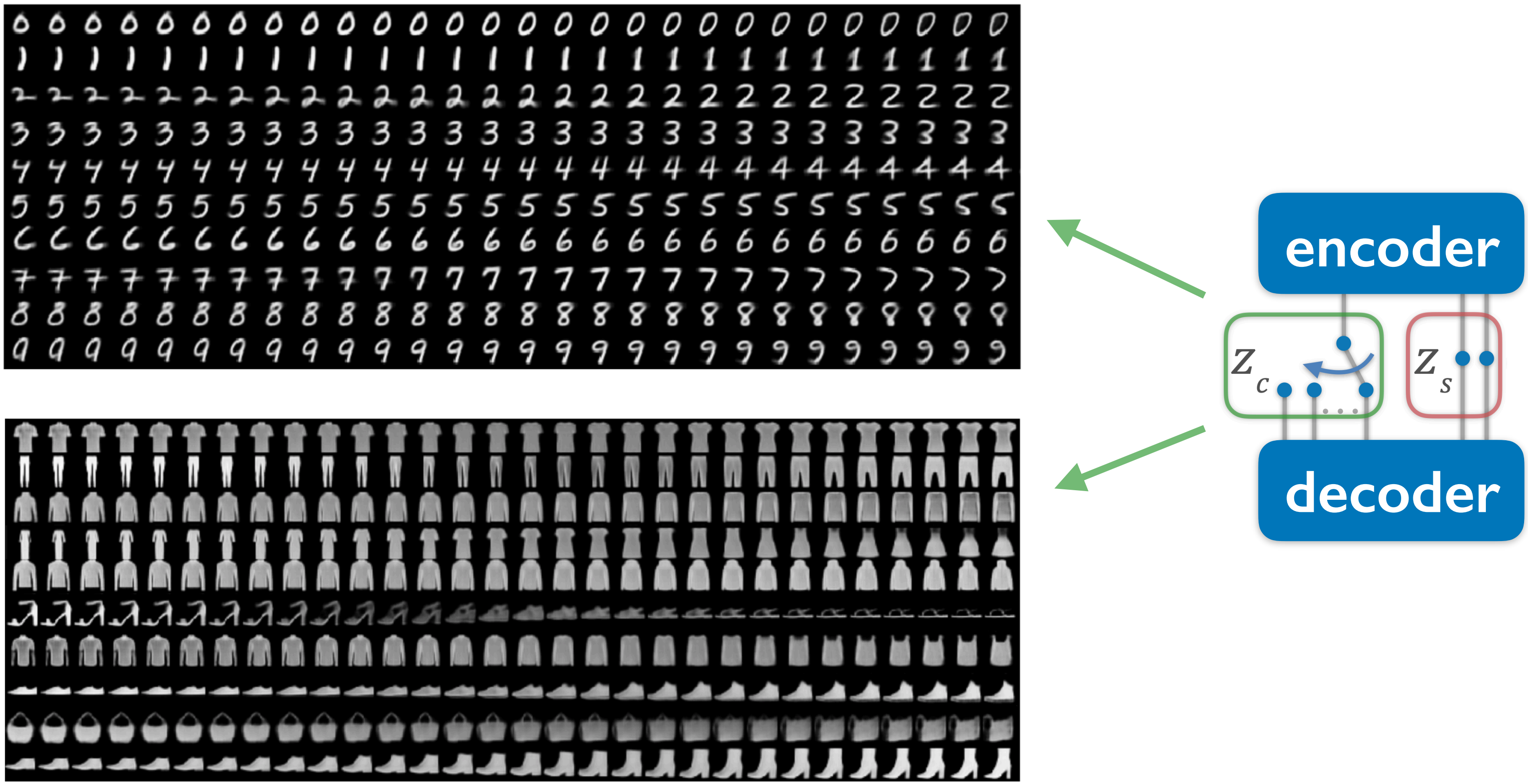


Shared Latent Space: Yale Ext B

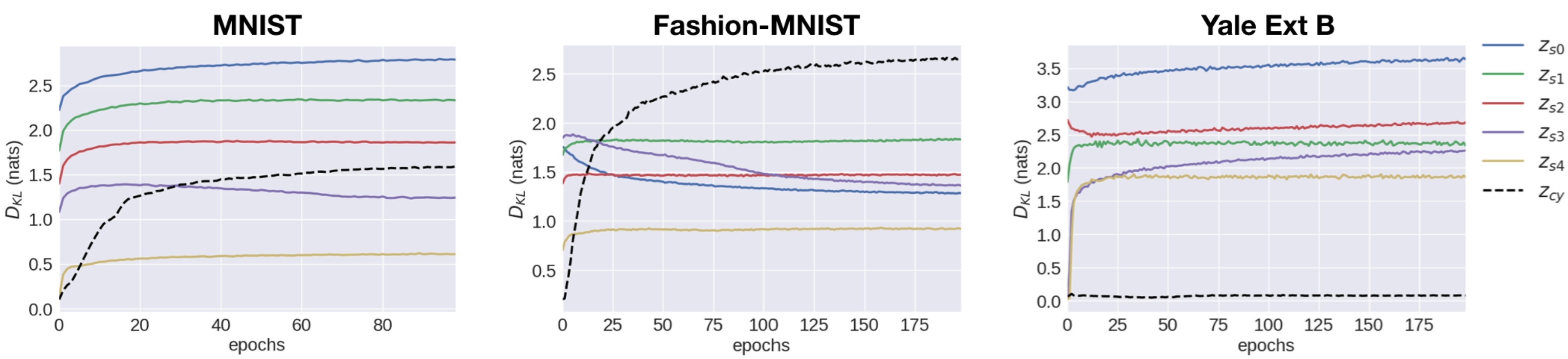
Illumination azimuth

Illumination elevation

Class-dependent Factors of Variability



Detecting Class-dependent Factors



- **KL term acts as a feature detector**

$$\mathcal{L}_{\beta_c} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|y,\mathbf{z})] \boxed{- D_{KL}(q_\phi(\mathbf{z}_s|\mathbf{x}) \parallel p(\mathbf{z}))}$$
$$\boxed{- \beta_c D_{KL}(q_\phi(\mathbf{z}_c|\mathbf{x}) \parallel p(\mathbf{z}))} + \log(q_\phi(y|\mathbf{x})) \qquad (7)$$

Ancestral Sampling from N-VAE

N-VAE samples with $\sigma = 1$

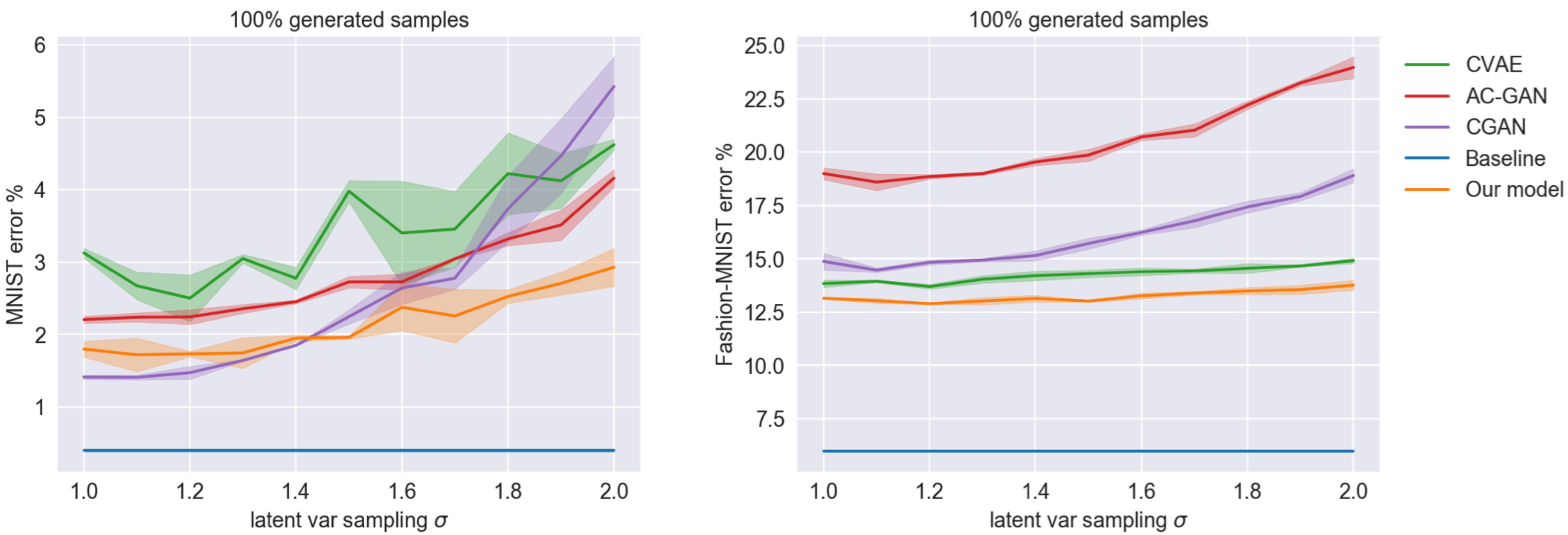
C: N-VAE SAMPLES WITH $\sigma = 1.4$

Training Discriminative Models with Artificial Data!

- **40% Artificial Data**

Training Discriminative Models with Artificial Data!

- **100% Artificial Data**



Summary

- The Natural Clustering inductive bias allows us to explain data better.
- N-VAE successfully disentangles latent factors in scenarios with class-related multimodality.
- N-VAE can be used for detecting and disentangling class-dependent factors of variability which are usually ignored by generative models.
- N-VAE’s aggregate posterior over latent variables better matches the prior, recovering the VAE’s ancestral sampling capabilities.
- The previous two characteristics result in a more expressive generative model.