β-Annealed Variational Autoencoder for glitches

Sivaramakrishnan Sankarapandian
Proscia Inc.
siva@proscia.com

Brian Kulis
Department of ECE
Boston University
bkulis@bu.edu

Abstract

Gravitational wave detectors such as LIGO and Virgo are susceptible to various types of instrumental and environmental disturbances known as glitches which can mask and mimic gravitational waves. While there are 22 classes of non-Gaussian noise gradients currently identified, the number of classes is likely to increase as these detectors go through commissioning between observation runs. Since identification and labelling new noise gradients can be arduous and time-consuming, we propose β-Annelead VAEs to learn representations from spectograms in an unsupervised way. Using the same formulation as [1], we view Bottleneck-VAEs [2] through the lens of information theory and connect them to β-VAEs [3]. Motivated by this connection, we propose an annealing schedule for the hyperparameter β in β-VAEs which has advantages of: 1) One fewer hyperparameter to tune, 2) Better reconstruction quality, while producing similar levels of disentanglement.

1 Introduction

Gravitational waves are a cosmic phenomenon that are a result of the collision of highly dense objects. Study of gravitational waves has become possible with ultra sensitive instruments such as the Laser Interferometer Gravitational-Wave Observatory (LIGO) [4] and Virgo [5]. These detectors can detect changes in length caused by gravitational waves less than the width of a proton [6]. Naturally, these hyper sensitive instruments are prone to instrumental and environmental disturbances such as non Gaussian transients known as *glitches* which can mimic gravitational waves. The frequency of occurrence of these glitches is so high that the chance of these glitches masking a gravitational wave is non-negligible [7]. It is important to identify these glitches and eliminate them for proper gravitational wave detection. Project *Gravity Spy* [8] is an effort to identify and categorize these glitches into different classes based on the morphology of spectrograms with help from citizen scientists. While these glitches can be categorized into 22 classes currently, there is a possibility that new classes of glitches might get added in the future as the detectors undergo commissioning before each observation run [9, 10].

There have been previous attempts [11, 12, 13] to classify glitches using supervised deep learning techniques, but in this work we take an unsupervised representation learning approach. Unsupervised representation learning can help alleviate the need for a large amount of labelled data and the need to identify new classes of glitches as they appear during the operation of LIGO. Disentangled representation learning, as a branch of unsupervised representation learning, has several advantages, as pointed out in [14]: invariance, transferability, interpretability, and conditioning and intervention. A large portion of recent literature on disentanglement learning is based on Variational Autoencoders (VAEs). Higgins et al. [3] introduce β-VAEs, which penalizes the KL divergence between the variational posterior and the prior using the hyperparameter β. Bottleneck-VAEs [2] increase the capacity of the information bottleneck as the training progresses thus offering better reconstructions than β-VAEs. In this work, we show that Bottleneck-VAEs and β-VAEs are closely connected and propose a decreasing schedule for the hyperparameter β in β-VAEs that controls the information

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

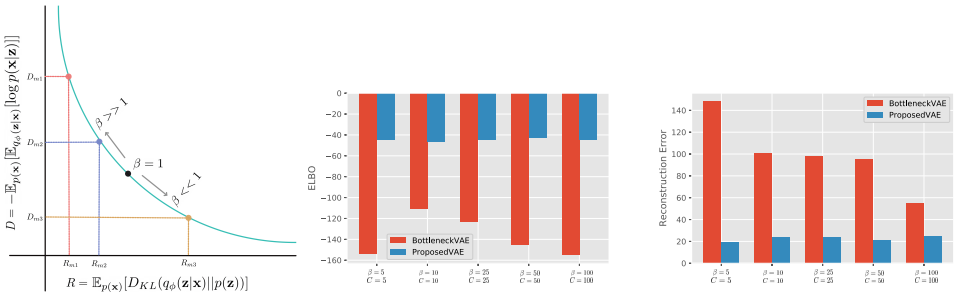

Figure 1: (**Left**) Training Bottleneck VAEs with different values of C equal to *Rate* $R_{m_1}, R_{m_2}, R_{m_3}$ (with $\gamma = 1$) corresponds to VAEs converging to points corresponding to *Distortion* equal to $D_{m_1}, D_{m_2}, D_{m_3}$ in the RD-curve. ELBO (**Center**) and reconstruction error (**Right**) for different hyperparameters in Bottleneck and Proposed VAEs on dSpirtes.

capacity similar to the hyperparameter C in the objective function of Bottleneck-VAEs. In addition, we provide experimental evidence on *Gravity Spy* dataset to show superior performance of our proposed VAE in unsupervised learning of glitches and advantages of using our proposed VAE when compared to Bottleneck-VAEs.

2 VAE, β-VAE and Bottleneck-VAE

The generative model of our data is defined as $p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}) = p(\boldsymbol{x}, \boldsymbol{z})$ where each observed datapoint $\boldsymbol{x}$ is assumed to be generated from its own latent variable $\boldsymbol{z}$. VAEs attempt to maximize the marginal likelihood of the data:

$$\log p_\theta(\boldsymbol{x}_1, ..., \boldsymbol{x}_N) = \sum_{i=1}^{N} \log p_\theta(\boldsymbol{x_i}).$$

Due to its intractability, a variational distribution q_ϕ is introduced to approximate the posterior $p(\boldsymbol{z}|\boldsymbol{x})$, which gives rise to the lower bound on the marginal likelihood called the *Evidence Lower Bound* (ELBO):

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

The integration in the first term is usually computed using samples from $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and backpropogation through the sampling process is done through the *reparametrization trick* [15]. In practice, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is assumed to be a Gaussian distribution with diagonal covariance and $p(\boldsymbol{z})$ to be $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

β-VAEs [3] are variants of regular VAEs that introduce a hyperparameter called β to the ELBO:

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})). \tag{2}$$

For $\beta > 1$, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is heavily constrained to be closer to the factorized prior $p(\boldsymbol{z})$. Heavy penalty on the D_{KL} encourages disentanglement, while at the same time leads to poor reconstruction quality. This is due to the fact that the latent factors are not able to encode enough information about the observations.

Burgess et al. [2] proposed an alternate objective with an additional hyperparameter C; we call this variant of VAE as Bottleneck-VAE:

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \gamma|D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) - C|. \tag{3}$$

When $C = 0$, the objective is the same as β-VAE, since $D_{KL} \geq 0$. In Bottleneck-VAEs, C is progressively increased during training (with γ kept constant, typically greater than one) to increase the amount of information stored about observations in the latent codes. This results in two effects: 1) As the information capacity is increased (through C) during training, the encoder learns to encode latent dimensions in the order of decreasing returns to log-likelihood. 2) This controlled capacity increase also encourages better reconstruction quality compared to β-VAEs, while achieving similar levels of disentanglement.

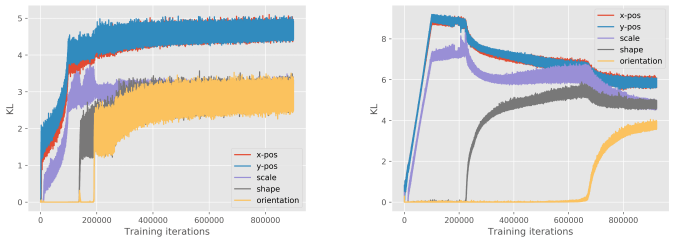

Figure 2: KL-divergence of each latent dimension with respect to a unit Gaussian during training on dSprites. **Left:** In β-Annealed VAE, β is decreased as the training progresses **Right:** In Bottleneck-VAE, C is increased as the training progresses to increase the information capacity

Table 1: Results on unsupervised representation learning of non Gaussian noise transients that occur in gravitational wave detectors

Model	Accuracy
β-VAE	61.26%
Bottleneck-VAE	80.01%
Proposed-VAE	**81.60%**

3 β-Annealed VAE

It is important to note that the objective functions corresponding to β-VAEs and Bottleneck-VAEs do not optimize the ELBO when $\beta > 1$ and $\gamma > 1, C > 0$, respectively. [1] offers an information theoretic perspective, in that β-VAEs try to find the optimal *distortion* (D) and *rate* (R) for a fixed $\beta = \frac{\partial D}{\partial R}$ by minimizing $\min_{q_\phi(\boldsymbol{z}|\boldsymbol{x}),p(\boldsymbol{z}),p_\theta(\boldsymbol{x}|\boldsymbol{z})} D + \beta R$, where $D = -\mathbb{E}_{p(\boldsymbol{x})}[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]]$ and $R = \mathbb{E}_{p(\boldsymbol{x})}[D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))]$. The inequality $H - D \leq \mathcal{I}(X;Z) \leq R$ from [1] shows the relationship between D, R, *data entropy* H ($-\mathbb{E}_{p(\boldsymbol{x})}[\log p(\boldsymbol{x})]$) and *mutual information* $\mathcal{I}$ ($D_{KL}(p(\boldsymbol{x},\boldsymbol{z})||p(\boldsymbol{x})p(\boldsymbol{z}))$). For a finite capacity encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, vanilla VAEs correspond to an operating point on the green curve (in Figure.1) with slope 1. Fixing the capacity of the encoder and decoder, if β is varied from greater than 1 to less than 1, the operating point shifts from $\uparrow D, \downarrow R$ to $\downarrow D, \uparrow R$ along the green curve (with $\uparrow$ & $\downarrow$ denoting high and low respectively).

We first view Bottleneck-VAE from an information theoretic standpoint. If we set $\gamma = 1$ and a constant $C = R_m$, optimizing (3) can be viewed as minimizing D for a constant $R = R_m$. From Figure 1, we can see that this corresponds to a point on the RD curve where $R = R_m$. For different increasing values of C, the point shifts to locations in the RD curve corresponding to increasing R. Concretely, increasing C corresponds to relaxing the constraint that $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ needs to be closer in terms of KL-divergence to the prior $p(\boldsymbol{z})$), and [2] showed that when $\gamma > 1$, this leads to better robust disentanglement and better reconstruction quality.

Since any point on the RD-curve for a fixed encoder and decoder is reachable through β, controllable information capacity can be achieved through β. Motivated by the monotonically increasing schedule of C in case of Bottleneck-VAEs, we propose a monotonically decreasing schedule of β for β-VAEs. The effect on distortion and rate while decreasing β in case of β-VAEs is very similar to the effect of increasing C in Bottleneck-VAEs. We formalize our claim in the following lemma (Proof in Appendix),

Lemma 3.1 *(For a fixed finite capacity encoder and decoder) Let $D^*_{C_1}, D^*_{C_2}$ and $R^*_{C_1}, R^*_{C_2}$ denote the optimal distortion and rate for a Bottleneck-VAE with C_1, C_2 respectively with a constant $\gamma \geq 0$. Similarly let $D^*_{\beta_1}, D^*_{\beta_2}$ and $R^*_{\beta_1}, R^*_{\beta_2}$ denote the optimal distortion and rate for a β-VAE with β_1, β_2 respectively. If $C_1 > C_2 \geq 0$, then $R^*_{C_1} > R^*_{C_2}$ and $D^*_{C_1} < D^*_{C_2}$. Similarly, with respect to β-VAEs, if $0 \leq \beta_1 < \beta_2$, then $R^*_{\beta_1} > R^*_{\beta_2}$ and $D^*_{\beta_1} < D^*_{\beta_2}$.*

If we want to replicate similar effects of linearly increasing C in the case of Bottleneck-VAEs, a β-VAE can be trained with monotonically decreasing β from $\beta \gg 1$ to $\beta \ll 1$. We use linearly decreasing schedule for β in all of our experiments. When compared to Bottleneck-VAEs, a linearly decreasing schedule of β in β-VAEs (which we call β-Annealed VAEs) offers advantages such as: 1) without having to set C, our proposed schedule have one less hyperparameter to tune; 2) in all of our experiments, we linearly decreasing β from $\beta \gg 1$ to $\beta = 1$ during training, which can be interpreted as β-VAEs are trained as vanilla VAEs during later stages in training leading to better reconstruction error.

Table 2: Quantitative assessment of disentanglement in dSprites

VAE Variant	Hyperparameter	BetaVAE Score	FactorVAE Score	MIG	DCI Disentanglement	Modularity	SAP
β-VAE [3]	β=1	0.851	0.685	0.072	0.127	0.790	0.052
	β=4	0.816	0.627	0.078	0.138	0.800	0.028
	β=16	0.742	0.546	0.141	0.277	0.809	0.010
Bottleneck-VAE [2]	C=5	0.868	0.596	**0.334**	0.402	0.791	**0.078**
	C=25	0.765	0.539	0.025	0.059	0.769	0.022
	C=100	0.625	0.369	0.014	0.022	0.746	0.007
Factor-VAE [16]	γ=10	0.862	0.706	0.144	0.221	0.781	0.068
	γ=30	0.878	**0.849**	0.190	0.328	0.796	0.068
	γ=100	0.862	0.792	0.312	**0.461**	0.820	0.062
β-TCVAE [17]	β=1	0.851	0.685	0.072	0.127	0.790	0.052
	β=4	0.875	0.830	0.226	0.347	0.805	0.064
	β=10	0.879	0.808	0.287	0.447	0.818	0.067
DIP-VAE-I [14]	λ_{od}=1	0.846	0.645	0.094	0.127	0.779	0.053
	λ_{od}=5	0.804	0.574	0.040	0.077	0.783	0.025
	λ_{od}=50	0.783	0.599	0.034	0.077	0.778	0.016
DIP-VAE-II [14]	λ_{od}=1	0.720	0.479	0.015	0.083	0.782	0.004
	λ_{od}=5	0.793	0.644	0.049	0.108	0.798	0.016
	λ_{od}=50	0.869	0.544	0.087	0.177	0.809	0.058
Proposed VAE	β=5	0.846	0.809	0.073	0.180	0.815	0.038
	β=25	0.739	0.599	0.111	0.233	0.790	0.034
	β=50	**0.902**	0.805	0.289	0.397	**0.832**	0.076

4 Results

We perform experiments to indicate β-Annealed VAEs behave similarly to Bottleneck-VAEs when the information capacity is increased. Then we compare β-Annealed VAEs and Bottleneck-VAEs in terms of ELBO, reconstruction error and disentanglement on the dSprites [18] (qualitative assessment of disentanglement can be found in Appendix).

We first show that the two effects of linearly increasing C in Bottleneck-VAEs (with $\gamma > 1$ kept constant) can be achieved using linearly decreasing β in β-VAEs. We use the same architecture of encoder and decoder used in [2] and trained a β-VAE with linearly decreasing β from 100 to 1 (with iteration threshold being 100000). Figure 2 shows the D_{KL} of each latent dimension $q(z|\boldsymbol{x})$ to its prior (standard normal distribution), we see that the generative factors are learned one at a time by the network in the order of decreasing returns to the log-likelihood, similar to Bottleneck-VAEs. To show that β-Annealed VAEs achieve better reconstruction error because they are trained as vanilla VAEs during the later stages of training (i.e after β is reduced to 1), we perform experiments with different values of β and C, and Figure 1 (*right*) shows the reconstruction error on dSpirtes for Bottleneck-VAEs and β-VAEs. We see that our proposed linearly decreasing schedule of β offers better reconstruction error than Bottleneck-VAEs. Figure 1 (*center*) also shows that β-Annealed VAEs achieve better ELBO than Bottleneck-VAEs. To quantitatively assess disentanglement offered by our proposed linear decreasing schedule of β in β-VAEs, we used the metrics *β-VAE metric* [3], *Factor VAE metric* [16], *Mutual Information Gap (MIG)* [17], *Modularity* [19], *DCI Disentanglement* [20] and *SAP score* [14], similar to [21]. We show the disentanglement performance of the proposed method with the following existing variants of VAEs: 1) β-VAE, 2) FactorVAE, 3) TCVAE, 4)DIP-VAE-I, 5)DIP-VAE-II, 6) Bottleneck-VAE in Table 2.

Further, we train β-VAE, Bottleneck-VAE and β-Annealed VAE on the *Gravity Spy* [22] dataset, which contains spectrogram samples of 22 different types of glitches. We check the quality of representations learnt by the encoders by training linear classifiers trained on top of latent representations and their performance are as shown in Table 1. We see that our proposed VAE learns better representations when compared to β-VAEs and Bottleneck-VAEs.

5 Conclusion

We introduce β-Annealed VAEs motivated by viewing Bottlenck-VAEs through the lens of information theory. We show that our proposed version of β-VAEs, with linearly decreasing β as the training progresses, offers similar robust disentanglement while having better reconstruction error. We prove its efficacy in learning representations of glitches in LIGO / Virgo detectors.

6 Broader Impact

Beyond unsupervised representation learning of glitches in gravitational wave detectors, β-Annealed VAEs can be used in applications requiring disentanglement of generative factors of data. Since our proposed VAEs have lower reconstruction errors, they can be used in applications where sample quality is important. We believe this work could encourage the ML community to delve into unsupervised learning techniques for the detection and study of glitches. We see research opportunities in devising specific types of VAEs after closely studying the characteristics of glitches and developing a standard benchmark to test different models.

References

[1] Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. *arXiv preprint arXiv:1711.00464*, 2017.

[2] 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.

[3] 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. *ICLR*, 2(5):6, 2017.

[4] Junaid Aasi, BP Abbott, Richard Abbott, Thomas Abbott, MR Abernathy, Kendall Ackley, Carl Adams, Thomas Adams, Paolo Addesso, RX Adhikari, et al. Advanced ligo. *Classical and quantum gravity*, 32(7):074001, 2015.

[5] F Acernese, M Agathos, K Agatsuma, D Aisa, N Allemandou, A Allocca, J Amarni, P Astone, G Balestri, G Ballardin, et al. Advanced virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity*, 32(2):024001, 2014.

[6] Daniel Sigg. The advanced ligo detectors in the era of first discoveries. In *Interferometry XVIII*, volume 9960, page 996009. International Society for Optics and Photonics, 2016.

[7] Kevin Crowston. Gravity spy: Humans, machines and the future of citizen science. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 163–166, 2017.

[8] Michael Zevin, Scott Coughlin, Sara Bahaadini, Emre Besler, Neda Rohani, Sarah Allen, Miriam Cabero, Kevin Crowston, Aggelos K Katsaggelos, Shane L Larson, et al. Gravity spy: integrating advanced ligo detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity*, 34(6):064003, 2017.

[9] Benjamin P Abbott, R Abbott, TD Abbott, MR Abernathy, F Acernese, K Ackley, M Adamo, C Adams, T Adams, P Addesso, et al. Characterization of transient noise in advanced ligo relevant to gravitational wave signal gw150914. *Classical and Quantum Gravity*, 33(13): 134001, 2016.

[10] Benjamin P Abbott, R Abbott, TD Abbott, MR Abernathy, K Ackley, C Adams, P Addesso, RX Adhikari, VB Adya, C Affeldt, et al. Calibration of the advanced ligo detectors for the discovery of the binary black-hole merger gw150914. *Physical Review D*, 95(6):062003, 2017.

[11] Hongyu Shen, Daniel George, Eliu Huerta, et al. Glitch classification and clustering for ligo with deep transfer learning. *APS*, 2018:L01–027, 2018.

[12] Robert E Colgan, K Rainer Corley, Yenson Lau, Imre Bartos, John N Wright, Zsuzsa Márka, and Szabolcs Márka. Efficient gravitational-wave glitch identification from environmental data through machine learning. *Physical Review D*, 101(10):102003, 2020.

[13] Sara Bahaadini, Vahid Noroozi, Neda Rohani, Scott Coughlin, Michael Zevin, Joshua R Smith, Vicky Kalogera, and A Katsaggelos. Machine learning for gravity spy: Glitch classification and dataset. *Information Sciences*, 444:172–186, 2018.

[14] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

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

[16] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[17] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[18] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[19] Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pages 185–194, 2018.

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

[21] 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.

[22] Michael Zevin, Scott Coughlin, Sara Bahaadini, Emre Besler, Neda Rohani, Sarah Allen, Miriam Cabero, Kevin Crowston, Aggelos K Katsaggelos, Shane L Larson, et al. Gravity spy: Integrating advanced ligo detector characterization, machine learning, and citizen science. *arXiv preprint arXiv:1611.04596*, 2016.

A Qualitative assessment

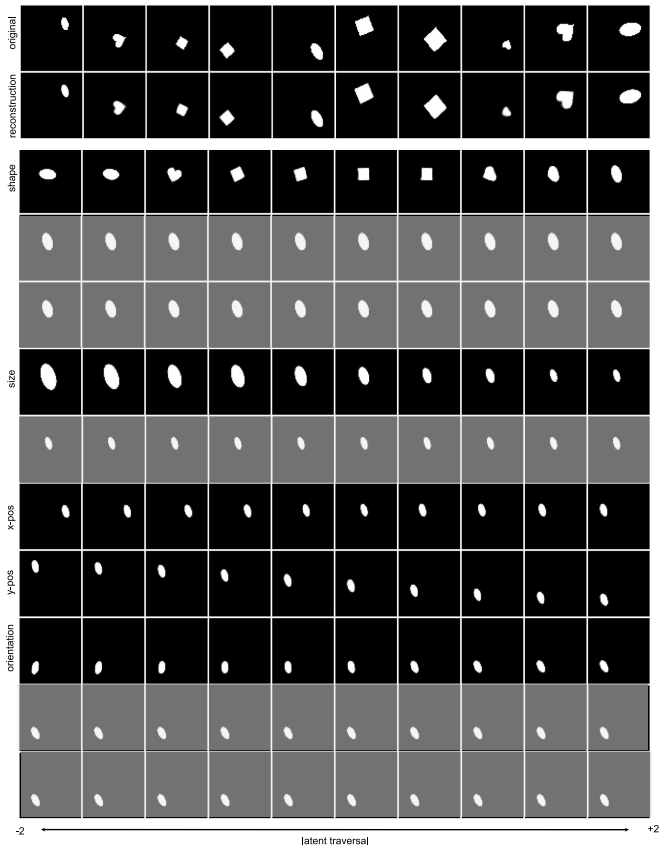

Figure 3: First two rows: Original data and their reconstructions **Other rows:** all 10 latent dimension traversals with captured attribute indicated in the sides. Greyed out rows indicate dead dimensions.

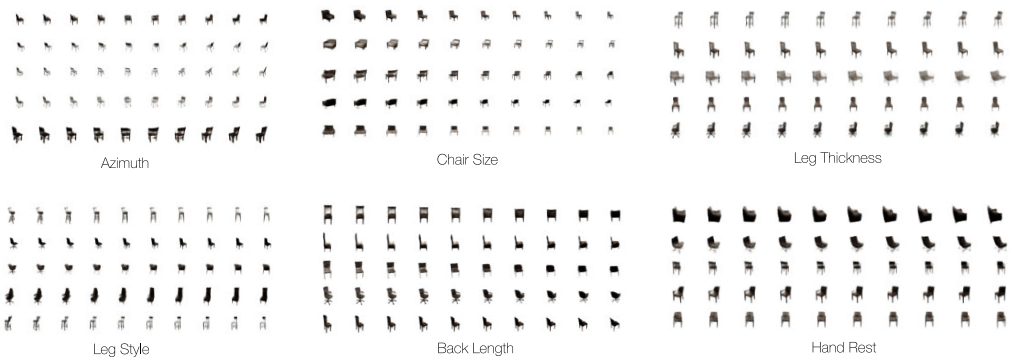

Figure 4: Latent traversals of different latent dimensions on 3DChairs dataset with traversals in the range [-2, 2] using β-Annealed VAEs with $\beta = 50$

B Proofs

Lemma B.1 *(For a fixed finite capacity encoder and decoder) Let $D^*_{C_1}, D^*_{C_2}$ and $R^*_{C_1}, R^*_{C_2}$ denote the optimal distortion and rate for a Bottleneck-VAE with C_1, C_2 respectively with a constant $\gamma \geq 0$. Similarly let $D^*_{\beta_1}, D^*_{\beta_2}$ and $R^*_{\beta_1}, R^*_{\beta_2}$ denote the optimal distortion and rate for a β-VAE with β_1, β_2 respectively. If $C_1 > C_2 \geq 0$, then $R^*_{C_1} > R^*_{C_2}$ and $D^*_{C_1} < D^*_{C_2}$, similarly with respect to β-VAEs, if $0 \leq \beta_1 < \beta_2$, then $R^*_{\beta_1} > R^*_{\beta_2}$ and $D^*_{\beta_1} < D^*_{\beta_2}$*

Proof. From the objective function of Bottleneck-VAEs,

$$\min_{q_\phi(\boldsymbol{z}|\boldsymbol{x}), p(\boldsymbol{z}), p(\boldsymbol{x}|\boldsymbol{z})} D + \gamma|R - C|$$

one can see that the optimal values for R, when $C = C_1$ is $R^*_{C_1} = C_1$ (similarly, when $C = C_2$, $R^*_{C_2} = C_2$). If $C_1 > C_2$ then $R^*_{C_1} > R^*_{C_2}$. Also,

$$\begin{aligned} H - D^*_{C_2} &\leq R^*_{C_2} < R^*_{C_1} \\ H - D^*_{C_2} &< H - D^*_{C_1} \\ D^*_{C_1} &< D^*_{C_2} \end{aligned}$$

For β-VAEs, if $\beta_1 < \beta_2$, then $R^*_{\beta_1} > R^*_{\beta_2}$ and $D^*_{\beta_1} < D^*_{\beta_2}$ is a direct result from [1]. β-VAEs with a fixed architecture and finite capacity can be used to interpolate between auto-encoding behaviour ($\uparrow D, \downarrow R$) to auto-decoding ($\downarrow D, \uparrow R$) behaviour by changing from $\beta << 1$ to $\beta >> 1$.