Visual Representation Learning Does Not Generalize Strongly Within the Same Domain

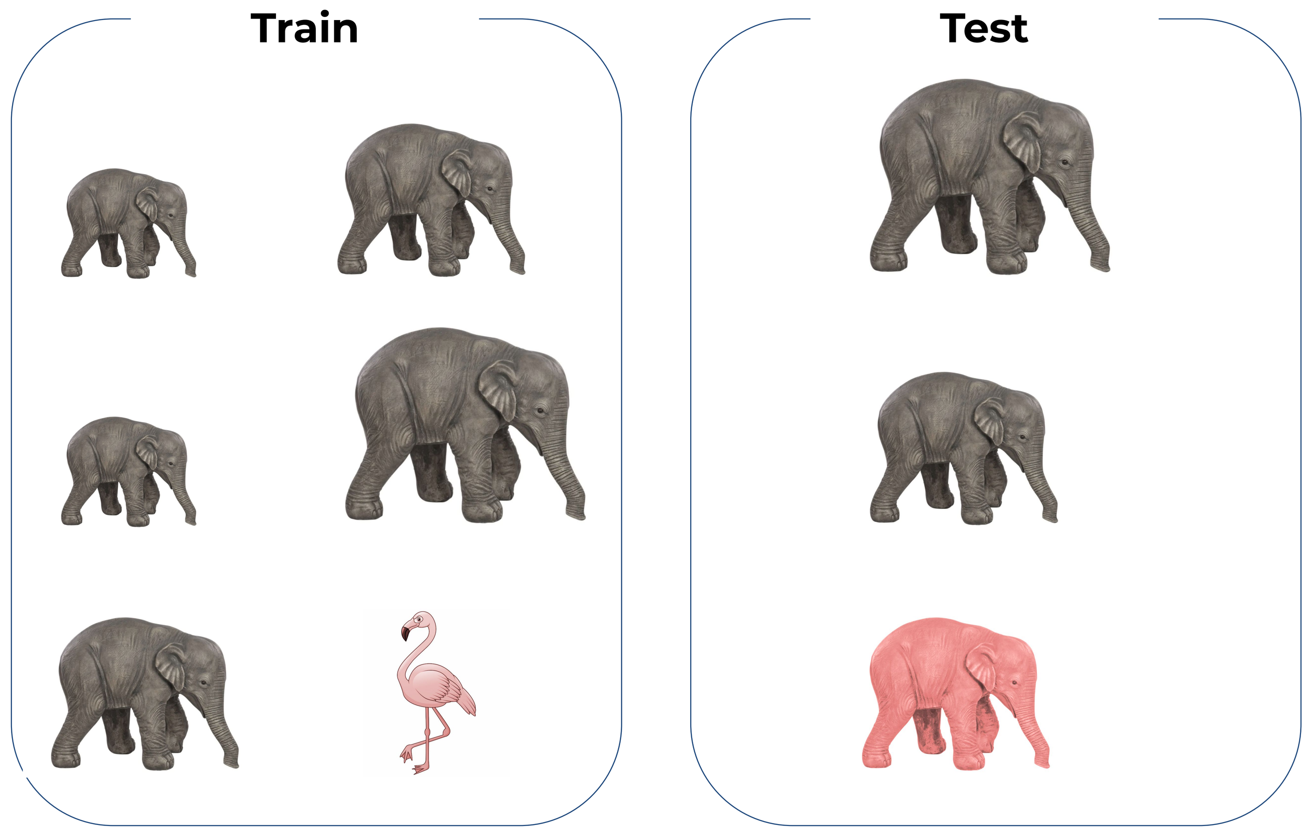
Lukas Schott[1,3], Julius von Kügelgen[2,3,4], Frederik Träuble[2,3], Peter Gehler[3], Chris Russell[3], Matthias Bethge[1,3], Bernhard Schölkopf[2,3], Francesco Locatello[3,‡], Wieland Brendel[1,‡]

(1) University of Tübingen, Germany (2) Max Planck Institute for Intelligent Systems (3) Amazon Web Services (4) University of Cambridge

‡ shared senior authorship

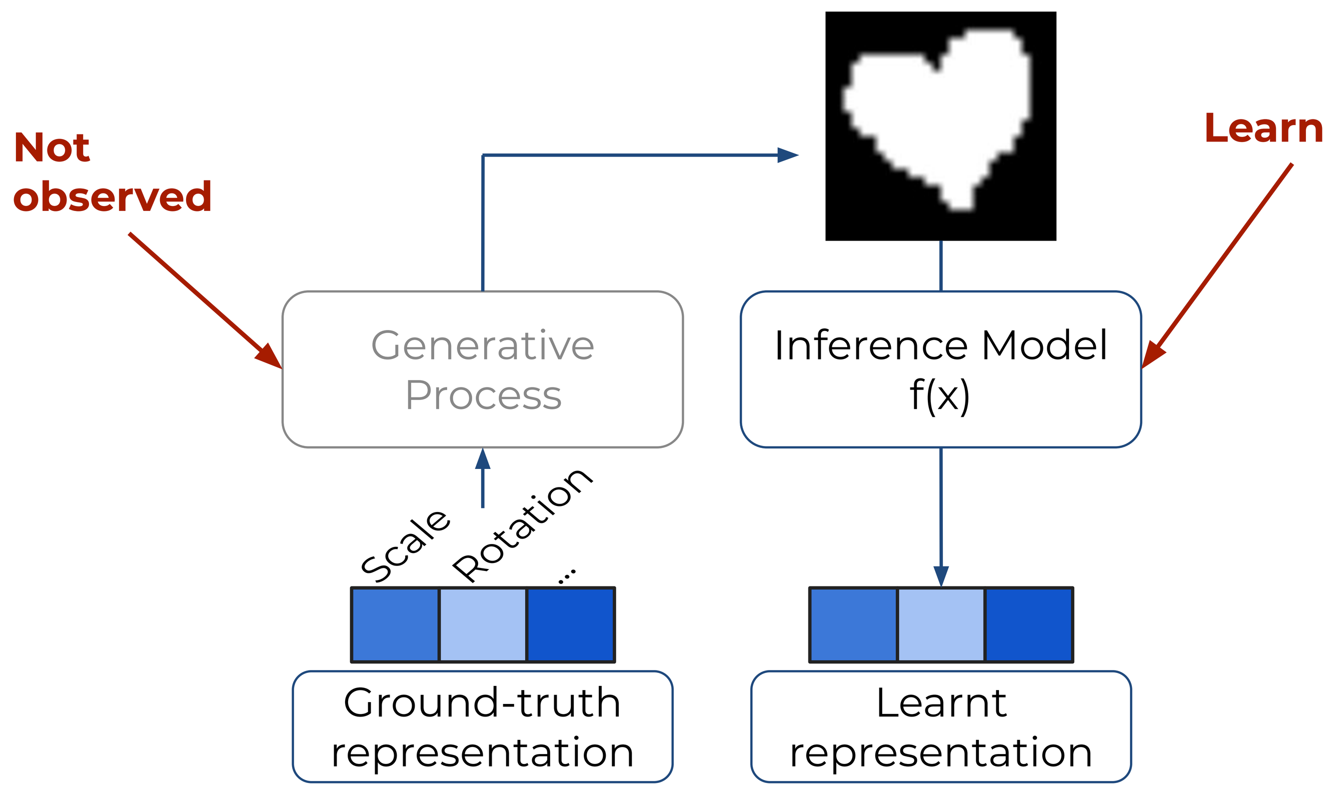
Motivation

Question: Can neural networks generalize factors of variation?
Do neural networks learn underlying mechanisms?

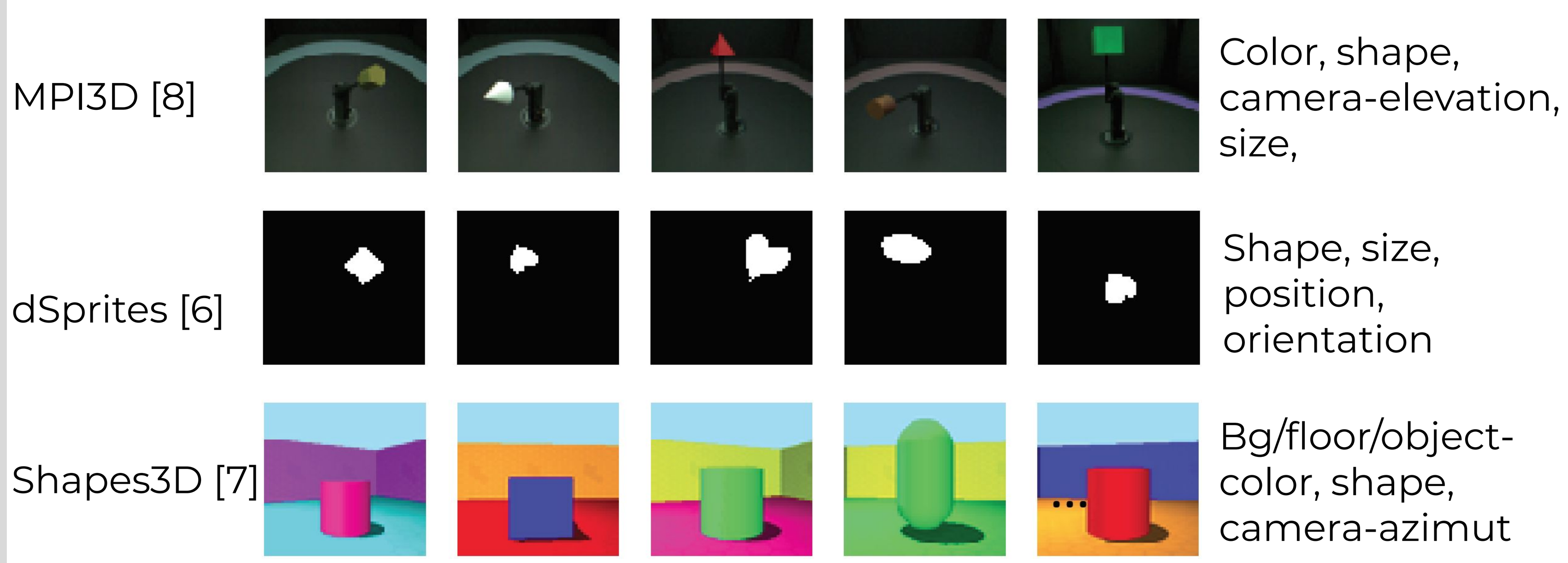


Goal: Test whether models learn the underlying mechanisms of a scene.

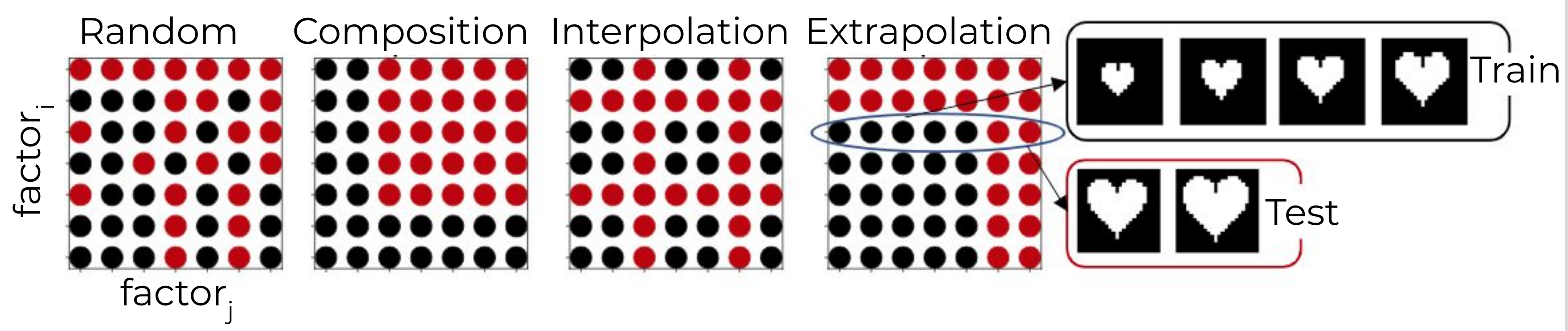
Problem Setting



Datasets



Setup



Factors of variation: Size, color, ...

Test-train-split: split all factors except shape: ~30%(train) : ~70%(test) samples

Training: Sample from the generative process on the training data in any way.

Un-/ weakly supervised: VAE, Ada-GVAE, SlowVAE, PCL
Fully Supervised: MLP, CNN, CoordConv, SetEncoder, Equivariant, ...
Transfer Learning: RN50 on IN-21k, RN101 on IN21-k, DenseNet on IN1-k

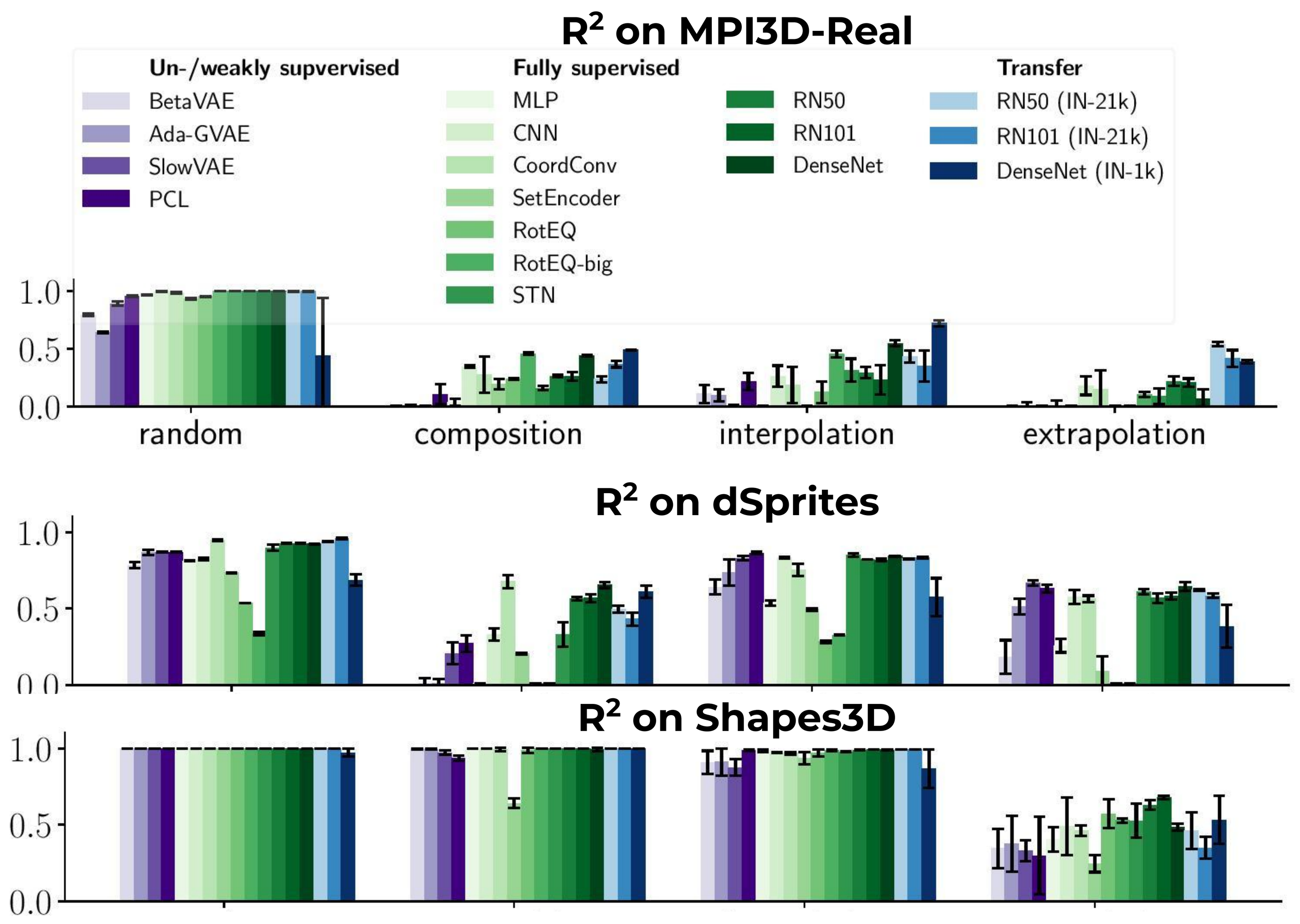
Evaluation

Regression of factors of variation:

$$R_i^2 = 1 - \frac{\text{MSE}_i}{\sigma_i^2} \qquad \text{MSE}_j = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{te}}\left[(\mathbf{y}_j - f_j(\mathbf{x}))^2\right]$$

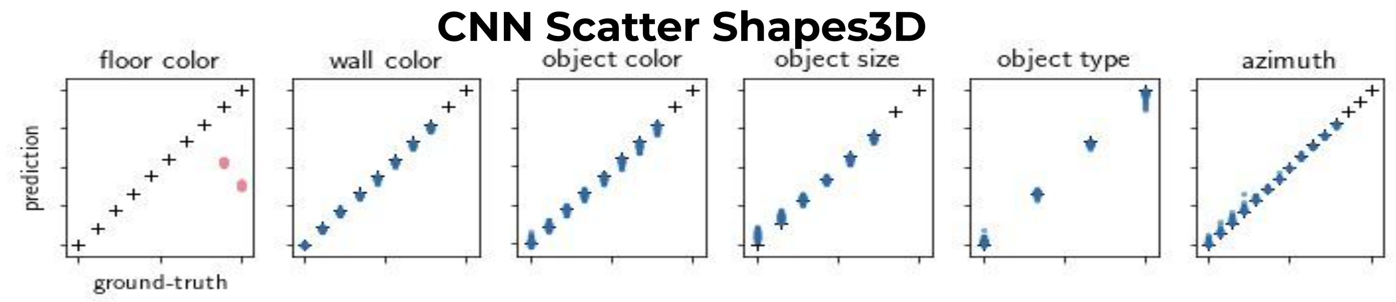
$R^2 = 1$ → perfect regression, 100% variance explained
$R^2 = 0$ → e.g. always predicting the mean
$R^2 < 0$ → worse than predicting the mean

Results - All Models



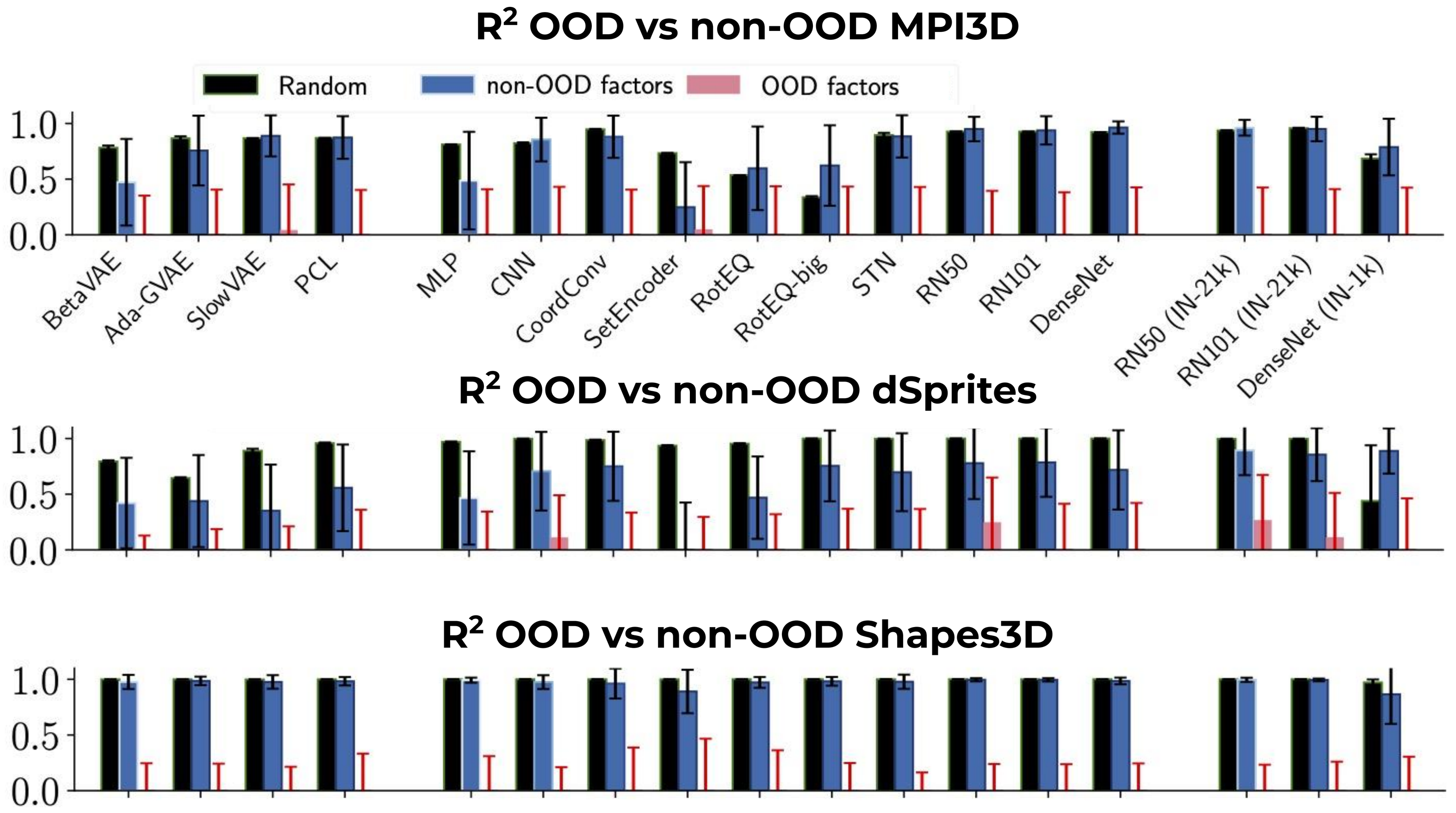
- On realistic dataset (MPI3D): Models do **not** learn underlying mechanism
- Shapes3D color composition/ interpolation works fairly well
- Extrapolation most difficult → investigate further

Extrapolation - Qualitative



→ OOD factors different behavior on non-OOD vs OOD factors
→ OOD factors tend towards the mean

Extrapolation - Error Analysis

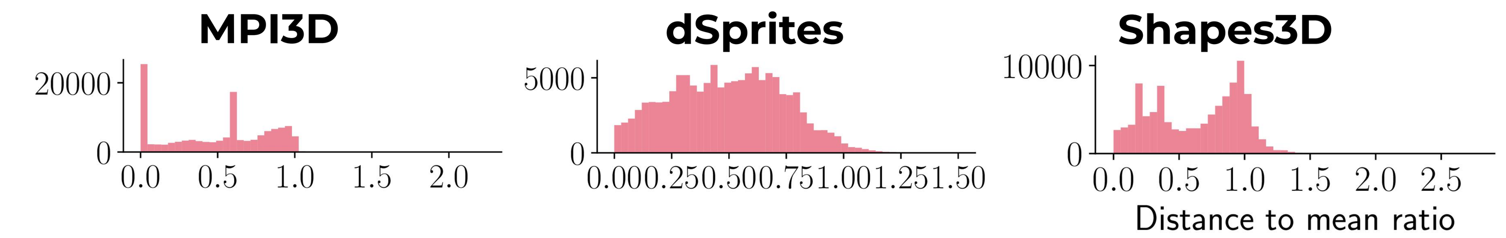


→ OOD factors R^2 low → Models do not extrapolate
→ non-OOD R^2 better → Models are surprisingly modular

Extrapolation Towards the Mean

$$\frac{|f(\mathbf{x}^i)_j - \bar{\mathbf{y}}_j|}{|\mathbf{y}_j^i - \bar{\mathbf{y}}_j|}$$

if in [0, 1] → prediction closer to mean than ground-truth
if > 1 → further away from mean



Values mostly in [0,1]→ Models tend to extrapolate towards the mean

Conclusions

- In more difficult settings, our tested ML models are unable to generalize and do not learn the underlying model.
- Models are surprisingly modular and tend to extrapolate towards the mean.
- On the non-artificial MPI3D we found transfer learning to be most helpful.

References

[1] Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, Jeffrey Bowers. The role of Disentanglement in Generalisation. ICLR2021
[2] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, Stefan Bauer. On Disentangled Representations Learned From Correlated Data. ArXiv, abs/2006.07886.
[3] Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, Bernhard Schölkopf. On the Transfer of Disentangled Representations in Realistic Settings. ICLR2021
[4]David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. ICML2018
[5] Xiaolin Wu, Xi Zhang, and Jun Du. Challenge of spatial cognition for deep learning. CoRR,624abs/1908.04396, 2019
[6] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017
[7] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. International Conference on Machine Learning, pages 2649–2658. PMLR, 2018.
[8] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. InAdvances in Neural Information Processing Systems, pages15714–15725, 2019.

lukas.schott@gmail.com