State Representations in Robotics: Identifying Relevant Factors of Variation using Weak Supervision

Constantinos Chamzas*[1], Martina Lippi*[2], Michael C. Welle*[3], Anastasiia Varava[3], Alessandro Marino[4], Lydia E. Kavraki[1] and Danica Kragic[3]

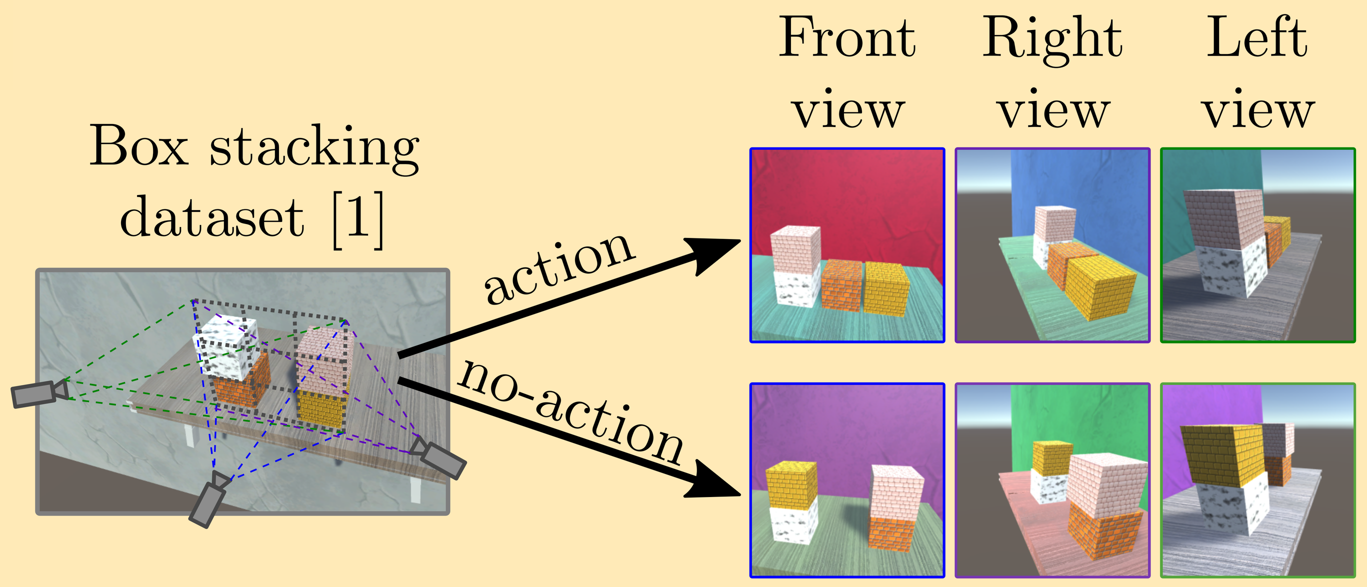
NEURAL INFORMATION PROCESSING SYSTEMS
3rd Robot Learning Workshop

1. Introduction

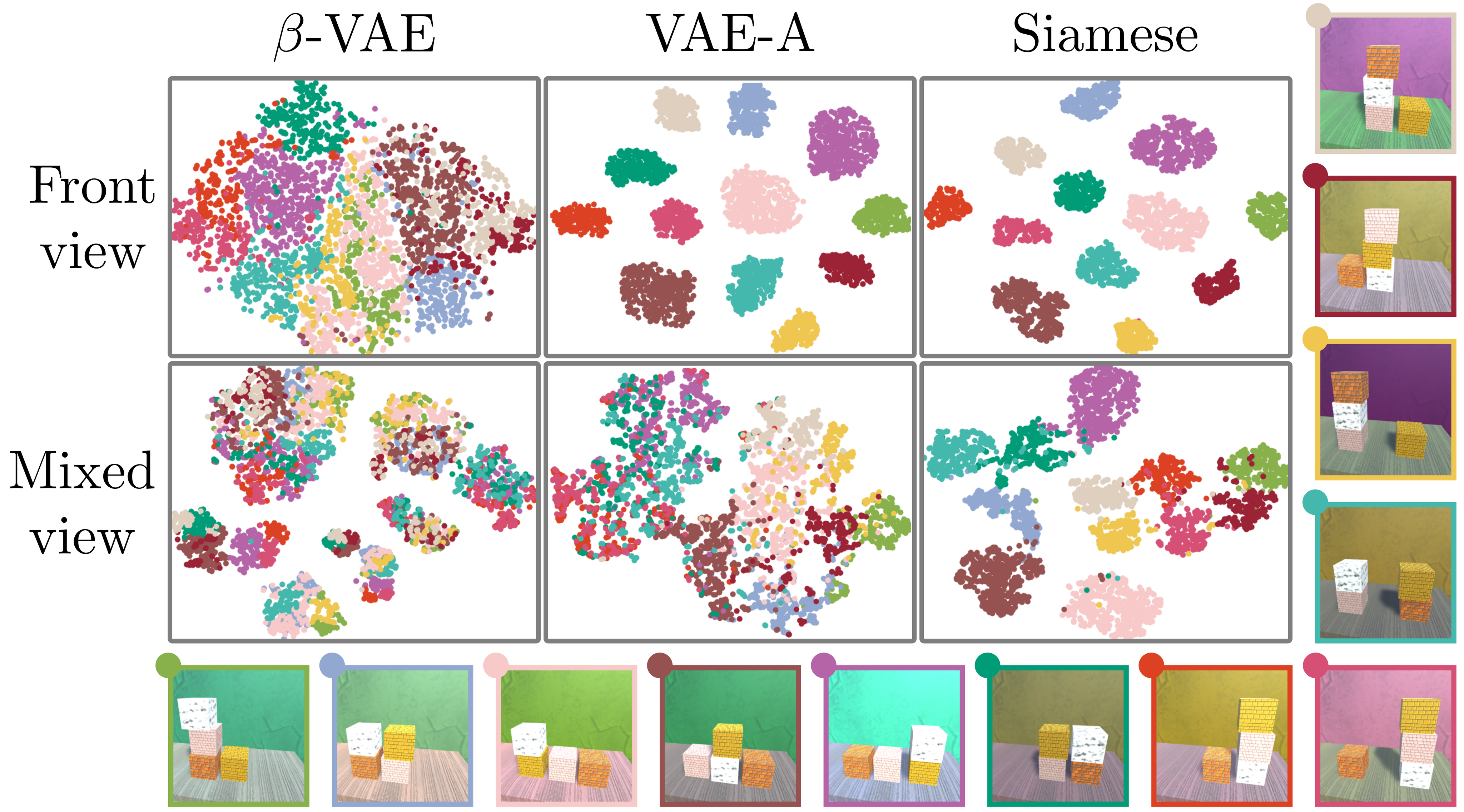
- State representation learning aims to infer **low-dimensional** features characterizing states of a dynamical system from high-dimensional data (such as images)
- Learn these features in an unsupervised manner is gaining increasing attention in robotics and reinforcement learning and, for this, latent space models, such as (variational) autoencoders (VAE), are commonly used
- Despite some successful applications, such approaches optimize the reconstruction loss and therefore are based on the following **assumption:** *Similar observations that are close in the image space correspond to similar states of the system*
- We argue that in general, very different images can correspond to the same underlying state when **task-irrelevant** factors of variation are present in the observations. Examples of these factors are changes in the lighting conditions, color changes of the background and camera position

2. Problem Description and Datasets

- A box stacking task in simulation is used as an example for a robotic manipulation task where the latent representations are built from images
- A dataset $\mathcal{D}$ consists of triplets $(I_1,\ I_2,\ a)$ containing an image I_1, its successor I_2 and a binary variable a denoting if an action took place ($a=1$) or not ($a=0$)
- Three viewpoints are used to acquire the images of the scene, based on which four datasets are defined: in $\mathcal{D}_f$, $\mathcal{D}_r$, $\mathcal{D}_l$ the images I_1, I_2 are taken from viewpoints "front", "right" and "left" respectively, in $\mathcal{D}_m$ the viewpoints are mixed



- Boxes are interchangeable
- Wall-color, table-color and Ilumination are randomized



3. Models

- **β-VAE:** (Reconstruction loss + KL divergence)

$$\mathcal{L}_{vae}(x) = E_{z\sim q(z|x)}[\log p(x|z)] + \beta \cdot D_{KL}\left(q(z|x)||p(z)\right)$$

- **VAE-A:** (β-VAE loss + contrastive loss)

$$\mathcal{L}_{vae-a}(x_i,x_j,a) = \frac{1}{2}(\mathcal{L}_{vae}(x_i) + \mathcal{L}_{vae}(x_j)) + \gamma\mathcal{L}_{action}(x_i,x_j,a)$$

$$\mathcal{L}_{action}(x_i,x_j,a) = \begin{cases} \max(0, d_m - ||z_i - z_j||_1) & \text{if } a=1 \\ ||z_i - z_j||_1 & \text{if } a=0 \end{cases}$$

- **Siamese:** (Contrastive loss)

$$\mathcal{L}_{margin}(x_i,x_j,a) = \frac{1}{2}\begin{cases} \max(0, m - ||(z_i - z_j||)^2 & \text{if } a=1 \\ ||(z_i - z_j||^2 & \text{if } a=0 \end{cases}$$

4. Results

Models	Dataset $\mathcal{D}_f$						Dataset $\mathcal{D}_m$					
	Clust. num.	Clust. hom.	Edges num.	Edge corr.	Paths scores % all	Paths scores % any	Clust. num.	Clust. hom.	Edges num.	Edge corr.	Paths scores % all	Paths scores % any
β-VAE	715	0.98	617	0.95	8.05	8.3	757	**0.95**	644	**0.93**	1.0	1.0
VAE-A	**12**	**1.0**	**24**	**1.0**	**100.0**	**100.0**	87	0.86	91	0.85	17.89	20.3
Siamese	**12**	**1.0**	25	0.96	94.21	98.1	**25**	**0.95**	**64**	0.88	**49.47**	**52.6**
Oracle	12	1.0	24	1.0	100.0	100.0	12	1.0	24	1.0	100.0	100.0

5. Discussion

- $\mathcal{D}_f$ fulfills the assumption that changes in the observations significantly correlate with changes in the system states
- β-VAE generates a fragmented latent space for both datasets (high number of clusters), achieving poor planning performance [1]
- VAE-A achieves perfect performance with dataset $\mathcal{D}_f$, as reconstruction loss is beneficial in this case, while a drop is recorded with $\mathcal{D}_m$
- Siamese achieves very good results with $\mathcal{D}_f$ and best performance with $\mathcal{D}_m$
- We conclude that:

  (i) performance improvement is achieved with weak supervision

  (ii) VAE-based encodings may be poorly structured when task-irrelevant factors of variation are present in the dataset

*Contributed equally, listed alphabeticaly

[1] Rice University, USA
[2] Roma Tre, Italy
[3] KTH Royal Institute of Technology, Sweden
[4] University of Cassino and Southern Lazio, Italy

chamzask@rice.edu, mlippi@unisa.it, mwelle@kth.se
varava@kth.se, al.marino@unicas.it
kavraki@rice.edu,dani@kth.se

[1] Lippi, M., Poklukar, P., Welle, M. C., Varava, A., Yin, H., Marino, A., & Kragic, D. (2020). Latent Space Roadmap for Visual Action Planning of Deformable and Rigid Object Manipulation. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems 2020 (IROS2020)