ON THE TRANSFER OF DISENTANGLED REPRESENTATIONS IN REALISTIC SETTINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning meaningful representations that disentangle the underlying structure of the data generating process is considered to be of key importance in machine learning. While disentangled representations were found to be useful for diverse tasks such as abstract reasoning and fair classification, their scalability and real-world impact remain questionable. We introduce a new high-resolution dataset with over 1M simulated images and 1k annotated real-world images of the same setup. In contrast to previous work, this new dataset exhibits correlations, a complex underlying structure, and allows to evaluate transfer to unseen simulated and real-world settings where the encoder i) remains in distribution or ii) is out of distribution. We propose new architectures in order to scale disentangled representation learning to realistic high-resolution settings and conduct a large-scale empirical study of disentangled representations on this dataset. We observe that disentanglement is a good predictor for out-of-distribution (OOD) task performance.

1 INTRODUCTION

Disentangled representations hold the promise of generalization to unseen scenarios (Higgins et al., 2017b), increased interpretability (Adel et al., 2018; Higgins et al., 2018) and allow for faster learning on downstream tasks (van Steenkiste et al., 2019; Locatello et al., 2019a). However, most of the focus in learning disentangled representations has been on small synthetic datasets whose ground truth factors exhibit perfect independence by design. More realistic settings remain largely unexplored. We hypothesize that this is because real-world scenarios present several challenges that have not been extensively studied to date. Important challenges are scaling (much higher resolution in observations and factors), occlusions, and correlation between factors. Consider, for instance, a robotic arm moving a cube: Here, the robot arm can occlude parts of the cube, and its end-effector position exhibits correlations with the cube's position and orientation. Another difficulty is that we typically have only limited access to ground truth labels in the real world, which requires robust frameworks for model selection when no or only weak labels are available.

The goal of this work is to provide a path towards disentanglement in realistic settings. First, we argue that this requires a new dataset that captures the challenges mentioned above. We propose a dataset consisting of simulated observations from a scene where a robotic arm interacts with a cube in a stage (see Fig. 1). This setting exhibits correlations and occlusions that are typical in real-world robotics. Second, we show how to scale the architecture of disentanglement methods to perform well on this dataset. Third, we extensively analyze the usefulness of disentangled representations in terms of out-of-distribution downstream generalization, both in terms of held-out factors of variations and sim2real transfer. In fact, our data set is based on the TriFinger robot from Wüthrich et al. (2020), which can be built to test the deployment of models in the real world. While the analysis in this paper focuses on the transfer and

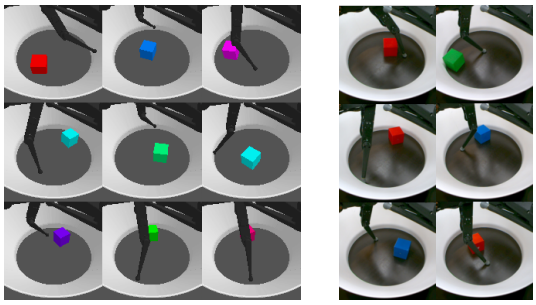


Figure 1: Left: Images from the simulated dataset. Right: Images from the real-world setup.

generalization of predictive models, we hope that our dataset may serve as a benchmark to explore the usefulness of disentangled representations in real-world control tasks.

The contributions of this paper can be summarized as follows:

- We propose a new dataset for disentangled representation learning, containing over 1M simulated high-resolution images from a robotic setup, with seven partly correlated factors of variation. Additionally, we provide a set of more than 1k annotated images from the corresponding real-world setup that can be used for challenging sim2real transfer tasks.
- We propose a new neural architecture to successfully scale VAE-based disentanglement learning approaches to complex datasets.
- We conduct a large-scale empirical study on generalization to various transfer scenarios on this challenging dataset. We train 1080 models from state-of-the-art disentanglement methods and and discover that disentanglement is a good predictor for out-of-distribution (OOD) task performance.

2 Related Work

Disentanglement methods. Most state-of-the-art disentangled representation learning approaches are based on the framework of variational autoencoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014). A (high-dimensional) observation $\boldsymbol{x}$ is assumed to be generated according to the latent variable model $p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ where the latent variables $\boldsymbol{z}$ have a fixed prior $p(\boldsymbol{z})$. The generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and the approximate posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ are typically parameterized by neural networks, which are optimized by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) \leq \log p(\boldsymbol{x}) \tag{1}$$

As the above objective does not enforce any structure on the latent space except for some similarity to $p(\boldsymbol{z})$, different regularization strategies have been proposed, along with evaluation metrics to gauge the disentanglement of the learned representations (Higgins et al., 2017a; Kim & Mnih, 2018; Burgess et al., 2018; Kumar et al., 2018; Chen et al., 2018; Eastwood & Williams, 2018). Recently, Locatello et al. (2019b, Theorem 1) showed that the purely unsupervised learning of disentangled representations is impossible. This limitation can be overcome without the need for explicitly labeled data by introducing weak labels (Locatello et al., 2020; Shu et al., 2019). Ideas related to disentangling the factors of variations date back to the non-linear ICA literature (Comon, 1994; Hyvärinen & Pajunen, 1999; Bach & Jordan, 2002; Jutten & Karhunen, 2003; Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019; Gresele et al., 2019). Recent work combines non-linear ICA with disentanglement (Khemakhem et al., 2020; Sorrenson et al., 2020; Klindt et al., 2020).

Evaluating disentangled representations. The *BetaVAE* (Higgins et al., 2017a) and *FactorVAE* (Kim & Mnih, 2018) scores measure disentanglement by performing an intervention on the factors of variation and predicting which factor was intervened on. The *Mutual Information Gap (MIG)* (Chen et al., 2018), *Modularity* (Ridgeway & Mozer, 2018), *DCI Disentanglement* (Eastwood & Williams, 2018) and *SAPscore* (Kumar et al., 2018) are based on matrices relating factors of variation and codes (e.g. pairwise mutual information, feature importance and predictability).

Datasets for disentanglement learning. *dSprites* (Higgins et al., 2017a), which consists of binary low-resolution 2D images of basic shapes, is one of the most commonly used synthetic datasets for disentanglement learning. *Color-dSprites*, *Noisy-dSprites*, and *Scream-dSprites* are slightly more challenging variants of dSprites. The *SmallNORB* dataset contains toy images rendered under different lighting conditions, elevations and azimuths (LeCun et al., 2004). *Cars3D* (Reed et al., 2015) exhibits different car models from Fidler et al. (2012) under different camera viewpoint. *3dshapes* is a popular dataset of simple shapes in a 3D scene (Kim & Mnih, 2018). Finally, Gondal et al. (2019) proposed *MPI3D*, containing images of physical 3D objects with seven factors of variation, such as object color, shape, size and position available in a simulated, simulated and highly realistic rendered simulated variant. Except MPI3D which has over 1M images, the size of the other datasets is limited with only $17,568$ to $737,280$ images. All of the above datasets exhibit perfect independence of all factors by design, the number of possible states is on the order of 1M or less and their context is embedded within a static setting that does not allow for any dynamic downstream task

such as reinforcement learning. In addition, except for SmallNORB, the image resolution is limited to 64x64 and there are no occlusions or correlations.

Other related work. Transfer of learned disentangled representations from simulation to the real world has been recently investigated by Gondal et al. (2019) on the MPI3D dataset, and previously by Higgins et al. (2017b) in the context of reinforcement learning. Locatello et al. (2020) probed the out-of-distribution generalization of downstream tasks trained on disentangled representations. However, these representations are trained on the entire dataset. Generalization and transfer performance especially for representation learning has likewise been studied in Dayan (1993); Muandet et al. (2013); Heinze-Deml & Meinshausen (2017); Rojas-Carulla et al. (2018); Suter et al. (2019); Li et al. (2018); Arjovsky et al. (2019); Krueger et al. (2020); Gowal et al. (2020). Moreover, sim2real transfer is of major interest in the robotic learning community, because of limited data and supervision in the real world (Tobin et al., 2017; Rusu et al., 2017; Peng et al., 2018; James et al., 2019; Yan et al., 2020; Andrychowicz et al., 2020).

3 Scaling Disentangled Representations to Complex Scenarios

A new challenging dataset. Simulated images in our dataset are derived from the trifinger robot platform introduced by Wüthrich et al. (2020). The motivation for choosing this setting is that (1) it is challenging due to occlusions, correlations, and other difficulties encountered in robotic settings, (2) it requires modeling of fine details such as tip links at high resolutions, and (3) it corresponds to a robotic setup, so that learned representations can be used for control and reinforcement learning in simulation and in the real world. The scene comprises a robot finger with three joints that can be controlled to manipulate a cube in a bowl-shaped stage. See Fig. 1 for examples of scenes from our dataset. The data is generated from 7 different factors of variation (FoV) listed in Table 1. Unlike in previous datasets, not all FoV are independent: The end-effector (the tip of the finger) can collide with the floor or cube, resulting in infeasible combinations of the joint angles. We argue that such correlations are a key feature in real-world data not present in current datasets. The high FoV resolution results in approximately 1.52 billion feasible states, but the dataset itself only contains 1 million realistically rendered samples from them (approximately 0.065% of all possible FoV combinations) with 128×128 resolution. Additionally, we recorded an annotated dataset under the same conditions in the real-world setup: we acquired camera images from the same viewpoint and recorded the labels of the seven underlying factors of variation. This dataset can be used for out-of-distribution evaluations, few-shot learning, and testing other sim2real aspects.

FoV	values
upper joint	30 values in $[-0.65, +0.65]$
middle joint	30 values in $[-0.5, +0.5]$
lower joint	30 values in $[-0.8, +0.8]$
cube position x	30 values in $[-0.11, +0.11]$
cube position y	30 values in $[-0.11, +0.11]$
cube rotation	10 values in $[0°, 81°]$
cube color hue	12 values in $[0°, 330°]$

Table 1: Factors of Variation in the new dataset. Values are linearly spaced in the specified intervals. The training set for the VAEs contains 8 randomly chosen color hues. Joint angles are in radians, cube positions in meters.

Model architecture. When scaling disentangled representation learning to more complex datasets, such as the one proposed here, one of the main bottlenecks in current VAE-based approaches is the flexibility of the encoder and decoder networks. In particular, using the architecture from Locatello et al. (2019b), none of the models we trained correctly captured all factors of variation or yielded high-quality reconstructions. While the increased image resolution already presents a challenge, the main practical issue in our new dataset is the level of detail that needs to be modeled. In particular, we identified the cube rotation and the lower joint position to be the factors of variation that were the hardest to capture. This is likely due to minimal pixel changes in images, corresponding to minimal changes in reconstruction error.

To overcome these issues, we propose neural architectures that are deeper and wider than those commonly used in the disentangled representation learning literature, where the encoder and decoder typically have 4 convolutional and 2 fully-connected layers. The encoder consists of a convolu-

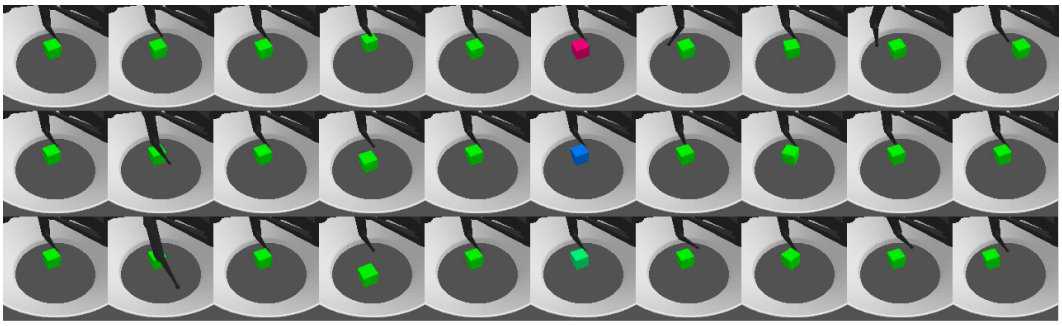


Figure 2: Latent traversals of a trained model that perfectly disentangles the dataset's FoV. In each column, all latent variables but one are fixed.

tional layer, 10 convolutional residual blocks, and 2 fully connected layers. Some residual blocks are followed by 1x1 convolutions that change the number of channels, or by average pooling that downsamples the tensors by a factor of 2 along the spatial dimensions. Each residual block consists of two 3x3 convolutions with a leaky ReLU nonlinearity, and a learnable scalar gating mechanism first introduced by Bachlechner et al. (2020). Overall, the encoder has 23 convolutional layers and 2 fully connected layers. The decoder loosely mirrors this architecture, with average pooling replaced by bilinear interpolation for upsampling. See Appendix A for further implementation details.

Experimental setup. We perform a large-scale empirical study on the simulated dataset introduced above, by training 1080 β-VAE models.[1] For further experimental details we refer the reader to Appendix A. The hyperparameter sweep is defined as follows:

- We train the models using either unsupervised learning or weakly supervised learning (Locatello et al., 2020). In the weakly supervised case, a model is trained with pairs of images that differ in k FoVs. Here we fix $k = 1$ as it was proven to lead to higher disentanglement. The dataset therefore consists of 500k pairs of images that differ in only one FoV.
- We vary the parameter β in $\{1, 2, 4\}$, and use linear deterministic warm-up (Bowman et al., 2015; Sønderby et al., 2016) over the first $\{0, 10000, 50000\}$ training steps.
- The latent space dimensionality is in $\{10, 25, 50\}$.
- Half of the models are trained with additive noise in the input image. This choice is motivated by the fact that adding noise to the input of neural networks is typically beneficial for out-of-distribution generalization (Sietsma & Dow, 1991; Bishop, 1995).
- Each of the 108 resulting configurations is trained with 10 random seeds.

Can we scale up disentanglement learning? Most of the trained VAEs in our empirical study fully capture all the elements of a scene, correctly model heavy occlusions, and generate detailed, high-quality samples and reconstructions (see Appendix B.4). From visual inspections such as the latent traversals in Fig. 2, we observe that many trained models fully disentangle the ground-truth factors of variation. This, however, appears to only be possible in the weakly supervised scenario. The fact that models trained without supervision learn entangled representations is in line with the impossibility result for the unsupervised learning of disentangled representations from Locatello et al. (2019b). Latent traversals from a selection of models with different degrees of disentanglement are presented in Appendix B.1. Interestingly, the high-disentanglement models seem to correct for correlations and interpolate infeasible states, i.e. the fingertip traverses through the cube or floor.

Summary: Our proposed model architecture can scale disentanglement learning to more realistic settings, but a form of weak supervision is necessary to consistently achieve high disentanglement.

How useful are commonly used disentanglement metrics in realistic scenarios? In the violin plot in Fig. 3 (left) we see that DCI and MIG measure high disentanglement under weak supervision and lower disentanglement in the unsupervised setting. This is consistent with our qualitative conclusion from visual inspection of the models, and with the aforementioned impossibility result.

[1] Reproducing these experiments requires approximately 2.8 GPU years (NVIDIA Tesla V100 PCIe 16 GB).

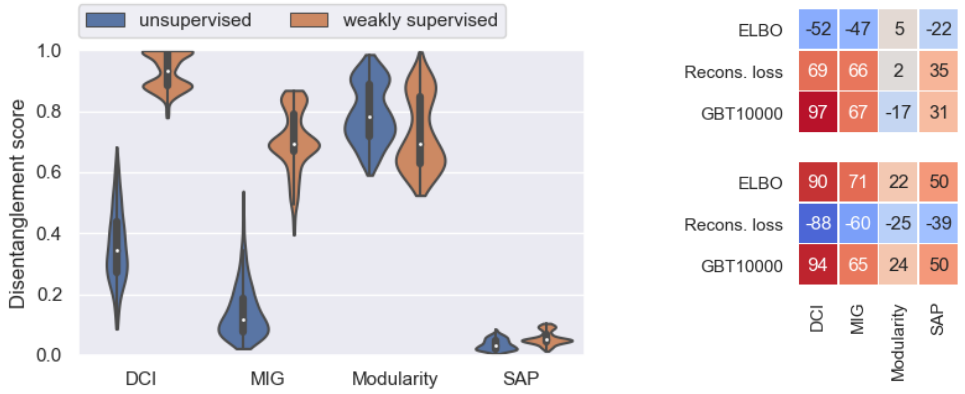


Figure 3: Left: Disentanglement metrics aggregating all hyperparameters except for supervision type. Right: Rank correlations (Spearman) of ELBO, Reconstruction loss and the test validation error of a GBT classifier trained on 10000 labelled data points with disentanglement metrics. The upper rank correlations correspond to the unsupervised models and the lower to the weakly supervised models.

Many of the models trained with weak supervision exhibit a very high DCI score (29% of them have $>$99% DCI, some of them up to 99.89%). SAP and Modularity appear to be ineffective at capturing disentanglement in this setting, a similar finding was also made by Locatello et al. (2019b). Finally, note that the BetaVAE and FactorVAE metrics are not straightforward to be evaluated on datasets that do not contain all possible combinations of factor values. According to Fig. 3 (right) there is a strong correlation of DCI and MIG with test accuracy of GBT classifiers predicting the FoVs. In the weakly supervised setting, these metrics are strongly correlated with the ELBO (positively) and with the reconstruction loss (negatively). Such correlations were also observed by Locatello et al. (2020) on significantly less complex datasets, and can be exploited for unsupervised model selection: these unsupervised metrics can be used as proxies for disentanglement metrics, which would require fully labeled data.

Summary: DCI and MIG appear to be useful disentanglement metrics in realistic scenarios, whereas other metrics seem to fall short at capturing disentanglement or can be difficult to compute. When using weak supervision, we can select disentangled models with unsupervised metrics.

4 Framework for the Evaluation of OOD Generalization

Previous work has focused on evaluating the usefulness of disentangled representations for various downstream tasks, such as predicting ground truth factors of variation, fair classification, and abstract reasoning. Here we propose a new framework for evaluating the out-of-distribution (OOD) generalization properties of representations. More specifically, we consider a downstream task – such as regression of ground truth factors – trained on a learned representation of the data, and evaluate the performance on a held-out test set. While the test set typically follows the same distribution as the training set (in-distribution generalization), we also consider test sets that follow a different distribution (out-of-distribution generalization). Our goal is to investigate to what extent, if at all, downstream tasks trained on disentangled representations exhibit a higher degree of OOD generalization than those trained on entangled representations.

Let D denote the training set for disentangled representation learning. To investigate OOD transfer, we consider two disjoint sets D_1 and D_2 such that the values of a specific factor of variation in D_1 are not observed in D_2, and vice versa. For example, if we were to split according to object color, we could define these datasets such that images in D_1 only contain a red object, and those in D_2 only contain an object of another set of colors that does not include red. In the following, we will always assume $D_1 \subset D$, i.e. the downstream task is trained on a subset of the training set for representation learning. We then differentiate between two scenarios: (1) the OOD test set is a

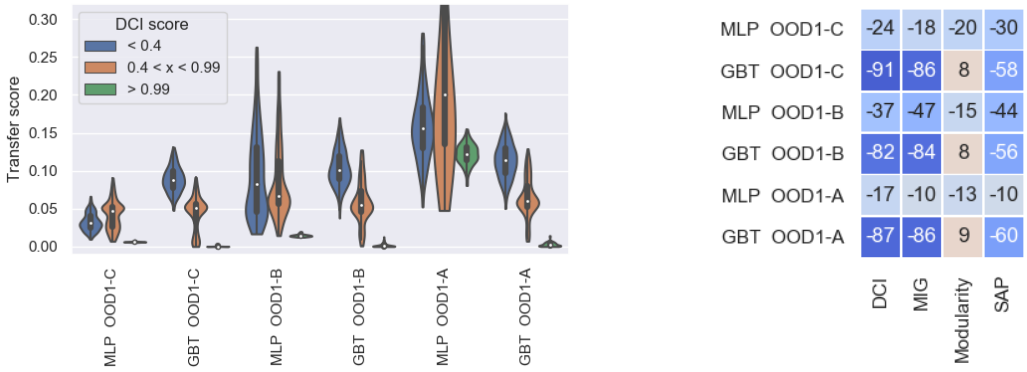


Figure 4: Higher disentanglement corresponds to better generalization across all OOD1 scenarios, as seen from the transfer scores (left). The transfer score is computed as the mean absolute prediction error of ground truth factor values (lower is better). This correlation is particularly evident in the GBT case, whereas MLPs appear to mostly benefit from very high disentanglement only. These results are mirrored in the Spearman rank correlations between transfer scores and disentanglement metrics (right).

subset of the dataset for representation learning, i.e. $D_2 \subset D$, and (2) the OOD test set is disjoint from that dataset, i.e. $D_2 \cap D = \varnothing$. These two scenarios will be denoted by OOD1 and OOD2, respectively. We stress that the OOD2 scenario, which is typically not studied extensively, is only possible because the representations are not trained on the entire dataset.

We start by training downstream regressors on D_1 to predict ground truth factor values from the learned image representation computed by the encoder. We independently train one predictor per factor. The regressors considered here are gradient-boosted trees (GBTs), random forests, and MLPs with $\{1, 2, 3\}$ layers. Since random forests exhibit similar behavior to GBTs, and all MLPs yield similar results to each other, we choose GBTs and the 2-layer MLP as representative models and only report results for those. We test the regressors on D_2, which differs distributionally from the training set D_1, as it either contains held-out values of the FoV or it consists of real-world images. In-distribution generalization would be evaluated by testing the regressors on held-out examples from D_1. To quantify prediction quality, we normalize the ground truth factor values to the range $[0, 1]$, and compute the mean absolute error (MAE). Since the values are normalized, we can take the average of the MAE over all factors (except for the FoV which is OOD).

5 Benefits and Transfer of Structured Representations

Experimental setup. We evaluate the transfer metric introduced in Section 4 across all 1080 trained models. To compute this metric, we train regressors to predict ground truth factors of variation, and test them under distributional shift. We report scores for two different regressors: a Gradient Boosted Tree (GBT) and a [256, 256] two-layer MLP. In the OOD1 setting, we have $D_2 \subset D$. Therefore, the encoder is in distribution – we are testing the regressor on representations that were in the training set of the representation learning algorithm – and we expect the representations to be meaningful. We consider three scenarios:

- OOD1-A: The regressors are trained on 1 color (red) and evaluated on the remaining 7.
- OOD1-B: The regressors are trained on 4 colors with high hue in the HSV space, and evaluated on 4 colors with low hue (extrapolation).
- OOD1-C: The regressors are again trained and evaluated on 4 colors, but the training and evaluation colors are now alternating along the hue dimension (interpolation).

In the more challenging setting where even the encoder goes out-of-distribution (OOD2, i.e. $D_2 \cap D = \varnothing$), we train the regressors on a subset of the training set D that includes all 8 colors, and we consider the following two scenarios:

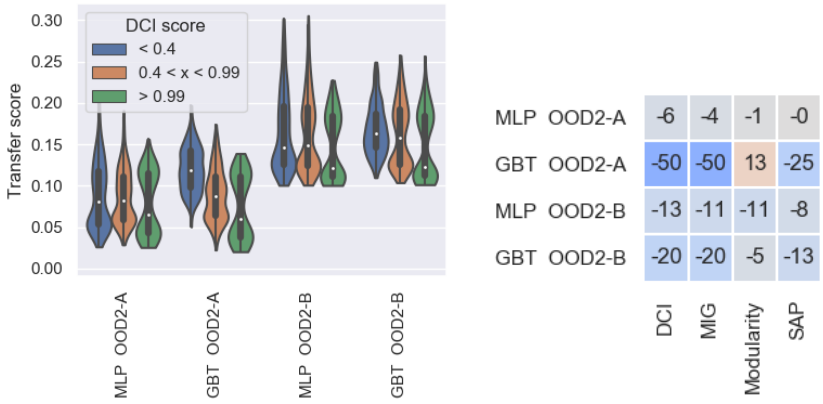


Figure 5: Disentanglement affects generalization across the OOD2 scenarios only minimally as seen from transfer scores (left) and corresponding rank correlations with disentanglement metrics (right).

- OOD2-A: The regressors are evaluated on simulated data, on 4 colors that are out of the encoder's training distribution.
- OOD2-B: The regressors are evaluated on real-world images of the robotic setup, without any adaptation or fine-tuning.

Is disentanglement beneficial for OOD1 generalization? In Fig. 4 we consistently observe a negative correlation between disentanglement and transfer error across all OOD1 settings. The correlation is mild when using MLPs, strong when using GBTs. This difference is expected, as GBTs have an axis-alignment bias whereas MLPs can – given enough data and capacity – disentangle an entangled representation more easily. Our results therefore suggest that **highly disentangled representations are useful for generalising out-of-distribution as long as the encoder remains in-distribution**. This is in line with the correlation found by Locatello et al. (2019b) between disentanglement and the GBT10000 metric. There, however, GBTs are tested on the same distribution as the training distribution, while here we test them under distributional shift. Given that the computation of disentanglement scores requires labels, this is of little benefit in the unsupervised setting. However, it can be exploited in the weakly supervised setting, where disentanglement was shown to correlate with reconstruction loss and ELBO (Section 3). Therefore, model selection for representations that transfer well in these scenarios is feasible based on the ELBO or reconstruction loss, when weak supervision is available. Note that, in absolute terms, the OOD generalization error with encoder in-distribution (OOD1) is very low in the high-disentanglement case (the only exception being the MLP in the OOD1-C case, with the 1-7 color split, which seems to overfit). This suggests that disentangled representations can be useful in downstream tasks even when transferring out of the training distribution.

Summary: Disentanglement seems to improve OOD generalization of downstream tasks, provided that the encoder remains in-distribution (OOD1). Since in the weakly supervised case disentanglement correlates with the ELBO and the reconstruction loss, model selection can be performed using these metrics as a proxy for disentanglement. These metrics have the advantage that they can be computed without labels, unlike the disentanglement metrics.

Is disentanglement beneficial for OOD2 generalization? As can be seen from Fig. 5, the negative correlation between disentanglement and GBT transfer error is weaker when the encoder is out of distribution (OOD2 case). Nonetheless, we observe a non-negligible correlation for GBTs in the OOD2-A case, where we investigate out-of-distribution generalization along one FoV, with observations in D_2 still generated from the same simulator. In the OOD2-B setting, where the observations are taken from cameras in the corresponding real world setting, there correlation between disentanglement and transfer performance appears to be minor at best. This scenario can be considered a variant of zero-shot sim2real generalization.
Summary: Disentanglement has a minor effect on out-of-distribution generalization outside of the training distribution of the encoder (OOD2).

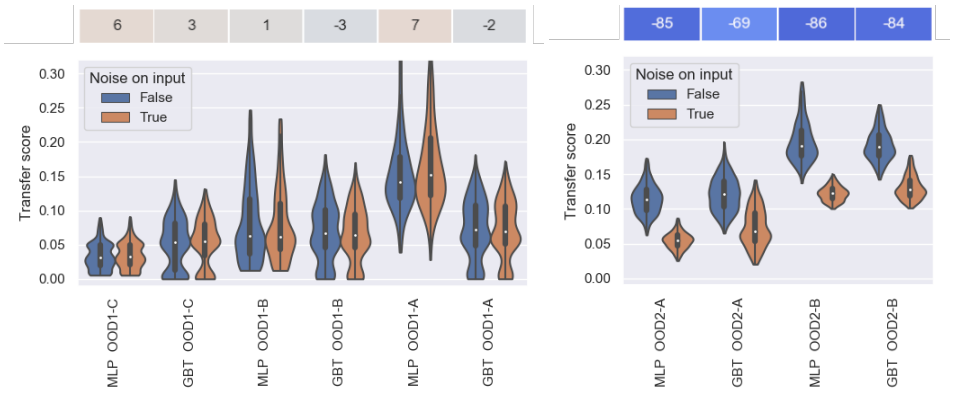


Figure 6: Noise improves generalization across the OOD2 scenarios and less so for the OOD1 scenarios as seen from the transfer scores. Top row: Spearman rank correlation coefficients between transfer metrics and presence of noise in the input.

What else matters for OOD2 generalization? Results in Fig. 6 suggest that adding Gaussian noise to the input during training as described in Section 3 leads to significantly better OOD2 generalization, and has no effect on OOD1 generalization. Adding noise to the input of neural networks is known to lead to better generalization (Sietsma & Dow, 1991; Bishop, 1995). This is in agreement with our results, since OOD1 generalization does not require generalization of the encoder, while OOD2 does. Interestingly, closer inspection reveals that the contribution of different ground truth factors to the generalization error can vary widely. See Appendix B.3 for further details. In particular, with noisy input, the position of the cube is predicted accurately even in real-world images (<5% mean absolute error on each axis). This is promising for robotics applications, where the true state of the joints is observable but inference of the cube position would have to rely on object tracking methods. Fig. 7 shows an example of real-world inputs and reconstructions of their simulated equivalents.

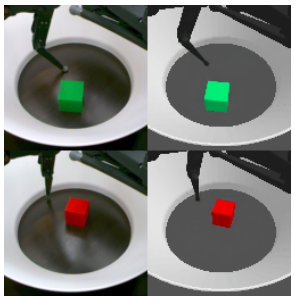


Figure 7: Zero-shot transfer to real-world observations of our models trained in simulation.

Summary: Adding input noise during training appears to be significantly beneficial for OOD2 generalization, while having no effect when the encoder is kept in its training distribution (OOD1).

6 Conclusion

Despite the growing importance of the field and the potential societal impact in the medical domain (Chartsias et al., 2018) or fair decision making (Locatello et al., 2019a), state-of-the-art approaches for learning disentangled representations have so far only been systematically evaluated on synthetic toy datasets. Here we introduced a new high-resolution dataset with over 1M simulated images and 1k annotated real-world images of the same setup. This dataset exhibits a number of challenges and features which are not present in previous datasets: it contains correlations between factors, occlusions, a complex underlying structure, and it allows for evaluation of transfer to unseen simulated and real-world settings. We proposed a new VAE architecture to scale disentangled representation learning to this realistic setting and conducted a large-scale empirical study of disentangled representations on this dataset. We discovered that disentanglement is a good predictor of OOD task performance and showed that, in the context of weak supervision, model selection for good OOD performance can be based on the ELBO or the reconstruction loss, which are accessible without explicit labels. Our setting allows for studying a wide variety of interesting downstream tasks in the future, such as reinforcement learning or learning a dynamics model of the environment. Finally, we believe that in the future it will be important to take further steps in the direction of this paper by considering settings with even more complex structures and stronger correlations between factors.

REFERENCES

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pp. 50–59, 2018.

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(7):1–48, 2002.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Huanru Henry Mao, Garrison W Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. *arXiv preprint arXiv:2003.04887*, 2020.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 490–498. Springer, 2018.

Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Sanja Fidler, Sven Dickinson, and Raquel Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *Advances in neural information processing systems*, pp. 611–619, 2012.

Muhammad Waleed Gondal, Manuel Wüthrich, Djordje Miladinović, 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. In *Advances in Neural Information Processing Systems*, 2019.

Sven Gowal, Chongli Qin, Po-Sen Huang, Taylan Cemgil, Krishnamurthy Dvijotham, Timothy Mann, and Pushmeet Kohli. Achieving robustness in the wild via adversarial mixing with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1211–1220, 2020.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017a.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, 2017b.

Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bošnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. In *International Conference on Learning Representations*, 2018.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, 2016.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 1999.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12627–12637, 2019.

Christian Jutten and Juha Karhunen. Advances in nonlinear blind source separation. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 245–256, 2003.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pp. 14611–14624, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019b.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1–8. IEEE, 2018.

Scott Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, 2018.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Andrei A Rusu, Matej Večerík, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *Conference on Robot Learning*, pp. 262–270, 2017.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Jocelyn Sietsma and Robert JF Dow. Creating artificial neural networks that generalize. *Neural networks*, 4(1):67–79, 1991.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pp. 3738–3746, 2016.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pp. 6056–6065. PMLR, 2019.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30. IEEE, 2017.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv:1905.12506*, 2019.

Manuel Wüthrich, Felix Widmaier, Felix Grimminger, Joel Akpo, Shruti Joshi, Vaibhav Agrawal, Bilal Hammoud, Majid Khadiv, Miroslav Bogdanovic, Vincent Berenz, et al. Trifinger: An open-source robot for learning dexterity. *arXiv preprint arXiv:2008.03596*, 2020.

Mengyuan Yan, Qingyun Sun, Iuri Frosio, Stephen Tyree, and Jan Kautz. How to close sim-real gap? transfer with segmentation! *arXiv preprint arXiv:2005.07695*, 2020.

A IMPLEMENTATION DETAILS

Here we will provide more details on the implementation and training of the models.

We train the β-VAEs by maximizing the objective function

$$\mathcal{L}_{VAE}^{\beta} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})) \leq \log p(\boldsymbol{x}) \tag{2}$$

with $\beta > 0$ using the Adam optimizer (Kingma & Ba, 2014) with default parameters. We use a batch size of 64 and train for 400k steps. The learning rate is initialized to 1e-4 and halved at 150k and 300k training steps. We clip the global gradient norm to 1.0 before each weight update. Following Locatello et al. (2019b), we use a Gaussian encoder with an isotropic Gaussian prior for the latent variable, and a Bernoulli decoder.

Architectural details are provided in Tables 2 to 4. In preliminary experiments we observed that batch normalization, layer normalization, and dropout did not significantly affect performance in terms of ELBO, model samples, and disentanglement scores, both in the unsupervised and weakly supervised settings. On the other hand, layer normalization before the posterior parameterization (last layer of the encoder) appeared to be beneficial for stability in early training. Training deep models typically benefits from a gradient flow from outputs to inputs that is as direct as possible. While using encoder and decoder architectures based on residual blocks leads to fast and effective convergence, in practice we observed that it is sometimes challenging to keep the gradients in check at the beginning of training. In order to solve this issue, we resorted to a simple scalar gating mechanism on the residual blocks (Bachlechner et al., 2020) which ensures that each residual block is initialized as the identity function. This instability may also be exacerbated in probabilistic models by the sampling step in latent space, where a large log variance causes the decoder input to take very large values. Intuitively, this might be a reason why layer normalization before latent space is beneficial for training stability.

Our implementation of weakly supervised learning is based on Ada-GVAE (Locatello et al., 2020), but uses a symmetrized KL divergence:

$$\tilde{D}_{\mathbf{KL}}(p, q) = \frac{1}{2} D_{\mathbf{KL}}(p\|q) + \frac{1}{2} D_{\mathbf{KL}}(q\|p) \tag{3}$$

to infer which latent dimensions should be aggregated.

The noise added to the encoder's input consists of two independent components, both iid Gaussian with zero mean: one is independent for each subpixel (RGB) and has standard deviation 0.03, the other is a 8×8 pixel-wise (greyscale) noise with standard deviation 0.15, bilinearly upsampled by a factor 16. The latter has been designed (by visual inspection) to approximately mimic observation noise in the real images due to complex lighting conditions.

Residual Block
Input: $H \times W \times C$
LeakyReLU(0.02)
Conv 3x3, C channels
LeakyReLU(0.02)
Conv 3x3, C channels
Scalar gate
Sum with input

Table 2: Architecture of one residual block. The scalar gate is implemented by multiplying the tensor by a learnable scalar parameter before adding it to the block input. Initializing the residual block to the identity by setting this parameter to a very small number has been proposed by Bachlechner et al. (2020). It is useful for stabilizing training in deep models using residual connections.

Operation	Output Shape
Input	$128 \times 128 \times C$
Conv 5x5, stride 2, 64 channels	$64 \times 64 \times 64$
LeakyReLU(0.02)	—
2x ResidualBlock(64)	—
Conv 1x1, 128 channels	$64 \times 64 \times 128$
AveragePool(2)	$32 \times 32 \times 128$
2x ResidualBlock(128)	—
AveragePool(2)	$16 \times 16 \times 128$
2x ResidualBlock(128)	—
Conv 1x1, 256 channels	$16 \times 16 \times 256$
AveragePool(2)	$8 \times 8 \times 256$
2x ResidualBlock(256)	—
AveragePool(2)	$4 \times 4 \times 256$
2x ResidualBlock(256)	—
Flatten	4096
LeakyReLU(0.02)	—
FC(512)	512
LeakyReLU(0.02)	—
LayerNorm	—
2x FC(d)	$2d$

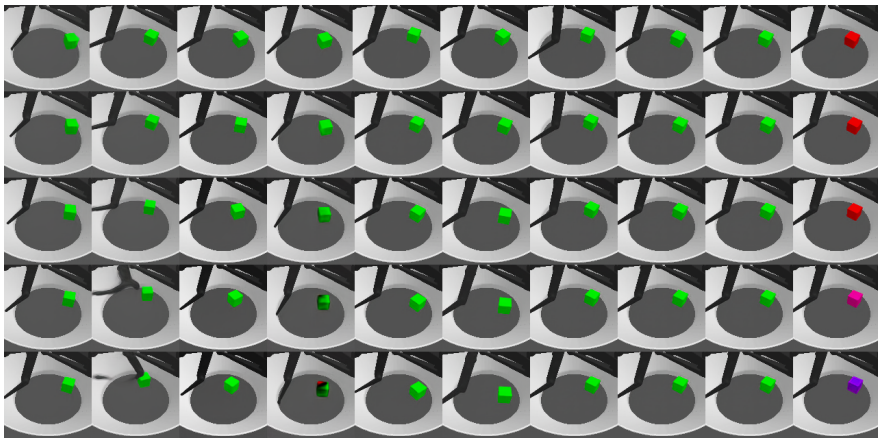
Table 3: Encoder architecture. Last line: the fully connected layer parameterizing the log variance of the approximate posterior distributions of the latent variables has custom initialization. The weights are initialized with $1/10$ standard deviation than the default value, and the biases are initialized to -1 instead of 0. Empirically, this together with (learnable) LayerNorm was beneficial for training stability at the beginning of training. The latent space dimensionality is denoted by d.

Operation	Output Shape
Input	d
FC(512)	512
LeakyReLU(0.02)	—
FC(4096)	4096
Reshape	$4 \times 4 \times 256$
2x ResidualBlock(256)	—
BilinearInterpolation(2)	$8 \times 8 \times 256$
2x ResidualBlock(256)	—
Conv 1x1, 128 channels	$8 \times 8 \times 128$
BilinearInterpolation(2)	$16 \times 16 \times 128$
2x ResidualBlock(128)	—
BilinearInterpolation(2)	$32 \times 32 \times 128$
2x ResidualBlock(128)	—
Conv 1x1, 64 channels	$32 \times 32 \times 64$
BilinearInterpolation(2)	$64 \times 64 \times 64$
2x ResidualBlock(64)	—
BilinearInterpolation(2)	$128 \times 128 \times 64$
LeakyReLU(0.02)	—
Conv 5x5, C channels	$128 \times 128 \times C$

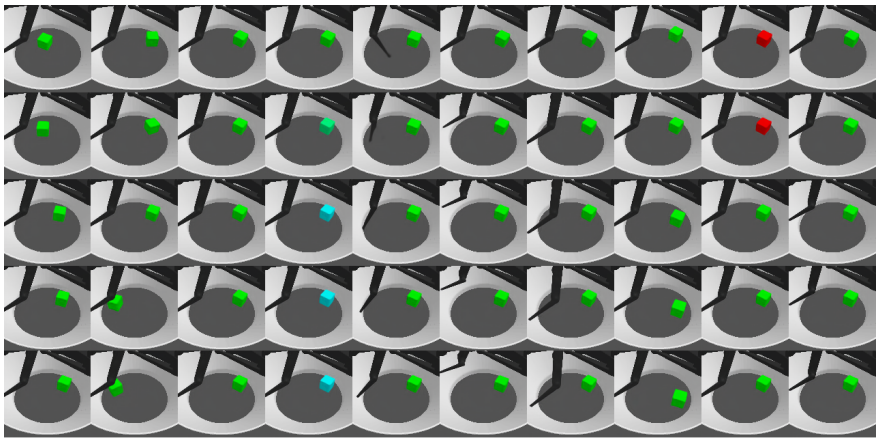
Table 4: Decoder architecture. The latent space dimensionality is denoted by d.

B ADDITIONAL RESULTS

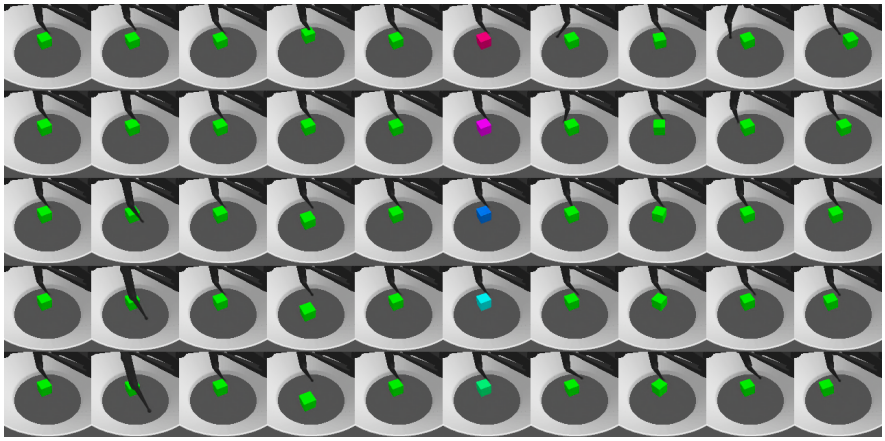
B.1 LATENT TRAVERSALS



(a) Low disentanglement



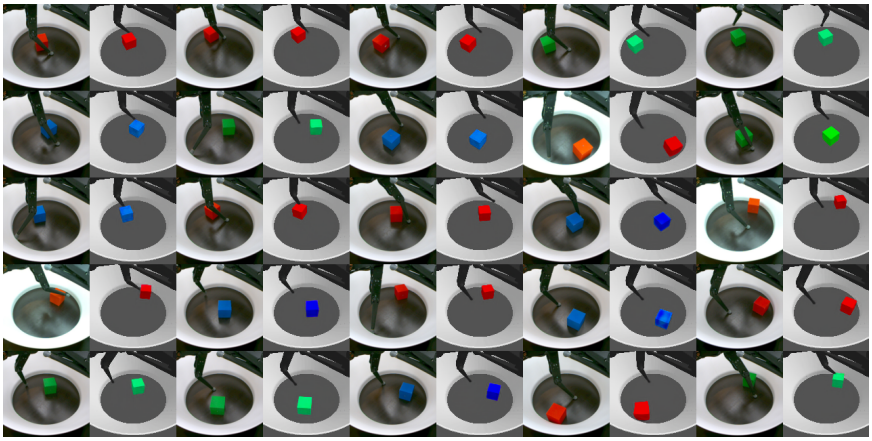
(b) Medium disentanglement



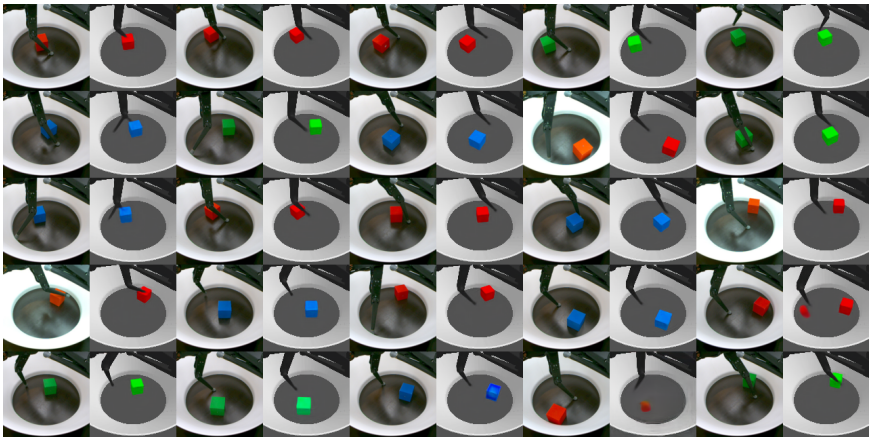
(c) High disentanglement

Figure 8: Latent traversals for a model with low DCI score (0.15) in (a), medium DCI score (0.5) in (b), and high DCI score (1.0) in (c)

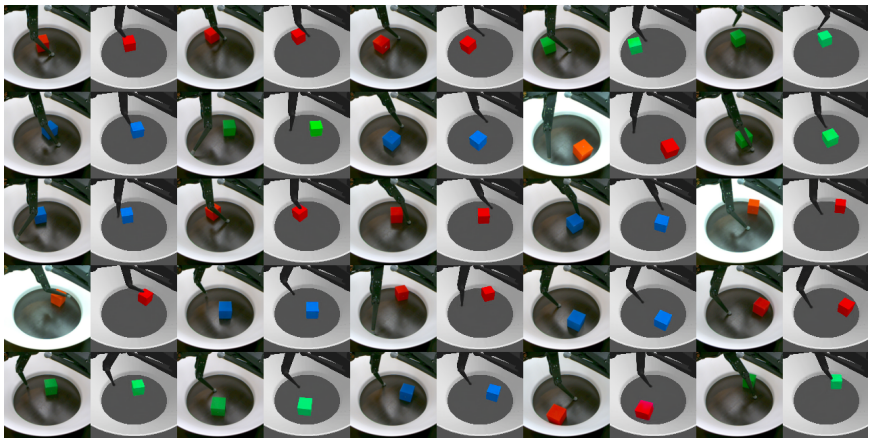
B.2 OUT-OF-DISTRIBUTION RECONSTRUCTIONS



(a) Low disentanglement

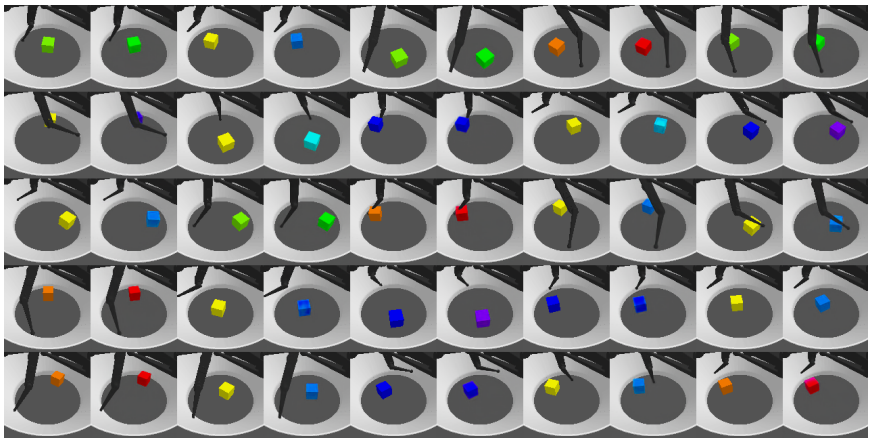


(b) Medium disentanglement

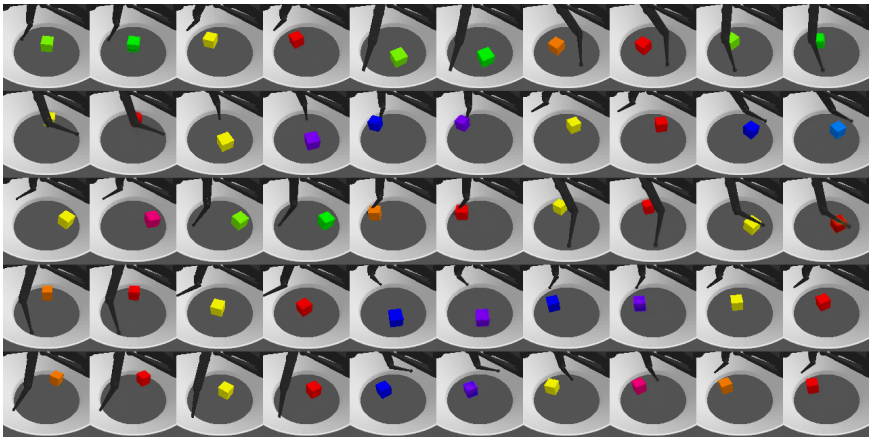


(c) High disentanglement

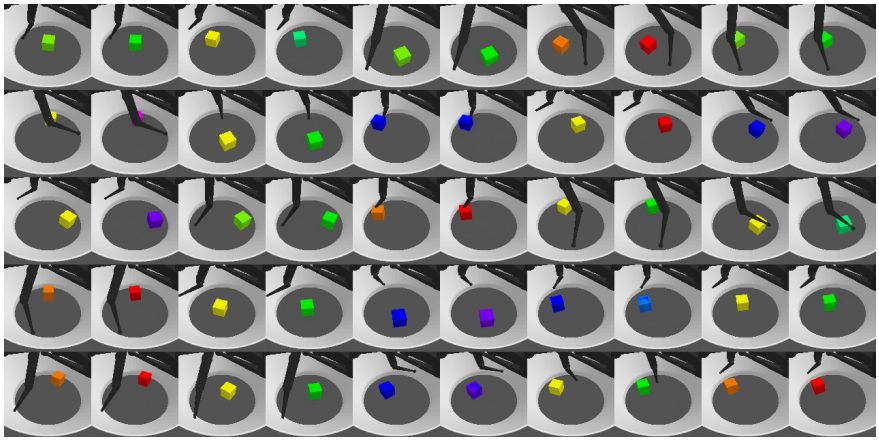
Figure 9: Reconstructions of real-world images (OOD2-B) for a model with low DCI score (0.15) in (a), medium DCI score (0.5) in (b), and high DCI score (1.0) in (c)



(a) Low disentanglement



(b) Medium disentanglement



(c) High disentanglement

Figure 10: Reconstructions of simulated images with encoder out-of-distribution colors (OOD2-A) for a model with low DCI score (0.15) in (a), medium DCI score (0.5) in (b), and high DCI score (1.0) in (c)

B.3 OUT-OF-DISTRIBUTION TRANSFER

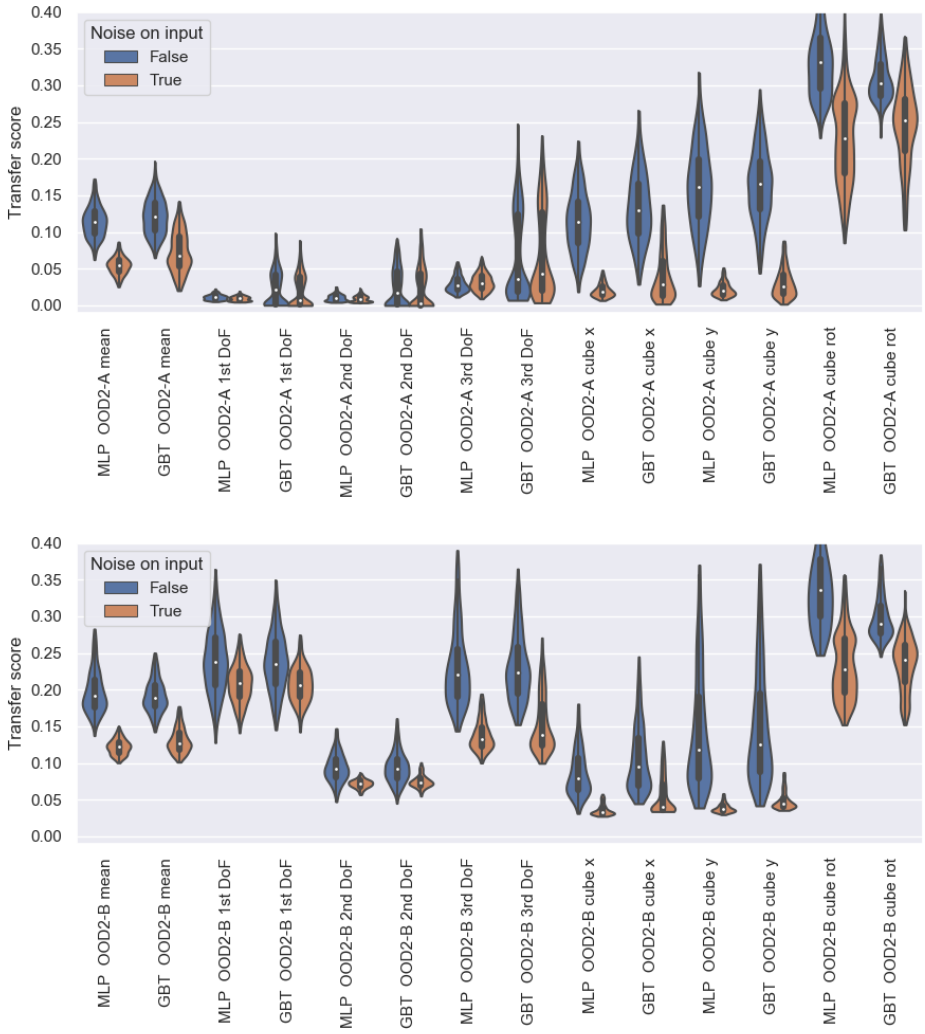


Figure 11: Transfer metric in OOD2-A (top) and OOD2-B (bottom) settings, decomposed according to the factor of variation and presence of input noise. When noise is added to the input during training, the inferred cube position error is relatively low (the scores are the mean absolute error, and they are normalized to $[0, 1)$. This is particularly useful in the OOD2-B setting (real world) where the joint state is anyway considered known, while object position has to be inferred with tracking methods.

B.4 SAMPLES AND RECONSTRUCTIONS

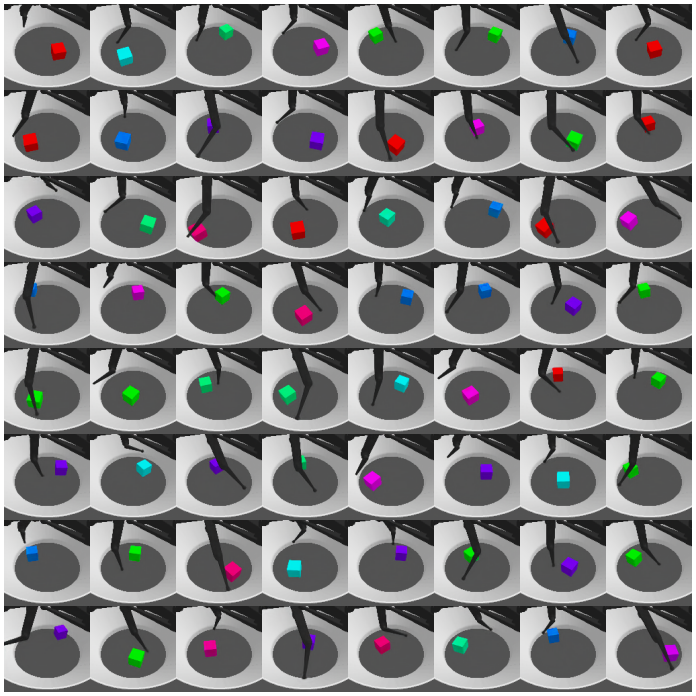


Figure 12: Examples of samples generated by a trained model. This model was selected based on the ELBO.

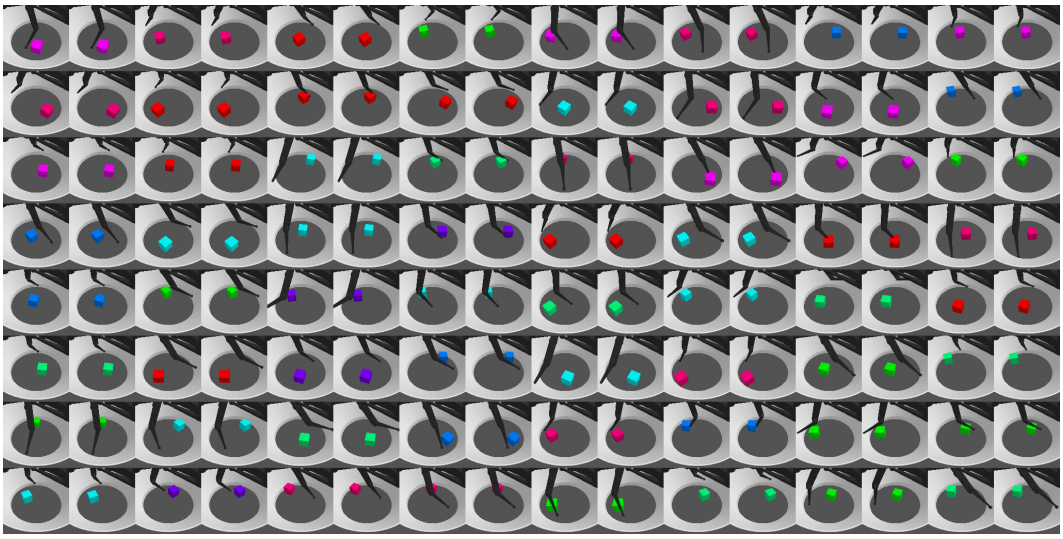


Figure 13: Examples of input reconstructions by a trained model. This model was selected based on the ELBO.

B.5 ADDITIONAL PLOTS

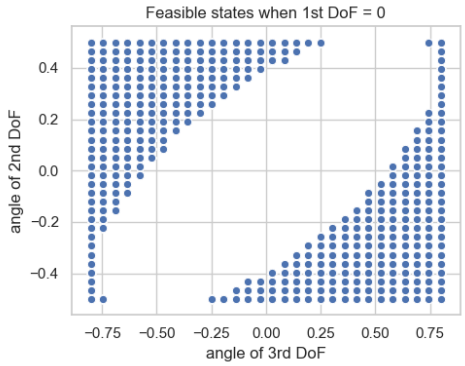


Figure 14: Feasible states of 2nd and 3rd DoF when the 1st DoF has a 0° angle.

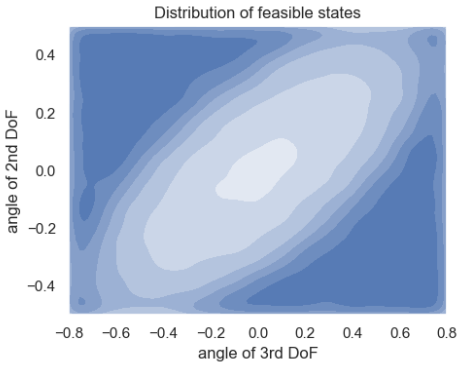


Figure 15: Distribution of feasible states of 2nd and 3rd DoF over the whole dataset.

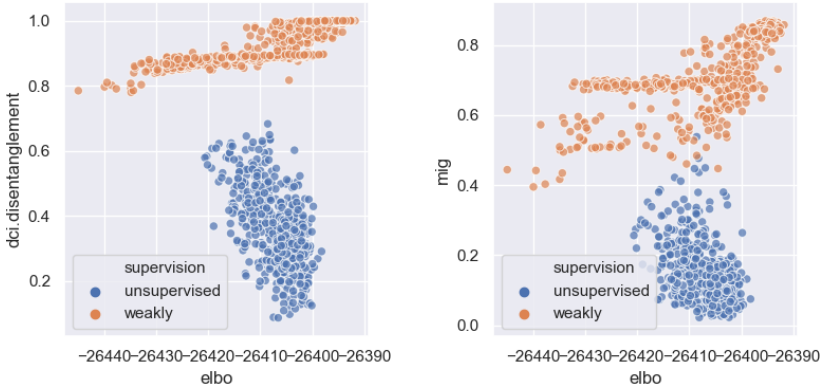


Figure 16: Scatter plots of ELBO vs disentanglement (DCI and MIG) for 1080 trained models, color-coded according to supervision.