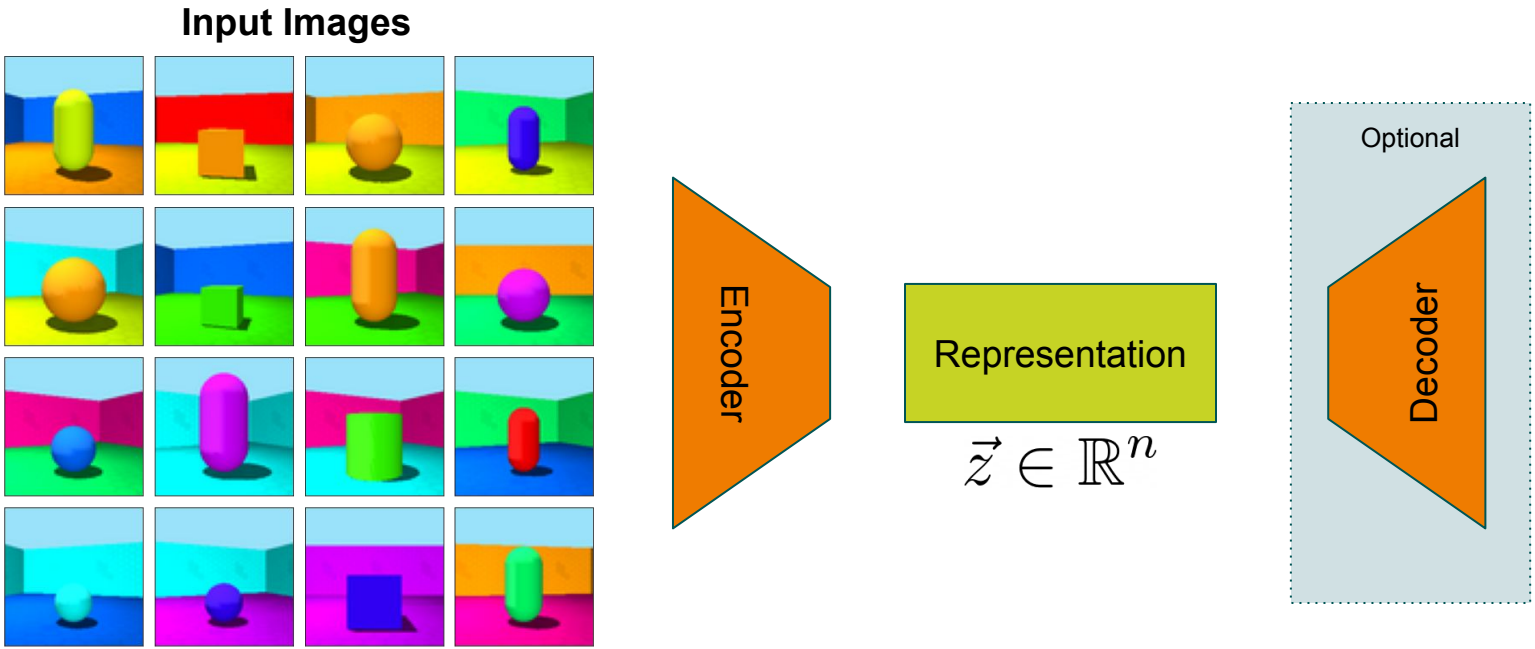
Demystifying Inductive Biases for **(β-)VAE** Based Architectures

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Dominik Zietlow, Michal Rolínek, Georg Martius

Unsupervised Disentangled Representation Learning: The Task

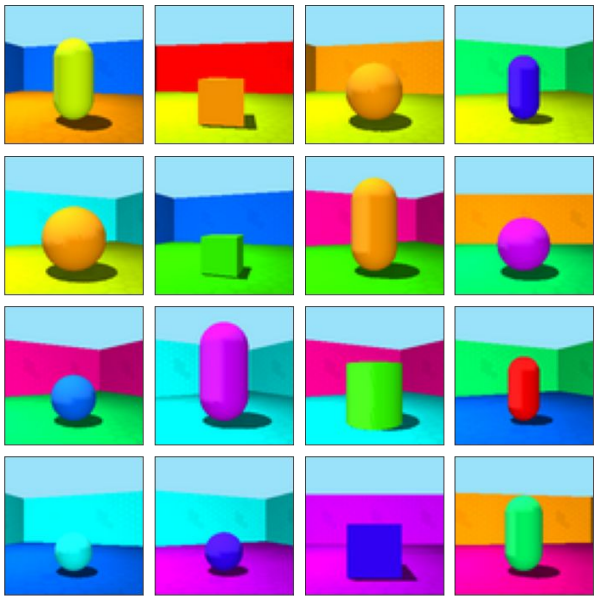


Each **image is fully described** by: floor color, wall color, object color, scale, shape, orientation

Burgess et al., 3D Shapes Dataset, 2018, https://github.com/deepmind/3dshapes-dataset/

Unsupervised Disentangled Representation Learning: The Task

Input Images



Encoder

Representation

$\vec{z} \in \mathbb{R}^n$

Goal

z_0 : floor color

z_1 : wall color

z_2 : object color

z_3 : scale

z_4 : shape

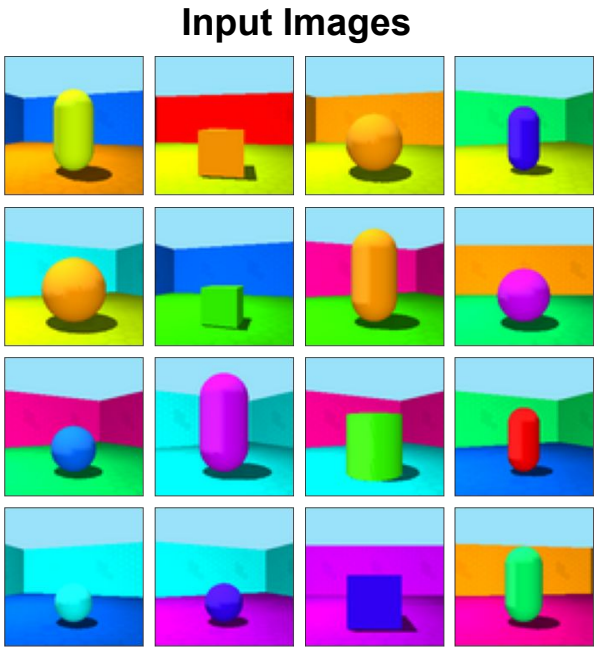
z_5 : orientation

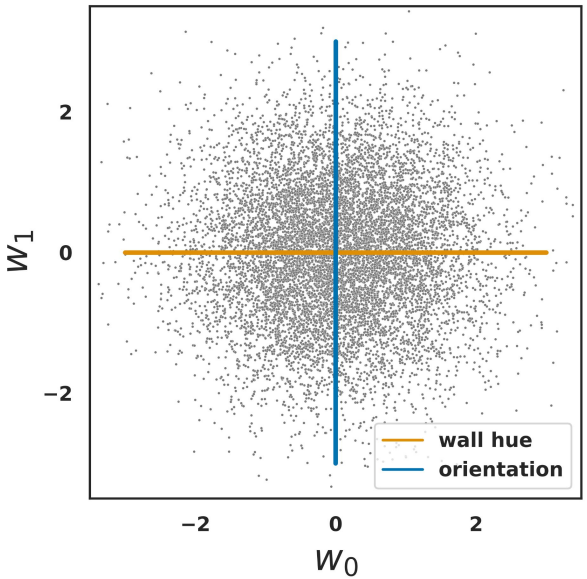
(up to permutations)

Each **image is fully described** by: floor color, wall color, object color, scale, shape, orientation

Burgess et al., 3D Shapes Dataset, 2018, https://github.com/deepmind/3dshapes-dataset/

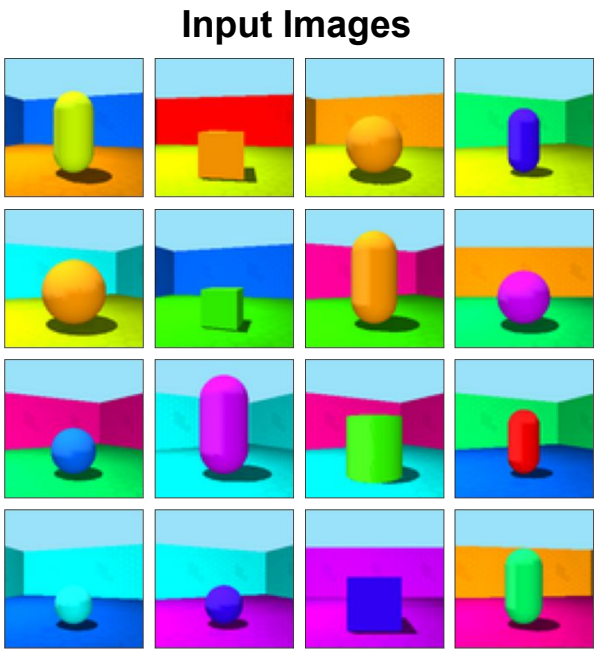
Unsupervised Disentangled Representation Learning: An Ill-posed Task

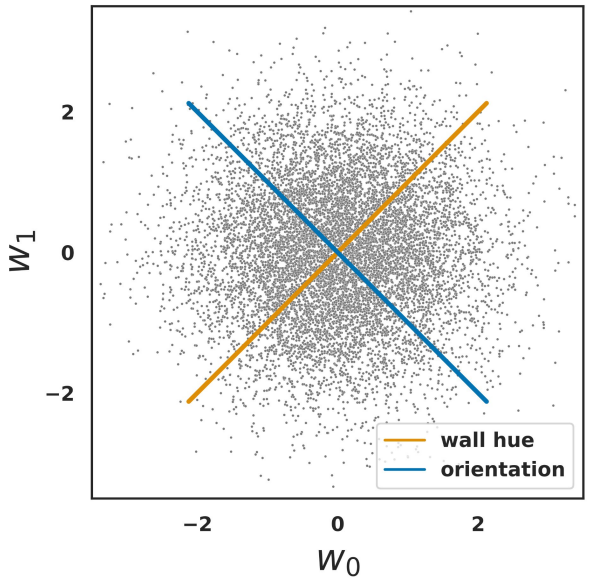




Burgess et al., 3D Shapes Dataset, 2018, https://github.com/deepmind/3dshapes-dataset/

Unsupervised Disentangled Representation Learning: An Ill-posed Task





Locatello et al., **Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations**, 2019, ICML

Burgess et al., 3D Shapes Dataset, 2018, https://github.com/deepmind/3dshapes-dataset/

Unsupervised Disentangled Representation Learning: An Ill-posed Task

	Shapes3d				
	AE	**β-VAE**	**Factor-VAE**	**β-TC-VAE**	**Slow-VAE**
MIG	0.06 ± 0.03	0.60 ± 0.31	0.27 ± 0.18	0.58 ± 0.20	0.53 ± 0.19

β-VAE: Higgins et al., β-VAE: Learning basic visual concepts with a constrained variational framework, 2017, **ICLR**

FactorVAE: Kim el al., Disentangling by factorising, 2018 **ICML**

β-TC-VAE: Chen et al., Isolating sources of disentanglement in variational autoencoders, 2018, **NeurIPS**

SlowVAE: Klindt et al., Towards non-linear disentanglement in natural data with temporal sparse coding, 2021, **ICLR**

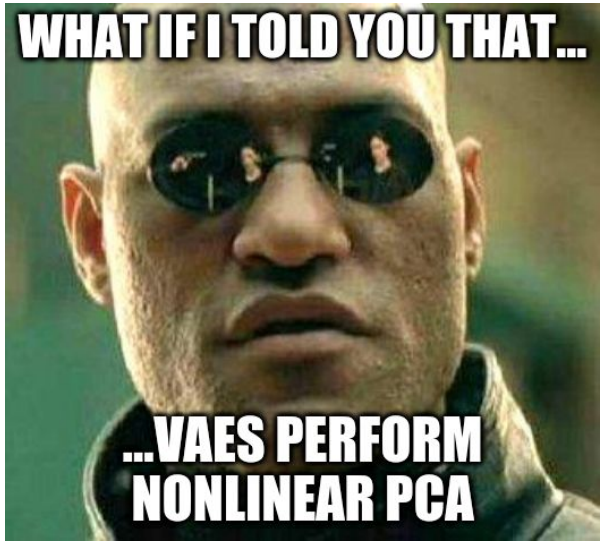
Why does it work?

What's the **inductive bias** in β-VAE based architectures and **datasets**?

VAE based methods have similarities to **PCA**!

- Rolinek, Zietlow, Martius: **Variational autoencoders pursue pca directions (by accident)**, 2019, CVPR

- Lucas et al.: **Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse**, 2019, NeurIPS



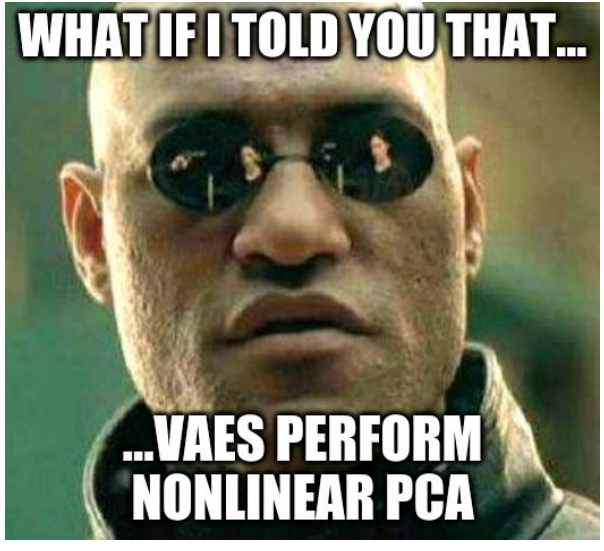
Why does it work?

What's the **inductive bias** in β-VAE based architectures and **datasets**?

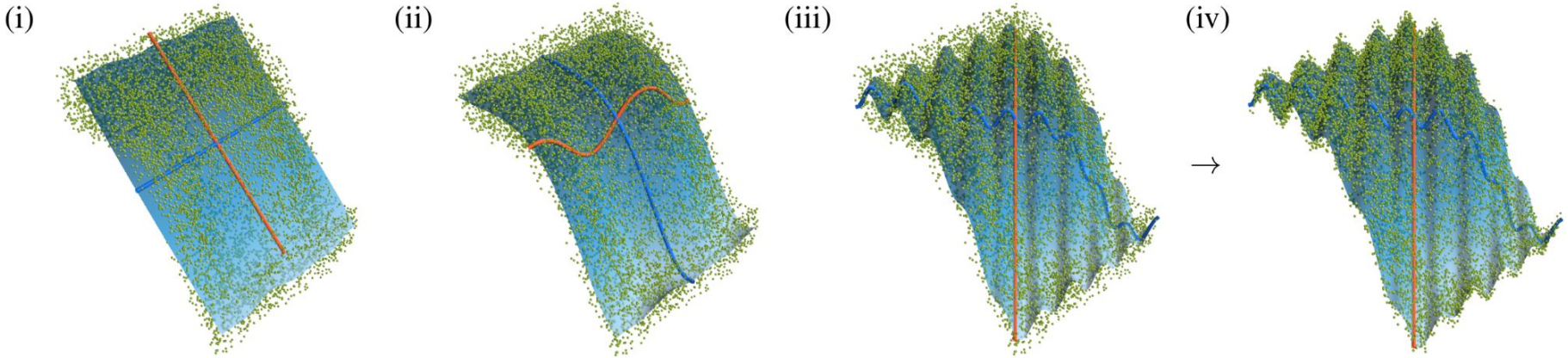
VAE based methods have similarities to **PCA**!

- Rolinek, Zietlow, Martius: **Variational autoencoders pursue pca directions (by accident)**, 2019, CVPR
- Lucas et al.: **Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse**, 2019, NeurIPS

What is the inductive bias in the data that aligns the non-linear PCA directions with the generating factors?

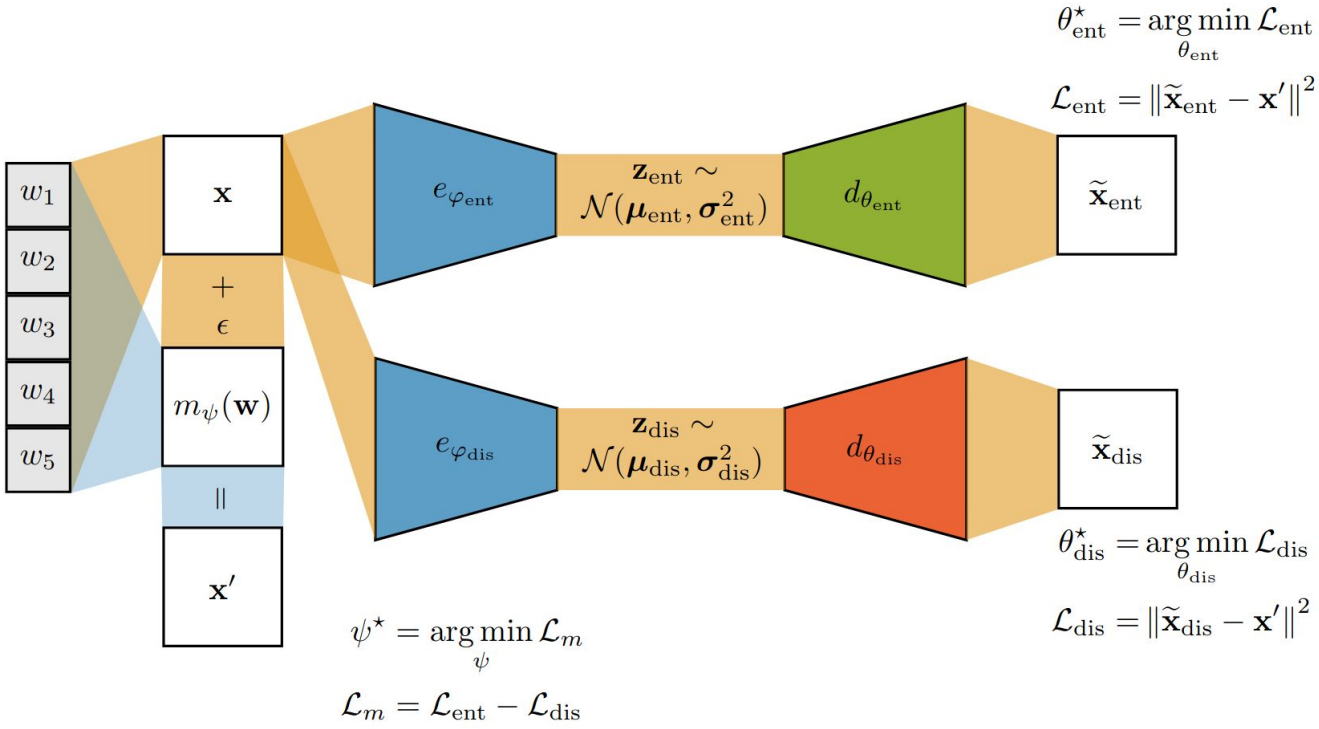


Can we alter the inductive bias in the data?

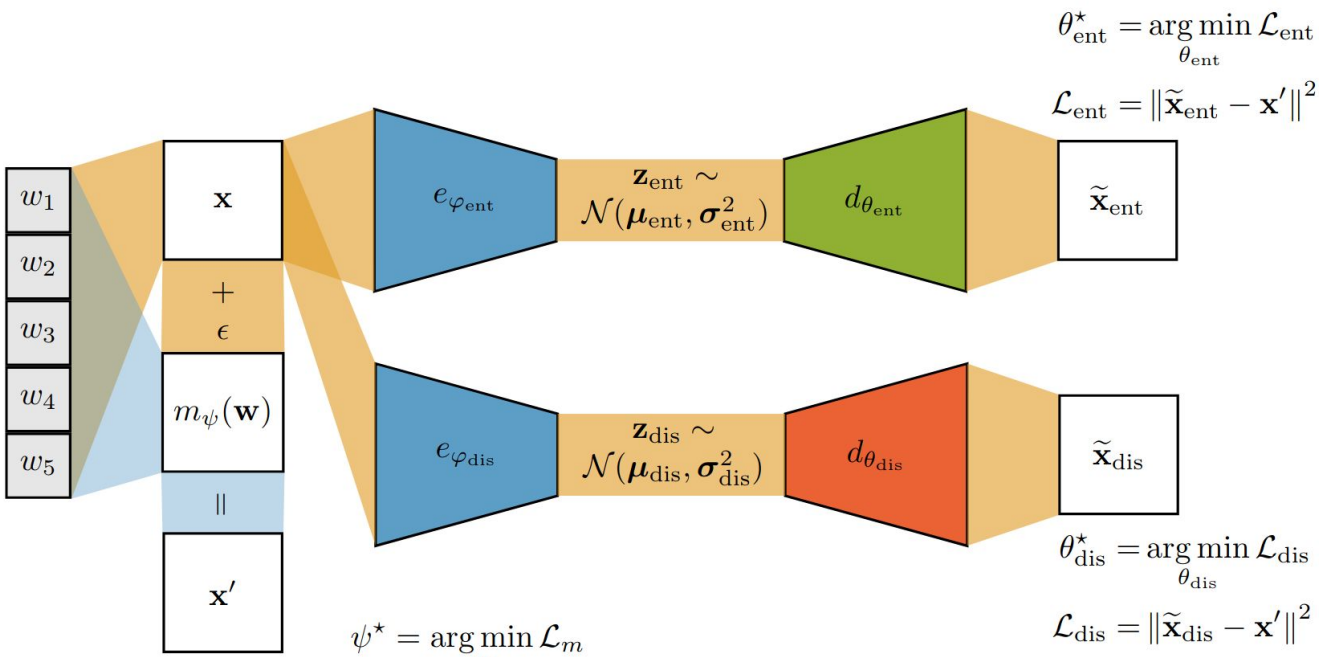


PCA **VAE** **Entangled Representation** **Adjusted Data**

Model Based Dataset Manipulation



Model Based Dataset Manipulation



$$\theta^\star_{\text{ent}} = \arg\min_{\theta_{\text{ent}}} \mathcal{L}_{\text{ent}}$$

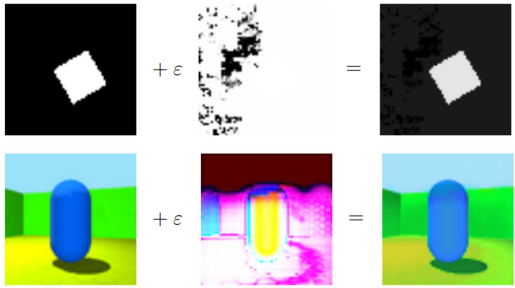
$$\mathcal{L}_{\text{ent}} = \|\widetilde{\mathbf{x}}_{\text{ent}} - \mathbf{x}'\|^2$$

$$\psi^\star = \arg\min_{\psi} \mathcal{L}_m$$

$$\mathcal{L}_m = \mathcal{L}_{\text{ent}} - \mathcal{L}_{\text{dis}}$$

$$\theta^\star_{\text{dis}} = \arg\min_{\theta_{\text{dis}}} \mathcal{L}_{\text{dis}}$$

$$\mathcal{L}_{\text{dis}} = \|\widetilde{\mathbf{x}}_{\text{dis}} - \mathbf{x}'\|^2$$



Small Data Perturbations Can Disrupt VAE Based Methods

	dSprites			Shapes3d		
	orig.	mod.	noise	orig.	mod.	noise
AE	0.09 ± 0.06	0.05 ± 0.02	0.06 ± 0.03	0.06 ± 0.03	0.05 ± 0.03	0.07 ± 0.03
β-VAE	0.23 ± 0.08	0.07 ± 0.09	0.14 ± 0.07	0.60 ± 0.31	0.09 ± 0.14	0.66 ± 0.05
Fac. VAE	0.27 ± 0.11	0.20 ± 0.12	0.16 ± 0.08	0.27 ± 0.18	0.07 ± 0.05	0.33 ± 0.20
TC-β-VAE	0.25 ± 0.08	0.14 ± 0.10	0.20 ± 0.04	0.58 ± 0.20	0.24 ± 0.16	0.60 ± 0.11
Slow-VAE	0.39 ± 0.08	0.27 ± 0.08	0.37 ± 0.09	0.53 ± 0.19	0.13 ± 0.08	0.60 ± 0.10
PCL	0.21 ± 0.03	0.24 ± 0.07	0.24 ± 0.07	0.44 ± 0.06	0.47 ± 0.08	0.40 ± 0.07
Weak sup. GAN	0.45 ± 0.05	0.36 ± 0.02	0.36 ± 0.01	0.69 ± 0.12	0.66 ± 0.12	0.77 ± 0.13

- Performance of VAE-based methods drop
- Non variational methods are robust against the dataset adjustments

Small Data Perturbations Can Disrupt VAE Based Methods

	dSprites			Shapes3d		
	orig.	mod.	noise	orig.	mod.	noise
AE	0.09 ± 0.06	0.05 ± 0.02	0.06 ± 0.03	0.06 ± 0.03	0.05 ± 0.03	0.07 ± 0.03
β-VAE	0.23 ± 0.08	0.07 ± 0.09	0.14 ± 0.07	0.60 ± 0.31	0.09 ± 0.14	0.66 ± 0.05
Fac. VAE	0.27 ± 0.11	0.20 ± 0.12	0.16 ± 0.08	0.27 ± 0.18	0.07 ± 0.05	0.33 ± 0.20
TC-β-VAE	0.25 ± 0.08	0.14 ± 0.10	0.20 ± 0.04	0.58 ± 0.20	0.24 ± 0.16	0.60 ± 0.11
Slow-VAE	0.39 ± 0.08	0.27 ± 0.08	0.37 ± 0.09	0.53 ± 0.19	0.13 ± 0.08	
PCL	0.21 ± 0.03	0.24 ± 0.07	0.24 ± 0.07	0.44 ± 0.06	0.47 ± 0.08	
Weak sup. GAN	0.45 ± 0.05	0.36 ± 0.02	0.36 ± 0.01	0.69 ± 0.12	0.66 ± 0.12	

- Performance of VAE-based methods drop
- Non variational methods are robust against the dataset adjustments
- Even identifiable architectures are relying heavily on the data bias