Supplementary Material
Demystifying Inductive Biases for (Beta-)VAE Based Architectures

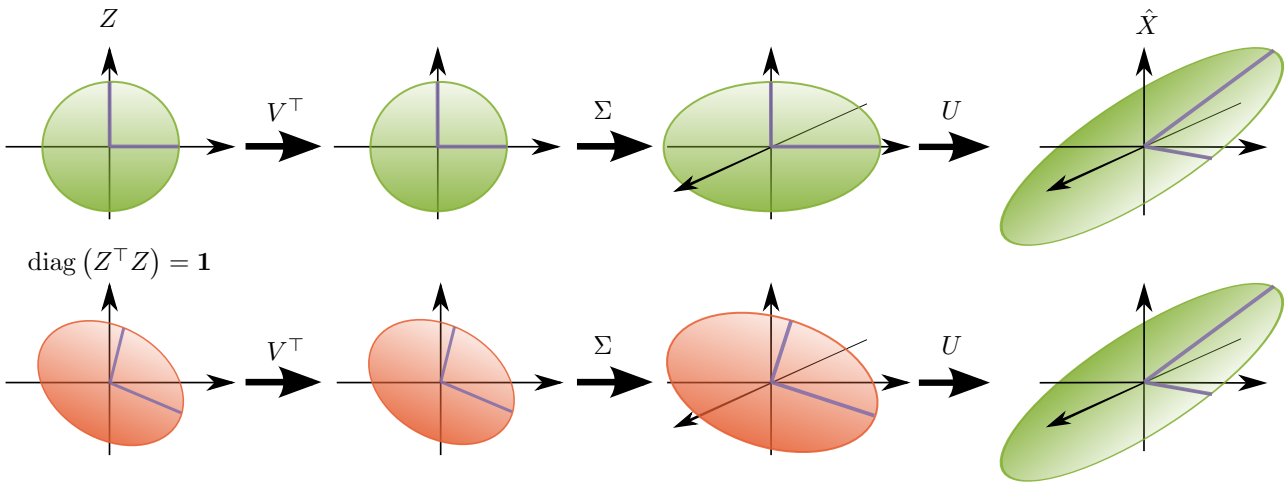

Figure 1: The SVD decomposition of a VAE decoder (top) and an alternative decoder (bottom) which decodes the same data $\hat{X}$, complies with $V = \mathcal{I}$, and also shares $\text{diag}\left(Z^\top Z\right) = \mathbf{1}$. The difference lies in the rotation induced by U, which for VAEs (and PCA) aligns the directions of largest variance in $\hat{X}$ with the cartesian axes.

1. Link to Datasets

https://dx.doi.org/10.17617/3.6i

2. Proofs

2.1. The Formal Setting

The simplified objective stated in this paper as

$$\min_{\Sigma,U,V} \mathbb{E}_i \left(\left\| U\Sigma V^\top \boldsymbol{\varepsilon}^{(i)} \right\|^2 \right) \tag{1}$$

$$\text{s.t.} \quad \mathbb{E}_i \left(\mathcal{L}^{(i)}_{\approx \text{KL}} \right) = c_{\approx \text{KL}}. \tag{2}$$

resembles the minimization problem (20) and (21) from Rolinek et al. (2019). They only optimize for distributing the latent noise $\boldsymbol{\sigma}^{(i)}$ and the orthogonal matrix V of the SVD decomposition of the whole linear decoder and conclude that for $M = U\Sigma V^\top$

> *In every global minimum, the columns of M are orthogonal.*

Which is equivalent to V being a signed permutation matrix (Proposition 1 of (Rolinek et al., 2019)). Without loss of generality, we assume $V = \mathcal{I}$ and rearrange the elements of Σ in ascending order and those of $\varepsilon^{(i)}$ in descending order with respect to ${\boldsymbol{\sigma}^{(i)}}^2$.

In the setting of Theorem (1), we consider the mean latent representation Z to be constrained only by the condition $\text{diag}\left(Z^\top Z\right) = \mathbf{1}$, which reads as "each active latent variable has unit variance". Even though, this statement is unsurprising in the context of VAEs, we offer a quick proof of how this follows directly from the KL loss in Lemma 1. Additionally, we fully fix the matrix $\hat{X}$, which contains the reconstruction of all data-points. The remaining freedom in U and Σ has the following nature: for each fixed $U^\top$ (which rotates $\hat{X}$), the nonzero singular values of Σ (scaling factors along individual axes in the latent space) are fully determined by the $\text{diag}\left(Z^\top Z\right) = \mathbf{1}$ requirement. We minimize objective (1) under these constraints.

Remark Notice that fixing the reconstructed data-points ensures that the observed effect is entirely independent of the deterministic loss. The deterministic loss, is known to have some PCA-like effects, as it is basically a MSE loss of a deterministic autoencoder. The additional (and in fact stronger) effects of the stochastic loss are precisely the novelty of the following theoretical derivations.

For technical reasons regarding the uniqueness of SVD, we additionally inherit the assumption of (Rolinek et al., 2019)

that the random variables $\boldsymbol{\varepsilon}^{(i)}$ have distinct variances.

Finally, the orthonormal matrix U acts isometrically and can be removed from the objective (1), even though it still plays a vital role in how the problem is constrained. The reduced objective is further conveniently rewritten as a trace as:

$$\min_{\Sigma} \mathbb{E}_i \left\| \Sigma \boldsymbol{\varepsilon}^{(i)} \right\|^2 = \min_{\Sigma} \mathbb{E}_i \operatorname{tr}\left(E\Sigma^\top \Sigma E\right), \tag{3}$$

where E is the diagonal matrix induced by the vector $\boldsymbol{\varepsilon}$.

A visualization of the role of U, Σ and V in the decoding process is illustrated in Fig. 1.

2.2. Proof of Theorem 1

We rewrite the objective in order to introduce U, $\hat{X}$, and Z and make use of the constraints $\operatorname{diag}(Z^\top Z) = \mathbf{1}$ and $\hat{X} = Z\Sigma U$. We have

$$E\Sigma^\top \Sigma E = E\Sigma^\top (Z^\top Z + M)\Sigma E, \tag{4}$$

where $M = \mathcal{I} - Z^\top Z$ is a matrix with $\operatorname{diag}(M) = 0$. Also, we can expand

$$\Sigma^\top Z^\top Z \Sigma = U\left(U^\top \Sigma^\top Z^\top\right)(Z\Sigma U)\, U^\top = U\hat{X}^\top \hat{X} U^\top \tag{5}$$

By combining (4) and (5), we learn that

$$E\Sigma^\top \Sigma E - EU\hat{X}^\top \hat{X} U^\top E = E\Sigma^\top M \Sigma E. \tag{6}$$

By repeating Lemma 2, we learn that $\operatorname{diag}(E\Sigma^\top M\Sigma E) = 0$, which allows us to use Lemma 2 yet again, this time on the left-hand side of (6) and obtain a key intermediate conclusion:

$$\operatorname{tr}\left(E\Sigma^\top \Sigma E\right) = \operatorname{tr}\left(EU\hat{X}^\top \hat{X} U^\top E\right) \tag{7}$$

This has a lower bound according to a classical trace inequality (see Proposition 1), as $EU\hat{X}^\top \hat{X}U^\top E$ is positive semi-definite.

$$\operatorname{tr}\left(EU\hat{X}^\top \hat{X} U^\top E\right) \geq n \det\left(EU\hat{X}^\top \hat{X} U^\top E\right)^{1/n} \tag{8}$$

$$= n \det\left(E\hat{X}^\top \hat{X} E\right)^{1/n} \tag{9}$$

with equality if and only if

$$EU\hat{X}^\top \hat{X} U^\top E = \lambda \mathcal{I}. \tag{10}$$

For the SVD decomposition $\hat{X} = U_X \Sigma_X V_X^\top$, we see that $\hat{X}^\top \hat{X} = V_X \Sigma_X^2 V_X^\top$ and with $U' = UV_X$ we arrive at

$$U'\Sigma_X^2 U'^\top = \lambda E^{-2}. \tag{11}$$

The left-hand side gives an SVD decomposition of the diagonal matrix E^{-2}. The SVD decomposition of a diagonal matrix is unique up to a signed permutation matrix. The conclusion of Theorem 1 now follows.

2.3. Auxiliary Statements

In the following lemma, the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ correspond to the mean latent $\boldsymbol{\mu}$ and the noise standard deviation $\boldsymbol{\sigma}$ respectively. We allow for scaling the latent space and find that the KL loss is minimal for unit standard deviation of the means.

Lemma 1. *For vectors $\boldsymbol{x} = (x_0, \ldots, x_n) \in \mathbb{R}^n$, $\boldsymbol{y} = (y_0, \ldots, y_n) \in \mathbb{R}^n$ and*

$$c = \arg\min_{c \in \mathbb{R}} \sum_i \left(c^2 {x_i}^2 - \log\left(c^2 {y_i}^2\right)\right),$$

it holds that

$$c = \sqrt{\sum_i \left({x_i}^2\right)} \tag{12}$$

Proof. It is easy to inspect that the minimum of $\sum_i \left(c^2 {x_i}^2 - \log\left(c^2 {y_i}^2\right)\right)$ with respect to c fulfils the statement. □

Proposition 1 (Trace Inequality)**.** *For a positive semi-definite $M \in \mathbb{R}^{n\times n}$, that is $M \succcurlyeq 0$, it holds that*

$$\operatorname{tr}(M) \geq n \det(M)^{1/n} \tag{13}$$

with equality if and only if $M = \lambda \cdot \mathcal{I}$ for some $\lambda \geq 0$.

Proof. Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of M, then $\operatorname{tr}(M) = \sum_i \lambda_i$ and $\det(M) = \prod_i \lambda_i$. Since $M \succcurlyeq 0$, we have $\lambda_i \geq 0$ for every $i = 1, \ldots, n$. Then, due to the classical AM-GM inequality, we have

$$\operatorname{tr}(M) = \sum_i \lambda_i \geq n \cdot \left(\prod_i \lambda_i\right)^{1/n} = n \det(M)^{1/n}, \tag{14}$$

with equality precisely if all eigenvalues are equal to the same value $\lambda \geq 0$. Then by the definition of eigenvalues, the $M - \lambda\mathcal{I}$ has zero rank, and equals to zero as required. □

Lemma 2 ("Empty diagonal absorbs")**.** *Let $D \in \mathbb{R}^{m\times m}$ be a diagonal matrix and let $M \in \mathbb{R}^{m\times m}$ be a matrix with zero elements on the diagonal, that is $\operatorname{diag}(M) = 0$. Then $\operatorname{diag}(MD) = \operatorname{diag}(DM) = 0$ and consequently also $\operatorname{tr}(MD) = \operatorname{tr}(DM) = 0$.*

Proof. Follows immediately from the definition of matrix multiplication. □

Architecture	dSprites	Shapes3D
β-VAE (β)	8	32
TC-β-VAE (β)	6	32
Factor-VAE (γ)	35	7
Slow-VAE (β)	1	1

Table 1: Primary hyperparameters, for other parameters we used the defaults in the Disentanglement Library or literature values.

3. Experimental Details

3.1. Architecture for $m(\boldsymbol{w})$

The model implemented for $m(\boldsymbol{w})$ has almost the same architecture as the CNN decoder as it is implemented in the Disentanglement Library (Locatello et al., 2019). The only differences lies in the input MLP which was extended by a single neuron hidden layer. This enforces a compression of the generating factors $\mathbf{w}^{(i)}$ to some scalar value based on which the modifications are rendered. Both m and the decoders were trained with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$) and 10^{-4} learning rate. To ensure training stability, we train the decoders on three times more batches as the manipulation network and reconstruct five latent samples per image to get a better estimate of the stochastic losses. We achieved a better result on Shapes3D when using an ensemble of four disentangling and four entangling encoder-decoder pairs instead of single models. In order to stay in the same value range as the original images, we ensured normalization of the manipulated images $\mathbf{x}'^{(i)} = \mathbf{x}^{(i)} + m(\mathbf{w}^{(i)})$ by $\mathbf{x}'^{(i)}_{\text{norm}} = \mathbf{x}^{(i)} - 2\text{ReLu}(\mathbf{x}^{(i)} - 1) + 2\text{ReLu}(-\mathbf{x}^{(i)})$.

4. Additional Experiments

4.1. Evaluation on Different Metrics

We have evaluated all architectures on three additional metrics. See Tables (2, 3, 4) for the resulting DCI-, FactorVAE- and SAP-Scores. Figures (4, 5, 6) show the scores for a line search of the primary hyperparameter of each architecture. The hyperparameters are listed in Table 1. We used the implementations of the Disentanglement Library. Additional information about the distribution of MIG scores on the modified datasets on Shapes3D are presented in the histograms of Figure 2. The individual MIG scores per generating factor for the β-VAE on Shapes3D are shown in Figure 3.

4.2. Inspection of Entangled and Disentangled Latent Embeddings

Over multiple restarts of β-VAE trainings on the unmodified dataset, we used the runs that achieved highest and lowest MIG scores. Exemplary, Fig. 7 and Fig. 8 show two

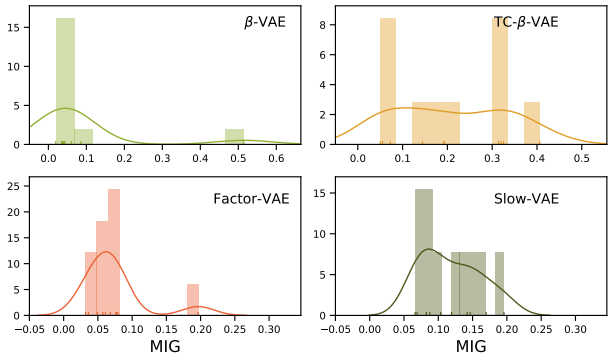

Figure 2: Histogram of MIG scores for the VAE based methods on the altered Shapes3D dataset.

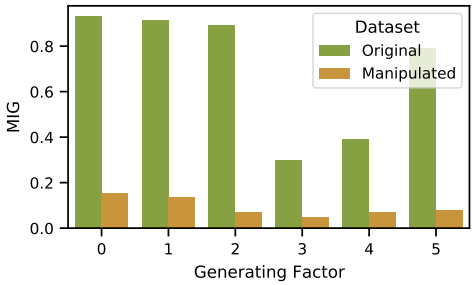

Figure 3: Individual MIG scores for β-VAE trained on the original and the altered Shapes3D dataset.

dimensional latent traversals of four disentangled and four entangled β-VAE representation respectively. The dimension of the latent traversal were hand-picked to encode for the wall hue and the orientation. Interestingly, the disentangled models reliably encode the color in the same way (e.g. starting from green to cyan). The entangled models reliably mix the two generating factors in a very similar way: The color is encoded as the angular component of the two latent dimensions and the orientation as the radial component.

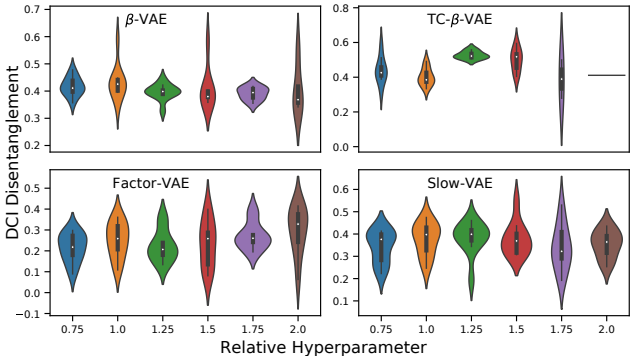

Figure 4: DCI scores for scaled literature hyperparameters over 10 restarts for Shapes3D. Overpruning runs with fewer active units than generating factors were discarded

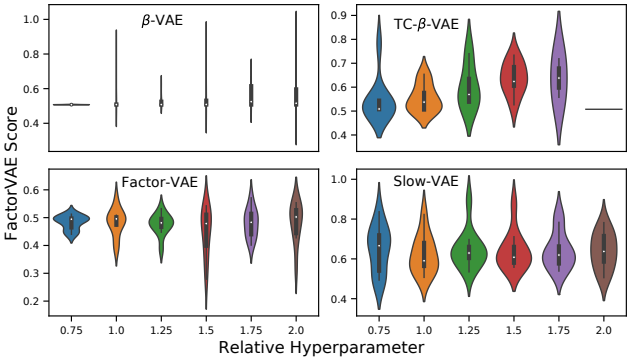

Figure 5: FactorVAE scores for scaled literature hyperparameters over 10 restarts for Shapes3D. Overpruning runs with fewer active units than generating factors were discarded

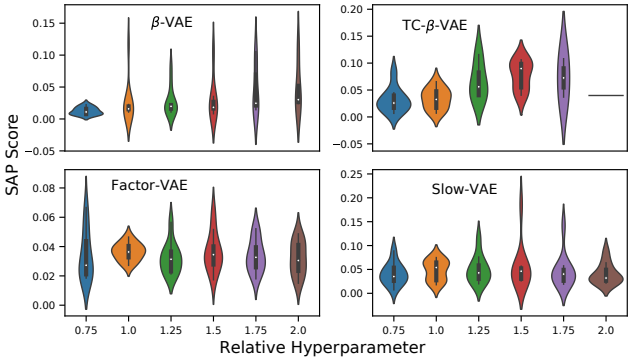

Figure 6: SAP scores for scaled literature hyperparameters over 10 restarts for Shapes3D. Overpruning runs with fewer active units than generating factors were discarded

Table 2: DCI Disentanglement Scores for unmodified, modified and noisy datasets. We report the mean and standard deviation over 10 distinct random seeds for each setting. PCL is the only disentangling non-variational model. The modification leads to a significant drop in all variational methods.

	dSprites			Shapes3d		
	orig.	**mod.**	**noise**	**orig.**	**mod.**	**noise**
β-VAE	0.11 ± 0.03	0.08 ± 0.11	0.14 ± 0.07	0.73 ± 0.14	0.43 ± 0.06	0.56 ± 0.06
Fac. VAE	0.37 ± 0.10	0.27 ± 0.11	0.24 ± 0.09	0.39 ± 0.18	0.25 ± 0.08	0.57 ± 0.20
TC-β-VAE	0.34 ± 0.06	0.19 ± 0.10	0.27 ± 0.03	0.67 ± 0.08	0.41 ± 0.05	0.59 ± 0.09
Slow-VAE	0.47 ± 0.07	0.40 ± 0.07	0.47 ± 0.08	0.65 ± 0.10	0.33 ± 0.08	0.73 ± 0.09
PCL	0.28 ± 0.03	0.30 ± 0.03	0.29 ± 0.06	0.70 ± 0.06	0.67 ± 0.09	0.71 ± 0.07

Table 3: FactorVAE Scores for unmodified, modified and noisy datasets. We report the mean and standard deviation over 10 distinct random seeds for each setting. PCL is the only disentangling non-variational model. The modification leads to a significant drop in all variational methods.

	dSprites			Shapes3d		
	orig.	**mod.**	**noise**	**orig.**	**mod.**	**noise**
β-VAE	0.47 ± 0.07	0.38 ± 0.13	0.50 ± 0.10	0.80 ± 0.17	0.54 ± 0.10	0.71 ± 0.06
Fac. VAE	0.67 ± 0.11	0.62 ± 0.14	0.60 ± 0.11	0.63 ± 0.15	0.48 ± 0.05	0.71 ± 0.15
TC-β-VAE	0.68 ± 0.09	0.53 ± 0.15	0.60 ± 0.12	0.76 ± 0.07	0.57 ± 0.07	0.71 ± 0.06
Slow-VAE	0.77 ± 0.03	0.77 ± 0.04	0.76 ± 0.07	0.87 ± 0.10	0.62 ± 0.06	0.85 ± 0.08
PCL	0.77 ± 0.09	0.82 ± 0.05	0.77 ± 0.08	0.80 ± 0.06	0.77 ± 0.07	0.80 ± 0.06

Table 4: SAP Scores for unmodified, modified and noisy datasets. We report the mean and standard deviation over 10 distinct random seeds for each setting. PCL is the only disentangling non-variational model. The modification leads to a significant drop in all variational methods.

	dSprites			Shapes3d		
	orig.	**mod.**	**noise**	**orig.**	**mod.**	**noise**
β-VAE	0.04 ± 0.01	0.02 ± 0.02	0.03 ± 0.03	0.16 ± 0.08	0.03 ± 0.03	0.09 ± 0.02
Fac. VAE	0.07 ± 0.03	0.06 ± 0.03	0.08 ± 0.01	0.07 ± 0.04	0.04 ± 0.01	0.08 ± 0.03
TC-β-VAE	0.08 ± 0.01	0.06 ± 0.03	0.05 ± 0.02	0.08 ± 0.02	0.04 ± 0.02	0.06 ± 0.03
Slow-VAE	0.08 ± 0.01	0.07 ± 0.01	0.07 ± 0.01	0.09 ± 0.04	0.04 ± 0.01	0.09 ± 0.05
PCL	0.07 ± 0.03	0.10 ± 0.03	0.10 ± 0.03	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01

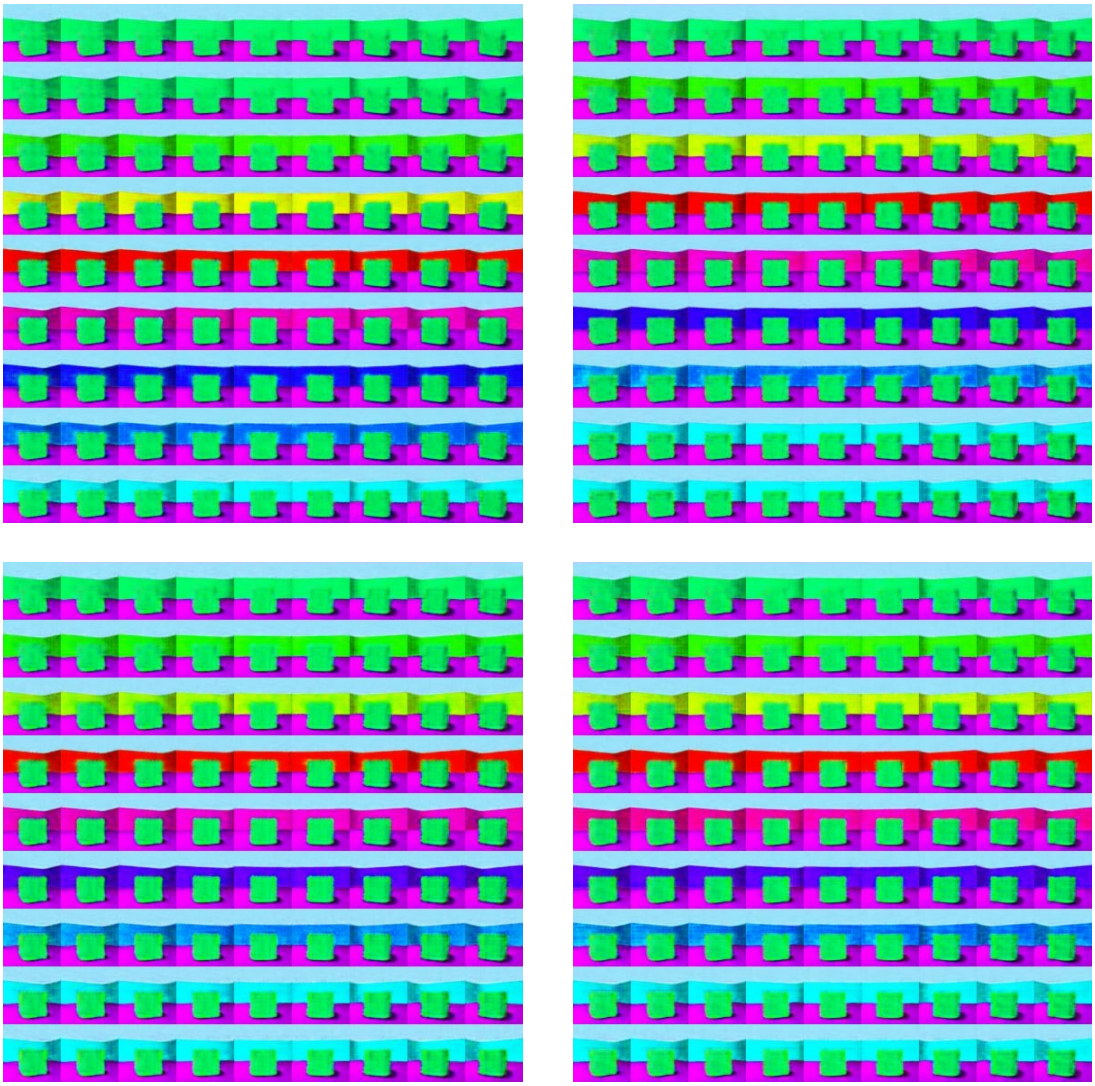

Figure 7: Latent traversals along two latent dimensions for four different disentangled representations. They encode the wall hue and orientation separately. The latent coordinates were flipped to match the same alignment.

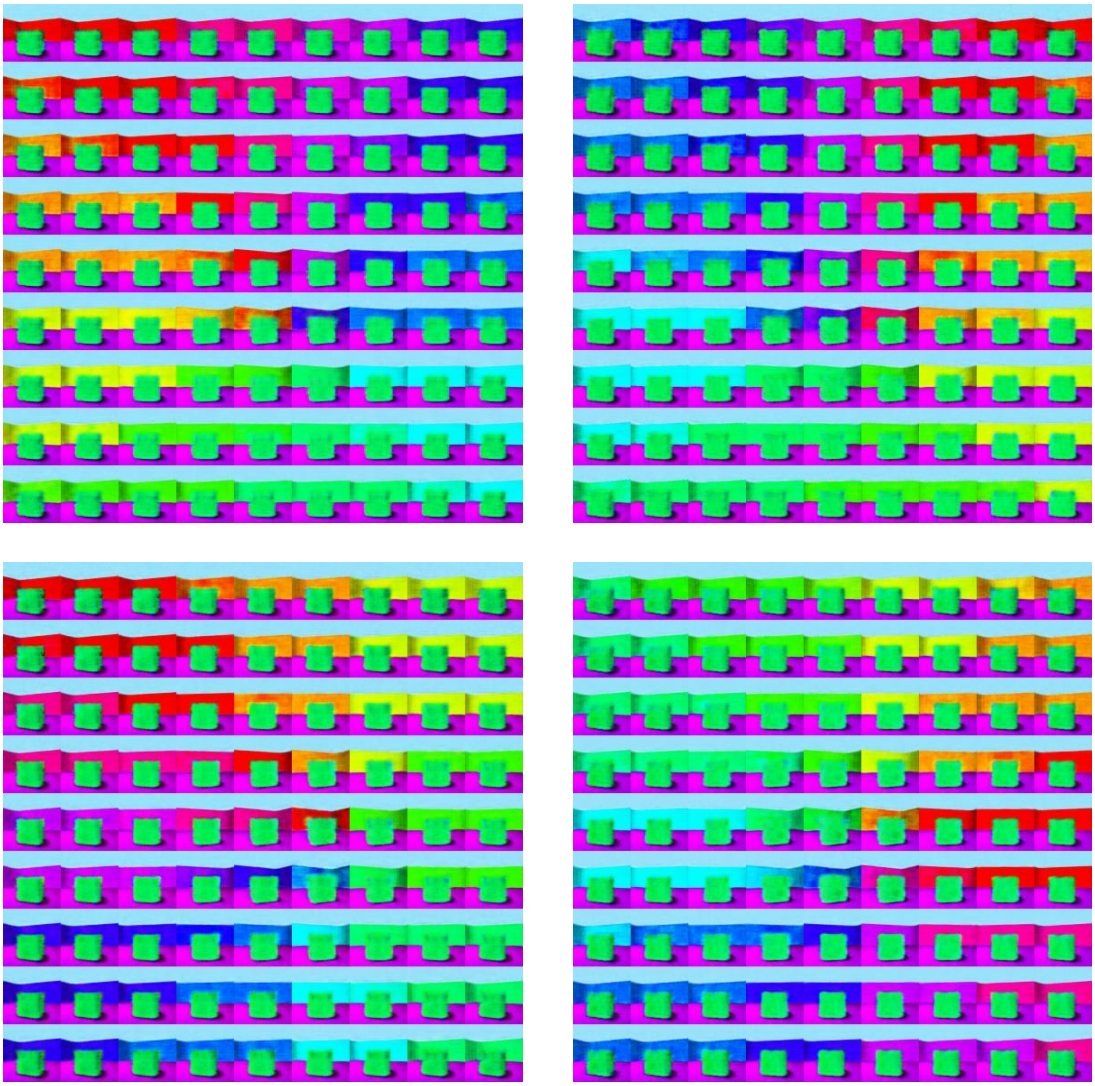

Figure 8: Latent traversals along two latent dimensions for four different disentangled representations. They encode a mixture of wall hue and orientation.