A Appendix

A.1 Latent Responses

In a similar setup as [55], we can extend the joint distribution $p(X, Z)$ to include the reconstruction and response as $p(X, Z, \hat{X}, \hat{Z})$, where crucial question is how the posterior $q(Z \mid X; \phi)$ relates to the response $q(\hat{Z} \mid \hat{X}; \phi)$ where Z and $\hat{Z}$ are related by (also shown in equation 7):

$$r(\hat{Z} \mid Z; \theta, \phi) = \int q(\hat{Z} \mid \hat{X}; \phi) p(\hat{X} \mid Z; \theta) \mathrm{d}\hat{X} \tag{12}$$

Note, that $r(\hat{Z} \mid Z; \theta, \phi)$ is equivalent to the transition kernel $\mathcal{Q}_{\mathrm{AVAE}}$ in [55], except that, crucially, we do not make two assumptions used to derive the AVAE objective. Most importantly, we do not assume that the decoder is a one-to-one mapping between latent samples and a corresponding generated sample. As also discussed in the paper, the contractive behavior observed in the latent space of autoencoders [59], implies a many-to-one mapping, which may be interpretted as the decoder filtering out useless exogenous information from the latent code. Consequently, we also do not treat $p(\hat{Z}; \theta, \phi) = \mathbb{E}_{p(Z)}[r(\hat{Z} \mid Z; \theta, \phi)]$ as a normal distribution, which would imply the encoder perfectly inverts the decoder. , and that we do not assume the decoder is a one-to-one mapping between latent samples and a corresponding generated sample.

Consider the reconstructions $\hat{X}$ of the maximally overfit encoder $q(z = Z \mid x_i = X; \tilde{\phi}) = \delta(s_i - z)$ (recall $s_i = f^{\phi}(x_i)$) and decoder $p(\hat{X} \mid Z; \tilde{\theta})$. Since the autoencoder is trained on the empirical generative process $\pi(X)$ rather than the true generative process $p(X)$, the optimal decoder generates samples from $p(\hat{X}; \tilde{\theta}) = \int p(\hat{X} \mid Z; \tilde{\theta}) p(Z) \mathrm{d}Z = \pi(X)$, which does not have continuous support. For such a decoder, all exogenous noise is completely removed and the decoder mapping is obviously many-to-one, and it follows that $r(\hat{Z} = \hat{z} \mid Z = z; \tilde{\theta}, \tilde{\phi}) = \delta(\hat{z} - s)$ (recall $z = s + u$).

Now consider the more desirable (and perhaps slightly more realistic) setting where the autoencoder extrapolates somewhat beyond $\pi(X)$ to resemble $p(X)$, in which case decoding the latent sample $z \sim q(Z \mid X = x; \phi)$ to generate $\hat{x}$ will not necessarily match the observation x, which implies a change in the endogenous information contained in z. When re-encoding to get $q(\hat{Z} \mid \hat{X} = \hat{x}; \phi)$, the changes in the endogenous information result in some width to the distribution over $\hat{Z}$.

A.2 Mean Curvature for Manifold Learning

The geometry of learned representations with a focus on the generalization ability of neural networks has been discussed in [72]. One key problem is that the standard Gaussian prior used in variational autoencoders relies on the usual Lebesgue measure which in turn, assumes a Euclidean structure over the latent space. This has been demonstrated to lead to difficulties in particular when interpolating in the latent space [25, 73, 74] due to a manifold mismatch [75, 76]. Given the complexity of the underlying data manifold, a viable alternative is based on riemanian geometry [77] which has previously been investigated for alternative probabilistic models like Gaussian Process regression [78].

These methods focus on the intrinsic curvature of the data manifold, which does not depend on the specific embedding of the manifold in the latent space. However, our focus is precisely on how the data manifold is embedded in the latent space, to (among other things) quantify the relationships between latent variables and how well the representation disentangles the true factors of variation. Consequently, we focus on the extrinsic curvature, and more specifically the mean curvature which can readily be estimated using the response maps.

As discussed in the main paper, $|u(z)| = |z - s|$ is interpretted as a distance where $|u(z)| = 0$ implies z is on the latent manifold and there is no exogenous noise. The gradient of this function $\nabla_z |u(z)|$, effectively projects any point in the latent space onto the manifold. Similarly, the mean curvature (equation 13) can be computed, which can be interpreted as identifying the regions in the latent space where the $|u(z)|$ converges and diverges. These gradients are estimated numerically by finite differencing.

$$H = -\frac{1}{2}\nabla_z \cdot \left(\frac{\nabla_z |u(z)|}{|\nabla_z |u(z)||} \right) = -\frac{1}{2}\nabla_z \cdot \frac{u(z)}{|u(z)|} \tag{13}$$

A.3 Double Helix Example Details

To illustrate how the latent response framework can be used to study the representation learned by a VAE, we show the process when learning a 2D representation for samples from a double helix embedded in $\mathbb{R}^3$, defined as:

$$x_i = [A_1 \cos(\pi(\omega t_i + n_i)), A_2 \sin(\pi(\omega t_i + n_i)), A_3 t_i]^T + \epsilon_i \tag{14}$$

where $t_i \sim \text{Uniform}(-1, 1)$, $n_i \sim \text{Bernoulli}(0.5)$, $\epsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$. For this experiment, we set $A_1 = A_2 = A_3 = \omega = 1$ and $\sigma = 0.1$.

Disregarding the additive noise ϵ_i, the data manifold has two degrees of freedom, which are the strand location t_i and the strand number n_i.

To provide the model sufficient capacity, we use four hidden layers with 32 units each for the encoder and decoder. We train until convergence (at most 5k steps) with $\beta = 0.05$ using an Adam optimizer on a total of $N = 1024$ training samples (see the supplementary code for the full training and evaluation details).

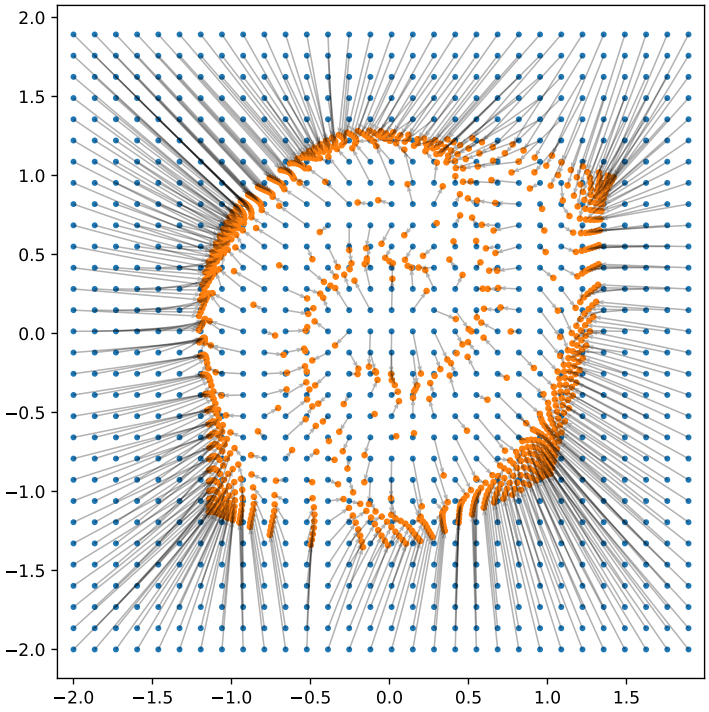


Figure 7: The response map of the representation trained on the double helix. Starting from the latent samples (blue dots), applying the decoder followed by the encoder (i.e. response function) results in the orange dots connected by the black arrows. Note that applying the response function effectively contracts points all over the latent space into a relatively small non-linear region, corresponding to endogenous information.

So we can interpret the blue region as approximating the region where the decoder can projection the latent sample confidently, while positive divergence regions are divergent as the decoding is less precise.

The path in red shows the shortest euclidean distance between the two points, however as is clearly seen in the maps in figure 3 and even the posterior density, the red path clearly does not remain on the learned manifold. Finally meanwhile the green path traces the natural path along one strand.

A.4 Architecture and Training Details

All our models are based on the same convolutional neural network architecture detailed in table 10 so that in total models have approximately 500k trainable parameters. For the smaller datasets MNIST and Fashion-MNIST, samples are upsampled to 32x32 pixels from their original 28x28 and the one convolutional block is removed from both the encoder and decoder.

The datasets are split into a 70-10-20 train-val-test split, and are optimized using Adam [70] with a learning rate of 0.0001, weight decay 0, and β_1, β_2 of 0.9 and 0.999 respectively. The models are trained for 100k iterations with a batch size of 64 (128 for MNIST and Fashion-MNIST).

Input 64x64x3 image
Conv Layer (64 filters, k=5x5, s=1x1, p=2x2)
Max pooling (filter 2x2, s=2x2)
Group Normalization (8 groups, affine)
ELU activation
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Max pooling (filter 2x2, s=2x2)
Group Normalization (8 groups, affine)
ELU activation
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Max pooling (filter 2x2, s=2x2)
Group Normalization (8 groups, affine)
ELU activation
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Max pooling (filter 2x2, s=2x2)
Group Normalization (8 groups, affine)
ELU activation
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Max pooling (filter 2x2, s=2x2)
Group Normalization (8 groups, affine)
ELU activation
Fully-connected Layer (256 units)
ELU activation
Fully-connected Layer (128 units)
ELU activation
Fully-connected Layer ($2d$ units)
Output posterior μ and $\log \sigma$

Figure 8: Encoder Architecture

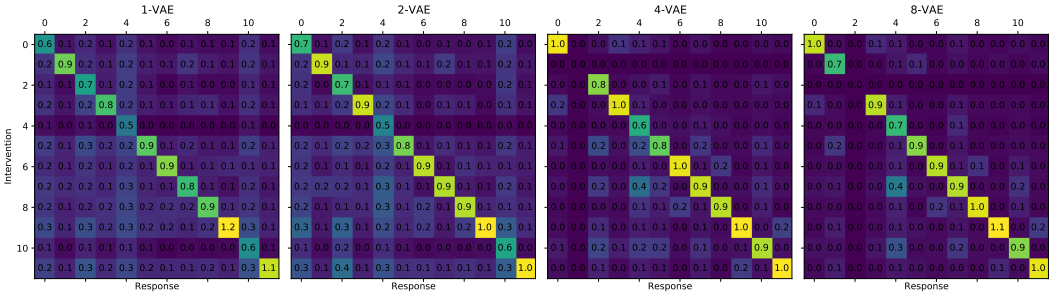
Input d latent vector
Fully-connected Layer (128 units)
ELU activation
Fully-connected Layer (256 units)
ELU activation
Fully-connected Layer (256 units)
ELU activation
Bilinear upsampling (scale 2x2)
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Group Normalization (8 groups, affine)
ELU activation
Bilinear upsampling (scale 2x2)
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Group Normalization (8 groups, affine)
ELU activation
Bilinear upsampling (scale 2x2)
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Group Normalization (8 groups, affine)
ELU activation
Bilinear upsampling (scale 2x2)
Conv Layer (64 filters, k=3x3, s=1x1, p=1x1)
Group Normalization (8 groups, affine)
ELU activation
Conv Layer (3 filters, k=3x3, s=1x1, p=1x1)
Sigmoid activation
Output 64x64x3 image

Figure 9: Decoder Architecture

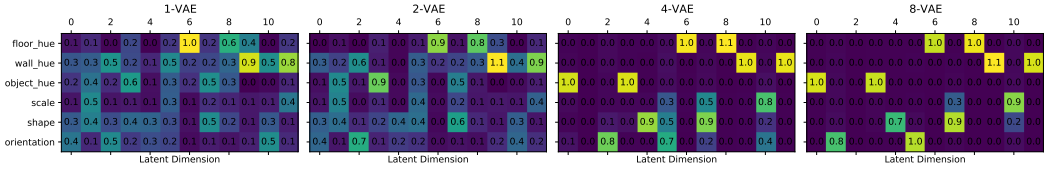
Figure 10: Model architectures where "k" is the kernel size, "s" is the stride, and "p" is the zero-padding

B Additional Results

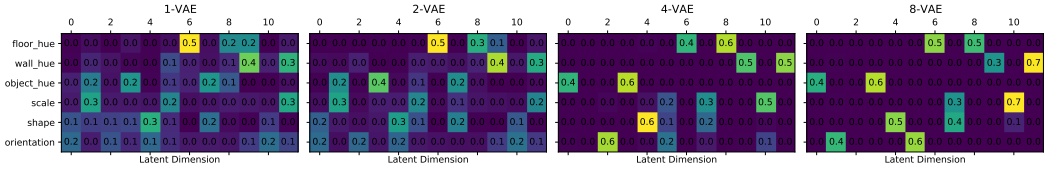
B.1 3D-Shapes



(a) Latent Response Matrices

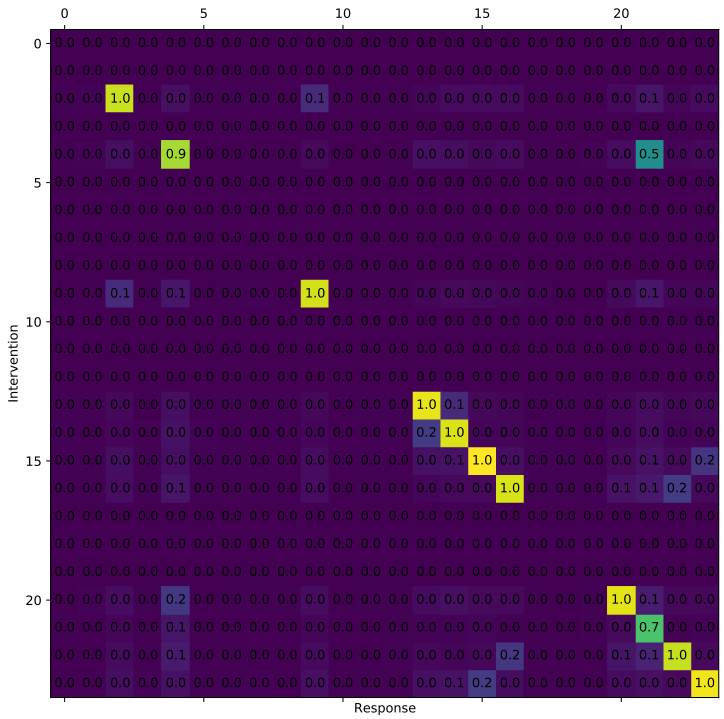


(b) Conditioned Response Matrices

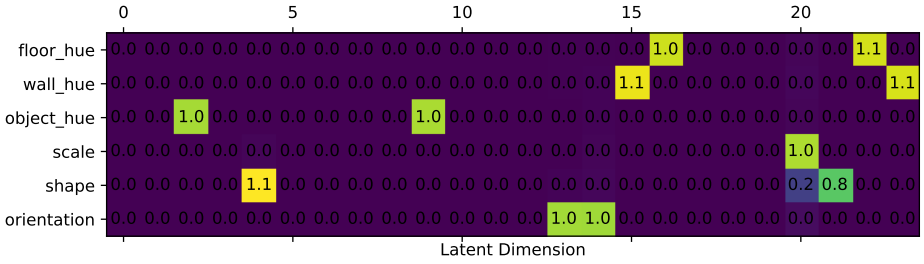


(c) DCI Responsibility Matrices

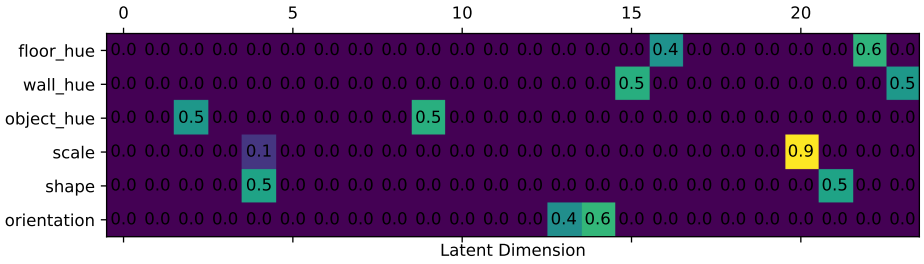
Figure 11: Response and Responsibility matrices for several VAEs ($d = 12$).



(a) Latent Response Matrix

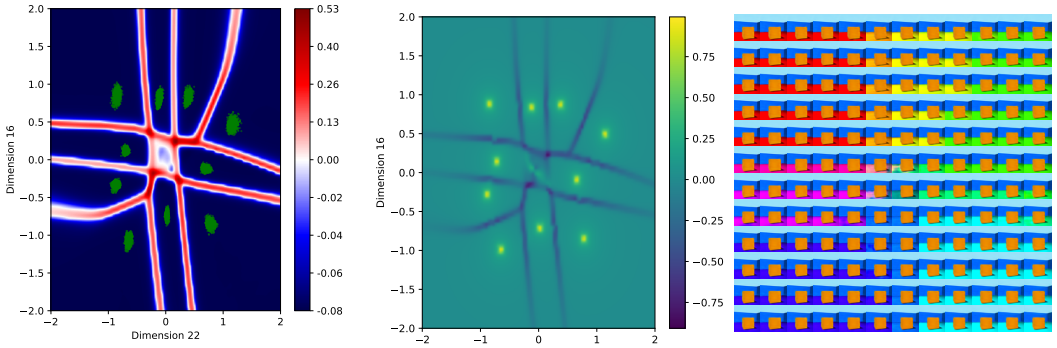


(b) Conditioned Response Matrix



(c) DCI Responsibility Matrix

Figure 12: Full response and responsibility matrices of the 4-VAE ($d = 24$) also shown in figure 6. Note how the Latent Response matrices (12a) shows a categorical difference between the latent dimensions where the diagonal element is close to zero (non-causal), compared to the dimensions with diagonal elements close to 1 (causal).

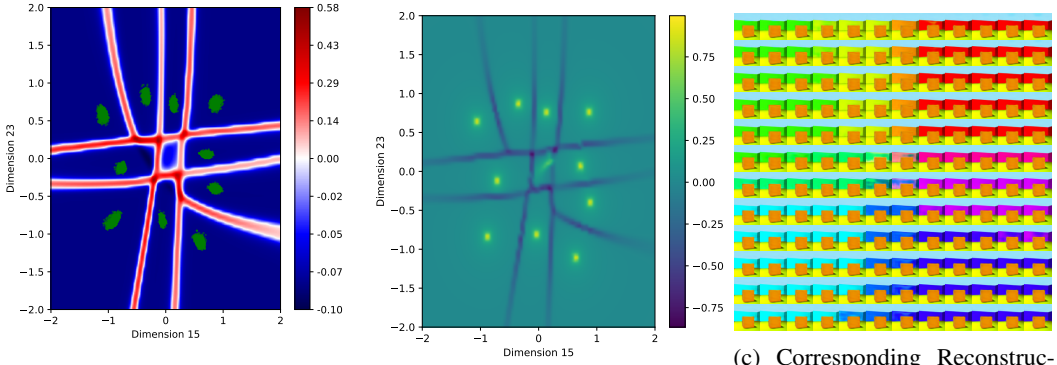


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 13: Visualization of the representation learned by a 4-VAE trained on 3D-Shapes (same model as in figure 12).

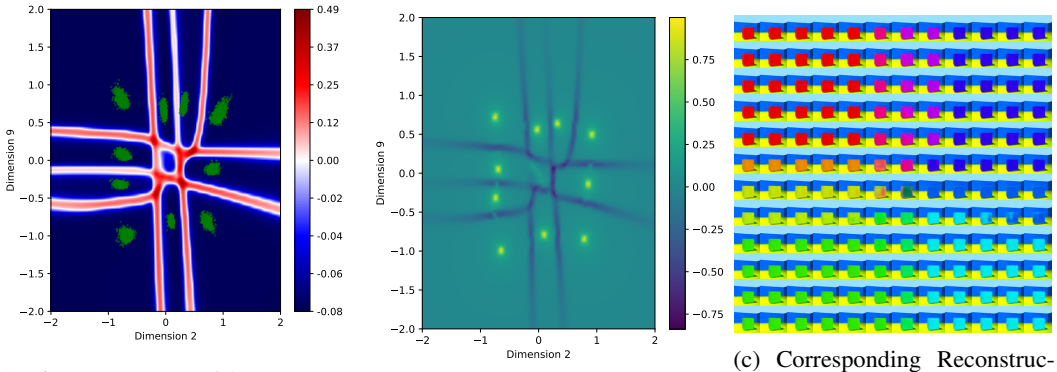


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 14: Visualization of the representation learned by a 4-VAE trained on 3D-Shapes (same model as in figure 12).

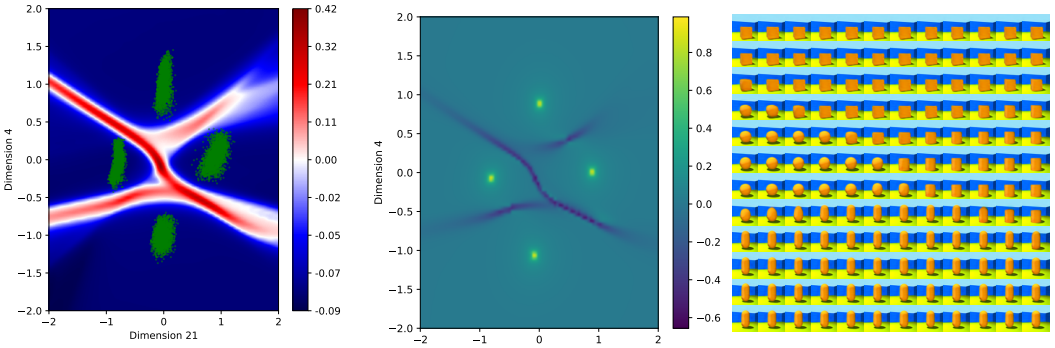


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 15: Visualization of the representation learned by a 4-VAE trained on 3D-Shapes (same model as in figure 12).

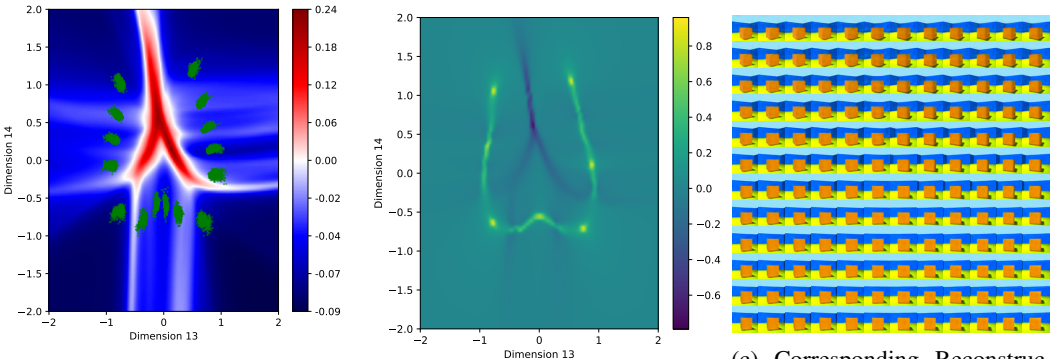


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 16: Visualization of the representation learned by a 4-VAE trained on 3D-Shapes (same model as in figure 12). This projection is particularly interesting as the information encoding shape is not exactly axis-aligned, leading to a slight mismatch between the aggregate posterior and the divergence maps. As our visualizations are presently confined to two dimensions, the structure can become significantly more obscured to us if the information is not disentangled and axis-aligned.



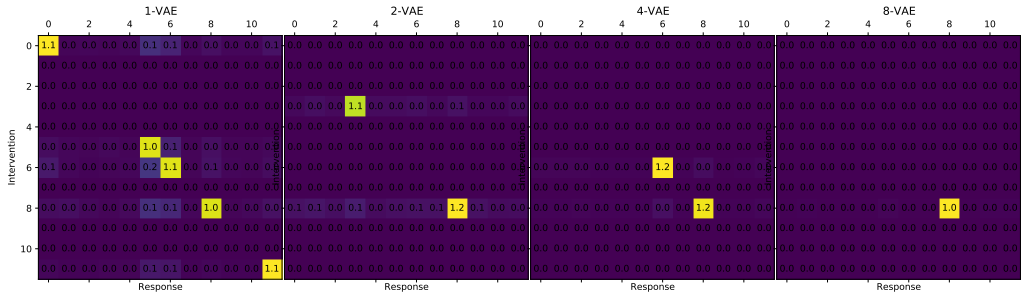
(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

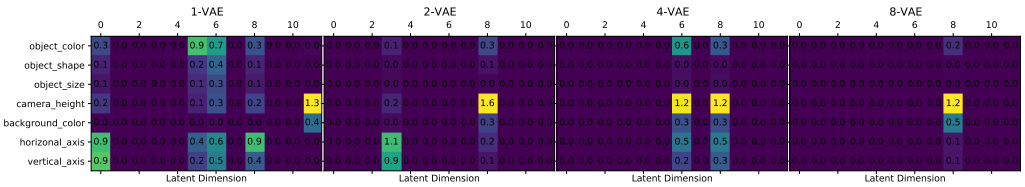
(c) Corresponding Reconstructions

Figure 17: Visualization of the representation learned by a 4-VAE trained on 3D-Shapes (same model as in figure 12).

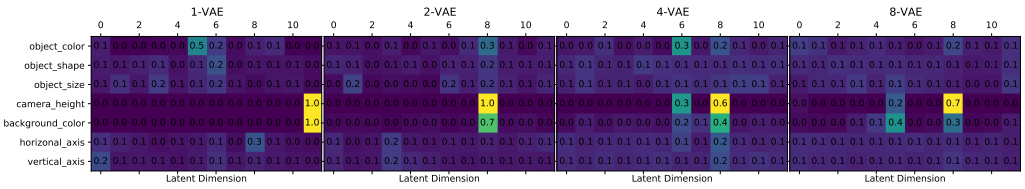
B.2 MPI3D



(a) Latent Response Matrices

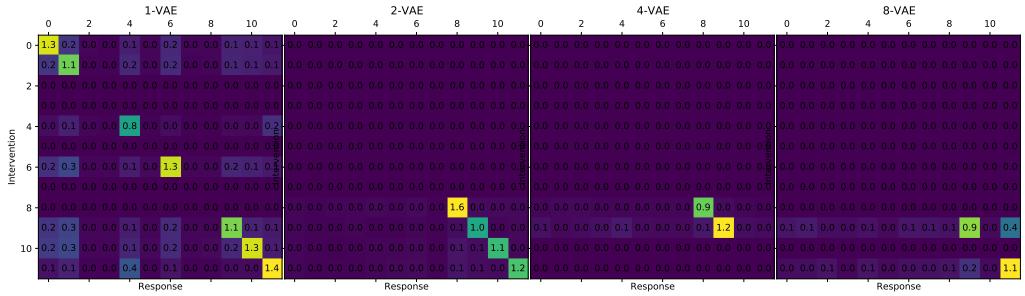


(b) Conditioned Response Matrices

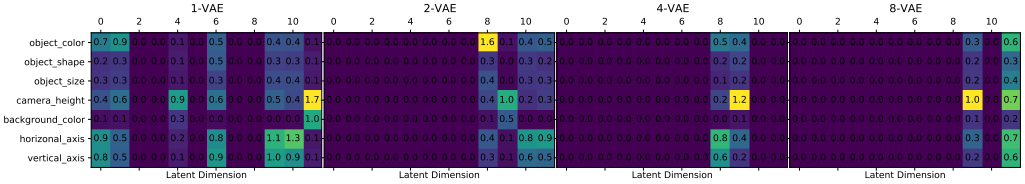


(c) DCI Responsibility Matrices

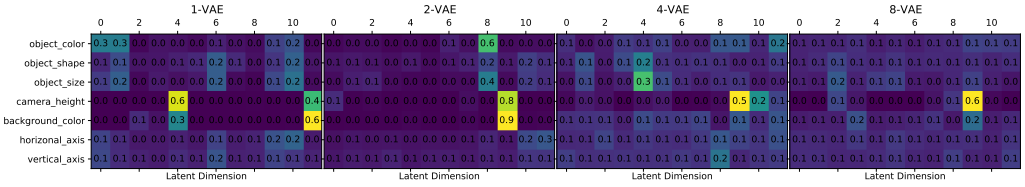
Figure 18: Response and Responsibility matrices for several VAEs ($d = 12$) trained on the MPI3D Toy dataset.



(a) Latent Response Matrices



(b) Conditioned Response Matrices

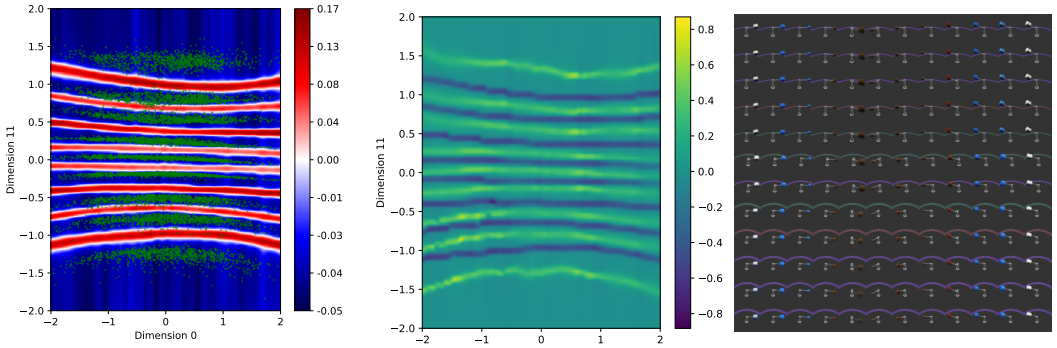


(c) DCI Responsibility Matrices

Figure 19: Response and Responsibility matrices for several VAEs ($d = 12$) trained on the MPI3D Real dataset.

Name	CDS	DCI-D	IRS	MIG
1-VAE	0.69	0.33	0.58	0.32
2-VAE	0.86	0.17	0.59	0.14
4-VAE	0.66	0.11	0.61	0.05
8-VAE	1	0.13	0.79	0.1
1-VAE	0.61	0.24	0.51	0.07
2-VAE	0.69	0.26	0.72	0.24
4-VAE	0.4	0.09	0.75	0.04
8-VAE	0.7	0.08	0.71	0.04

Table 2: disentanglement scores for the MPI3D Toy (first four rows) and Real (last four rows).

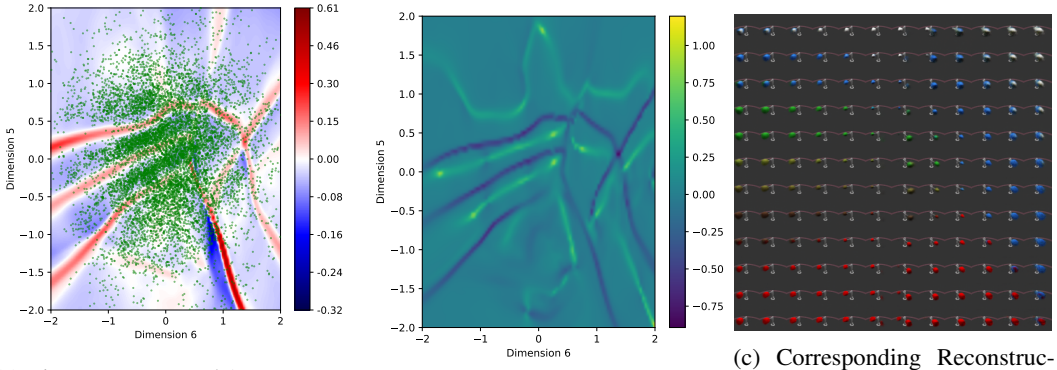


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 20: Visualization of the representation learned by the 1-VAE trained on MPI3D Toy.

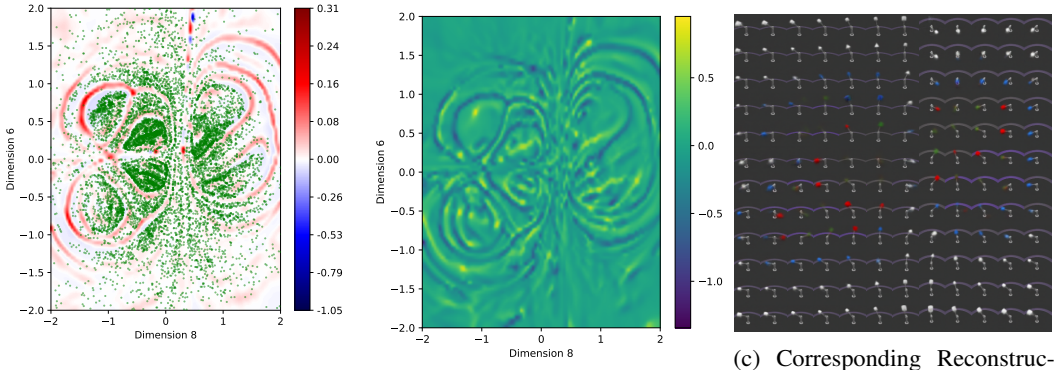


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 21: Visualization of the representation learned by the 1-VAE trained on MPI3D Toy.

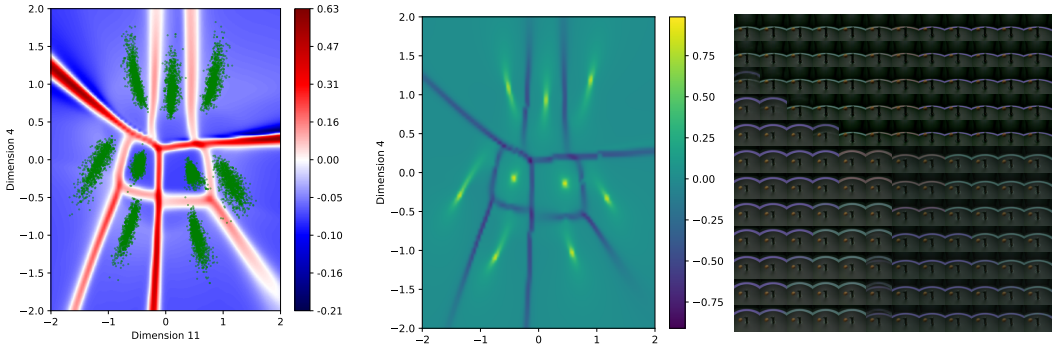


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 22: Visualization of the representation learned by the 4-VAE trained on MPI3D Toy. Note that due to posterior collapse, the full latent manifold is contained in this projection (see the corresponding response matrix in figure 18).

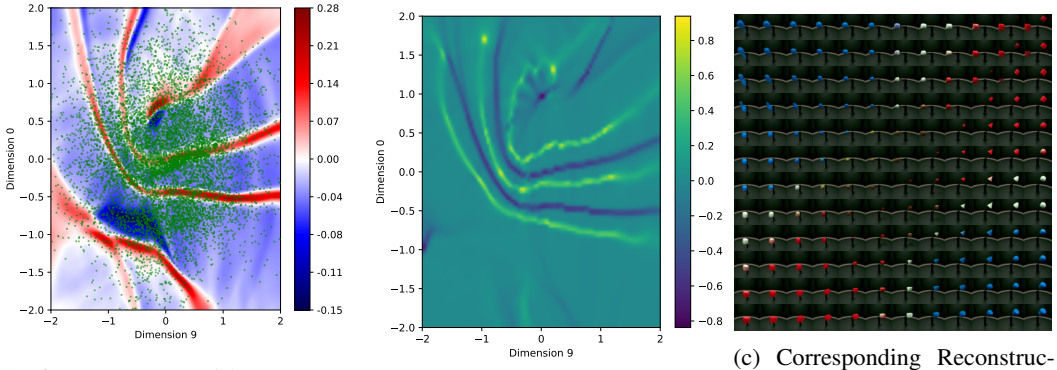


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 23: Visualization of the representation learned by the 1-VAE trained on MPI3D Real.

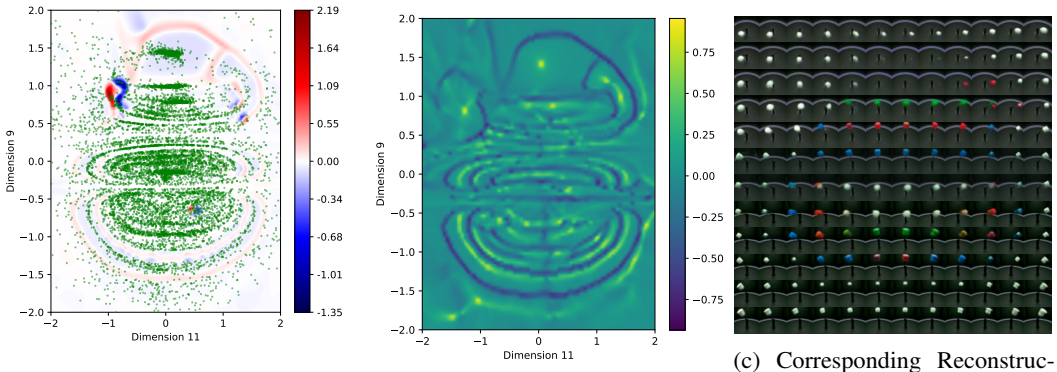


(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

(c) Corresponding Reconstructions

Figure 24: Visualization of the representation learned by the 1-VAE trained on MPI3D Real.



(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map

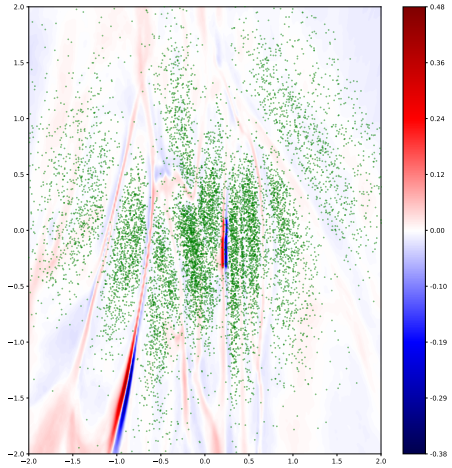
(c) Corresponding Reconstructions

Figure 25: Visualization of the representation learned by the 8-VAE trained on MPI3D Real. Note that due to posterior collapse, the full latent manifold is contained in this projection (see the corresponding response matrix in figure 19).

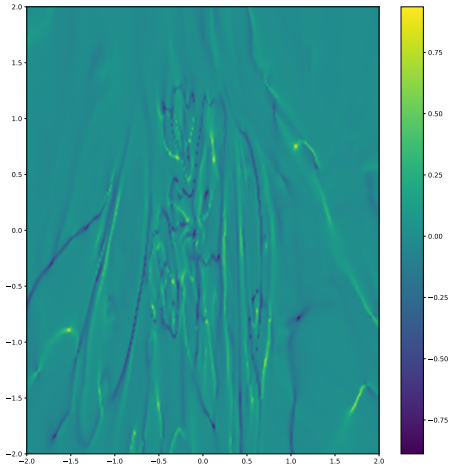
B.3 MNIST

Due to the computational cost of evaluating the response function over a dense grid, we focus our visualizations to 2D projections of the latent space. However, for MNIST and Fashion-MNIST, we train several VAE models to embed the whole representation into two dimensions $d = 2$, so that we can visualize the full representation. While the resulting divergence and curvature maps do not demonstrate as intuitive structure as in the disentangled representations for 3D-Shapes or MPI-3D, we can nevertheless appreciate the learned manifold beyond qualitatively observing reconstructions.

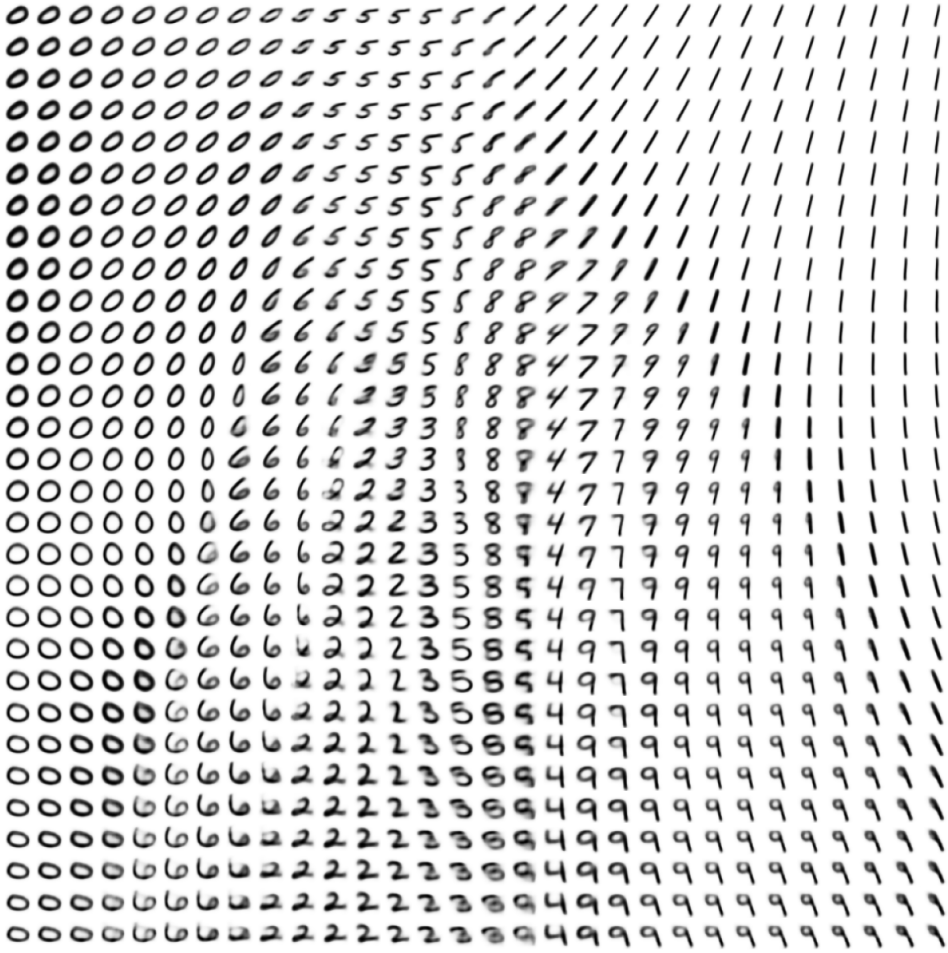
B.3.1 MNIST



(a) Divergence Map and Aggregate Posterior

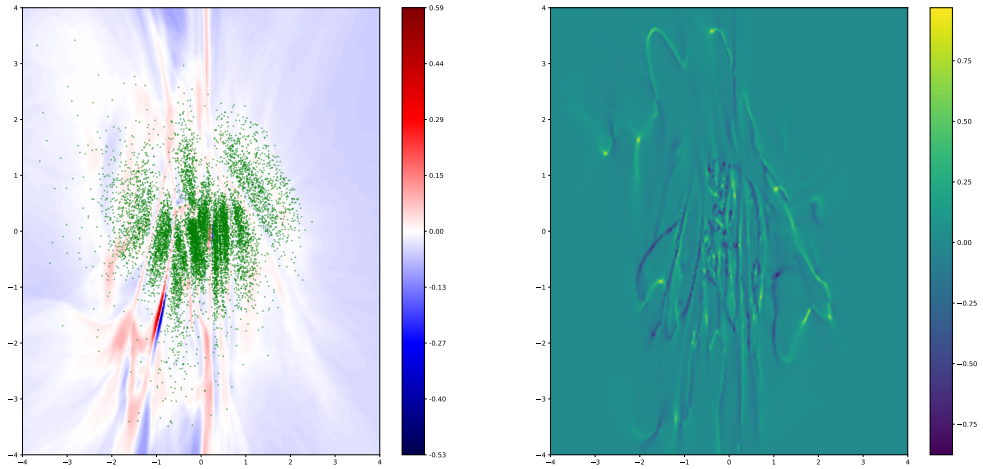


(b) Mean Curvature Map



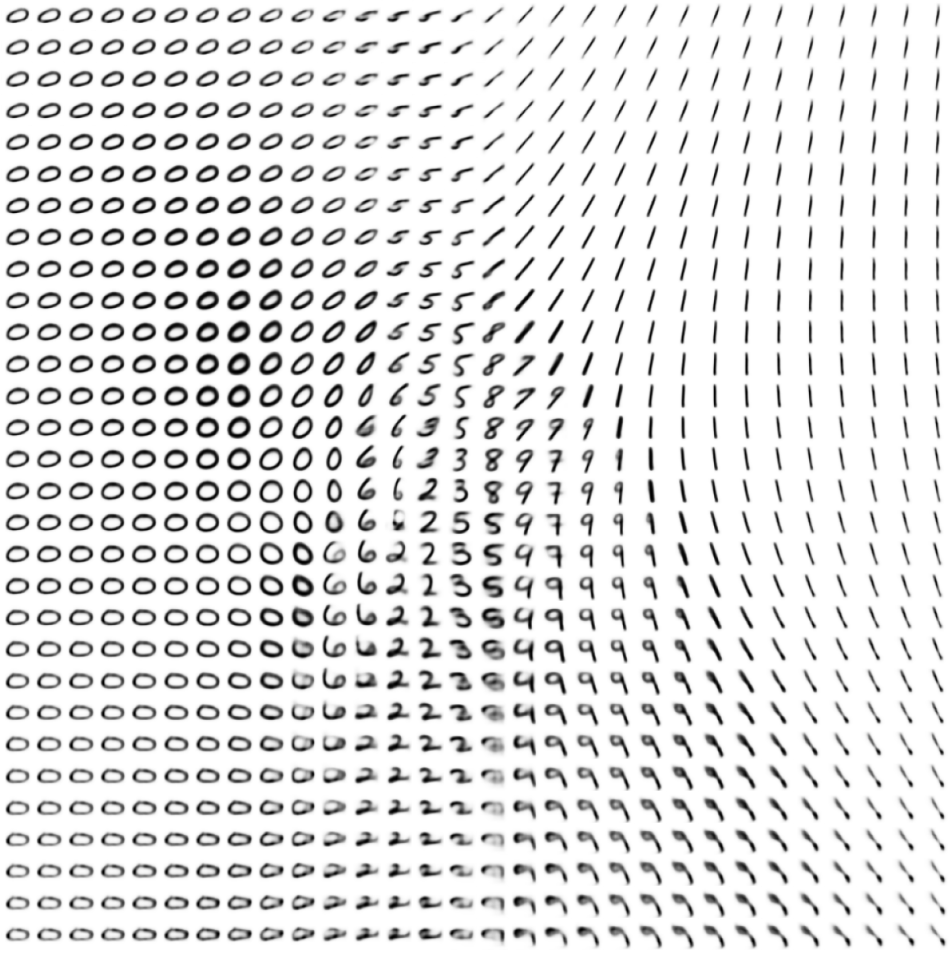
(c) Corresponding Reconstructions

Figure 26: The full latent space for a VAE ($d = 2$) model trained on MNIST. 26a shows the computed divergence of the response field in blue and red while the green points are samples from the aggregate posterior. 26b shows the resulting mean curvature, which identifies 10 points where the curvature spikes and the boundaries between the regions corresponding to different clusters in the posterior. Finally 26c shows the reconstructions over the same region. Note how the high divergence (red) regions correspond to boundaries between significantly different samples (such as changing digit value or stroke thickness).



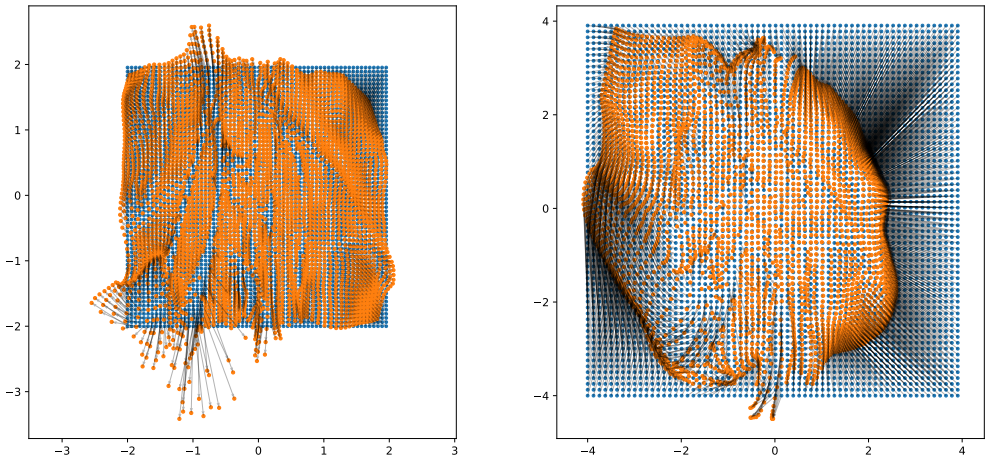
(a) Divergence Map and Aggregate Posterior

(b) Mean Curvature Map



(c) Corresponding Reconstructions

Figure 27: Same plot and model as figure 26, except over a larger range of the latent space $[-4, 4]$. Note that even though the posterior (green dots) is concentrated near the prior (standard normal), reconstructions far away (along the edges of the figure) still look recognizable, demonstrating the exceptional robustness of VAEs to project unexpected latent vectors back onto the learned manifold.

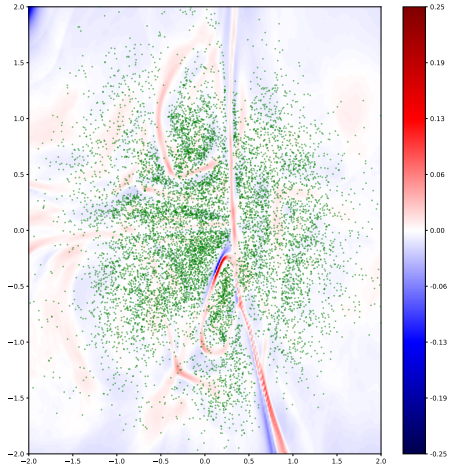


(a) Response field $[-2, 2]$

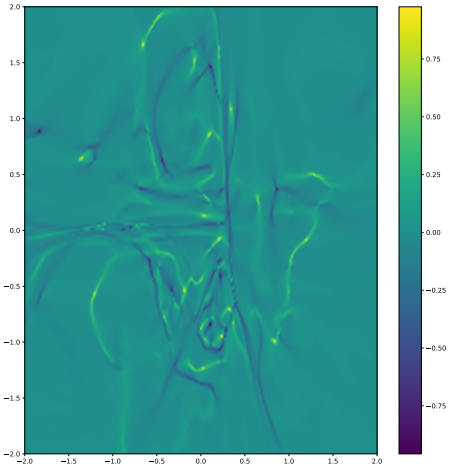
(b) Response field $[-4, 4]$

Figure 28: Response fields for the same model analyzed in figures 26 and 27. The blue dots show the initial latent samples, and the orange dots connected by the black arrows show the corresponding responses (the latent sample after decoding and reencoding).

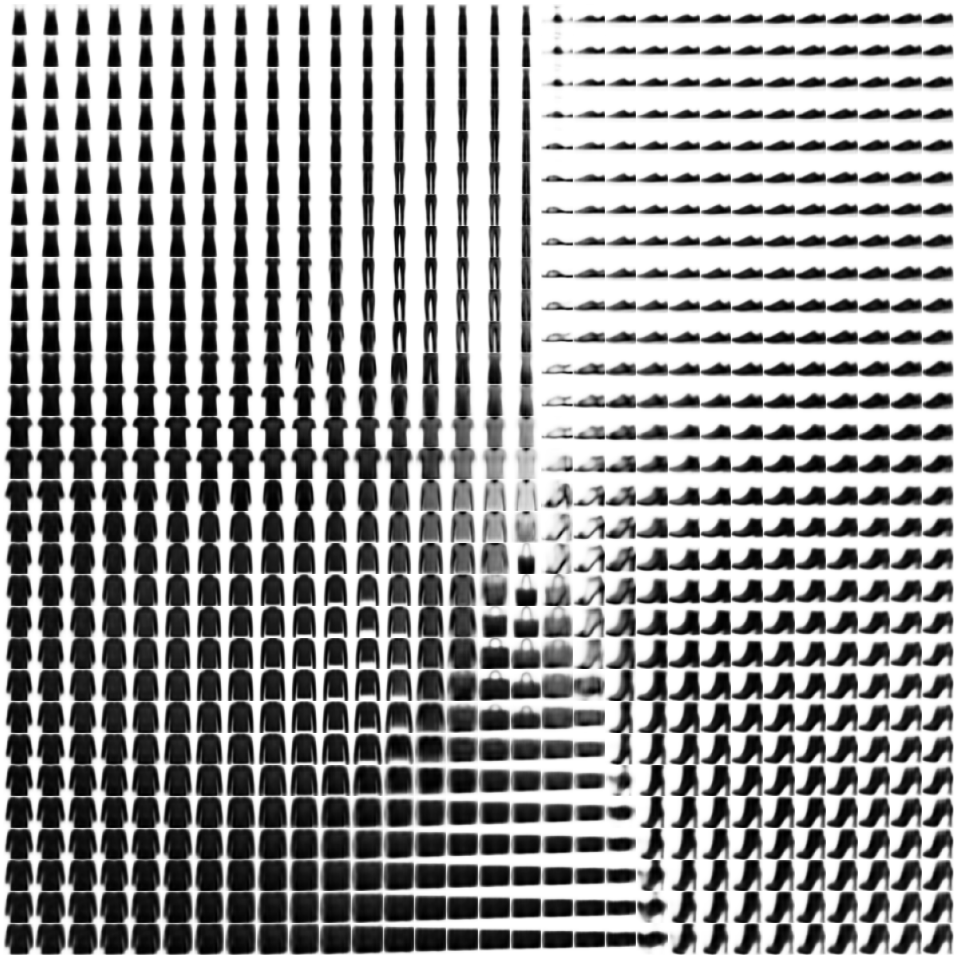
B.3.2 Fashion-MNIST



(a) Divergence Map and Aggregate Posterior

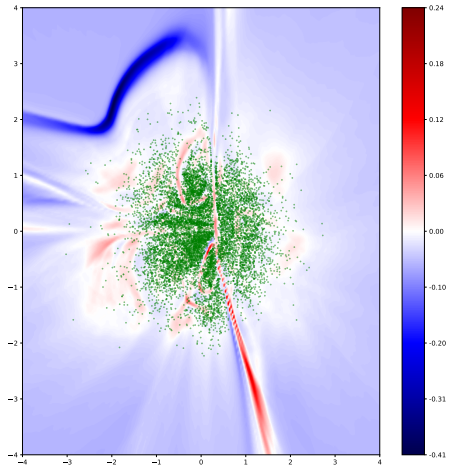


(b) Mean Curvature Map

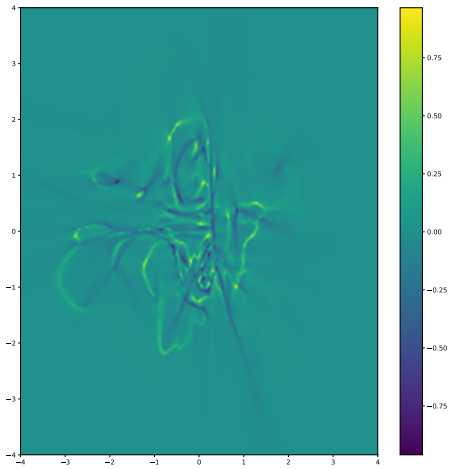


(c) Corresponding Reconstructions

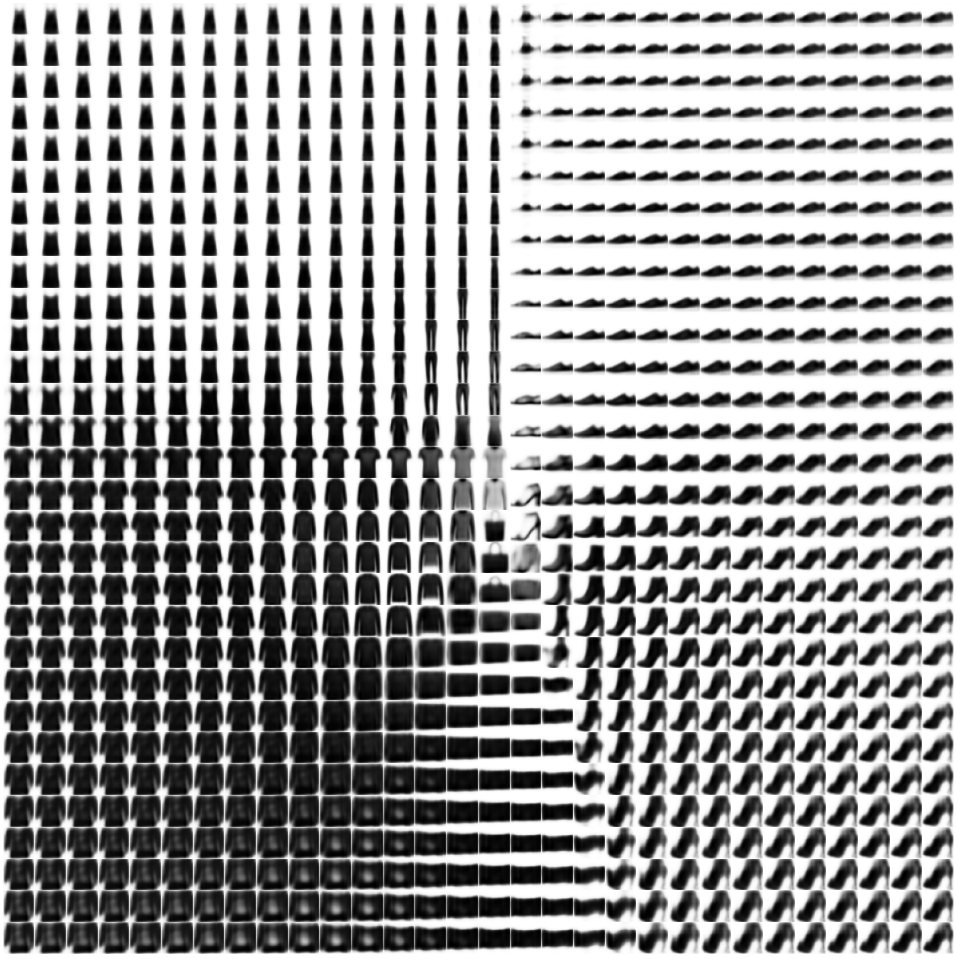
Figure 29: The full latent space for a 8-VAE ($d = 2$) model trained on Fashion-MNIST. 29a shows the computed divergence of the response field in blue and red while the green points are samples from the aggregate posterior. 29b shows the resulting mean curvature, which identifies 10 points where the curvature spikes and the boundaries between the regions corresponding to different clusters in the posterior. Finally 29c shows the reconstructions over the same region. Note how the high divergence (red) regions correspond to boundaries between significantly different samples (such as changing digit value or stroke thickness).



(a) Divergence Map and Aggregate Posterior

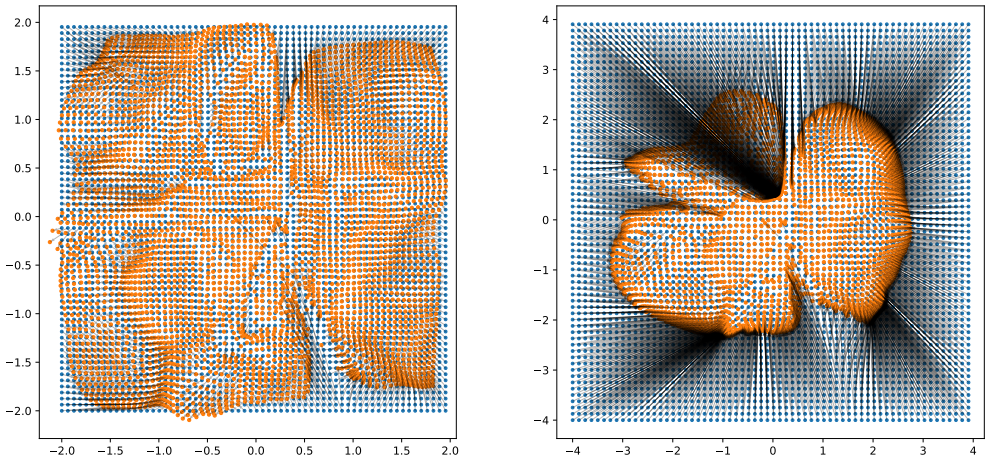


(b) Mean Curvature Map



(c) Corresponding Reconstructions

Figure 30: Same plot and model as figure 29, except over a larger range of the latent space $[-4, 4]$. Note that even though the posterior (green dots) is concentrated near the prior (standard normal), reconstructions far away (along the edges of the figure) still look recognizable, demonstrating the exceptional robustness of VAEs to project unexpected latent vectors back onto the learned manifold.



(a) Response field $[-2, 2]$

(b) Response field $[-4, 4]$

Figure 31: Response fields for the same model analyzed in figures 29 and 30. The blue dots show the initial latent samples, and the orange dots connected by the black arrows show the corresponding responses (the latent sample after decoding and reencoding).

References

[1] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[2] Dana H Ballard. Modular learning in neural networks. In *AAAI*, pages 279–284, 1987.

[3] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes, 2013. URL `arXiv:1312.6114`.

[4] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *arXiv preprint arXiv:1906.00446*, 2019.

[5] James Townsend, Tom Bird, and David Barber. Practical lossless compression with latent variables using bits back coding. *arXiv preprint arXiv:1901.04866*, 2019.

[6] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

[7] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015.

[8] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

[9] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

[10] Ligong Han, Sri Harsha Musunuri, Martin Renqiang Min, Ruijiang Gao, Yu Tian, and Dimitris Metaxas. Ae-stylegan: Improved training of style-based auto-encoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3134–3143, 2022.

[11] Đorđe Miladinović, Aleksandar Stanić, Stefan Bauer, Jürgen Schmidhuber, and Joachim M Buhmann. Spatial dependency networks: Neural layers for improved generative image modeling. *International Conference on Learning Representations (ICLR)*, 2021.

[12] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[13] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[14] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.

[15] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2012. URL `arXiv:1206.5538`.

[16] Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. *arXiv preprint arXiv:2010.14407*, 2020.

[17] Aravind Srinivas, Michael Laskin, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

[18] Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*, 2021.

[19] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[20] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

[21] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019.

[22] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8642–8651, 2020.

[23] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[24] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602, 2021.

[25] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[26] Tao Yang, Georgios Arvanitidis, Dongmei Fu, Xiaogang Li, and Søren Hauberg. Geodesic clustering in deep generative models. *arXiv preprint arXiv:1809.04747*, 2018.

[27] Marissa Connor, Gregory Canal, and Christopher Rozell. Variational autoencoder with learned latent structure. In *International Conference on Artificial Intelligence and Statistics*, pages 2359–2367. PMLR, 2021.

[28] Clément Chadebec, Clément Mantoux, and Stéphanie Allassonnière. Geometry-aware hamiltonian variational auto-encoder. *arXiv preprint arXiv:2010.11518*, 2020.

[29] Nutan Chen, Alexej Klushyn, Francesco Ferroni, Justin Bayer, and Patrick Van Der Smagt. Learning flat latent manifolds with vaes. *arXiv preprint arXiv:2002.04881*, 2020.

[30] Dimitris Kalatzis, Johan Ziruo Ye, Jesper Wohlert, and Søren Hauberg. Multi-chart flows. *arXiv preprint arXiv:2106.03500*, 2021.

[31] Mike Yan Michelis and Quentin Becker. On linear interpolation in the latent space of deep generative models. *arXiv preprint arXiv:2105.03663*, 2021.

[32] Luis A Pérez Rey, Vlado Menkovski, and Jacobus W Portegies. Diffusion variational autoencoders. *arXiv preprint arXiv:1901.08991*, 2019.

[33] Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick Smagt. Metrics for deep generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1550. PMLR, 2018.

[34] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34 (4):18–42, 2017.

[35] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124, 2017.

[36] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[37] 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*, pages 14584–14597, 2019.

[38] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep learning applications in medical image analysis. *Ieee Access*, 6:9375–9389, 2017.

[39] Xiaoran Chen, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. Deep generative models in the real-world: An open challenge from medical imaging. *arXiv preprint arXiv:1806.05452*, 2018.

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[42] Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, 2019.

[43] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. 2018.

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

[45] William F Whitney, Min Jae Song, David Brandfonbrener, Jaan Altosaar, and Kyunghyun Cho. Evaluating representations by the complexity of learning low-loss predictors. *arXiv preprint arXiv:2009.07368*, 2020.

[46] Jakub Tomczak and Max Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018.

[47] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34, 2021.

[48] Tom White. Sampling generative networks. *arXiv preprint arXiv:1609.04468*, 2016.

[49] Shujian Yu and Jose C Principe. Understanding autoencoders with information theoretic concepts. *Neural Networks*, 117:104–123, 2019.

[50] Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[51] James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Understanding posterior collapse in generative latent variable models. 2019.

[52] Danilo Jimenez Rezende and Fabio Viola. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

[53] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dj Dvijotham, and Pushmeet Kohli. Adversarially robust representations with smooth encoders. In *International Conference on Learning Representations*, 2019.

[54] Felix Leeb, Guilia Lanzillotta, Yashas Annadani, Michel Besserve, Stefan Bauer, and Bernhard Schölkopf. Structure by architecture: Disentangled representations without regularization. *arXiv preprint arXiv:2006.07796*, 2020.

[55] A Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. Autoencoding variational autoencoder. *arXiv preprint arXiv:2012.03715*, 2020.

[56] Steve Dias Da Cruz, Bertram Taetz, Thomas Stifter, and Didier Stricker. Autoencoder attractors for uncertainty estimation. *arXiv preprint arXiv:2204.00382*, 2022.

[57] Zijun Zhang, Ruixiang Zhang, Zongpeng Li, Yoshua Bengio, and Liam Paull. Perceptual generative autoencoders. In *International Conference on Machine Learning*, pages 11298–11306. PMLR, 2020.

[58] Giulia Lanzillotta, Felix Leeb, Stefan Bauer, and Bernhard Schölkopf. On the interventional consistency of autoencoders. 2021.

[59] Adityanarayanan Radhakrishnan, Karren Yang, Mikhail Belkin, and Caroline Uhler. Memorization in overparameterized autoencoders. *arXiv preprint arXiv:1810.10333*, 2018.

[60] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Icml*, 2011.

[61] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. *arXiv preprint arXiv:1903.05789*, 2019.

[62] Jan Stühmer, Richard Turner, and Sebastian Nowozin. Independent subspace analysis for unsupervised learning of disentangled representations. In *International Conference on Artificial Intelligence and Statistics*, pages 1200–1210. PMLR, 2020.

[63] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.

[64] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[65] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019.

[66] Holger Theisel. *Vector field curvature and applications*. PhD thesis.

[67] Chris Burgess and Hyunjik Kim. 3d shapes dataset, 2018.

[68] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[69] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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

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

[72] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

[73] Georgios Arvanitidis, Soren Hauberg, Philipp Hennig, and Michael Schober. Fast and robust shortest paths on manifolds learned from data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1506–1515. PMLR, 2019.

[74] Dimitris Kalatzis, David Eklund, Georgios Arvanitidis, and Søren Hauberg. Variational autoencoders with riemannian brownian motion priors. *arXiv preprint arXiv:2002.05227*, 2020.

[75] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

[76] Luca Falorsi, Pim de Haan, Tim R Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S Cohen. Explorations in homeomorphic variational auto-encoding. *arXiv preprint arXiv:1807.04689*, 2018.

[77] Tong Lin and Hongbin Zha. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809, 2008.

[78] Alessandra Tosi, Søren Hauberg, Alfredo Vellido, and Neil D Lawrence. Metrics for probabilistic geometries. *arXiv preprint arXiv:1411.7432*, 2014.