Explicit Disentanglement of Appearance and Perspective in Generative Models

Nicki S. Detlefsen[1] nsde@dtu.dk and Søren Hauberg[1] sohau@dtu.dk

[1]Technical University of Denmark

Introduction

Disentangled Representation Learning (DRL) approaches assume that an AI agent can benefit from separating out (disentangle) the underlying structure of data into disjointed parts of its representation.

Problem: DRL is fundamentally impossible without an inductive bias in the disentanglement model [1]. Reasearch in this area has been limited.

Our approach: Incorporate a spatial transformer layer into a variational autoencoder (VAE) [2], introducing a inductive bias towards specific transformations in data.

Architecture

Definition: We define the *appearance* as being the factors of data that are left after transforming x by its *perspective*.

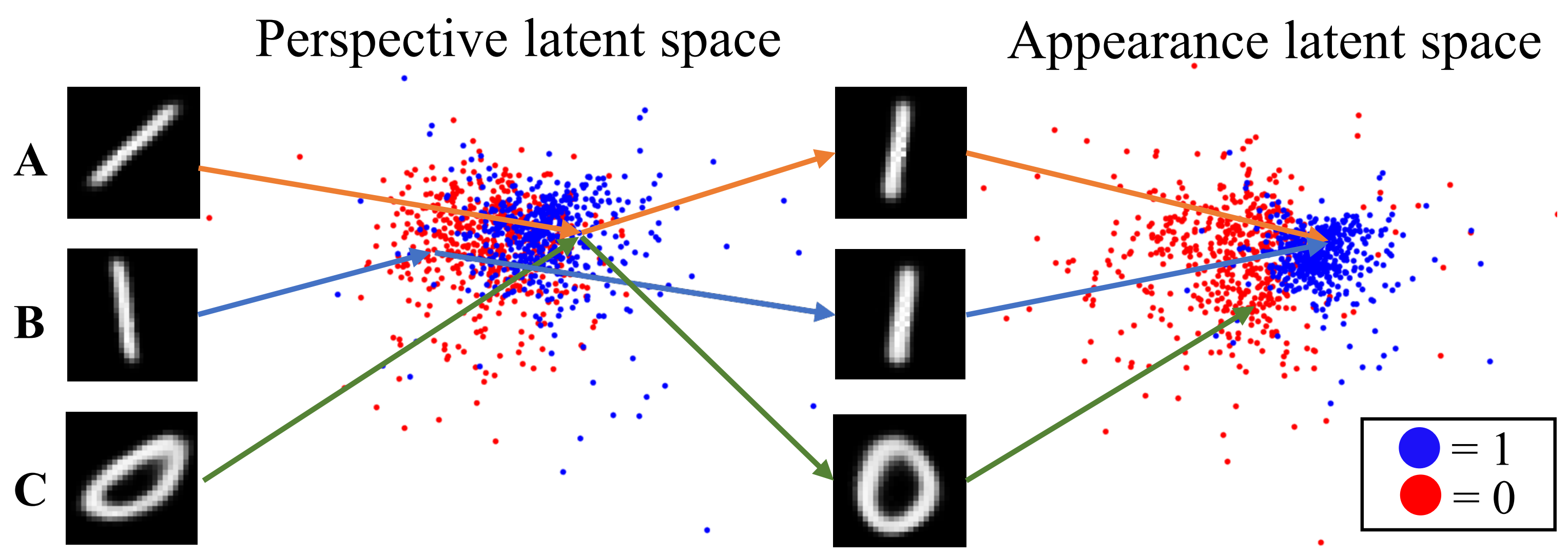


Figure: Disentangle of data into *appearance* and *perspective* factors.

Our generative model:

$$p(\boldsymbol{x}) = \iint p(\boldsymbol{x}|\boldsymbol{z}_A, \boldsymbol{z}_P)p(\boldsymbol{z}_A)p(\boldsymbol{z}_P)\mathrm{d}\boldsymbol{z}_A\mathrm{d}\boldsymbol{z}_P$$

$$p(\boldsymbol{z}) = \mathcal{N}(0, \mathbb{I}_d)$$

$$p(\boldsymbol{x}|\boldsymbol{z}_A, \boldsymbol{z}_P) = \mathcal{N}(\boldsymbol{x}|\mu_p(\boldsymbol{z}_A), \sigma_p^2(\boldsymbol{z}_P))$$

Sampling process:

1. Sample $\boldsymbol{z}_A$ and $\boldsymbol{z}_P$ from $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$.
2. Decode both samples $\tilde{\boldsymbol{x}} \sim p(\boldsymbol{x}|\boldsymbol{z}_A)$, $\boldsymbol{\gamma} \sim p(\boldsymbol{x}|\boldsymbol{z}_P)$.
3. Transform $\tilde{\boldsymbol{x}}$ with parameters $\boldsymbol{\gamma}$ using a spatial transformer layer: $\boldsymbol{x} = \mathcal{T}_\gamma(\tilde{\boldsymbol{x}})$.

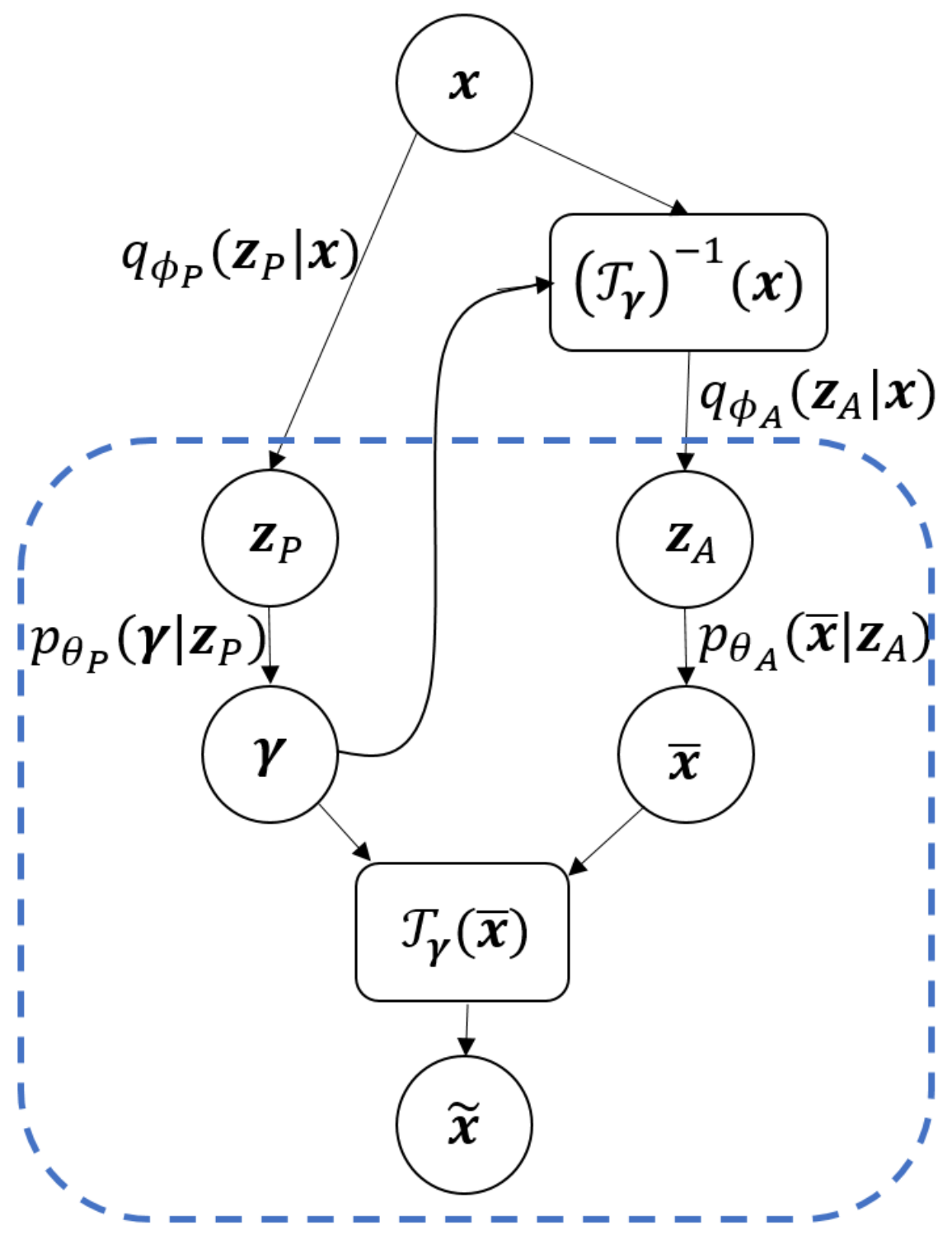

$\boldsymbol{x}$
$q_{\phi_P}(\boldsymbol{z}_P|\boldsymbol{x})$
$(\mathcal{T}_\gamma)^{-1}(\boldsymbol{x})$
$q_{\phi_A}(\boldsymbol{z}_A|\boldsymbol{x})$
$\boldsymbol{z}_P$
$\boldsymbol{z}_A$
$p_{\theta_P}(\boldsymbol{\gamma}|\boldsymbol{z}_P)$
$p_{\theta_A}(\overline{\boldsymbol{x}}|\boldsymbol{z}_A)$
$\boldsymbol{\gamma}$
$\overline{\boldsymbol{x}}$
$\mathcal{T}_\gamma(\overline{\boldsymbol{x}})$
$\tilde{\boldsymbol{x}}$

Figure: Our VITAE architecture

Training details:

- Optimized using variational inference
- Use inverse transform in encoder to mimic the inverse of the decoder
- Inference of appearance then becomes dependent on perspective

$$p(\boldsymbol{z}_P|\boldsymbol{x}) \approx q_P(\boldsymbol{z}_P|\boldsymbol{x}) \text{ and } p(\boldsymbol{z}_A|\boldsymbol{x}) \approx q_A(\boldsymbol{z}_A|\boldsymbol{x}, \boldsymbol{z}_P)$$

By explicit incorporating a spatial transformer in the architecture, the model gains a inductive bias towards transformations parametrized by $\mathcal{T}_\gamma$. We denote this model for *Variational Inferred Transformational AutoEncoder (VITAE).*

Inductive bias

The inductive bias is restricted by the used transformation. By requirements of the architecture the transformations needs to have three properties:'

1. Differentiable
2. Invertible
3. Differentiable invertible

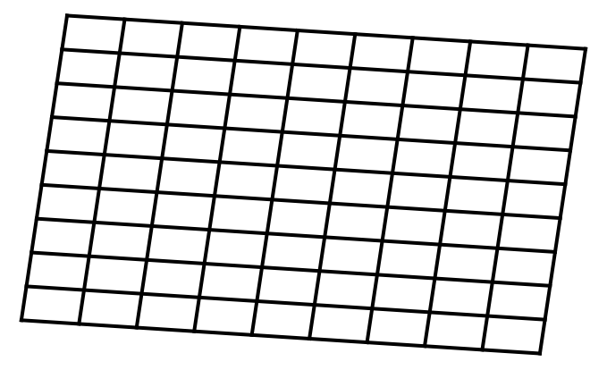
} Diffiomorphic!

We investigate two choices:

Affine

Parametrize the velocity field of the transformation using the matrix exponential operator

$$\mathcal{T}_\gamma(\boldsymbol{x}) = \textbf{expm}\left(\begin{bmatrix}\gamma_{11} & \gamma_{12} & \gamma_{13}\\ \gamma_{21} & \gamma_{22} & \gamma_{14}\\ 0 & 0 & 0\end{bmatrix}\right)\begin{bmatrix}x\\ y\\ 1\end{bmatrix}.$$

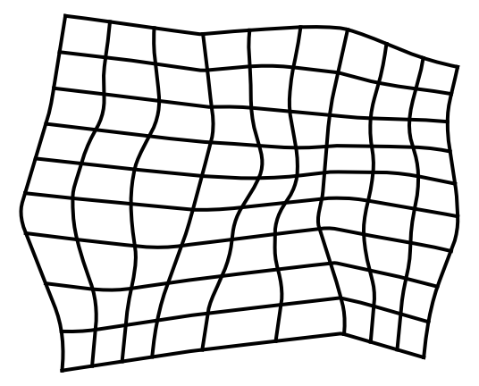


CPAB

Use highly flexible transformation [3] as black box transformation when no prior knowledge exist.

$$\mathcal{T}_\gamma(\boldsymbol{x}) = \boldsymbol{x} + \int_0^1 v^{\boldsymbol{\theta}}(\phi^{\boldsymbol{\theta}}(\boldsymbol{x};\tau))\,d\tau$$

with $v^{\boldsymbol{\theta}}$ begin a continues vector field.



For both transformation it holds that $\mathcal{T}_\gamma^{-1} = \mathcal{T}_{-\gamma}$. Note: diffiomorphic transformations forms a smooth group, required by [1]. We are the first to for fill this requirement.

MNIST

Below we try to disentangle digit (appearance) from writing style (perspective) on randomly translated and rotated MNIST images.

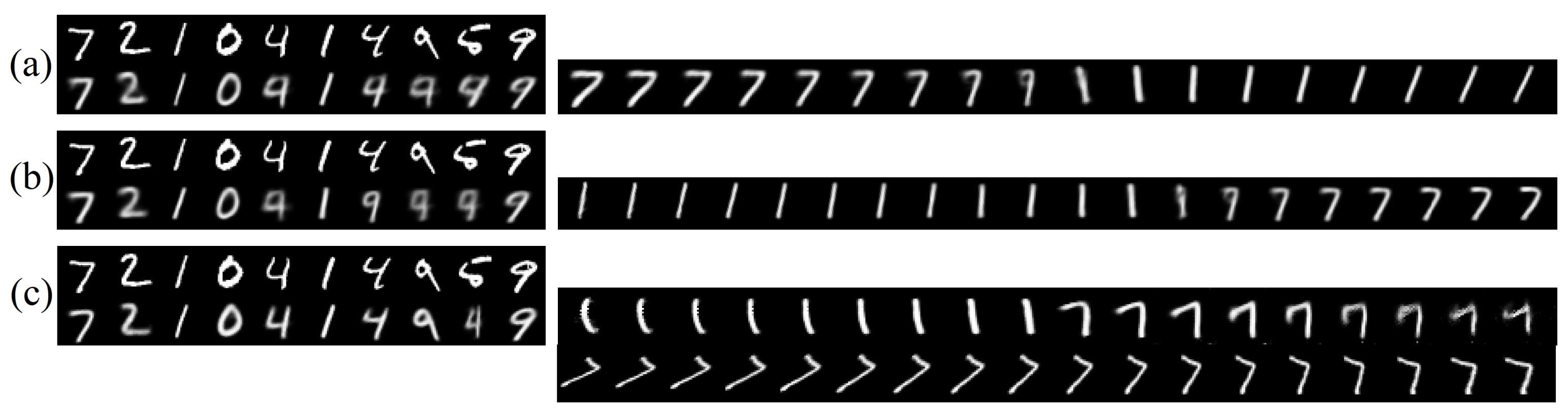
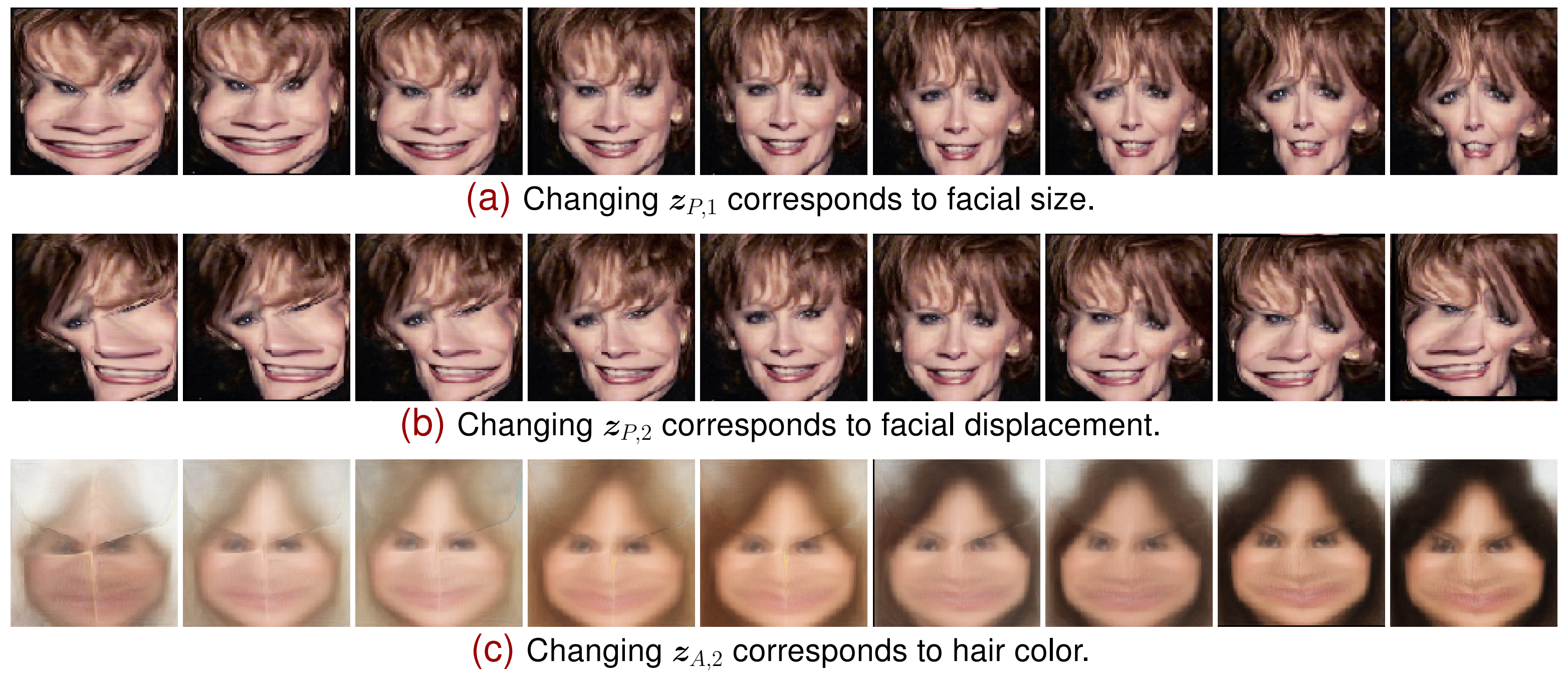


Figure: Reconstructions (left images) and manipulation of latent codes (right images) on MNIST for the three different models: VAE (a), β-VAE (b) and C-VITAE (c).

CelebA

Train model to disentangle facial shape (perspective) from facial features (appearance) on CelebA dataset.



(a) Changing $\boldsymbol{z}_{P,1}$ corresponds to facial size.

(b) Changing $\boldsymbol{z}_{P,2}$ corresponds to facial displacement.

(c) Changing $\boldsymbol{z}_{A,2}$ corresponds to hair color.

Figure: Traversal in latent space shows, that our model can disentangle complex factors such as facial size, facial position and hair color.

Quantitative study

	dSprite			MNIST			SMPL		
	ELBO	$\log p(\boldsymbol{x})$	D_{score}	ELBO	$\log p(\boldsymbol{x})$	D_{score}	ELBO	$\log p(\boldsymbol{x})$	D_{score}
VAE [2]	-47.05	-49.32	0.05	-169	-172	0.579	-8.62×10^3	-8.62×10^3	0.485
β-VAE [4]	-79.45	-81.38	0.18	-150	-152	0.653	-8.62×10^3	-8.60×10^3	0.525
β-TCVAE [5]	-66.48	-68.12	0.30	-141	-144	0.679	-8.62×10^3	-8.56×10^3	0.651
DIP-VAE-II [6]	**-46.32**	**-48.92**	0.12	-140	-155	0.733	-8.62×10^3	-8.54×10^3	0.743
U-VITAE	-55.25	-57.29	0.22	-142	-143	0.782	-8.62×10^3	-8.55×10^3	0.673
C-VITAE	-68.26	-70.49	**0.38**	**-139**	**-141**	**0.884**	$\mathbf{-8.62\times10^3}$	$\mathbf{-8.52\times10^3}$	**0.943**

Table: Quantitative results on three datasets. For each dataset we report the ELBO, test set log likelihood and disentanglement score D_{score}. Bold marks best results.

Human pose and shape disentanglement

Learn model that can disentangle body pose (appearance) and shape (perspective) on generated human bodies (SMPL framework).

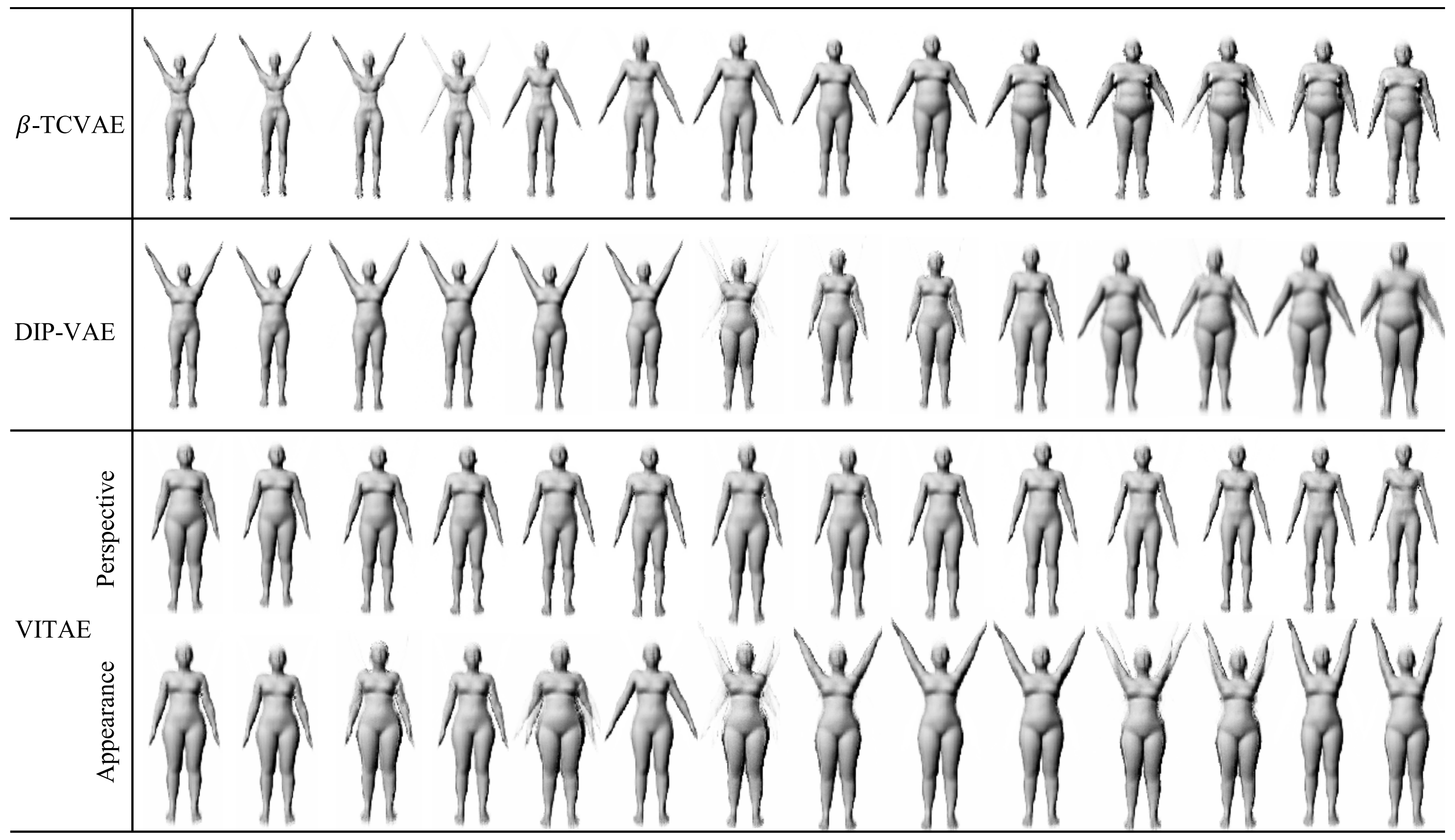


Figure: Latent traversals

β-TCVAE and DIP-VAE-II both changes the body shape and body position when latent space is traversed. Our proposed VITAE model, better disentangle these two factors, but not perfectly.

Our proposed VITAE architecture gives the model a short-cut to learn highly complex transformations of the input, that otherwise would required overparametrized neural networks.

Conclusion

- We provide an explicit framework for disentanglement of appearance and perspective factors
- State-of-the-art disentanglement of style and content on MNIST images and similar of pose and shape on generated human bodies
- We (unsurprisingly) find that in situations where prior knowledge about the generative factors are known, incorporating these into the model give better results than ignoring such information.

References

[1] F. Locatello et. al, Challenging common assumptions in the unsupervised learning of disentangled representations, 2018.

[2] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes, 2013.

[3] O. Freifeld et. al, Highly-expressive spaces of well-behaved transformations: Keeping it simple, 2015.

[4] L. Higgins, et. al, β-vae: Learning basic visual concepts with a constrained variational framework, 2017.

[5] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders, 2018

[6] A. Kumar, P. Sattigeri, and A. Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, 2017

[7] C. Eastwood and C. K. I. Williams. A Framework for the Quantitative Evaluation of Disentangled Representations, 2019