Deep Generative Models I

Variational Autoencoders – Part II

Machine Perception

Otmar Hilliges

23 April 2020

Advanced Interactive Technologies

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Last week(s)

Intro to generative modelling

VAE

Derivation of the ELBO

This week

More generative models

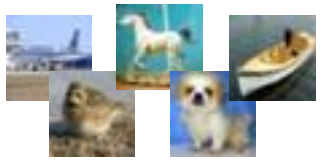
Step I: Variatonal autoencoders:

- Training
- Representation learning & disentanglement
- Applications

Step II: Generative adverserial networks

Generative Modelling

Given training data, generate new samples, drawn from “same” distribution

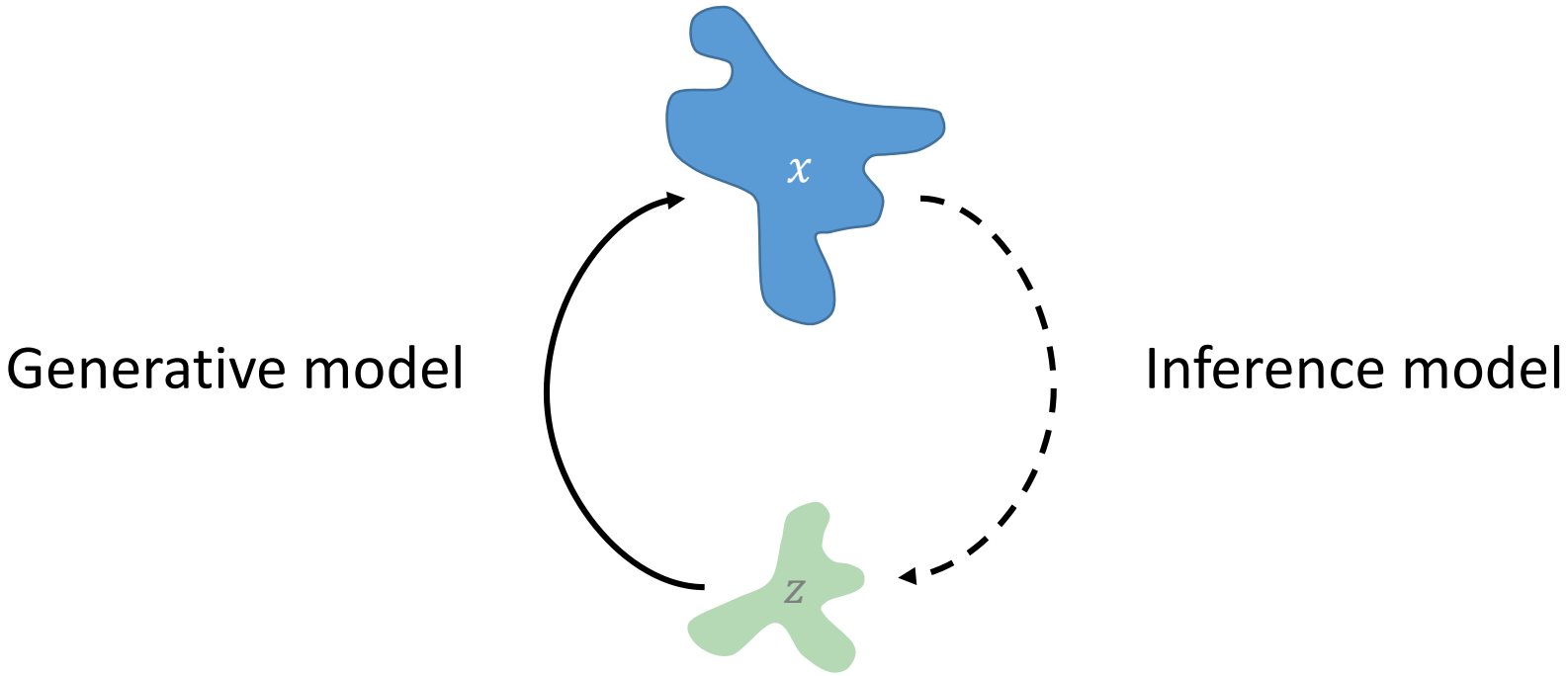


Training samples $\sim p_{data}(x)$

Generated samples $\sim p_{model}(x)$

We want $p_{model}(x)$ to be similar to $p_{data}(x)$

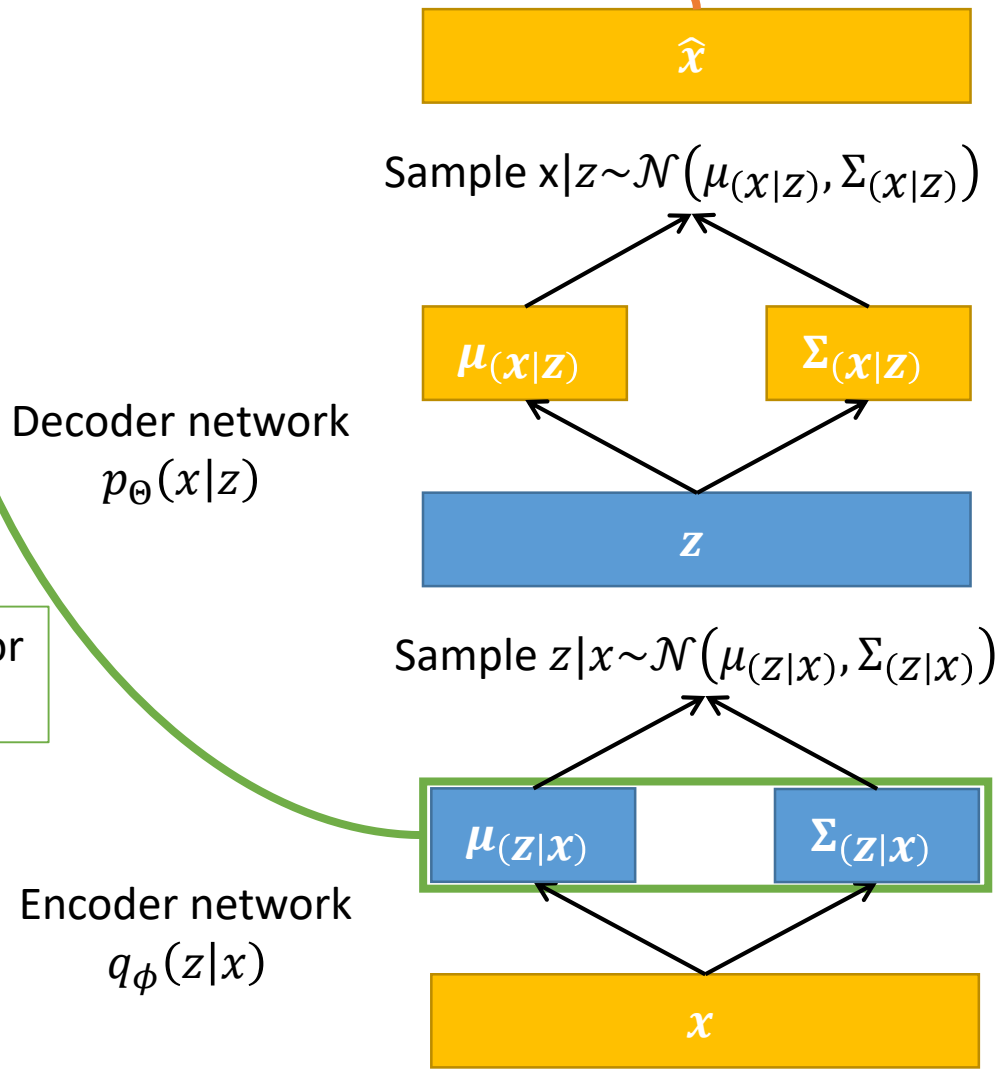
Variational autoencoder



Recap: ELBO

$$\underbrace{\mathrm{E}_z\left[\log p_\Theta\left(x^{(i)}\middle|\, z\right)\right] - D_{KL}\left(q_\phi(z|x^{(i)}) \parallel p_\Theta\left(z^{(i)}\right)\right)}_{\mathcal{L}(x^{(i)}, \Theta, \phi)}$$

Maximize reconstruction likelihood



Decoder network
$p_\Theta(x|z)$

Make approximate posterior
Similar to prior

Encoder network
$q_\phi(z|x)$

Training VAEs

Training of VAEs is just backprop.

But wait, how do gradients flow through z?

Or any other random operation (requiring sampling)?

(see blackboard)

re parametrization trick:

Given

$$z \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \text{value changes every time we sample}$$

Assume $\exists$ underlying random variable

$$\epsilon \sim \mathcal{N}(0, 1)$$

$$z = \mu + \sigma \epsilon \leftarrow \text{still random but does not depend on } \mu \text{ or } \sigma$$

$$z = f(x, \epsilon, \theta)$$

$\Rightarrow$ <u>can</u> take deriv wrt μ, σ.

tells us how a small change to μ, σ affects output if we keep ϵ fixed

Data log-likelihood

$$\log p_\theta(x^{(i)}) =$$

$$= \mathbb{E}_{z \sim q_\phi(z|x^{(i)})}\left[\log p_\theta(x^{(i)})\right] \quad (p_\theta(\cdot) \text{ doesn't depend on } z)$$

$$= \mathbb{E}_z\left[\log \frac{p_\theta(x^{(i)}|z)\, p_\theta(z)}{p_\theta(z|x^{(i)})}\right] \quad (\text{Bayes' rule})$$

$$= \mathbb{E}_z\left[\log \frac{p_\theta(x^{(i)}|z)\, p_\theta(z)}{p_\theta(z|x^{(i)})} \frac{q_\phi(z|x^{(i)})}{q_\phi(z|x^{(i)})}\right] \quad (\text{mult by constant})$$

$$= \mathbb{E}_z\left[\log p_\theta(x^{(i)}|z)\right] + \mathbb{E}_z\left[\log \frac{p_\theta(z)}{q_\phi(z|x^{(i)})}\right] + \mathbb{E}_z\left[\log \frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}\right] \quad (\text{product rule logarithms})$$

(continued on next slide)

$$= E_z[\log p_\theta(x^{(i)}|z)] - E_z\left[\log \frac{q_\phi(z|x^{(i)})}{p_\theta(z)}\right] + E_z\left[\log \frac{q_\phi(z|x^{(i)})}{p_\theta(z|x^{(i)})}\right] \quad \left(\log \frac{x}{y} = -\log \frac{y}{x}\right)$$

✓ (with sampling) ✓ closed-form if Gaussian ✓ ✗

$$= \underbrace{E_z[\log p_\theta(x^{(i)}|z)]}_{\text{reconstruction}} - D_{KL}\left(\underbrace{q_\phi(z|x^{(i)})}_{\text{tractable}} \| p_\theta(z)\right) + \underbrace{D_{KL}\left(q_\phi(z|x^{(i)}) \| p_\theta(z|x^{(i)})\right)}_{\geq 0 \text{ (see proof earlier)}}$$

NN tractable

"make approx. posterior as similar as possible to prior"

reconstruction

goes high if our sample is similar to true distribution

目的是: opt. only tractable term: q_ϕ, p_θ

不想要处理 $p(x)$, $p(z|x)$

loss function $\mathcal{L}(x^{(i)}; \theta, \phi)$

$$\log(p_\theta(x^{(i)})) \geq \mathcal{L}$$

evidence lower bound "ELBO"

⇒ data is at least as likely as $\mathcal{L}$

Traing procedure:

$$\theta^*, \phi^* = \text{argmax} \sum_1^N \mathcal{L}(x^{(i)}; \theta, \phi)$$

Gradient computation for backprop

$$\nabla_{\theta,\phi}\, \mathbb{E}_{z\sim q_\phi}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x^{(i)})}\right]$$

$$= \mathbb{E}_z\left[\log p_\theta(x^{(i)}|z)\right] + \mathbb{E}_z\left[\log \frac{p_\theta(z)}{q_\phi(z|x^{(i)})}\right]$$

$$= \nabla_{\theta,\phi}\, \mathbb{E}_{\epsilon\sim N(0,1)}\left[\log \frac{p_\theta(x, f(x,\epsilon,\theta))}{q_\phi(f(x,\epsilon,\phi)|x)}\right]$$

f is NN (deterministic)

$$= \mathbb{E}_\epsilon\left[\nabla_{\theta,\phi} \log \frac{p_\theta(x,f)}{q_\phi(f|x)}\right]$$

$= \omega$

deterministic NN

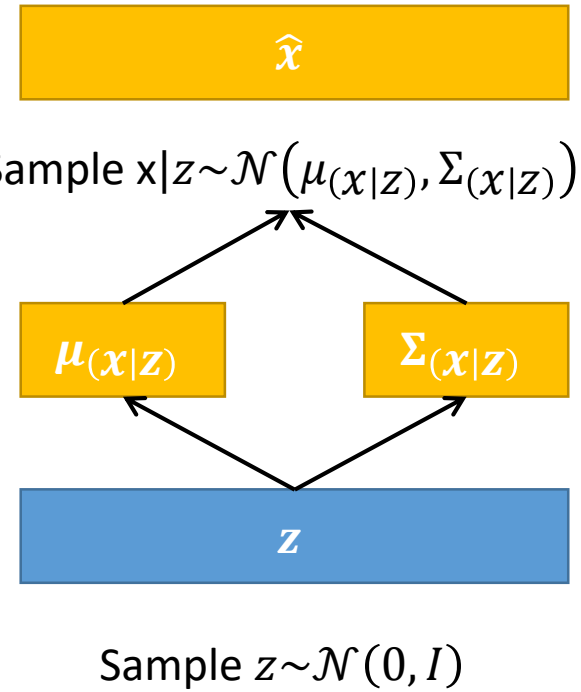
$$\approx \frac{1}{k}\sum_{i=1}^{k} \nabla_{\theta,\phi} \log \omega(x,f,\theta,\phi)$$

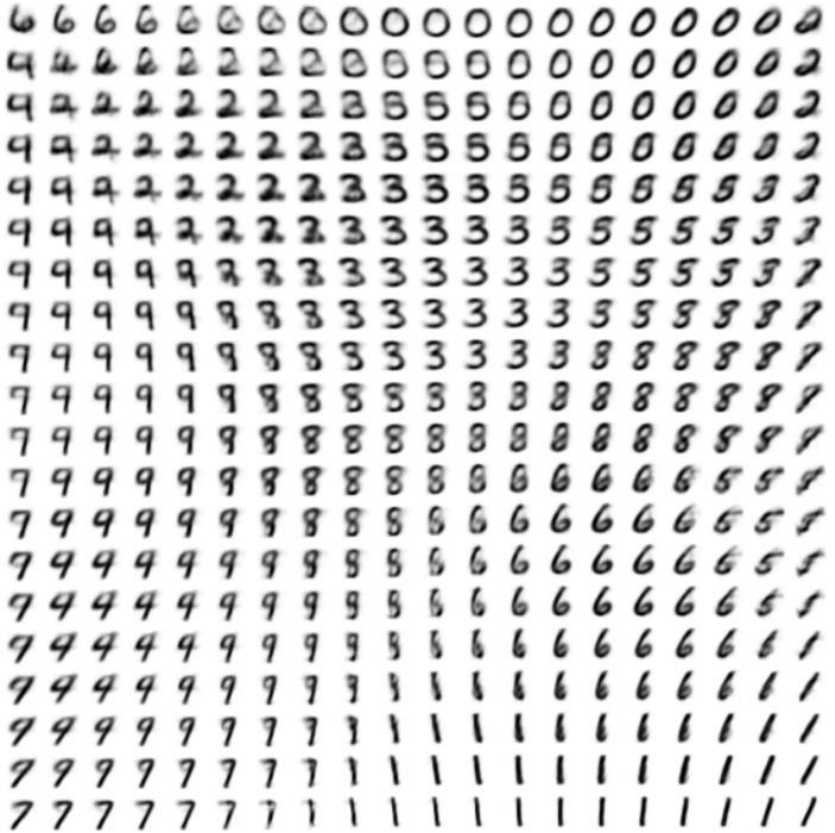
[Kingma and Welling, “Auto-Encoding Variational Bayes”, ICLR 2014]

Generating Data

At runtime only use decoder network. Sample z from prior.

Decoder network
$p_\Theta(x|z)$





Cross-modal Deep Variational Hand Pose Estimation

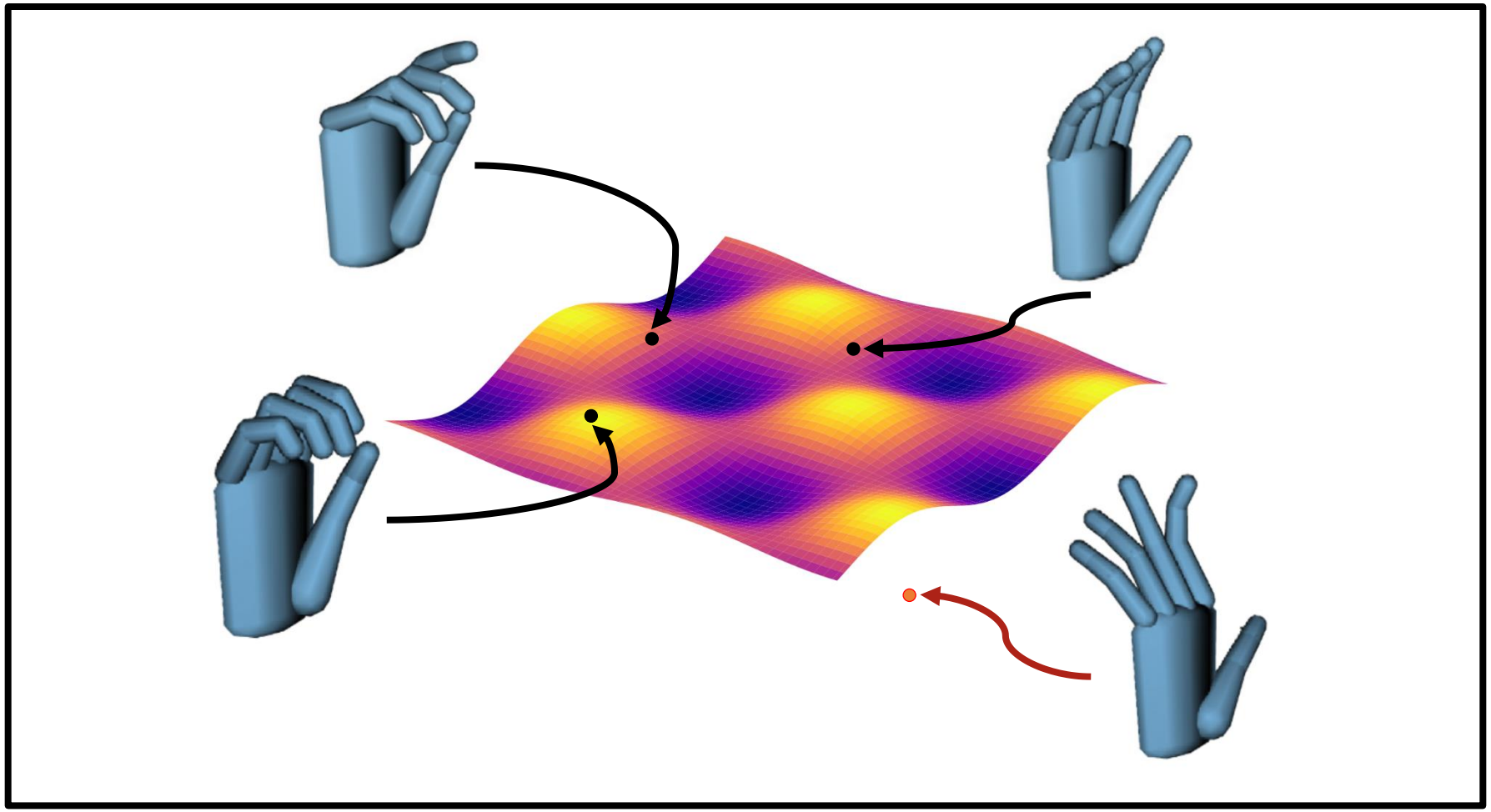
Adrian Spurr, Jie Song, Seonwook Park, Otmar Hilliges

CVPR 2018

RGB hand pose estimation

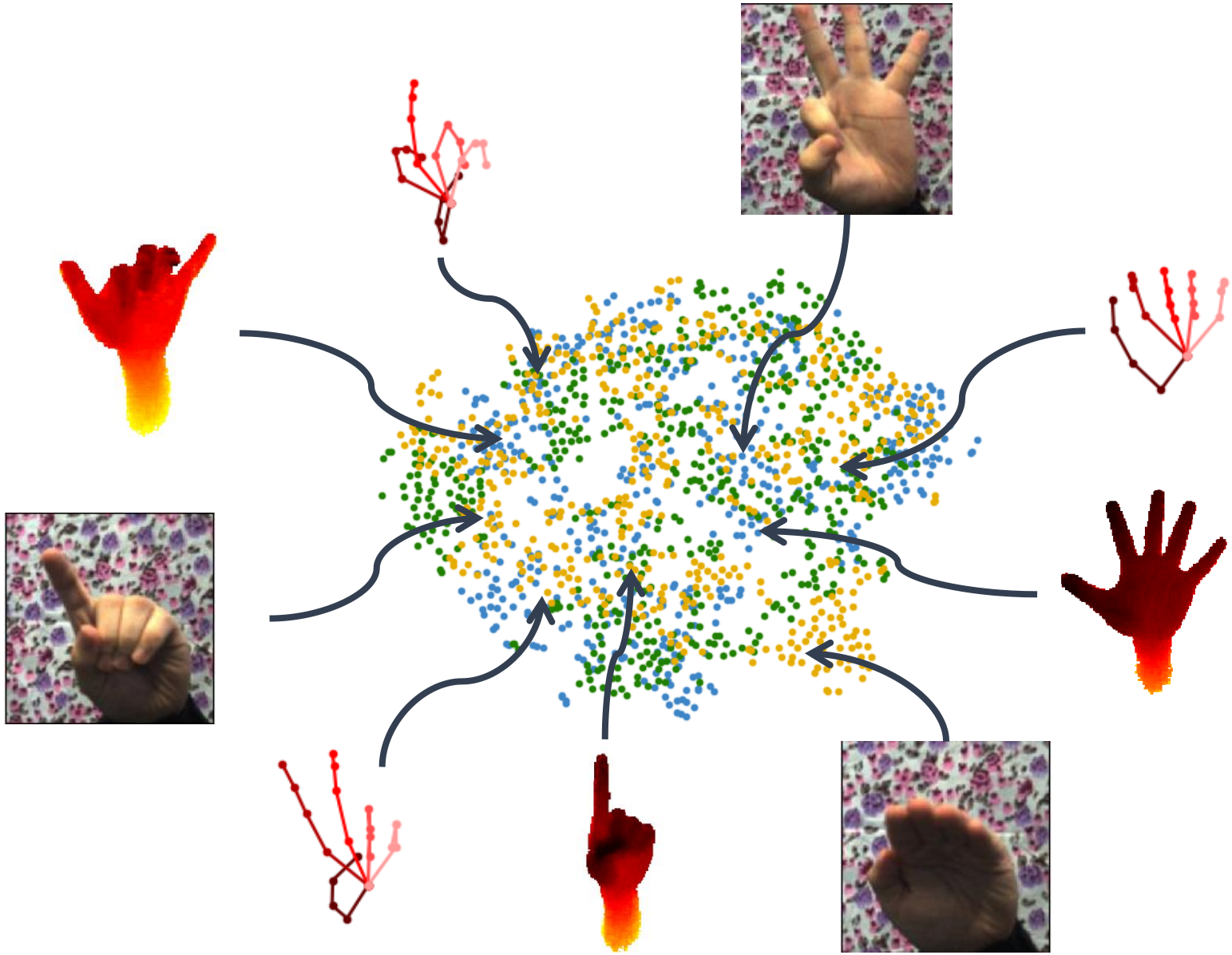
Assumption: Manifold of valid hand poses

exist a low dimensional space that contains all valid hand pose

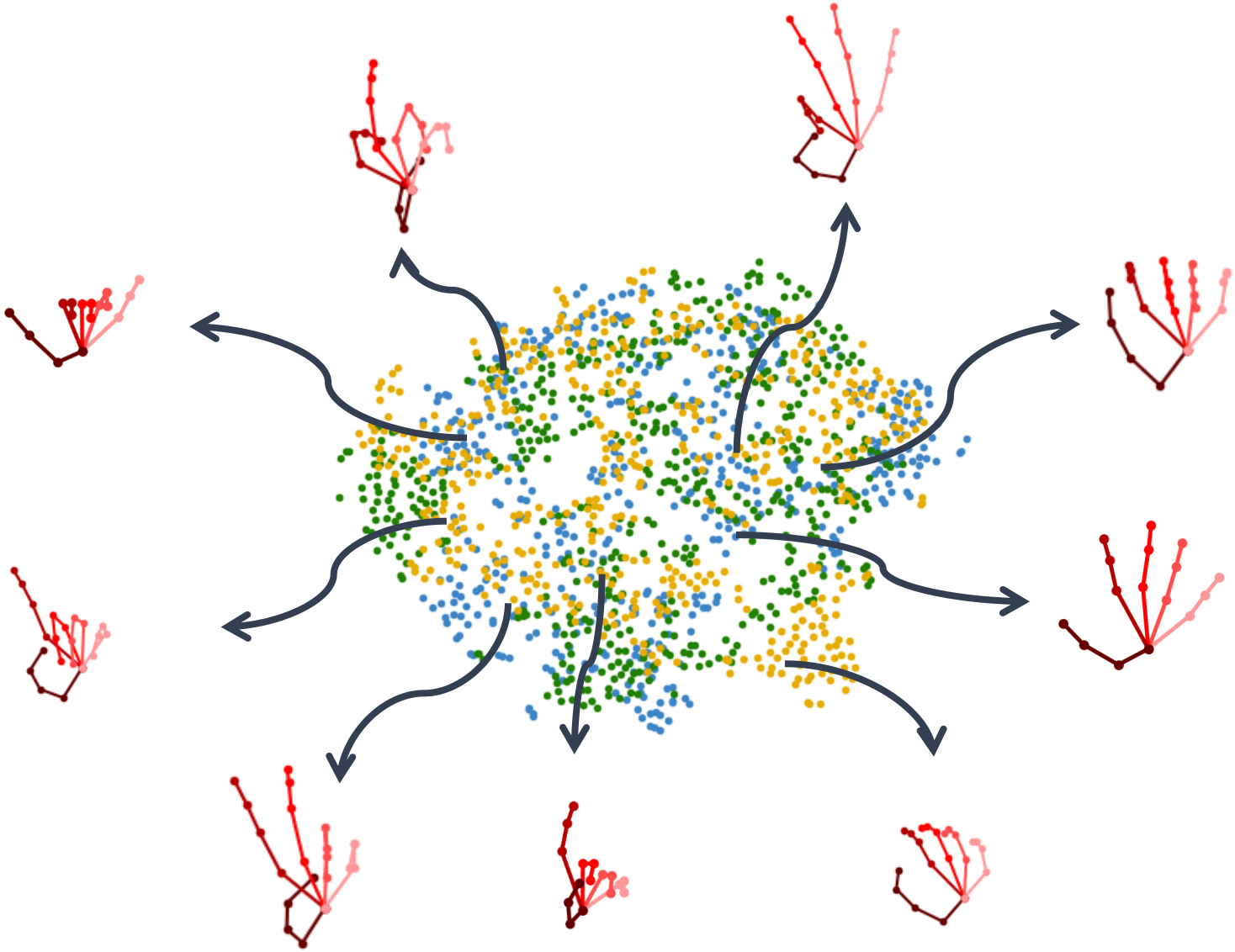


Adopted from [Tagliasacchi et al. 2015]

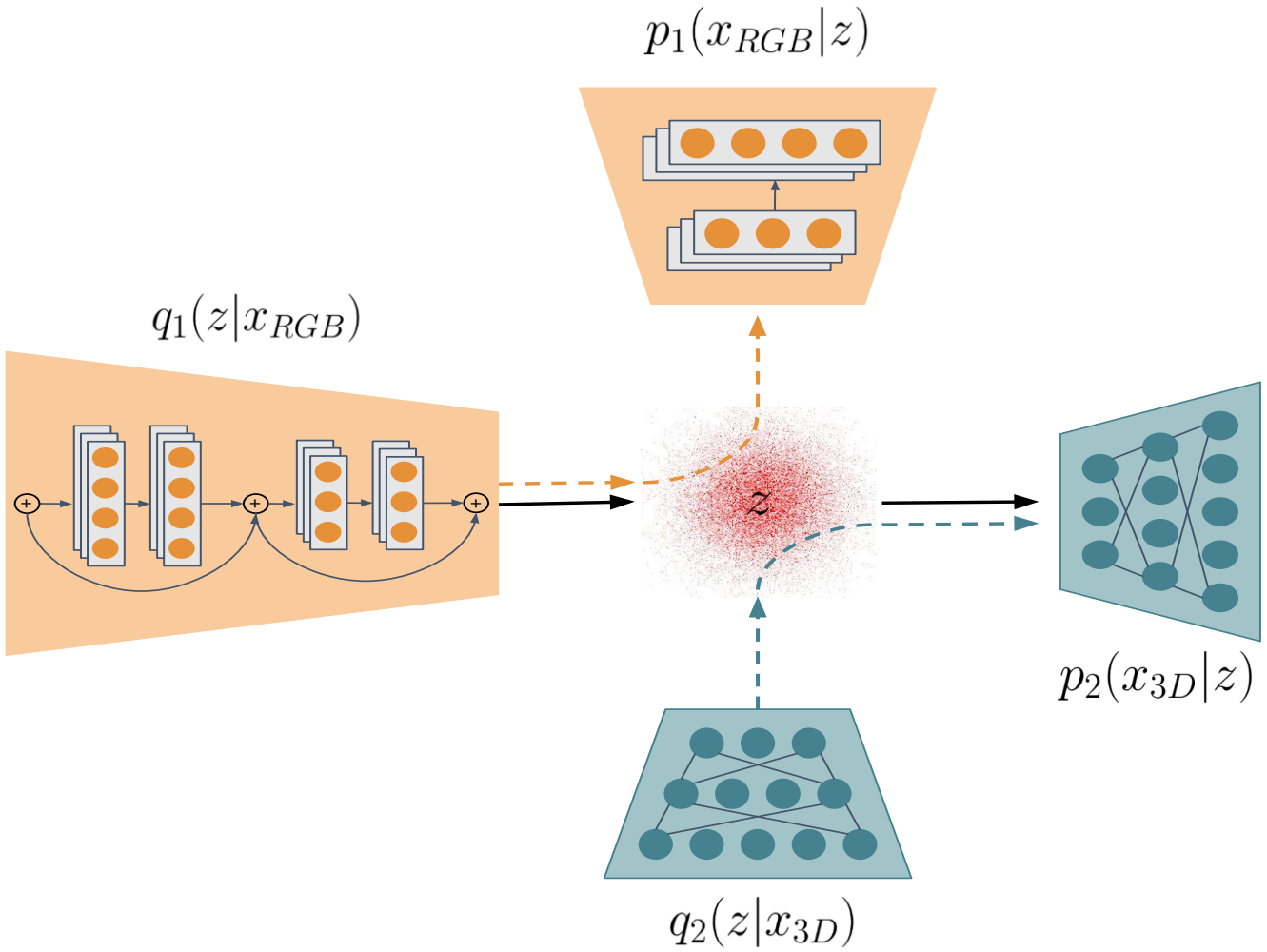
Cross-modal latent space – Inference model



Cross-modal latent space – Generative model



[Spurr et al.: CVPR '18]



[Spurr et al.: CVPR '18]

Cross-modal training objective

Original VAE considers only one modality → need to modify to consider multiple modalities

Consider x_i and x_t samples from different modalities (e.g RGB and 3D joint positions)

Important: Both describe inherently the same thing. In our case, the hand pose.

Task: maximize $\log p_\theta(x_t)$ under x_i

Cross-modal training objective derivation

$$\log p(x_t) = \int_z q(z|x_i) \log p(x_t) dz = \int_z q(z|x_i) \log \frac{p(x_t)p(z|x_t)q(z|x_i)}{p(z|x_t)q(z|x_i)} dz \qquad \text{(Bayes' Rule)}$$

$$= \int_z q(z|x_i) \log \frac{q(z|x_i)}{p(z|x_t)} dz + \int_z q(z|x_i) \log \frac{p(x_t)p(z|x_t)}{q(z|x_i)} dz \qquad \text{(Logarithms)}$$

$$= D_{KL}(q(z|x_i)||p(z|x_t)) + \int_z q(z|x_i) \log \frac{p(x_t|z)p(z)}{q(z|x_i)} dz \qquad \text{(Definition KL)}$$

$$\geq \int_z q(z|x_i) \log p(x_t|z)\, dz - \int_z q(z|x_i) \log \frac{q(z|x_i)}{p(z)} dz \qquad \left(\log \frac{x}{y} = -\log \frac{y}{x}\right)$$

$$= \mathbb{E}_{z\sim q(z|x_i)}[\log p(x_t|z)] - D_{KL}(q(z|x_i)||p(z)) \qquad (\mathbb{E} \text{ over z, KL})$$

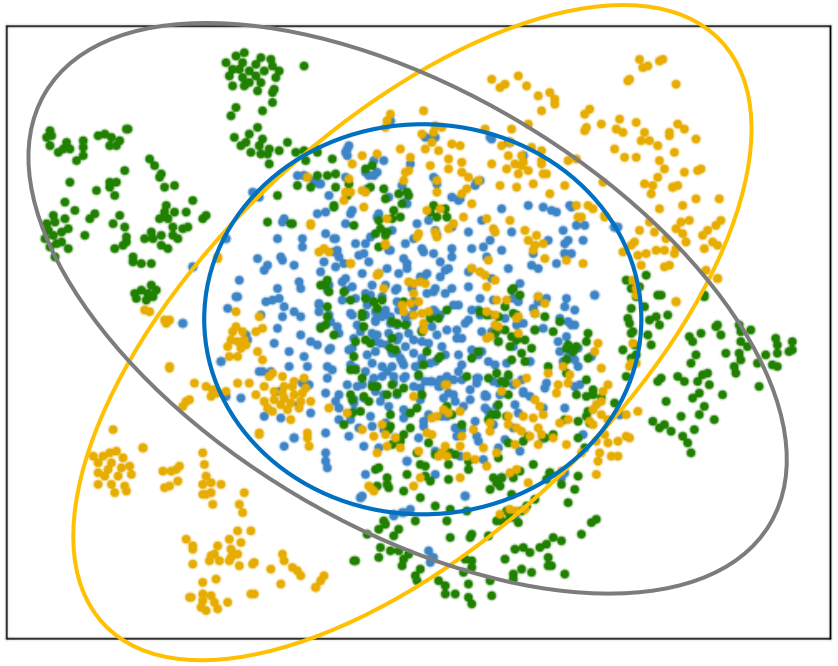
[Spurr et al.: CVPR '18]

t-SNE embedding comparisons

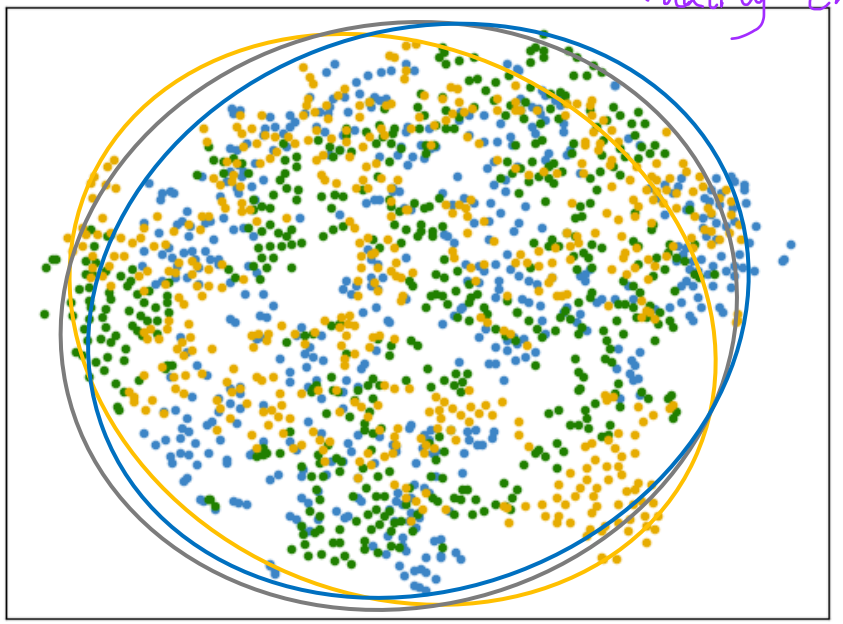
overlapping

one latent space similar configurations, indep. of their modality end up being close together

Trained individually



Trained cross-modal



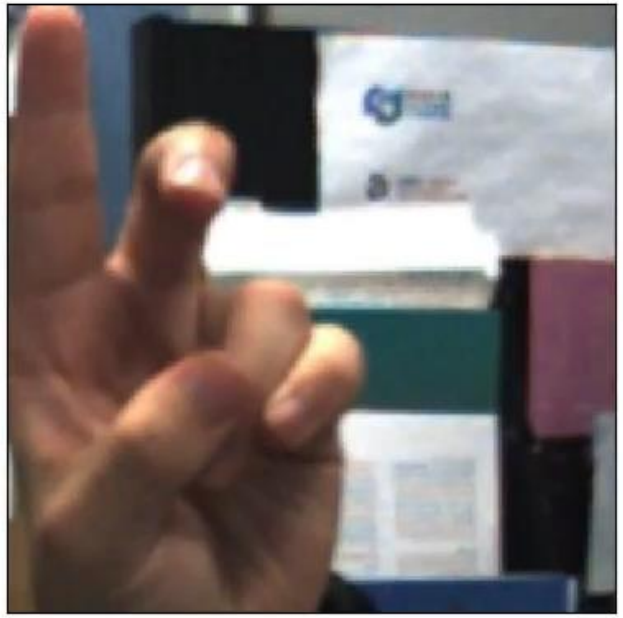
non-overlapping partially overlapping

latent space

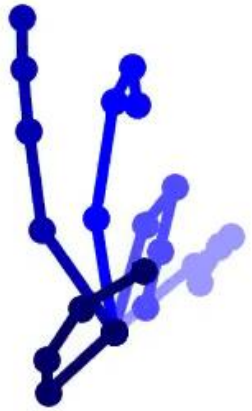
[Spurr et al.: CVPR ‘18]

Posterior estimates: RGB (real) to 3D joints

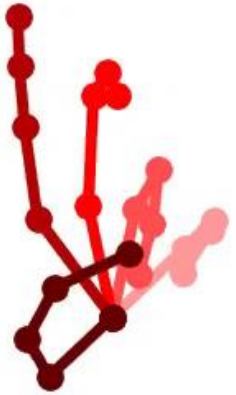
RGB Input



Ground Truth



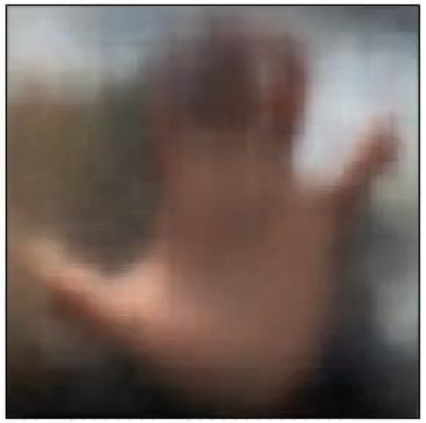
Prediction

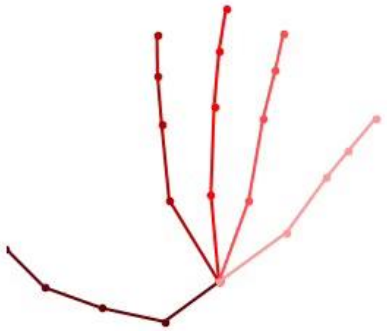


[Spurr et al.: CVPR '18]

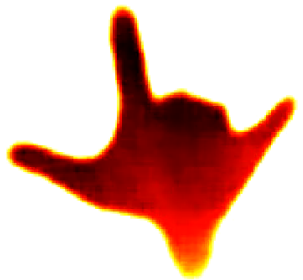
Synthetic hands: Latent Space Walk

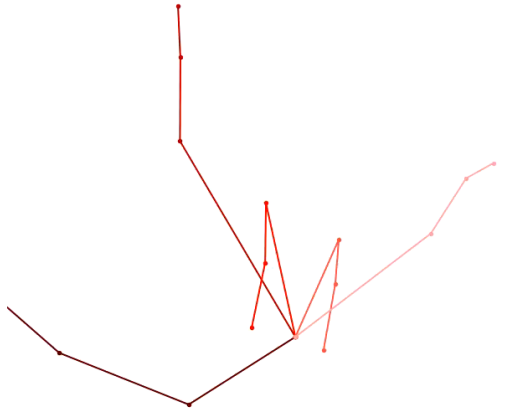
generated images





[Synthetic samples RGB]





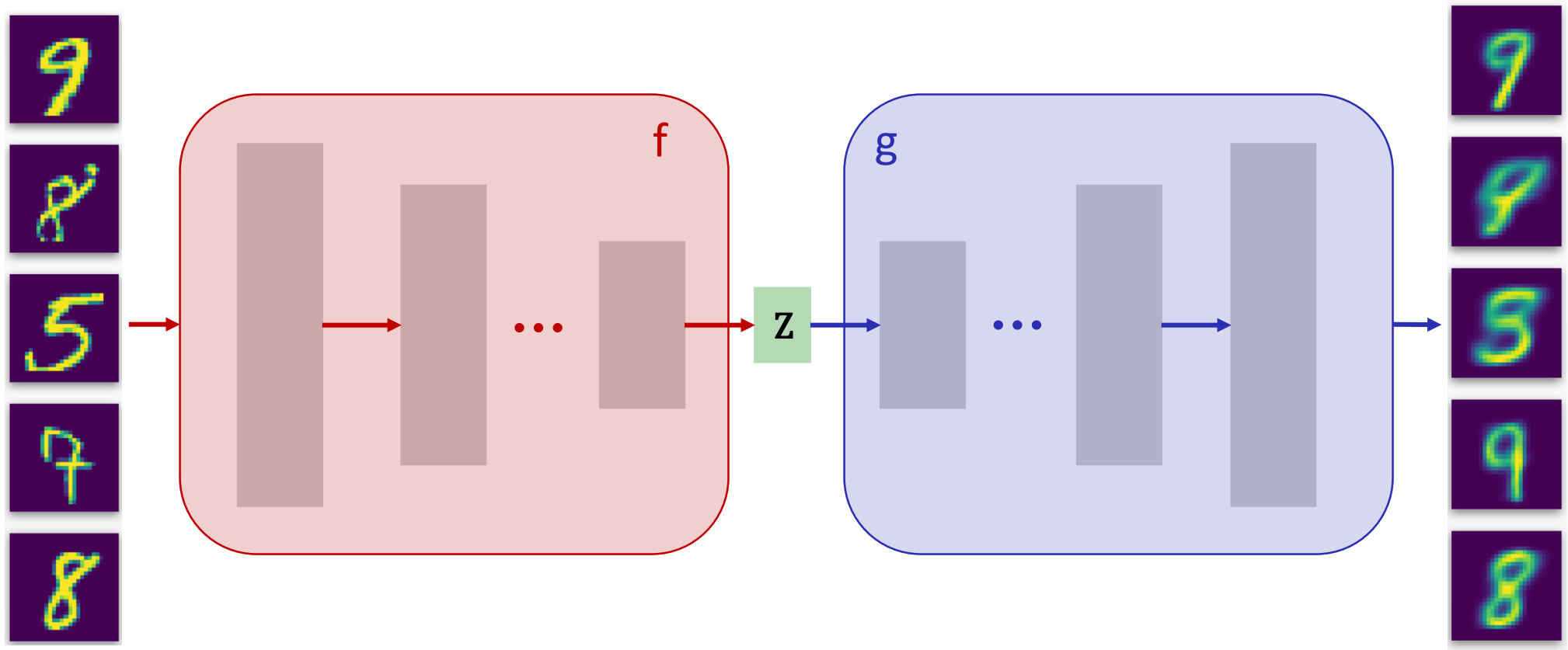
[Synthetic samples depth]

Learning useful representations

Goal: Learn features that capture similarities and dissimilarities

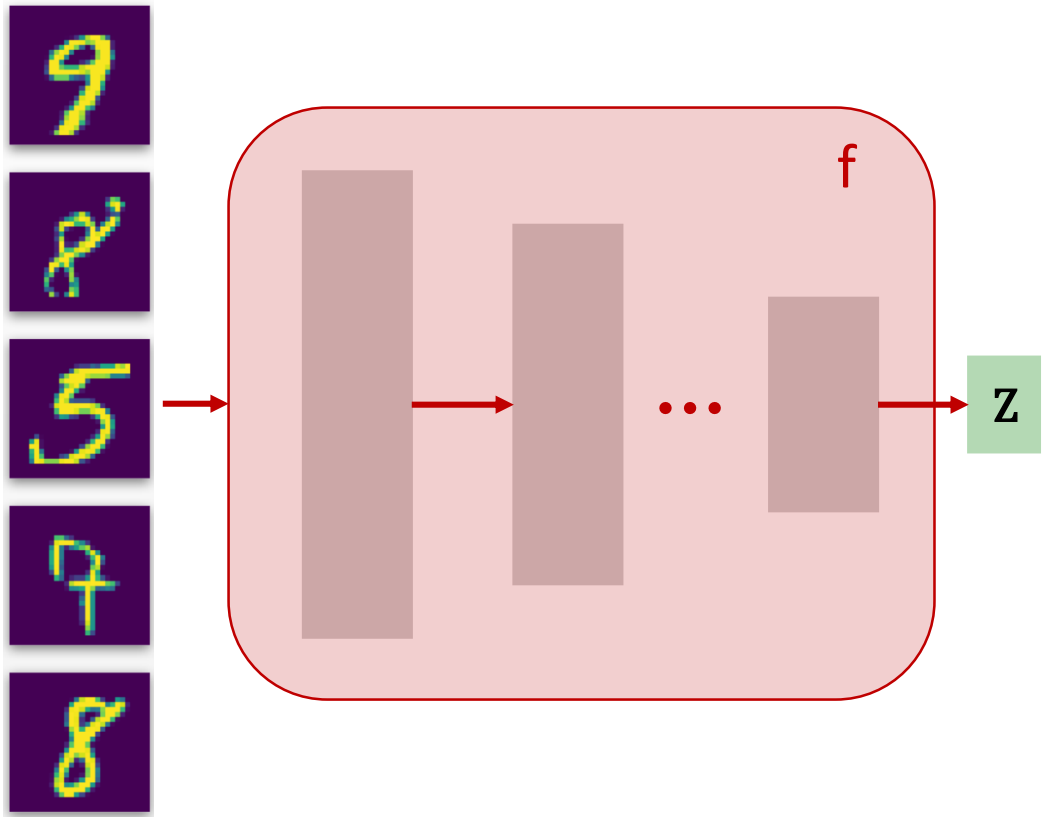
Requirement: Objective that defines notion of utility (task-dependent)

Autoencoders

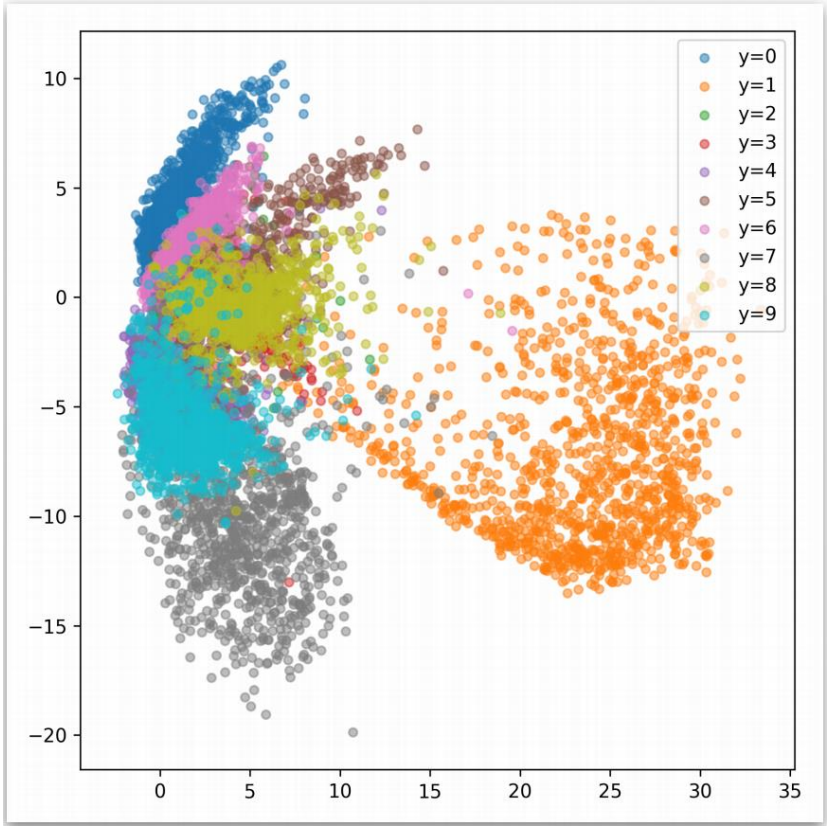


Notion of utility: Ability to reconstruct pixels from code (features are a compressed representation of original data)

Autoencoders



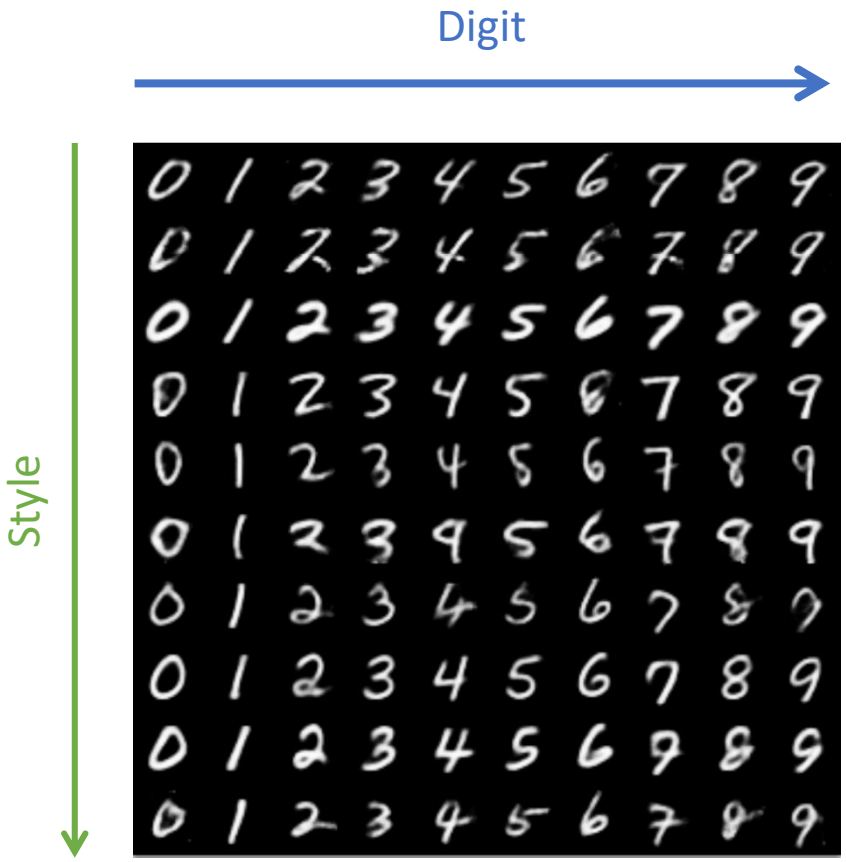
Learned representation

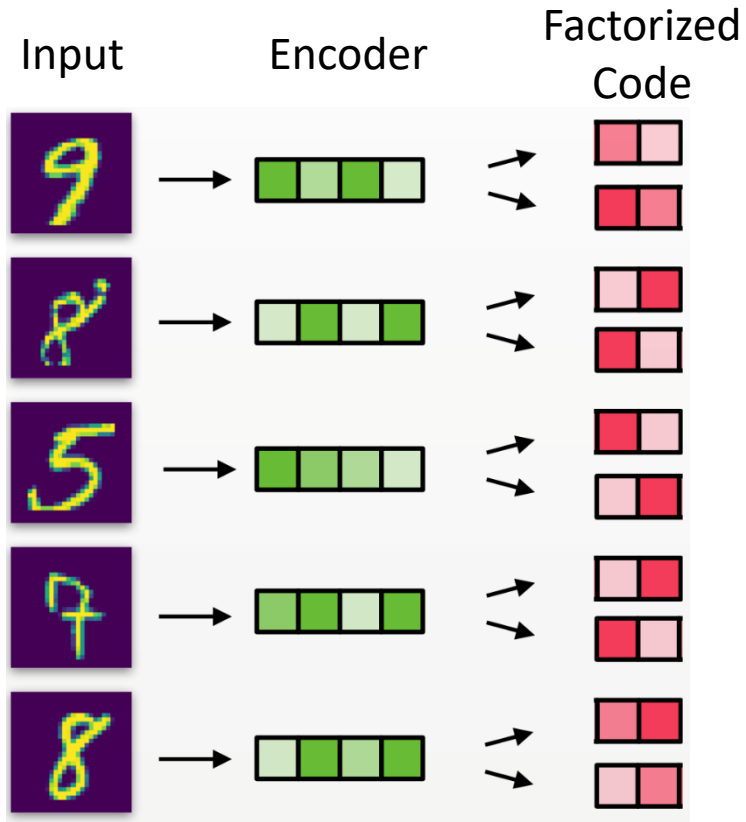


纠缠表示

Entangled Representation: Individual dimensions in code encode some unknown combination of features in the data.

Goal: Disentangled Representations

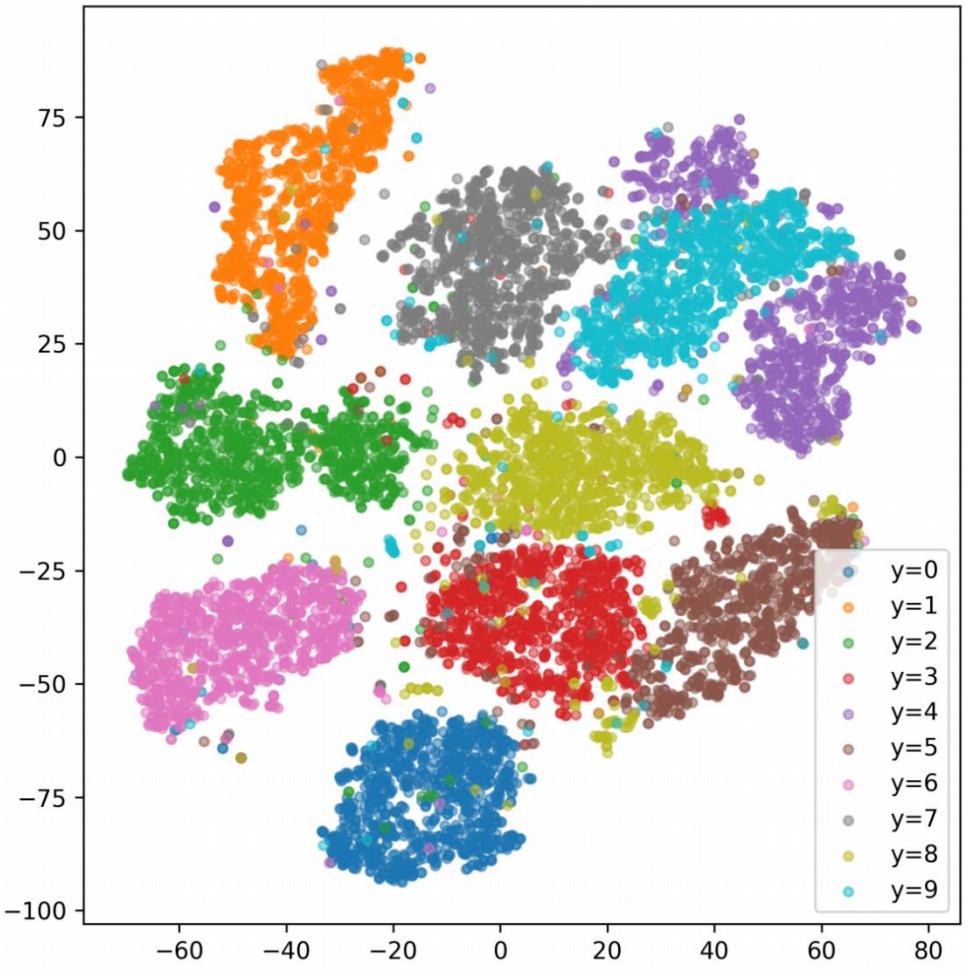




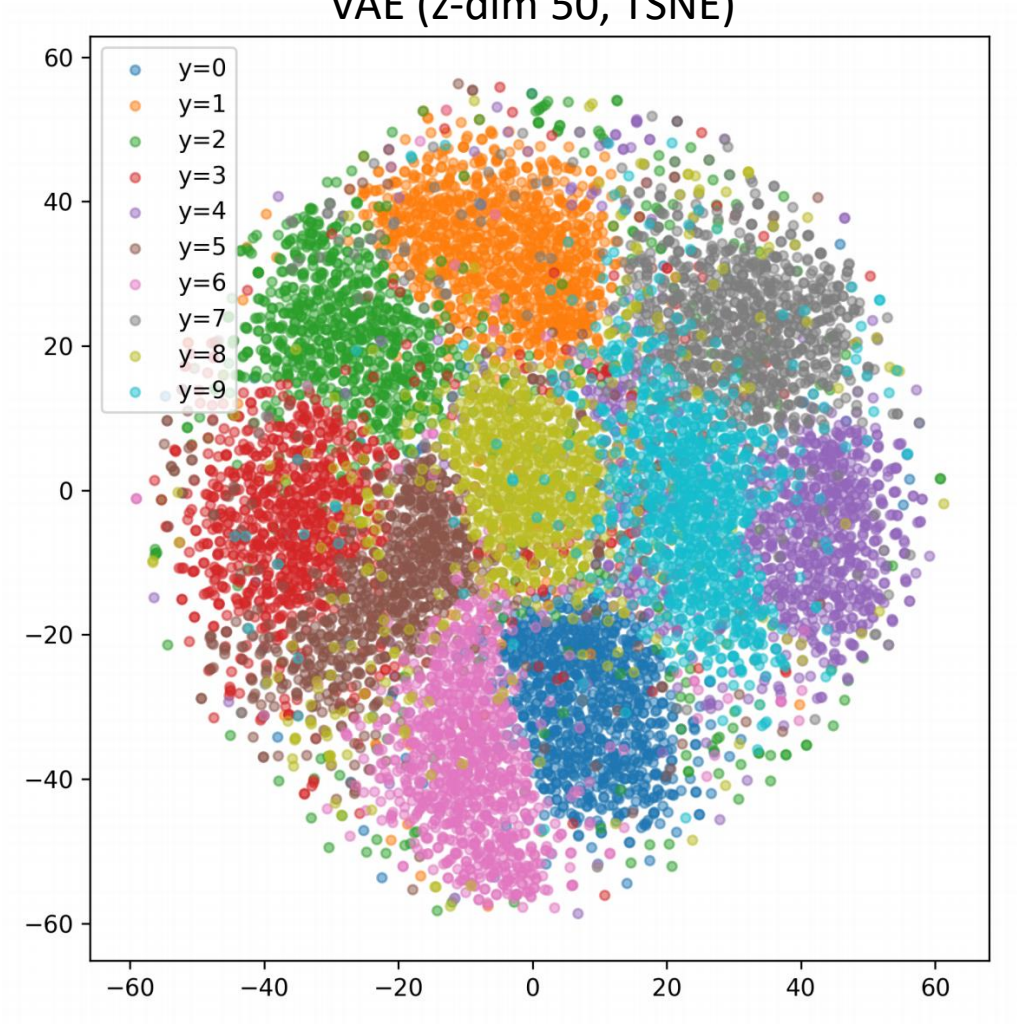
Goal: Learn features that correspond to distinct factors of variation
One **Notion of Utility**: statistical independence

Regular vs Variational Autoencoders

AE (z-dim 50, TSNE)



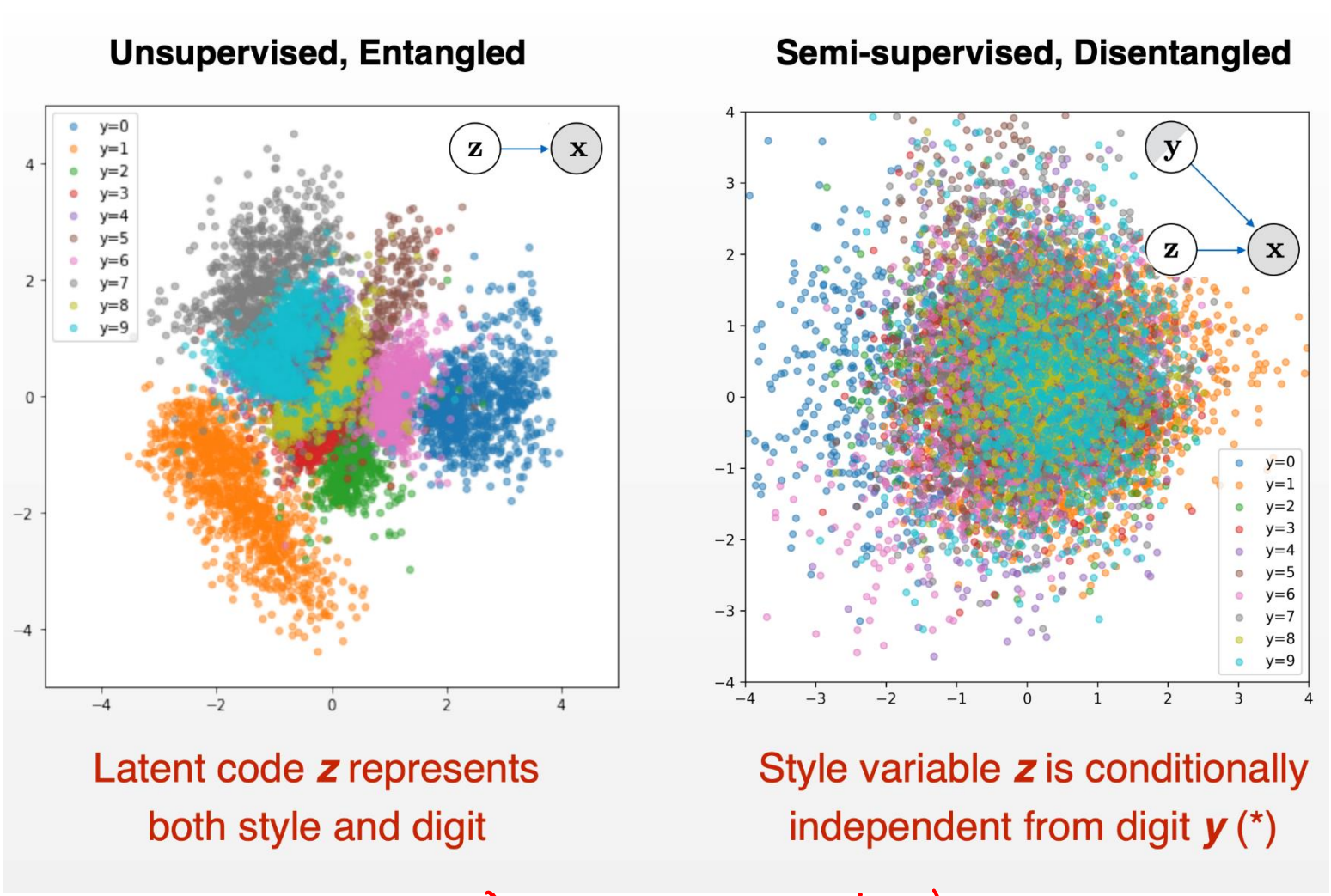
VAE (z-dim 50, TSNE)



Issue: KL-Divergence regularizes values of z but Representations are still entangled

Unsupervised vs Semi-supervised

[Kingma et al. Semi-Supervised Learning with Deep Generative Models. NIPS 2014]



$p_\theta(x|y,z)$ $q_\phi(y,z|x)$

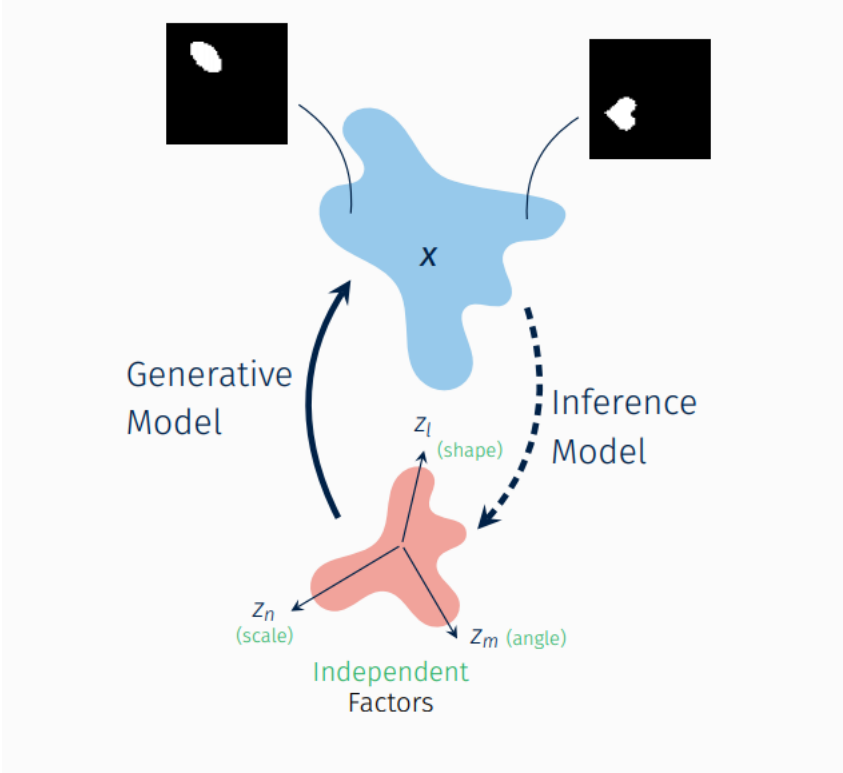
[image source: Babak Esmaeli]

Digit

Style

Separate label y from "nuisance" variable z

Learning factorized representations (unsupervised)



[image source: Emile Mathieu]

β-VAE

[Higgins et al., "Learning Basic Visual Concepts with a Constrained Variational Framework", ICLR 2017]

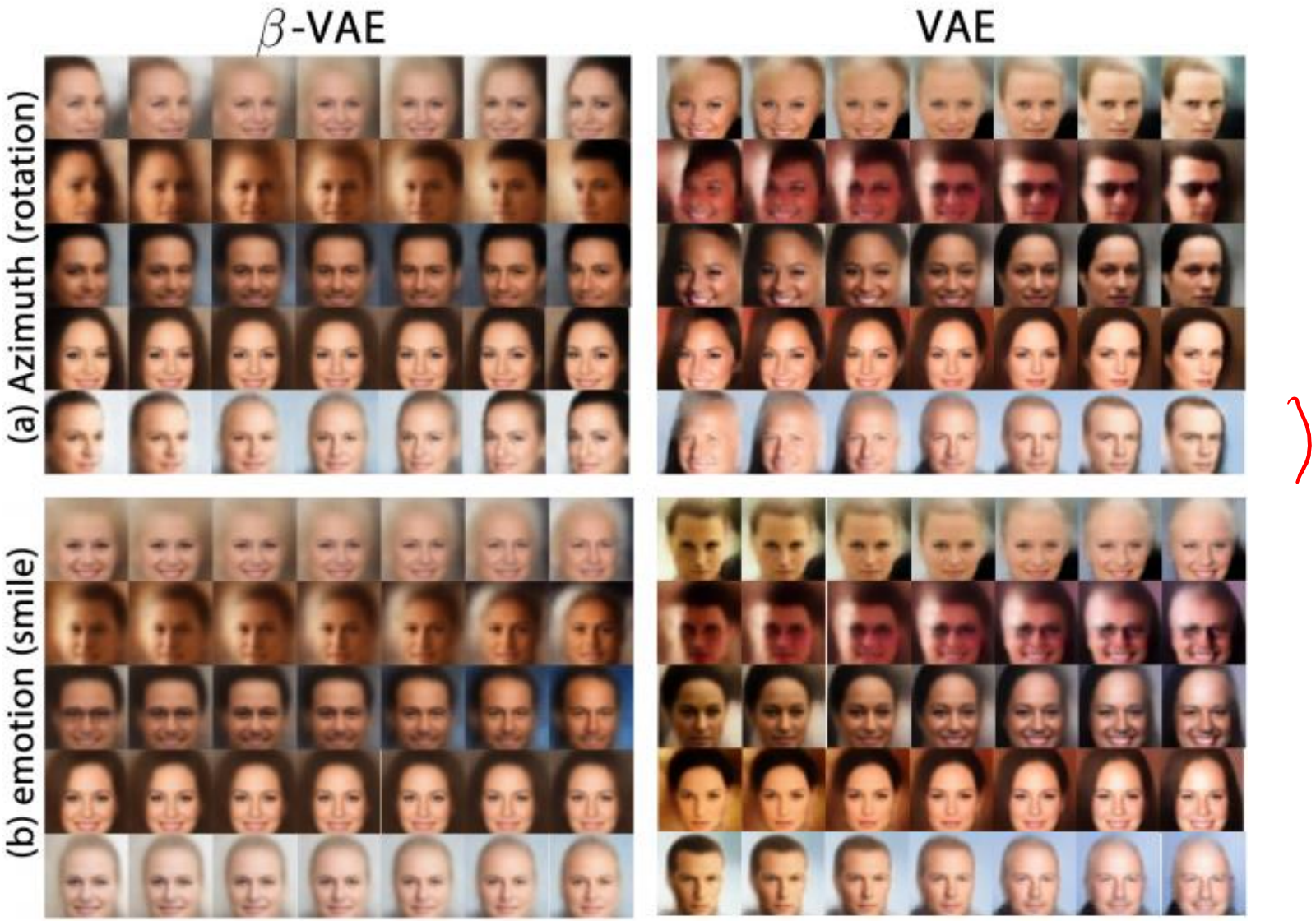
Goal: Learn disentangled representation *without* supervision

Idea: Provide framework for automated discovery of interpretable factorised latent representations

Approach: modification of the VAE, introducing an adjustable hyperparameter beta that balances latent channel capacity and independence constraints with reconstruction accuracy.

β-VAE

[Higgins et al., "Learning Basic Visual Concepts with a Constrained Variational Framework", ICLR 2017]



β-VAE

[Higgins et al., "Learning Basic Visual Concepts with a Constrained Variational Framework", ICLR 2017]

$$p(\mathbf{x}|\mathbf{z}) \approx p(\mathbf{x}|\mathbf{v}, \mathbf{w})$$

Conditionally independent factors $\mathbf{v} \in \mathbb{R}^K$, where $\log p(\mathbf{v}|\mathbf{x}) = \sum_k \log p(v_k|\mathbf{x})$

Conditionally dependent factors $\mathbf{w} \in \mathbb{R}^H$

$\mathbf{z} \in \mathbb{R}^M$, where $M \geq K$

model : which factor $\in v$, which $\in w$

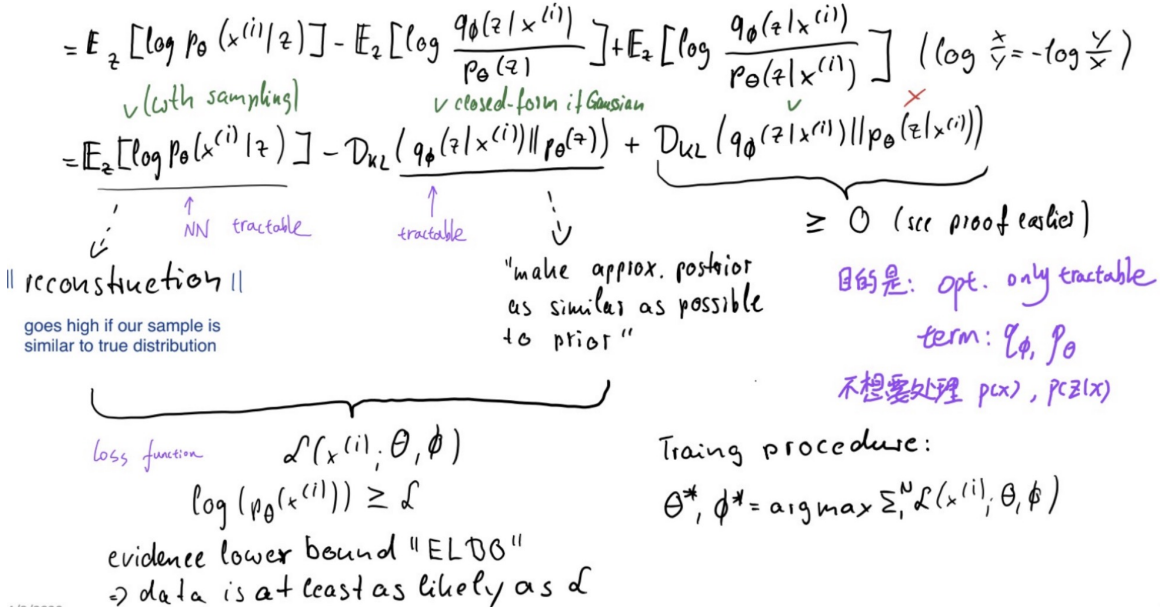
Assume that samples x are generated by controlling the generative factors v and w.

β-VAE – Training Objective

[Higgins et al., "Learning Basic Visual Concepts with a Constrained Variational Framework", ICLR 2017]

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z)\right]$$

$$\text{subject to } \mathrm{D}_{\mathrm{KL}}\left(q_\phi(z|x) \parallel p_\theta(z)\right) < \delta$$



β-VAE – Training Objective

[Higgins et al., "Learning Basic Visual Concepts with a Constrained Variational Framework", ICLR 2017]

$\beta = 1$, => ELBO

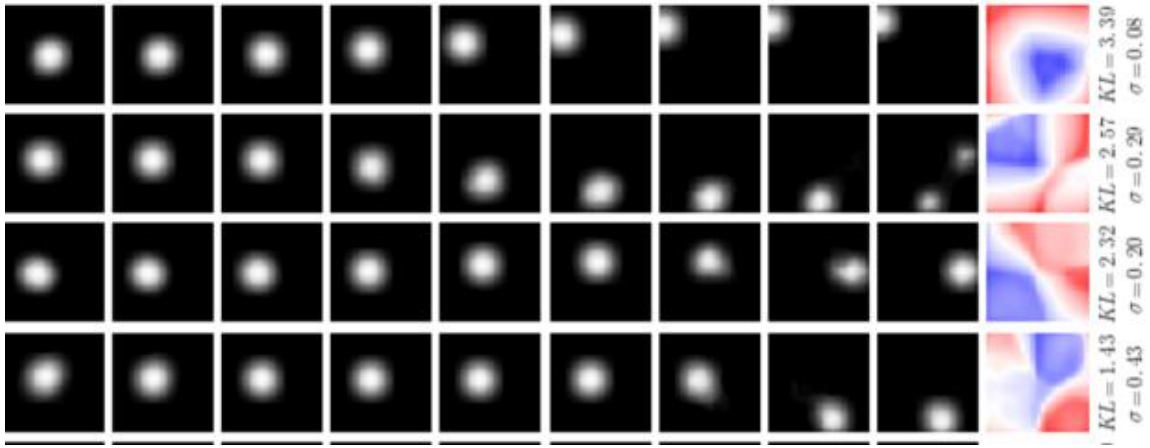
$\beta > 1$, => put more pressure on KL

$$\begin{aligned}
\mathcal{F}(\theta, \phi, \beta) &= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - \beta(D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) - \delta) \\
&= \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) + \beta\delta \\
&\geq \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})) \qquad ; \text{Because } \beta,\delta \geq 0
\end{aligned}$$

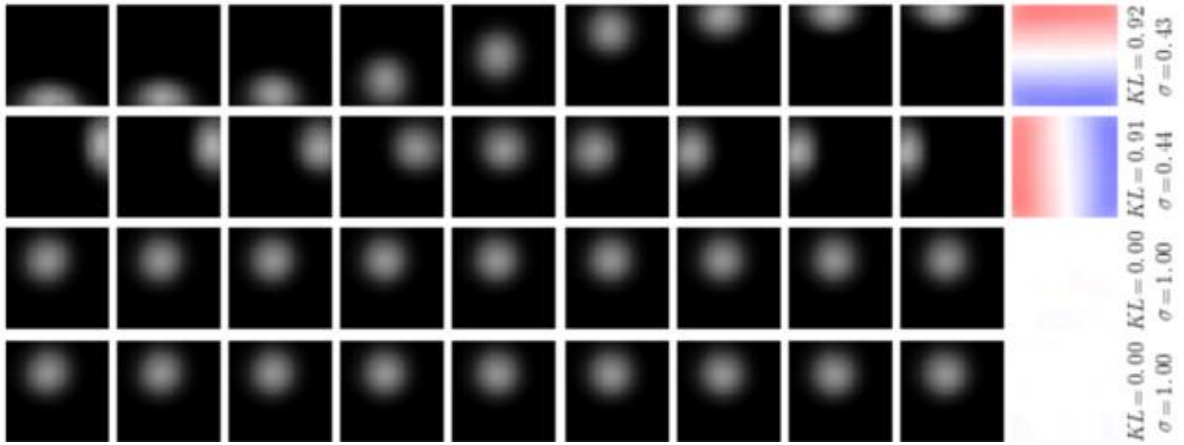
$$L_{\mathrm{BETA}}(\phi, \beta) = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) + \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}))$$

Entangled versus Disentangled Representations

vanilla VAE



β-VAE



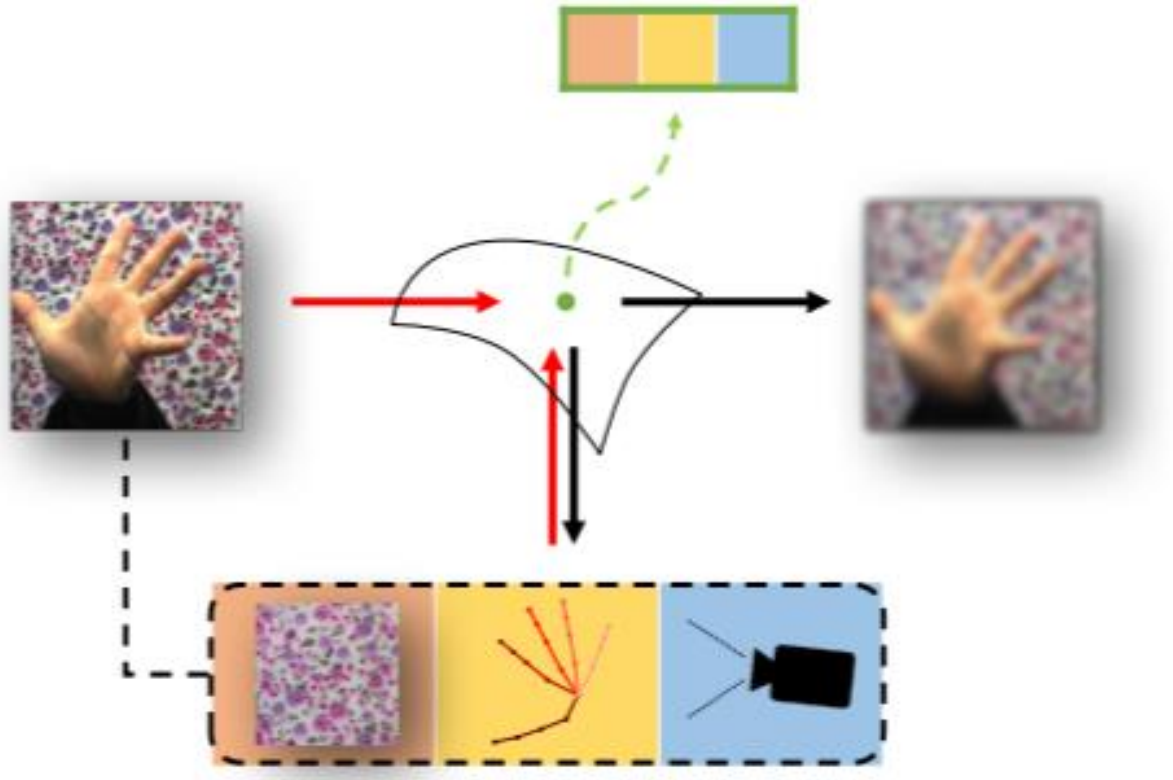
Disentangling Latent Hands for Image Synthesis and Pose Estimation

Linlin Yang, Angela Yao

CVPR 2019

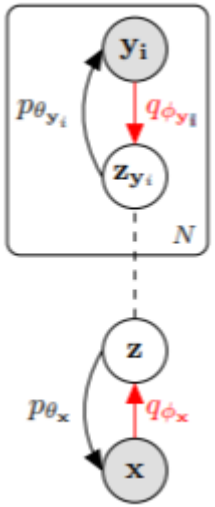
[Yang et al.: CVPR ‘19]

Disentangling latent factors



Disentangling latent factors

$$L(\phi_x, \phi_{y_1}, \phi_{y_2}, \theta_x, \theta_{y_1}, \theta_{y_2}) = ELBO_{dis}(x, y_1, y_2, \phi_{y_1}, \phi_{y_2}, \theta_x, \theta_{y_1}, \theta_{y_2}) + ELBO_{emb}(x, y_1, y_2, \phi_x).$$



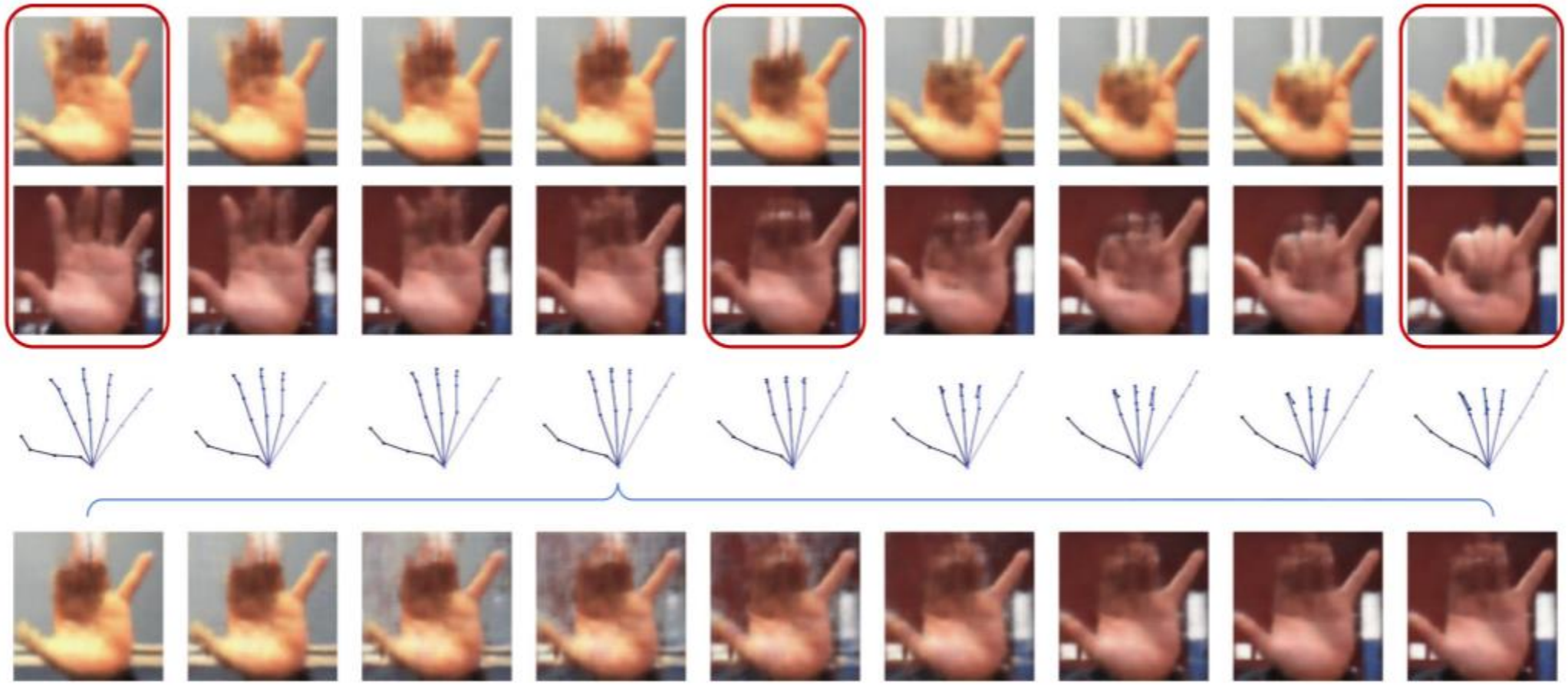
where

$$\begin{aligned}\log p(x, y1, y2) \geq ELBO_{dis}(x, y1, y2, \phi_{y_1}, \phi_{y_2}, \theta_x, \theta_{y_1}, \theta_{y_2}) \\ = \lambda_x \mathbb{E}_{z \sim q_{\phi_{y_1}, \phi_{y_2}}}[\log p_{\theta_x}(x|z)] \\ + \lambda_{y_1} \mathbb{E}_{z_{y_1} \sim q_{\phi_{y_1}}}\left[\log p_{\theta_{y_1}}(y_1|z_{y_1})\right] \\ + \lambda_{y_2} \mathbb{E}_{z_{y_2} \sim q_{\phi_{y_2}}}\left[\log p_{\theta_{y_2}}(y_2|z_{y_2})\right] \\ - \beta D_{KL}\left(q_{\phi_{y_1}, \phi_{y_2}}(z|y_1, y_2)||p(z)\right)\end{aligned}$$

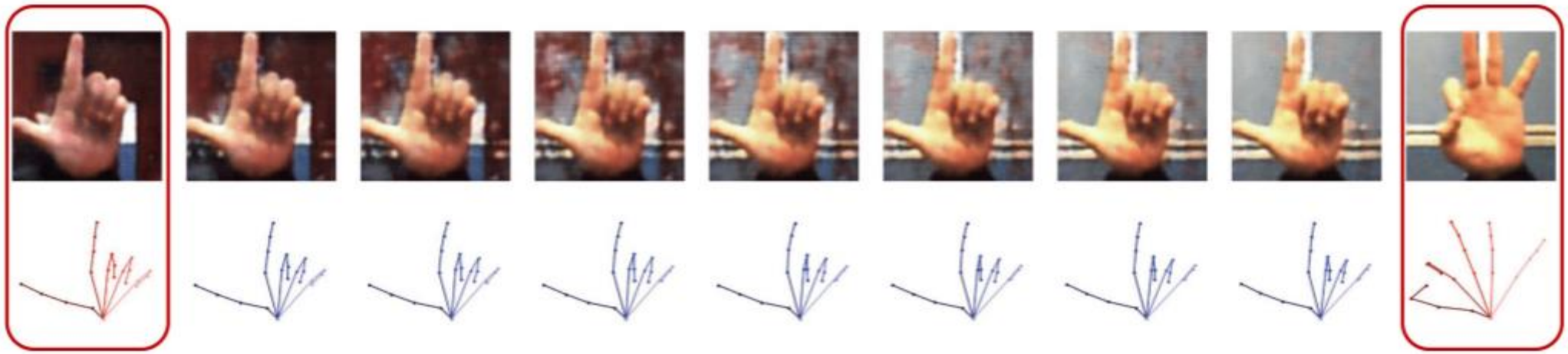
and

$$\begin{aligned}ELBO_{emb}(x, y_1, y_2, \phi_x) = \lambda'_x \mathbb{E}_{z \sim q_{\phi_x}}[\log p_{\theta_x}(x|z)] \\ + \lambda'_{y_1} \mathbb{E}_{z_{y_1} \sim q_{\phi_x}}\left[\log p_{\theta_{y_1}}(y_1|z_{y_1})\right] \\ + \lambda'_{y_2} \mathbb{E}_{z_{y_2} \sim q_{\phi_x}}\left[\log p_{\theta_{y_2}}(y_2|z_{y_2})\right] \\ - \beta' D_{KL}\left(q_{\phi_x}(z|x)||p(z)\right)\end{aligned}$$

Latenspace walk with disentangled z



Latenspace walk with disentangled z



Limitations of VAEs

Tendency to generate blurry images.

Believed to be due to injected noise and weak inference models (Gaussian assumption of latent samples to simplistic, model capacity to weak)

More expressive model → substantially better results (e.g Kingma et al. 2016, *Improving variational inference with inverse autoregressive flow*)

Next

Deep Generative Modelling II:

Generative Adversarial Networks (GANs)