β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, Alexander Lerchner
Google DeepMind
{irinah,lmatthey,arkap,cpburgess,glorotx,
botvinick,shakir,lerchner}@google.com

ABSTRACT

Learning an interpretable factorised representation of the independent data generative factors of the world without supervision is an important precursor for the development of artificial intelligence that is able to learn and reason in the same way that humans do. We introduce β-VAE, a new state-of-the-art framework for automated discovery of interpretable factorised latent representations from raw image data in a completely unsupervised manner. Our approach is a modification of the variational autoencoder (VAE) framework. We introduce an adjustable hyperparameter β that balances latent channel capacity and independence constraints with reconstruction accuracy. We demonstrate that β-VAE with appropriately tuned $\beta > 1$ qualitatively outperforms VAE ($\beta = 1$), as well as state of the art unsupervised (InfoGAN) and semi-supervised (DC-IGN) approaches to disentangled factor learning on a variety of datasets (*celebA*, *faces* and *chairs*). Furthermore, we devise a protocol to quantitatively compare the degree of disentanglement learnt by different models, and show that our approach also significantly outperforms all baselines quantitatively. Unlike InfoGAN, β-VAE is stable to train, makes few assumptions about the data and relies on tuning a single hyperparameter β, which can be directly optimised through a hyperparameter search using weakly labelled data or through heuristic visual inspection for purely unsupervised data.

1 INTRODUCTION

The difficulty of learning a task for a given machine learning approach can vary significantly depending on the choice of the data representation. Having a representation that is well suited to the particular task and data domain can significantly improve the learning success and robustness of the chosen model (Bengio et al., 2013). It has been suggested that learning a disentangled representation of the generative factors in the data can be useful for a large variety of tasks and domains (Bengio et al., 2013; Ridgeway, 2016). A disentangled representation can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors (Bengio et al., 2013). For example, a model trained on a dataset of 3D objects might learn independent latent units sensitive to single independent data generative factors, such as object identity, position, scale, lighting or colour, thus acting as an inverse graphics model (Kulkarni et al., 2015). In a disentangled representation, knowledge about one factor can generalise to novel configurations of other factors. According to Lake et al. (2016), disentangled representations could boost the performance of state-of-the-art AI approaches in situations where they still struggle but where humans excel. Such scenarios include those which require knowledge transfer, where faster learning is achieved by reusing learnt representations for numerous tasks; zero-shot inference, where reasoning about new data is enabled by recombining previously learnt factors; or novelty detection.

Unsupervised learning of a disentangled posterior distribution over the underlying generative factors of sensory data is a major challenge in AI research (Bengio et al., 2013; Lake et al., 2016). Most previous attempts required a priori knowledge of the number and/or nature of the data generative factors (Hinton et al., 2011; Rippel & Adams, 2013; Reed et al., 2014; Zhu et al., 2014; Yang et al., 2015; Goroshin et al., 2015; Kulkarni et al., 2015; Cheung et al., 2015; Whitney et al., 2016; Karaletsos et al., 2016). This is not always feasible in the real world, where the newly initialised learner may be exposed to complex data where no a priori knowledge of the generative factors exists, and little to no supervision for discovering the factors is available. Until recently purely unsupervised

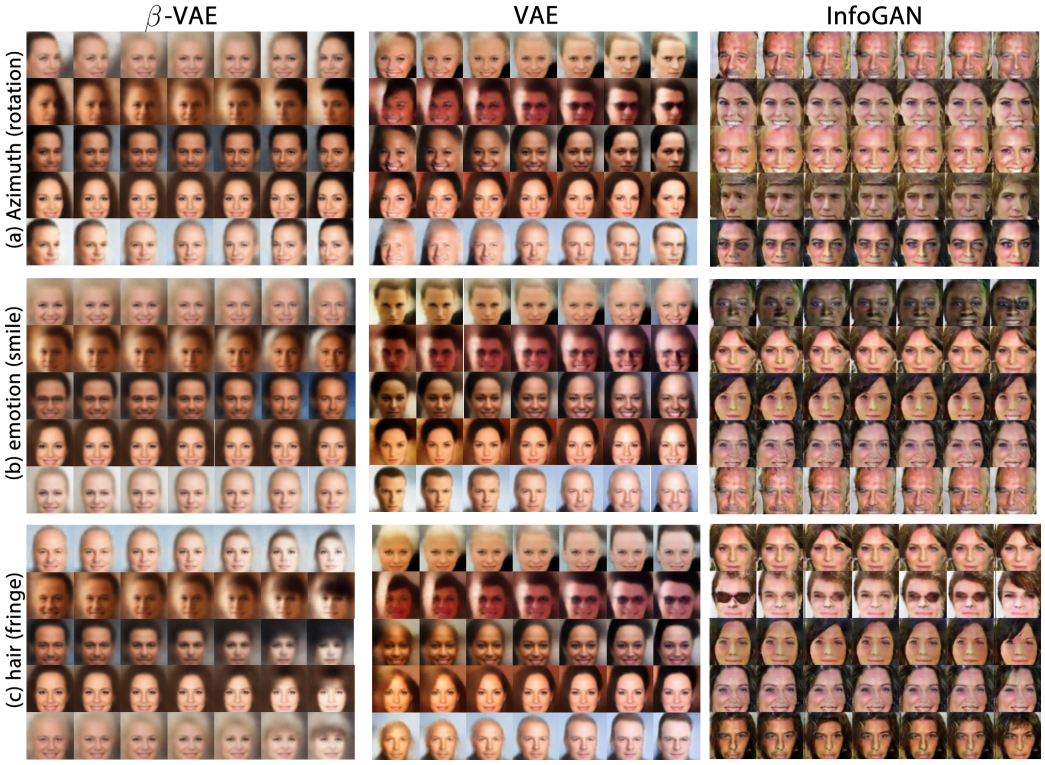

Figure 1: **Manipulating latent variables on celebA:** Qualitative results comparing disentangling performance of β-VAE ($\beta = 250$), VAE (Kingma & Welling, 2014) ($\beta = 1$) and InfoGAN (Chen et al., 2016). *In all figures of latent code traversal each block corresponds to the traversal of a single latent variable while keeping others fixed to either their inferred (β-VAE, VAE and DC-IGN where applicable) or sampled (InfoGAN) values. Each row represents a different seed image used to infer the latent values in the VAE-based models, or a random sample of the noise variables in InfoGAN. β-VAE and VAE traversal is over the [-3, 3] range.* InfoGAN traversal is over ten dimensional categorical latent variables. Only β-VAE and InfoGAN learnt to disentangle factors like azimuth (a), emotion (b) and hair style (c), whereas VAE learnt an entangled representation (e.g. azimuth is entangled with emotion, presence of glasses and gender). InfoGAN images adapted from Chen et al. (2016). Reprinted with permission.

approaches to disentangled factor learning have not scaled well (Schmidhuber, 1992; Desjardins et al., 2012; Tang et al., 2013; Cohen & Welling, 2014; 2015).

Recently a scalable unsupervised approach for disentangled factor learning has been developed, called InfoGAN (Chen et al., 2016). InfoGAN extends the generative adversarial network (GAN) (Goodfellow et al., 2014) framework to additionally maximise the mutual information between a subset of the generating noise variables and the output of a recognition network. It has been reported to be capable of discovering at least a subset of data generative factors and of learning a disentangled representation of these factors. The reliance of InfoGAN on the GAN framework, however, comes at the cost of training instability and reduced sample diversity. Furthermore, InfoGAN requires some a priori knowledge of the data, since its performance is sensitive to the choice of the prior distribution and the number of the regularised noise variables. InfoGAN also lacks a principled inference network (although the recognition network can be used as one). The ability to infer the posterior latent distribution from sensory input is important when using the unsupervised model in transfer learning or zero-shot inference scenarios. Hence, while InfoGAN is an important step in the right direction, we believe that further improvements are necessary to achieve a principled way of using unsupervised learning for developing more human-like learning and reasoning in algorithms as described by Lake et al. (2016).

Finally, there is currently no general method for quantifying the degree of learnt disentanglement. Therefore there is no way to quantitatively compare the degree of disentanglement achieved by different models or when optimising the hyperparameters of a single model.

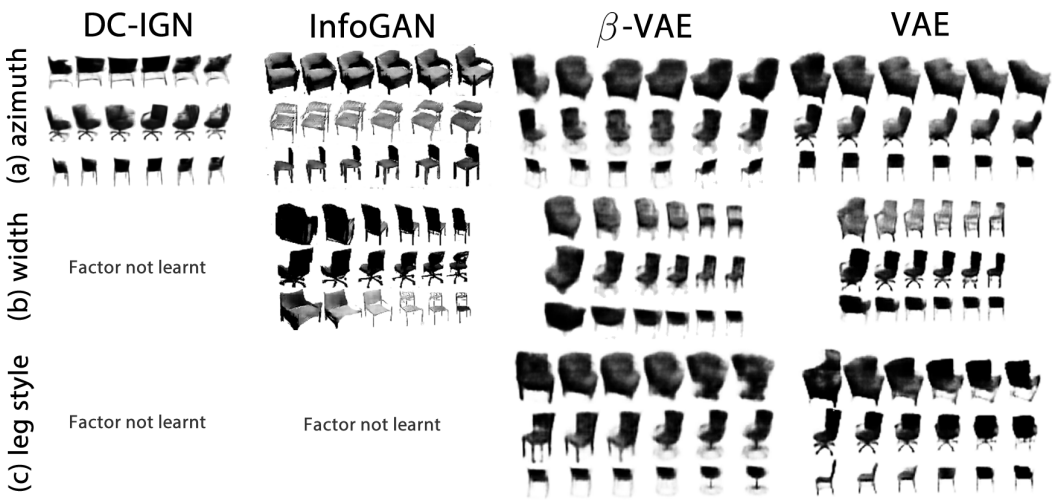

Figure 2: **Manipulating latent variables on 3D chairs:** Qualitative results comparing disentangling performance of β-VAE ($\beta = 5$), VAE (Kingma & Welling, 2014) ($\beta = 1$), InfoGAN (Chen et al., 2016) and DC-IGN (Kulkarni et al., 2015). InfoGAN traversal is over the [-1, 1] range. VAE always learns an entangled representation (e.g. chair width is entangled with azimuth and leg style (b)). All models apart from VAE learnt to disentangle the labelled data generative factor, azimuth (a). InfoGAN and β-VAE were also able to discover unlabelled factors in the dataset, such as chair width (b). Only β-VAE, however, learnt about the unlabelled factor of chair leg style (c). InfoGAN and DC-IGN images adapted from Chen et al. (2016) and Kulkarni et al. (2015), respectively. Reprinted with permission.

In this paper we attempt to address these issues. We propose β-VAE, a deep unsupervised generative approach for disentangled factor learning that can automatically discover the independent latent factors of variation in unsupervised data. Our approach is based on the variational autoencoder (VAE) framework (Kingma & Welling, 2014; Rezende et al., 2014), which brings scalability and training stability. While the original VAE work has been shown to achieve limited disentangling performance on simple datasets, such as FreyFaces or MNIST (Kingma & Welling, 2014), disentangling performance does not scale to more complex datasets (e.g. Aubry et al., 2014; Paysan et al., 2009; Liu et al., 2015), prompting the development of more elaborate semi-supervised VAE-based approaches for learning disentangled factors (e.g. Kulkarni et al., 2015; Karaletsos et al., 2016).

We propose augmenting the original VAE framework with a single hyperparameter β that modulates the learning constraints applied to the model. These constraints impose a limit on the capacity of the latent information channel and control the emphasis on learning statistically independent latent factors. β-VAE with $\beta = 1$ corresponds to the original VAE framework (Kingma & Welling, 2014; Rezende et al., 2014). With $\beta > 1$ the model is pushed to learn a more efficient latent representation of the data, which is disentangled if the data contains at least some underlying factors of variation that are independent. We show that this simple modification allows β-VAE to significantly improve the degree of disentanglement in learnt latent representations compared to the unmodified VAE framework (Kingma & Welling, 2014; Rezende et al., 2014). Furthermore, we show that β-VAE achieves state of the art disentangling performance against both the best unsupervised (InfoGAN: Chen et al., 2016) and semi-supervised (DC-IGN: Kulkarni et al., 2015) approaches for disentangled factor learning on a number of benchmark datasets, such as CelebA (Liu et al., 2015), chairs (Aubry et al., 2014) and faces (Paysan et al., 2009) using qualitative evaluation. Finally, to help quantify the differences, we develop a new measure of disentanglement and show that β-VAE significantly outperforms all our baselines on this measure (ICA, PCA, VAE Kingma & Ba (2014), DC-IGN Kulkarni et al. (2015), and InfoGAN Chen et al. (2016)).

Our main contributions are the following: 1) we propose β-VAE, a new unsupervised approach for learning disentangled representations of independent visual data generative factors; 2) we devise a protocol to quantitatively compare the degree of disentanglement learnt by different models; 3) we demonstrate both qualitatively and quantitatively that our β-VAE approach achieves state-of-the-art disentanglement performance compared to various baselines on a variety of complex datasets.

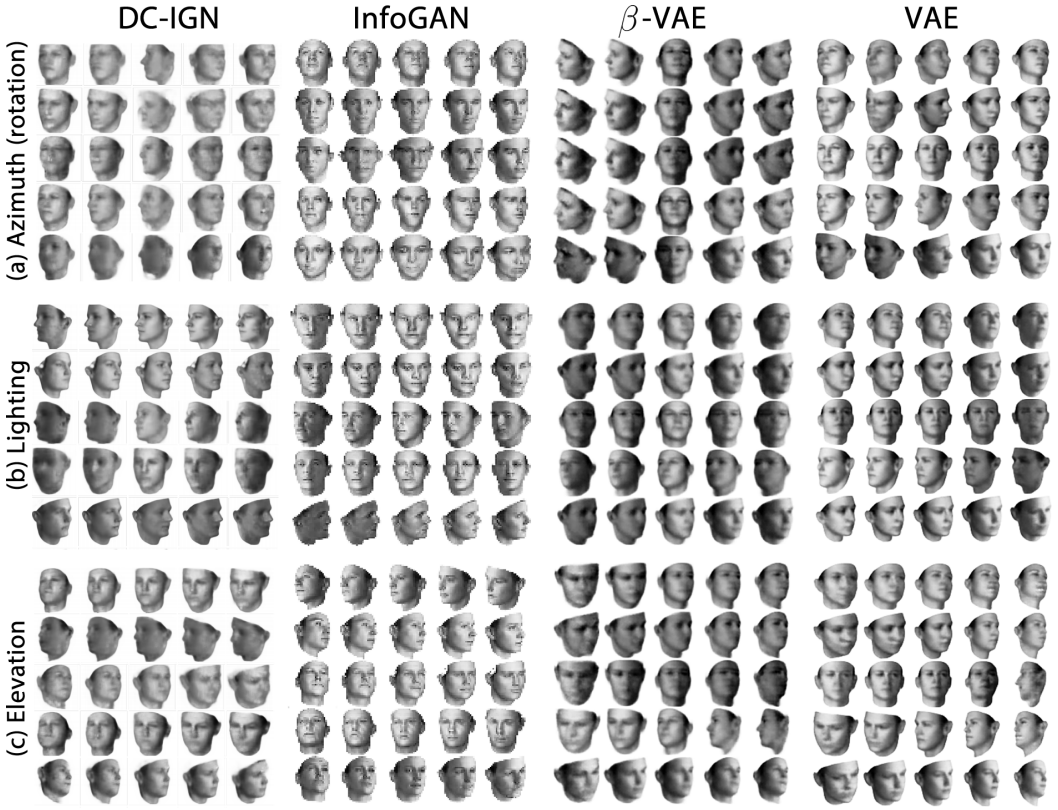

Figure 3: **Manipulating latent variables on 3D faces:** Qualitative results comparing disentangling performance of β-VAE ($\beta = 20$), VAE (Kingma & Welling, 2014) ($\beta = 1$), InfoGAN (Chen et al., 2016) and DC-IGN (Kulkarni et al., 2015). InfoGAN traversal is over the [-1, 1] range. All models learnt to disentangle lighting (b) and elevation (c). DC-IGN and VAE struggled to continuously interpolate between different azimuth angles (a), unlike β-VAE, which additionally learnt to encode a wider range of azimuth angles than other models. InfoGAN and DC-IGN images adapted from Chen et al. (2016) and Kulkarni et al. (2015), respectively. Reprinted with permission.

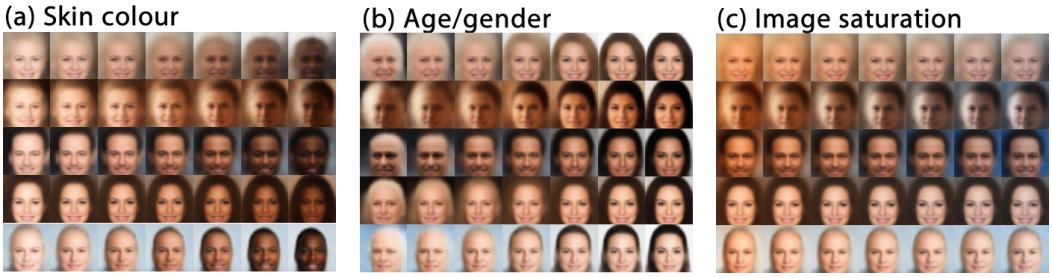

Figure 4: **Latent factors learnt by β-VAE on celebA:** traversal of individual latents demonstrates that β-VAE discovered in an unsupervised manner factors that encode skin colour, transition from an elderly male to younger female, and image saturation.

2 β-VAE FRAMEWORK DERIVATION

Let $\mathcal{D} = \{X, V, W\}$ be the set that consists of images $\mathbf{x} \in \mathbb{R}^N$ and two sets of ground truth data generative factors: conditionally independent factors $\mathbf{v} \in \mathbb{R}^K$, where $log\ p(\mathbf{v}|\mathbf{x}) = \sum_k log\ p(v_k|\mathbf{x})$; and conditionally dependent factors $\mathbf{w} \in \mathbb{R}^H$. We assume that the images $\mathbf{x}$ are generated by the true world simulator using the corresponding ground truth data generative factors: $p(\mathbf{x}|\mathbf{v}, \mathbf{w}) = \mathbf{Sim}(\mathbf{v}, \mathbf{w})$.

We want to develop an unsupervised deep generative model that, using samples from X only, can learn the joint distribution of the data $\mathbf{x}$ and a set of generative latent factors $\mathbf{z}$ ($\mathbf{z} \in \mathbb{R}^M$, where $M \geq K$) such that $\mathbf{z}$ can generate the observed data $\mathbf{x}$; that is, $p(\mathbf{x}|\mathbf{z}) \approx p(\mathbf{x}|\mathbf{v}, \mathbf{w}) = \mathbf{Sim}(\mathbf{v}, \mathbf{w})$. Thus a suitable objective is to maximise the marginal (log-)likelihood of the observed data $\mathbf{x}$ in expectation over the whole distribution of latent factors $\mathbf{z}$:

$$\max_{\theta} \mathbb{E}_{p_\theta(\mathbf{z})}[p_\theta(\mathbf{x}|\mathbf{z})] \tag{1}$$

For a given observation $\mathbf{x}$, we describe the inferred posterior configurations of the latent factors $\mathbf{z}$ by a probability distribution $q_\phi(\mathbf{z}|\mathbf{x})$. Our aim is to ensure that the inferred latent factors $q_\phi(\mathbf{z}|\mathbf{x})$ capture the generative factors $\mathbf{v}$ in a disentangled manner. The conditionally dependent data generative factors $\mathbf{w}$ can remain entangled in a separate subset of $\mathbf{z}$ that is not used for representing $\mathbf{v}$. In order to encourage this disentangling property in the inferred $q_\phi(\mathbf{z}|\mathbf{x})$, we introduce a constraint over it by trying to match it to a prior $p(\mathbf{z})$ that can both control the capacity of the latent information bottleneck, and embodies the desiderata of statistical independence mentioned above. This can be achieved if we set the prior to be an isotropic unit Gaussian ($p(\mathbf{z}) = \mathcal{N}(0, I)$), hence arriving at the constrained optimisation problem in Eq. 2, where ϵ specifies the strength of the applied constraint.

$$\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}} \left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \quad \text{subject to } D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon \tag{2}$$

Re-writing Eq. 2 as a Lagrangian under the KKT conditions (Kuhn & Tucker, 1951; Karush, 1939), we obtain:

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \left(D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - \epsilon\right) \tag{3}$$

where the KKT multiplier β is the regularisation coefficient that constrains the capacity of the latent information channel $\mathbf{z}$ and puts implicit independence pressure on the learnt posterior due to the isotropic nature of the Gaussian prior $p(\mathbf{z})$. Since $\beta, \epsilon \geq 0$ according to the complementary slackness KKT condition, Eq. 3 can be re-written to arrive at the β-VAE formulation - as the familiar variational free energy objective function as described by Jordan et al. (1999), but with the addition of the β coefficient:

$$\mathcal{F}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{4}$$

Varying β changes the degree of applied learning pressure during training, thus encouraging different learnt representations. β-VAE where $\beta = 1$ corresponds to the original VAE formulation of (Kingma & Welling, 2014). We postulate that in order to learn disentangled representations of the conditionally independent data generative factors $\mathbf{v}$, it is important to set $\beta > 1$, thus putting a stronger constraint on the latent bottleneck than in the original VAE formulation of Kingma & Welling (2014). These constraints limit the capacity of $\mathbf{z}$, which, combined with the pressure to maximise the log likelihood of the training data $\mathbf{x}$ under the model, should encourage the model to learn the most efficient representation of the data. Since the data $\mathbf{x}$ is generated using at least some conditionally independent ground truth factors $\mathbf{v}$, and the D_{KL} term of the β-VAE objective function encourages conditional independence in $q_\phi(\mathbf{z}|\mathbf{x})$, we hypothesise that higher values of β should encourage learning a disentangled representation of $\mathbf{v}$. The extra pressures coming from high β values, however, may create a trade-off between reconstruction fidelity and the quality of disentanglement within the learnt latent representations. Disentangled representations emerge when the right balance is found between information preservation (reconstruction cost as regularisation) and latent channel capacity restriction ($\beta > 1$). The latter can lead to poorer reconstructions due to the loss of high frequency details when passing through a constrained latent bottleneck. Hence, the log likelihood of the data under the learnt model is a poor metric for evaluating disentangling in β-VAEs. Instead we propose a quantitative metric that directly measures the degree of learnt disentanglement in the latent representation.

Since our proposed hyperparameter β directly affects the degree of learnt disentanglement, we would like to estimate the optimal β for learning a disentangled latent representation directly. However, it is not possible to do so. This is because the optimal β will depend on the value of ϵ in Eq.2. Different datasets and different model architectures will require different optimal values of ϵ. However, when optimising β in Eq. 4, we are indirectly also optimising ϵ for the best disentanglement (see Sec.A.7 for details), and while we can not learn the optimal value of β directly, we can instead estimate it using either our proposed disentanglement metric (see Sec. 3) or through visual inspection heuristics.

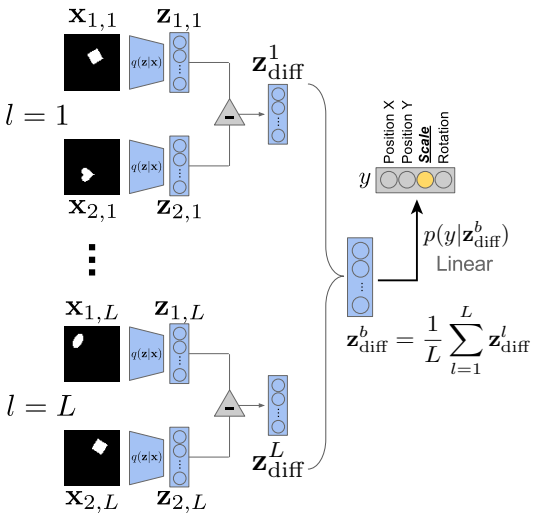

Figure 5: Schematic of the proposed disentanglement metric: over a batch of L samples, each pair of images has a fixed value for one target generative factor y (here $y = scale$) and differs on all others. A linear classifier is then trained to identify the target factor using the average pairwise difference $\mathbf{z}^b_{\text{diff}}$ in the latent space over L samples.

3 DISENTANGLEMENT METRIC

It is important to be able to quantify the level of disentanglement achieved by different models. Designing a metric for this, however, is not straightforward. We begin by defining the properties that we expect a disentangled representation to have. Then we describe our proposed solution for quantifying the presence of such properties in a learnt representation.

As stated above, we assume that the data is generated by a ground truth simulation process which uses a number of data generative factors, some of which are conditionally *independent*, and we also assume that they are *interpretable*. For example, the simulator might sample independent factors corresponding to object shape, colour and size to generate an image of a *small green apple*. Because of the independence property, the simulator can also generate small *red* apples or *big* green apples. A representation of the data that is disentangled with respect to these generative factors, i.e. which encodes them in separate latents, would enable robust classification even using very simple linear classifiers (hence providing *interpretability*). For example, a classifier that learns a decision boundary that relies on object shape would perform as well when other data generative factors, such as size or colour, are varied.

Note that a representation consisting of *independent* latents is not necessarily disentangled, according to our desiderata. Independence can readily be achieved by a variety of approaches (such as PCA or ICA) that learn to project the data onto independent bases. Representations learnt by such approaches do not in general align with the data generative factors and hence may lack *interpretability*. For this reason, a simple cross-correlation calculation between the inferred latents would not suffice as a disentanglement metric.

Our proposed disentangling metric, therefore, measures both the *independence* and *interpretability* (due to the use of a simple classifier) of the inferred latents. To apply our metric, we run inference on a number of images that are generated by fixing the value of one data generative factor while randomly sampling all others. If the independence and interpretability properties hold for the inferred representations, there will be less variance in the inferred latents that correspond to the fixed generative factor. We use a low capacity linear classifier to identify this factor and report the accuracy value as the final disentanglement metric score. Smaller variance in the latents corresponding to the target factor will make the job of this classifier easier, resulting in a higher score under the metric. See Fig. 5 for a representation of the full process.

More formally, we start from a dataset $\mathcal{D} = \{X, V, W\}$ as described in Sec. 2, assumed to contain a balanced distribution of ground truth factors $(\mathbf{v}, \mathbf{w})$, where images data points are obtained using a ground truth simulator process $\mathbf{x} \sim \mathbf{Sim}(\mathbf{v}, \mathbf{w})$. We also assume we are given labels identifying a subset of the independent data generative factors $\mathbf{v} \in V$ for at least some instances.
We then construct a batch of B vectors $\mathbf{z}^b_{\text{diff}}$, to be fed as inputs to a linear classifier as follows:

1. Choose a factor $y \sim Unif[1...K]$ (e.g. $y = scale$ in Fig. 5).

2. For a batch of L samples:
   (a) Sample two sets of latent representations, $\mathbf{v}_{1,l}$ and $\mathbf{v}_{2,l}$, enforcing $[\mathbf{v}_{1,l}]_k = [\mathbf{v}_{2,l}]_k$ if $k = y$ (so that the value of factor $k = y$ is kept *fixed*).
   (b) Simulate image $\mathbf{x}_{1,l} \sim \mathbf{Sim}(\mathbf{v_{1,l}})$, then infer $\mathbf{z}_{1,l} = \mu(\mathbf{x}_{1,l})$, using the encoder $q(\mathbf{z}|\mathbf{x}) \sim N(\mu(\mathbf{x}), \sigma(\mathbf{x}))$.
   Repeat the process for $\mathbf{v}_{2,l}$.
   (c) Compute the difference $\mathbf{z}^l_{\text{diff}} = |\mathbf{z}_{1,l} - \mathbf{z}_{2,l}|$, the absolute linear difference between the inferred latent representations.
3. Use the average $\mathbf{z}^b_{\text{diff}} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{z}^l_{\text{diff}}$ to predict $p(y|\mathbf{z}^b_{\text{diff}})$ (again, $y = scale$ in Fig. 5) and report the accuracy of this predictor as **disentangement metric score**.

The classifier's goal is to predict the index y of the generative factor that was kept fixed for a given $\mathbf{z}^b_{\text{diff}}$. The accuracy of this classifier over multiple batches is used as our disentanglement metric score. We choose a linear classifier with low VC-dimension in order to ensure it has no capacity to perform nonlinear disentangling by itself. We take differences of two inferred latent vectors to reduce the variance in the inputs to the classifier, and to reduce the conditional dependence on the inputs $\mathbf{x}$. This ensures that on average $\left[\mathbf{z}^b_{\text{diff}}\right]_y < \left[\mathbf{z}^b_{\text{diff}}\right]_{\{\setminus y\}}$. See Equations 5 in Appendix A.4 for more details of the process.

4 EXPERIMENTS

In this section we first qualitatively demonstrate that our proposed β-VAE framework consistently discovers more latent factors and disentangles them in a cleaner fashion that either unmodified VAE (Kingma & Welling, 2014) or state of the art unsupervised (InfoGAN: Chen et al., 2016) and semi-supervised (DC-IGN: Kulkarni et al., 2015) solutions for disentangled factor learning on a variety of benchmarks. We then quantify and characterise the differences in disentangled factor learning between our β-VAE framework and a variety of benchmarks using our proposed new disentangling metric.

4.1 QUALITATIVE BENCHMARKS

We trained β-VAE (see Tbl. 1 for architecture details) on a variety of datasets commonly used to evaluate disentangling performance of models: celebA (Liu et al., 2015), chairs (Aubry et al., 2014) and faces (Paysan et al., 2009). Figures 1-3 provide a qualitative comparison of the disentangling performance of β-VAE, VAE ($\beta = 1$) (Kingma & Welling, 2014), InfoGAN (Chen et al., 2016) and DC-IGN (Kulkarni et al., 2015) as appropriate.

It can be seen that across all datasets β-VAE is able to automatically discover and learn to disentangle all of the factors learnt by the semi-supervised DC-IGN (Kulkarni et al., 2015): azimuth (Fig. 3a, Fig. 2a), lighting and elevation (Fig. 3b,c)). Often it acts as a more convincing inverse graphics network than DC-IGN (e.g. Fig. 3a) or InfoGAN (e.g. Fig. 2a, Fig. 1a-c or Fig. 3a). Furthermore, unlike DC-IGN, β-VAE requires no supervision and hence can learn about extra unlabelled data generative factors that DC-IGN can not learn by design, such as chair width or leg style (Fig. 2b,c). The unsupervised InfoGAN (Chen et al., 2016) approach shares this quality with β-VAE, and the two frameworks tend to discover overlapping, but not necessarily identical sets of data generative factors. For example, both β-VAE and InfoGAN (but not DC-IGN) learn about the width of chairs (Fig. 2b). Only β-VAE, however, learns about the chair leg style (Fig. 2c). It is interesting to note how β-VAE is able to generate an armchair with a round office chair base, even though such armchairs do not exist in the dataset (or, perhaps, reality). Furthermore, only β-VAE is able to discover all three factors of variation (chair azimuth, width and leg style) within a single model, while InfoGAN learns to allocate its continuous latent variable to *either* azimuth *or* width. InfoGAN sometimes discovers factors that β-VAE does not precisely disentangle, such as the presence of sunglasses in celebA. β-VAE does, however, discover numerous extra factors such as skin colour, image saturation, and age/gender that are not reported in the InfoGAN paper (Chen et al., 2016) (Fig. 4). Furthermore, β-VAE latents tend to learn a smooth continuous transformation over a wider range of factor values than InfoGAN (e.g. rotation over a wider range of angles as shown in Figs. 1-3a).

Overall β-VAE tends to consistently and robustly discover more latent factors and learn cleaner disentangled representations of them than either InfoGAN or DC-IGN. This holds even on such challenging datasets as celebA. Furthermore, unlike InfoGAN and DC-IGN, β-VAE requires no design decisions or assumptions about the data, and is very stable to train.

When compared to the unmodified VAE baseline ($\beta = 1$) β-VAE consistently learns significantly more disentangled latent representations. For example, when learning about chairs, VAE entangles chair width with leg style (Fig. 2b). When learning about celebA, VAE entangles azimuth with emotion and gender (Fig. 1a); emotion with hair style, skin colour and identity (Fig. 1b); while the VAE fringe latent also codes for baldness and head size (Fig. 1c). Although VAE performs relatively well on the faces dataset, it still struggles to learn a clean representation of azimuth (Fig. 3a). This, however, suggests that a continuum of disentanglement quality exists, and it can be traversed by varying β within the β-VAE framework. While increasing β often leads to better disentanglement, it may come at the cost of blurrier reconstructions and losing representations for some factors, particularly those that correspond to only minor changes in pixel space.

4.2 Quantitative benchmarks

In order to quantitatively compare the disentangling performance of β-VAE against various baselines, we created a synthetic dataset of 737,280 binary 2D shapes (heart, oval and square) generated from the Cartesian product of the shape and four independent generative factors v_k defined in vector graphics: position X (32 values), position Y (32 values), scale (6 values) and rotation (40 values over the 2π range). To ensure smooth affine object transforms, each two subsequent values for each factor v_k were chosen to ensure minimal differences in pixel space given 64x64 pixel image resolution. This dataset was chosen because it contains no confounding factors apart from its five independent data generative factors (identity, position X, position Y, scale and rotation). This gives us knowledge of the ground truth for comparing the disentangling performance of different models in an objective manner.

We used our proposed disentanglement metric (see Sec. 3) to quantitatively compare the ability of β-VAE to automatically discover and learn a disentangled representation of the data generative factors of the synthetic dataset of 2D shapes described above with that of a number of benchmarks (see Tbl. 1 in Appendix for model architecture details). The table in Fig. 6 (left) reports the classification accuracy of the disentanglement metric for 5,000 test samples. It can be seen that β-VAE ($\beta = 4$) significantly outperforms all baselines, such as an untrained VAE and the original VAE formulation of Kingma & Welling (2014) ($\beta = 1$) with the same architecture as β-VAE, the top ten PCA or ICA components of the data (see Sec. A.3 for details), or when using the raw pixels directly. β-VAE also does better than InfoGAN. Remarkably, β-VAE performs on the same level as DC-IGN despite the latter being semi-supervised and the former wholly unsupervised. Furthermore, β-VAE achieved similar classification accuracy as the ground truth vectors used for data generation, thus suggesting that it was able to learn a very good disentangled representation of the data generative factors.

We also examined qualitatively the representations learnt by β-VAE, VAE, InfoGAN and DC-IGN on the synthetic dataset of 2D shapes. Fig. 7A demonstrates that after training, β-VAE with $\beta = 4$ learnt a good (while not perfect) disentangled representation of the data generative factors, and its decoder learnt to act as a rendering engine. Its performance was comparative to that of DC-IGN (Fig. 7C), with the difference that DC-IGN required a priori knowledge about the quantity of the data generative factors, while β-VAE was able to discover them in an unsupervised manner. The most informative latent units z_m of β-VAE have the highest KL divergence from the unit Gaussian prior ($p(z) = \mathcal{N}(0, I)$), while the uninformative latents have KL divergence close to zero. Fig. 7A demonstrates the selectivity of each latent z_m to the independent data generating factors: $z_m^\mu = f(v_k)\ \forall v_k \in \{v_{positionX}, v_{positionY}, v_{scale}, v_{rotation}\}$ (top three rows), where z_m^μ is the learnt Gaussian mean of latent unit z_m. The effect of traversing each latent z_m on the resulting reconstructions is shown in the bottom five rows of Fig. 7A. The latents z_6 and z_2 learnt to encode X and Y coordinates of the objects respectively; unit z_1 learnt to encode scale; and units z_5 and z_7 learnt to encode rotation. The frequency of oscillations in each rotational latent corresponds to the rotational symmetry of the corresponding object (2π for heart, π for oval and $\pi/2$ for square). Furthermore, the two rotational latents seem to encode $\cos$ and $\sin$ rotational coordinates, while the positional latents align with the Cartesian axes. While such alignment with intuitive factors for humans is not guaranteed, empirically we found it to be very common. Fig. 7B demonstrates that the unmodified

Model	Disentanglement metric score
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3 \pm 0.1}\%$
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23 \pm 0.1}\%$

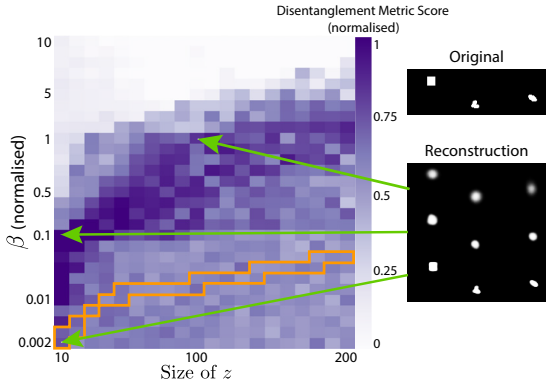

Figure 6: Disentanglement metric classification accuracy for 2D shapes dataset. **Left:** Accuracy for different models and training regimes **Right:** Positive correlation is present between the size of $\mathbf{z}$ and the optimal *normalised* values of β for disentangled factor learning for a fixed β-VAE architecture. β values are normalised by latent $\mathbf{z}$ size m and input $\mathbf{x}$ size n. Note that β values are not uniformly sampled. Orange approximately corresponds to *unnormalised* $\beta = 1$. Good reconstructions are associated with entangled representations (lower disentanglement scores). Disentangled representations (high disentanglement scores) often result in blurry reconstructions.

VAE baseline ($\beta = 1$) is not able to disentangle generative factors in the data as well as β-VAE with appropriate learning pressures. Instead each latent $\mathbf{z}$ (apart from $\mathbf{z}_9$, which learnt rotation) encodes at least two data generative factors. InfoGAN also achieved a degree of disentangling (see Fig. 7D), particularly for positional factors. However, despite our best efforts to train InfoGAN, we were not able to achieve the same degree of disentangling in other factors, such as rotation, scale and shape. We also found its ability to generate the different shapes in the dataset to be inaccurate and unstable during training, possibly due to reported limitations of the GAN framework, which can struggle to learn the full data distribution and instead will often learn a small subset of its modes (Salimans et al., 2016; Zhao et al., 2016).

Understanding the effects of β We hypothesised that constrained optimisation is important for enabling deep unsupervised models to learn disentangled representations of the independent data generative factors (Sec. 2). In the β-VAE framework this corresponds to tuning the β coefficient. One way to view β is as a mixing coefficient (see Sec. A.6 for a derivation) for balancing the magnitudes of gradients from the reconstruction and the prior-matching components of the VAE lower bound formulation in Eq. 4 during training. In this context it makes sense to normalise β by latent $\mathbf{z}$ size m and input $\mathbf{x}$ size n in order to compare its different values across different latent layer sizes and different datasets ($\beta_{norm} = \frac{\beta M}{N}$). We found that larger latent $\mathbf{z}$ layer sizes m require higher constraint pressures (higher β values), see Fig. 6 (Right). Furthermore, the relationship of β for a given m is characterised by an inverted U curve. When β is too low or too high the model learns an entangled latent representation due to either too much or too little capacity in the latent $\mathbf{z}$ bottleneck. We find that in general $\beta > 1$ is necessary to achieve good disentanglement. However if β is too high and the resulting capacity of the latent channel is lower than the number of data generative factors, then the learnt representation necessarily has to be entangled (as a low-rank projection of the true data generative factors will compress them in a non-factorial way to still capture the full data distribution well). We also note that VAE reconstruction quality is a poor indicator of learnt disentanglement. Good disentangled representations often lead to blurry reconstructions due to the restricted capacity of the latent information channel $\mathbf{z}$, while entangled representations often result in the sharpest reconstructions. We therefore suggest that one should not necessarily strive for perfect reconstructions when using β-VAEs as unsupervised feature learners - though it is often possible to find the right β-VAE architecture and the right value of β to have both well disentangled latent representations and good reconstructions.

We proposed a principled way of choosing β for datasets with at least weak label information. If label information exists for at least a small subset of the independent data generative factors of variation, one can apply the disentanglement metric described in Sec. 3 to approximate the level of learnt disentanglement for various β choices during a hyperparameter sweep. When such labelled information is not available, the optimal value of β can be found through visual inspection of what

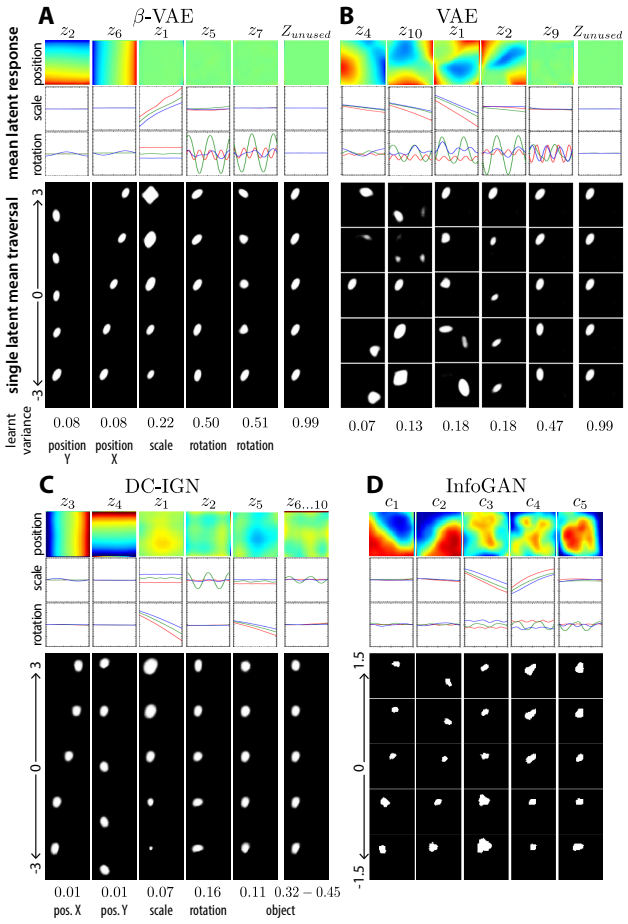

Figure 7: **A:** Representations learnt by a β-VAE ($\beta = 4$). Each column represents a latent z_i, ordered according to the learnt Gaussian variance (last row). Row 1 (position) shows the mean activation (red represents high values) of each latent z_i as a function of all 32x32 locations averaged across objects, rotations and scales. Row 2 and 3 show the mean activation of each unit z_i as a function of scale (respectively rotation), averaged across rotations and positions (respectively scales and positions). *Square* is red, *oval* is green and *heart* is blue. Rows 4-8 (second group) show reconstructions resulting from the traversal of each latent z_i over three standard deviations around the unit Gaussian prior mean while keeping the remaining 9/10 latent units fixed to the values obtained by running inference on an image from the dataset. **B:** Similar analysis for VAE ($\beta = 1$). **C:** Similar analysis for DC-IGN, clamping a single latent each for scale, positions, orientation and 5 for shape. **D:** Similar analysis for InfoGAN, using 5 continuous latents regularized using the mutual information cost, and 5 additional unconstrained noise latents (not shown).

effect the traversal of each single latent unit z_m has on the generated images $(\mathbf{x}|\mathbf{z})$ in pixel space (as shown in Fig. 7 rows 4-8). For the 2D shapes dataset, we have found that the optimal values of β as determined by visual inspection match closely the optimal values as determined by the disentanglement metric.

5 Conclusion

In this paper we have reformulated the standard VAE framework (Kingma & Welling, 2014; Rezende et al., 2014) as a constrained optimisation problem with strong latent capacity constraint and independence prior pressures. By augmenting the lower bound formulation with the β coefficient that regulates the strength of such pressures and, as a consequence, the qualitative nature of the representations learnt by the model, we have achieved state of the art results for learning disentangled representations of data generative factors. We have shown that our proposed β-VAE framework significantly outperforms both qualitatively and quantitatively the original VAE (Kingma & Welling, 2014), as well as state-of-the-art unsupervised (InfoGAN: Chen et al., 2016) and semi-supervised (DC-IGN: Kulkarni et al., 2015) approaches to disentangled factor learning. Furthermore, we have shown that β-VAE consistently and robustly discovers more factors of variation in the data, and it learns a representation that covers a wider range of factor values and is disentangled more cleanly than other benchmarks, all in a completely unsupervised manner. Unlike InfoGAN and DC-IGN, our approach does not depend on any *a priori* knowledge about the number or the nature of data generative factors. Our preliminary investigations suggest that the performance of the β-VAE framework may depend on the sampling density of the data generative factors within a training dataset (see Appendix A.8 for more details). It appears that having more densely sampled data generative factors results in better disentangling performance of β-VAE, however we leave a more principled investigation of this effect to future work.

β-VAE is robust with respect to different architectures, optimisation parameters and datasets, hence requiring few design decisions. Our approach relies on the optimisation of a single hyperparameter β, which can be found directly through a hyperparameter search if weakly labelled data is available to calculate our new proposed disentangling metric. Alternatively the optimal β can be estimated heuristically in purely unsupervised scenarios. Learning an interpretable factorised representation of the independent data generative factors in a completely unsupervised manner is an important precursor for the development of artificial intelligence that understands the world in the same way that humans do (Lake et al., 2016). We believe that using our approach as an unsupervised pretraining stage for supervised or reinforcement learning will produce significant improvements for scenarios such as transfer or fast learning.

6 ACKNOWLEDGEMENTS

We would like to thank Charles Blundell, Danilo Rezende, Tejas Kulkarni and David Pfau for helpful comments that improved the manuscript.

REFERENCES

M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. In *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv*, 2016.

Brian Cheung, Jesse A. Levezey, Arjun K. Bansal, and Bruno A. Olshausen. Discovering hidden factors of variation in deep networks. In *Proceedings of the International Conference on Learning Representations, Workshop Track*, 2015.

T. Cohen and M. Welling. Transformation properties of learned visual representations. In *ICLR*, 2015.

Taco Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. *arXiv*, 2014.

G. Desjardins, A. Courville, and Y. Bengio. Disentangling factors of variation via generative entangling. *arXiv*, 2012.

Carl Doersch. Tutorial on variational autoencoders. *arxiv*, 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, pp. 2672–2680, 2014.

Ross Goroshin, Michael Mathieu, and Yann LeCun. Learning to linearize under uncertainty. *NIPS*, 2015.

G. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. *International Conference on Artificial Neural Networks*, 2011.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *ICLR*, 2016.

W. Karush. Minima of Functions of Several Variables with Inequalities as Side Constraints. Master's thesis, Univ. of Chicago, Chicago, Illinois, 1939.

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

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of 2nd Berkeley Symposium*, pp. 481–492, 1951.

Tejas Kulkarni, William Whitney, Pushmeet Kohli, and Joshua Tenenbaum. Deep convolutional inverse graphics network. *NIPS*, 2015.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *arXiv*, 2016.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *ICCV*, 2015.

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. *AVSS*, 2009.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2011.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*, 2015.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. *ICML*, 2014.

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

Karl Ridgeway. A survey of inductive biases for factorial Representation-Learning. *arXiv*, 2016. URL http://arxiv.org/abs/1612.05299.

Oren Rippel and Ryan Prescott Adams. High-dimensional probability estimation with deep density models. *arXiv*, 2013.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *arXiv*, 2016. URL http://arxiv.org/abs/1606.03498.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–869, 1992.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-Time single image and video Super-Resolution using an efficient Sub-Pixel convolutional neural network. *arXiv*, 2016.

Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Tensor analyzers. In *Proceedings of the 30th International Conference on Machine Learning, 2013, Atlanta, USA*, 2013.

William F. Whitney, Michael Chang, Tejas Kulkarni, and Joshua B. Tenenbaum. Understanding visual concepts with continuation learning. *arXiv*, 2016. URL http://arxiv.org/pdf/1602.06822.pdf.

Jimei Yang, Scott Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. *NIPS*, 2015.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv*, 2016. URL http://arxiv.org/abs/1609.03126.

Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *Advances in Neural Information Processing Systems 27*. 2014.

A Appendix

A.1 Model architecture details

A summary of all model architectures used in this paper can be seen in Tbl 1.

A.2 InfoGAN training

To train the InfoGAN network described in Tbl. 1 on the 2D shapes dataset (Fig. 7), we followed the training paradigm described in Chen et al. (2016) with the following modifications. For the mutual information regularised latent code, we used 5 continuous variables c_i sampled uniformly from $(-1, 1)$. We used 5 noise variables z_i, as we found that using a reduced number of noise variables improved the quality of generated samples for this dataset. To help stabilise training, we used the *instance noise* trick described in Shi et al. (2016), adding Gaussian noise to the discriminator inputs (0.2 standard deviation on images scaled to $[-1, 1]$). We followed Radford et al. (2015) for the architecture of the convolutional layers, and used batch normalisation in all layers except the last in the generator and the first in the discriminator.

Dataset	Optimiser	Architecture	
2D shapes (VAE)	Adagrad 1e-2	Input	4096 (flattened 64x64x1).
		Encoder	FC 1200, 1200. ReLU activation.
		Latents	10
		Decoder	FC 1200, 1200, 1200, 4096. Tanh activation. Bernoulli.
2D shapes (DC-IGN)	rmsprop (as in Kulkarni et al., 2015)	Input	64x64x1.
		Encoder	Conv 96x3x3, 48x3x3, 48x3x3 (padding 1). ReLU activation and Max pooling 2x2.
		Latents	10
		Decoder	Unpooling, Conv 48x3x3, 96x3x3, 1x3x3. ReLU activation, Sigmoid.
2D shapes (InfoGAN)	Adam 1e-3 (gen) 2e-4 (dis)	Generator	FC 256, 256, Deconv 128x4x4, 64x4x4 (stride 2). Tanh.
		Discriminator	Conv and FC reverse of generator. Leaky ReLU activation. FC 1. Sigmoid activation.
		Recognition	Conv and FC shared with discriminator. FC 128, 5. Gaussian
		Latents	10: $z_{1...5} \sim Unif(-1, 1)$, $c_{1...5} \sim Unif(-1, 1)$
Chairs (VAE)	Adam 1e-4	Input	64x64x1.
		Encoder	Conv 32x4x4 (stride 2), 32x4x4 (stride 2), 64x4x4 (stride 2), 64x4x4 (stride 2), FC 256. ReLU activation.
		Latents	32
		Decoder	Deconv reverse of encoder. ReLU activation. Bernoulli.
CelebA (VAE)	Adam 1e-4	Input	64x64x3.
		Encoder	Conv 32x4x4 (stride 2), 32x4x4 (stride 2), 64x4x4 (stride 2), 64x4x4 (stride 2), FC 256. ReLU activation.
		Latents	32
		Decoder	Deconv reverse of encoder. ReLU activation. Gaussian.
3DFaces (VAE)	Adam 1e-4	Input	64x64x1.
		Encoder	Conv 32x4x4 (stride 2), 32x4x4 (stride 2), 64x4x4 (stride 2), 64x4x4 (stride 2), FC 256. ReLU activation.
		Latents	32
		Decoder	Deconv reverse of encoder. ReLU activation. Bernoulli.

Table 1: Details of model architectures used in the paper. The models were trained using either adagrad (Duchi et al., 2011) or adam (Kingma & Ba, 2014) optimisers.

A.3 ICA AND PCA BASELINES

In order to calculate the ICA benchmark, we applied fastICA (Pedregosa et al., 2011) algorithm to the whitened pixel data. Due to memory limitations we had to apply the algorithm to pairwise combinations of the subsets of the dataset corresponding to the transforms of each of the three 2D object identities. We calculated the disentangling metric for all three ICA models trained on each of the three pairwise combinations of 2D objects, before presenting the average of these scores in Fig. 6.

We performed PCA on the raw and whitened pixel data. Both approaches resulted in similar disentangling metric scores. Fig. 6 reports the PCA results calculated using whitened pixel data for more direct comparison with the ICA score.

A.4 DISENTANGLEMENT METRIC DETAILS

We used a linear classifier to learn the identity of the generative factor that produced $\mathbf{z}^b_{\text{diff}}$ (see Equations (5) for the process used to obtain samples of $\mathbf{z}^b_{\text{diff}}$). We used a fully connected linear

classifier to predict $p(y|\mathbf{z}^b_{\text{diff}})$, where y is one of four generative factors (position X, position Y, scale and rotation). We used softmax output nonlinearity and a negative log likelihood loss function. The classifier was trained using the Adagrad (Duchi et al., 2011) optimisation algorithm with learning rate of 1e-2 until convergence.

$$\mathcal{D} = \{V \in \mathbb{R}^K, W \in \mathbb{R}^H, X \in \mathbb{R}^N\},\ y \sim Unif[1...K]$$

Repeat for $b = 1 \dots B$:

$$\begin{aligned}
&\mathbf{v}_{1,l} \sim p(\mathbf{v}),\ \mathbf{w}_{1,l} \sim p(\mathbf{w}),\ \mathbf{w}_{2,l} \sim p(\mathbf{w}),\ [\mathbf{v}_{2,l}]_k = \begin{cases} [\mathbf{v}_{1,l}]_k\,, & \text{if } k = y \\ \sim p(v_k), & \text{otherwise} \end{cases} \\
&\mathbf{x}_{1,l} \sim \mathbf{Sim}(\mathbf{v_{1,l}}, \mathbf{w_{1,l}}),\ \mathbf{x_{2,l}} \sim \mathbf{Sim}(\mathbf{v_{2,l}}, \mathbf{w_{2,l}}), \\
&q(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mu(\mathbf{x}), \sigma(\mathbf{x})),\ \mathbf{z}_{1,l} = \mu(\mathbf{x}_{1,l}),\ \mathbf{z}_{2,l} = \mu(\mathbf{x}_{2,l}) \\
&\mathbf{z}^l_{\text{diff}} = |\mathbf{z}_{1,l} - \mathbf{z}_{2,l}|, \quad \mathbf{z}^b_{\text{diff}} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{z}^l_{\text{diff}}
\end{aligned} \tag{5}$$

All disentanglement metric score results reported in the paper were calculated in the following manner. Ten replicas of each model with the same hyperparameters were trained using different random seeds to obtain disentangled representations. Each of the ten trained model replicas was evaluated three times using the disentanglement metric score algorithm, each time using a different random seed to initialise the linear classifier. We then discarded the bottom 50% of the thirty resulting scores and reported the remaining results. This was done to control for the outlier results from the few experiments that diverged during training.

The results reported in table in Fig. 6 (left) were calculated using the following data. Ground truth uses independent data generating factors $\mathbf{v}$ (our dataset did not contain any correlated data generating factors $\mathbf{w}$). PCA and ICA decompositions keep the first ten components (PCA components explain 60.8% of variance). β-VAE ($\beta = 4$), VAE ($\beta = 1$) and VAE untrained have the same fully connected architecture with ten latent units $\mathbf{z}$. InfoGAN uses "inferred" values of the five continuous latents that were regularised with the mutual information objective during training.

A.5 Classifying the ground truth data generative factors values

In order to further verify the validity of our proposed disentanglement metric we ran an extra quantitative test: we trained a linear classifier to predict the ground truth value of each of the five data generative factors used to generate the 2D shapes dataset. While this test does not measure disentangling directly (since it does not measure independence of the latent representation), a disentangled representation should make such a classification trivial. It can be seen in Table 2 that the representation learnt by β-VAE is on average the best representation for factor classification across all five factors. It is closely followed by DC-IGN. It is interesting to note that ICA does well only at encoding object identity, while PCA manages to learn a very good representation of object position.

Model	**Classification accuracy**					
	id	scale	rotation	position X	position Y	**average**
PCA	43.38	36.08	5.96	**60.66**	**60.15**	41.25
ICA	**59.6**	34.4	7.61	25.96	25.12	30.54
DC-IGN	44.82	**45.92**	15.89	47.64	45.88	40.03
InfoGAN	44.47	40.91	6.39	27.51	23.73	28.60
VAE untrained	39.44	25.33	6.09	16.69	14.39	20.39
VAE	41.55	24.07	8	16.5	18.72	21.77
β-VAE	50.08	43.03	**20.36**	52.25	49.5	**43.04**

Table 2: Linear classifier classification accuracy for predicting the ground truth values for each data generative factor from different latent representations. Each factor could take a variable number of possible values: 3 for id, 6 for scale, 40 for rotation and 32 for position X or Y. Best performing model results in each column are printed in bold.

A.6 INTERPRETING NORMALISED β

We start with the β-VAE constrained optimisation formulation that we have derived in Sec. 2.

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta \, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{6}$$

We make the assumption that every pixel n in $\mathbf{x} \in \mathbb{R}^N$ is conditionally independent given $\mathbf{z}$ (Doersch, 2016). The first term of Eq. 6 then becomes:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log \prod_n p_\theta(x_n|\mathbf{z})] = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\sum_n \log p_\theta(x_n|\mathbf{z})] \tag{7}$$

Dividing both sides of Eq. 6 by N produces:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) \propto \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\mathbb{E}_n[\log p_\theta(x_n|\mathbf{z})] - \frac{\beta}{N} \, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \tag{8}$$

We design β-VAE to learn conditionally independent factors of variation in the data. Hence we assume conditional independence of every latent z_m given x (where $m \in 1...M$, and M is the dimensionality of $\mathbf{z}$). Since our prior $p(\mathbf{z})$ is an isotropic unit Gaussian, we can re-write the second term of Eq. 6 as:

$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = \int_z q_\phi(\mathbf{z}|\mathbf{x}) log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} = \sum_m \int_{z_m} q_\phi(z_m|\mathbf{x}) log \frac{q_\phi(z_m|\mathbf{x})}{p(z_m)} \tag{9}$$

Multiplying the second term in Eq. 8 by a factor $\frac{M}{M}$ produces:

$$\begin{aligned} \mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) &\propto \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\mathbb{E}_n[\log p_\theta(x_n|\mathbf{z})] - \frac{\beta M}{N} \, \mathbb{E}_m \int_{z_m} [q_\phi(z_m|\mathbf{x}) log \frac{q_\phi(z_m|\mathbf{x})}{p(z_m)}] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\mathbb{E}_n[\log p_\theta(x_n|\mathbf{z})] - \frac{\beta M}{N} \, \mathbb{E}_m[D_{KL}(q_\phi(z_m|\mathbf{x})||p(z_m))] \end{aligned} \tag{10}$$

Hence using

$$\beta_{norm} = \frac{\beta M}{N}$$

in Eq. 10 is equivalent to optimising the original β-VAE formulation from Sec. 2, but with the additional independence assumptions that let us calculate data log likelihood and KL divergence terms in expectation over the individual pixels x_n and individual latents z_m.

A.7 RELATIONSHIP BETWEEN β AND ϵ

For a given ϵ we can solve the constrained optimisation problem in Eq. 3 (find the optimal $(\theta^*, \phi^*, \beta^*)$, such that $\Delta\mathcal{F}(\theta^*, \phi^*, \beta^*) = 0$). We can then re-write our optimal solution to the original optimisation problem in Eq. 2 as a function of ϵ:

$$\mathcal{G}(\theta^*(\epsilon), \phi^*(\epsilon)) = \mathbb{E}_{q_{\phi^*(\epsilon)}(\mathbf{z}|\mathbf{x})}[\log p_{\theta^*(\epsilon)}(\mathbf{x}|\mathbf{z})] \tag{11}$$

Now β can be interpreted as the rate of change of the optimal solution (θ^*, ϕ^*) to $\mathcal{G}$ when varying the constraint ϵ:

$$\frac{\delta\mathcal{G}}{\delta\epsilon} = \beta^*(\epsilon) \tag{12}$$

A.8 DATA CONTINUITY

We hypothesise that data continuity plays a role in guiding unsupervised models towards learning the correct data manifolds. To test this idea we measure how the degree of learnt disentangling changes with reduced continuity in the 2D shapes dataset. We trained a β-VAE with $\beta = 4$ (Figure 7A) on subsamples of the original 2D shapes dataset, where we progressively decreased the generative factor sampling density. Reduction in data continuity negatively correlates with the average pixel wise (Hamming) distance between two consecutive transforms of each object (normalised by the average number of pixels occupied by each of the two adjacent transforms of an object to account for object

scale). Figure 8 demonstrates that as the continuity in the data reduces, the degree of disentanglement in the learnt representations also drops. This effect holds after additional hyperparameter tuning and can not solely be explained by the decrease in dataset size, since the same VAE can learn disentangled representations from a data subset that preserves data continuity but is approximately 55% of the original size (results not shown).

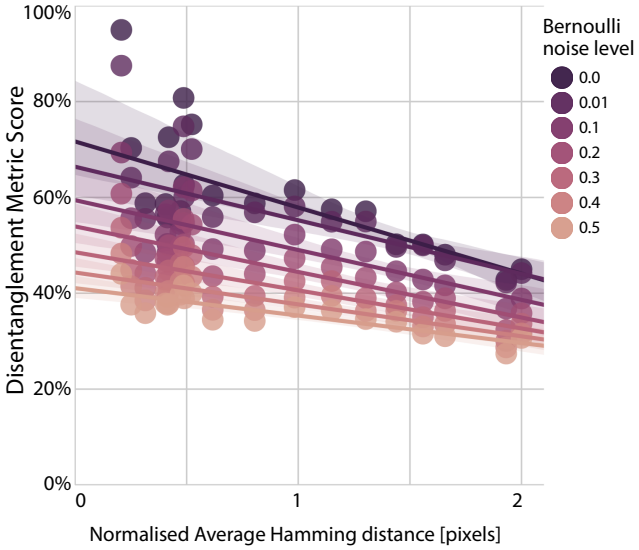

Figure 8: Negative correlation between data transform continuity and the degree of disentangling achieved by β-VAE. Abscissa is the average normalized Hamming distance between each of the two consecutive transforms of each object. Ordinate is disentanglement metric score. Disentangling performance is robust to Bernoulli noise added to the data at test time, as shown by slowly degrading classification accuracy up to 10% noise level, considering that the 2D objects occupy on average between 2-7% of the image depending on scale. Fluctuations in classification accuracy for similar Hamming distances are due the different nature of subsampled generative factors (i.e. symmetries are present in rotation but are lacking in position).

A.9 β-VAE SAMPLES

Samples from β-VAE that learnt disentangled ($\beta = 4$) and entangled ($\beta = 1$) representations can be seen in Figure 9.

A.10 EXTRA β-VAE TRAVERSAL PLOTS

We present extra latent traversal plots from β-VAE that learnt disentangled representations of 3D chairs (Figures 10-11) and CelebA (Figures 12-14) datasets. Here we show traversals from all informative latents from a large number of seed images.

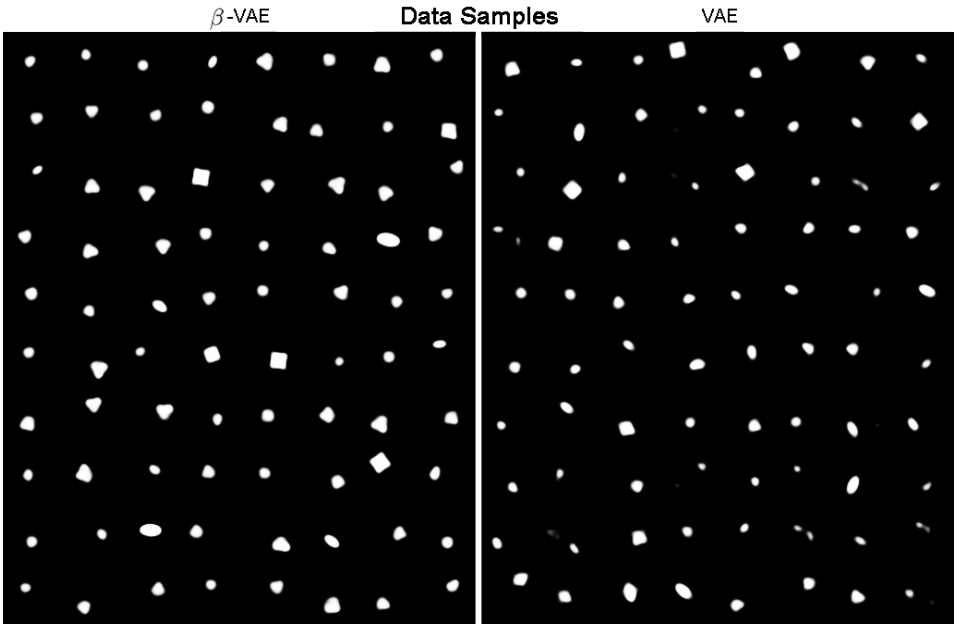

Figure 9: Samples from β-VAE trained on the dataset of 2D shapes that learnt either a disentangled (left, $\beta = 4$) or an entangled (right, $\beta = 1$) representation of the data generative factors. It can be seen that sampling from an entangled representation results in some unrealistic looking samples. A disentangled representation that inverts the original data generation process does not suffer from such errors.

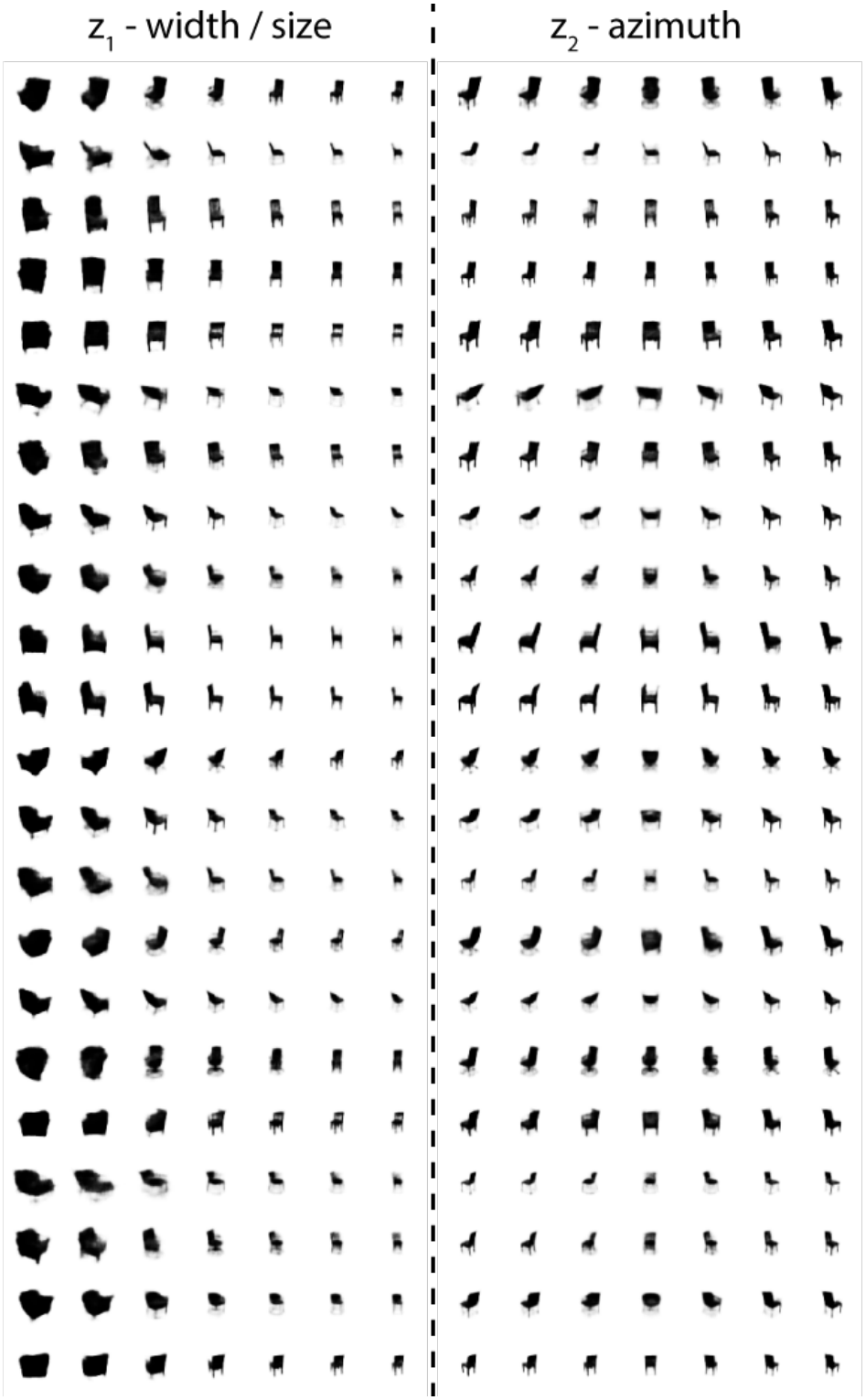

Figure 10: Latent traversal plots from β-VAE that learnt disentangled representations on the 3D chairs dataset.

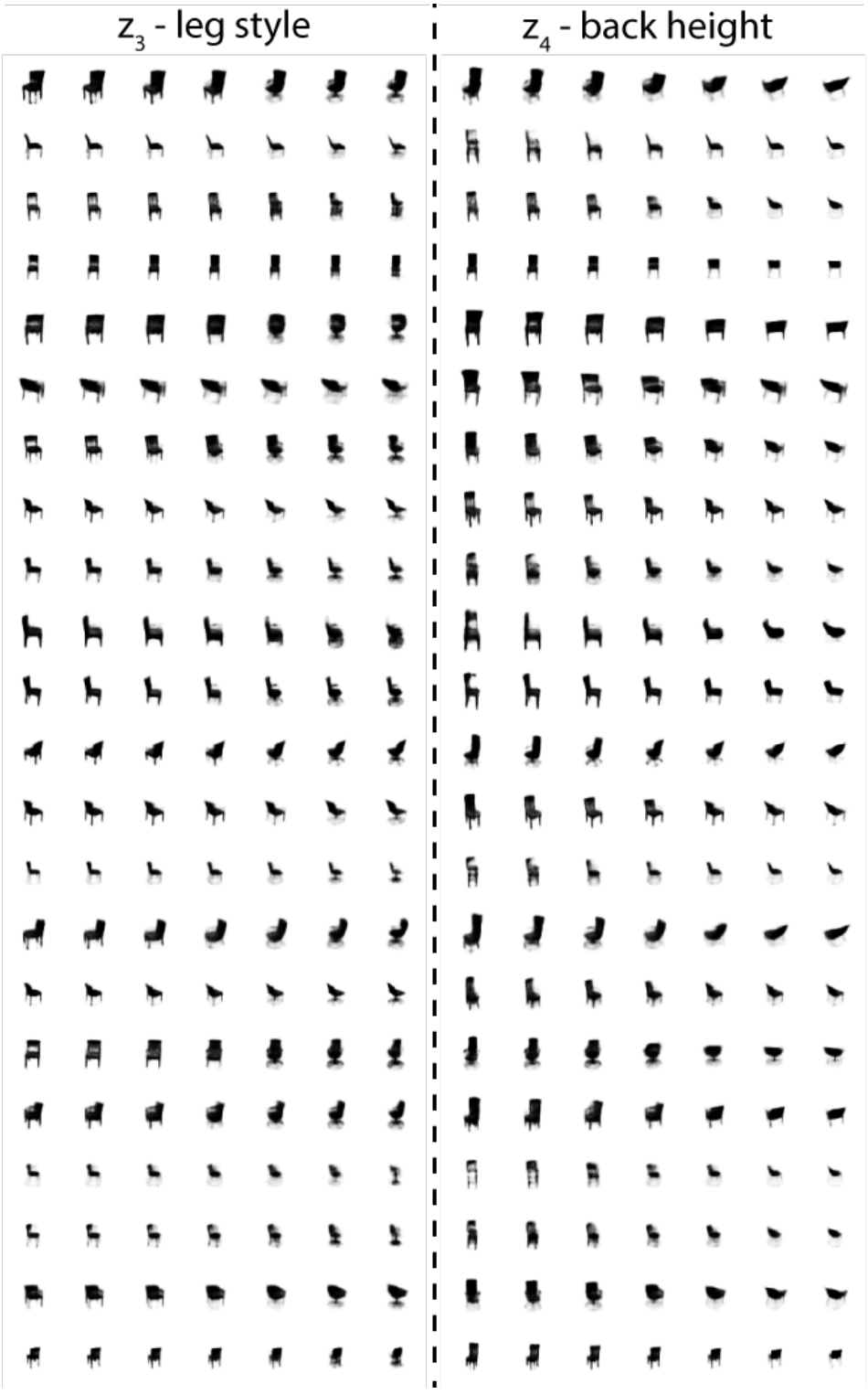

Figure 11: Latent traversal plots from β-VAE that learnt disentangled representations on the 3D chairs dataset.

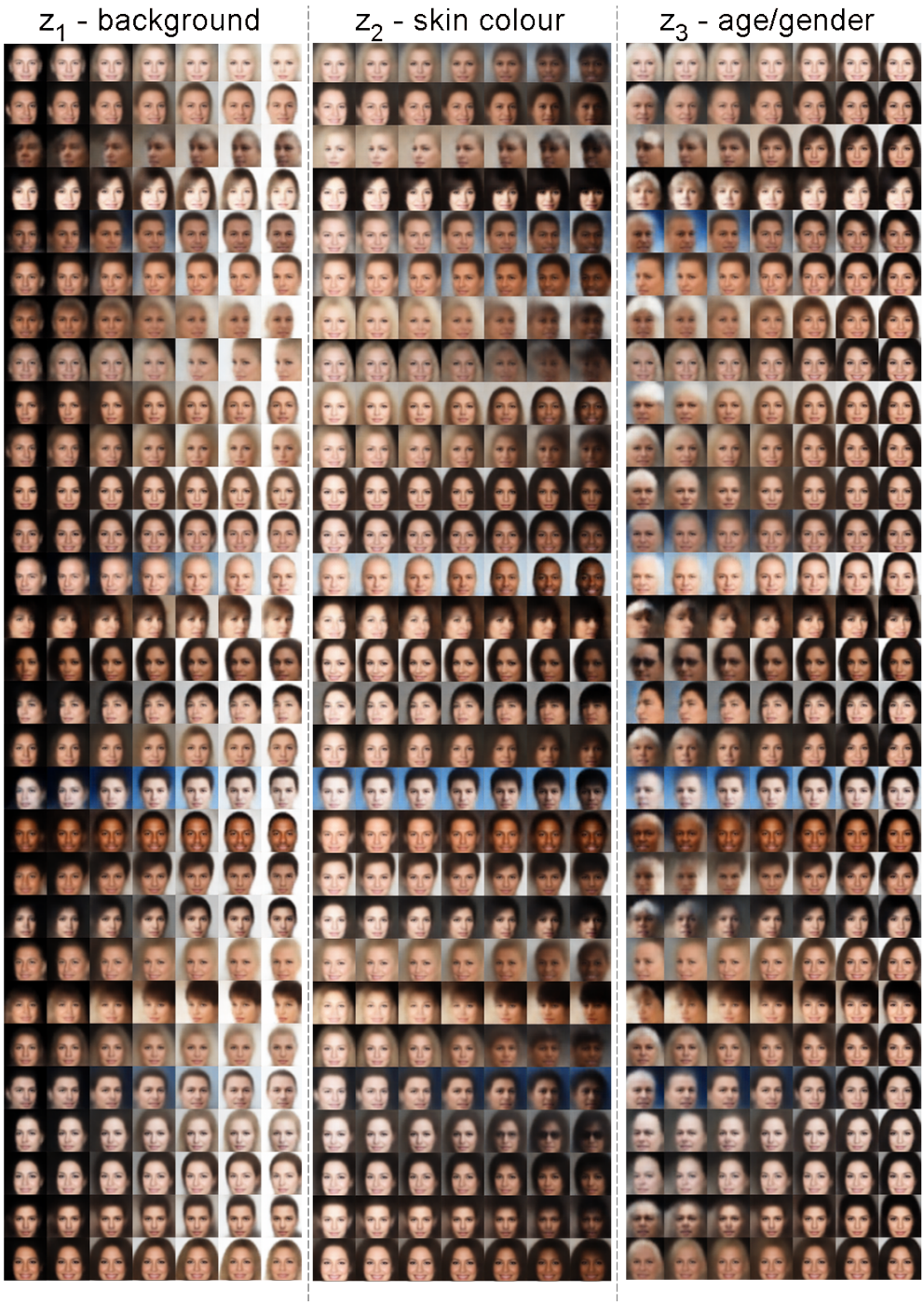

Figure 12: Latent traversal plots from β-VAE that learnt disentangled representations on the CelebA dataset.

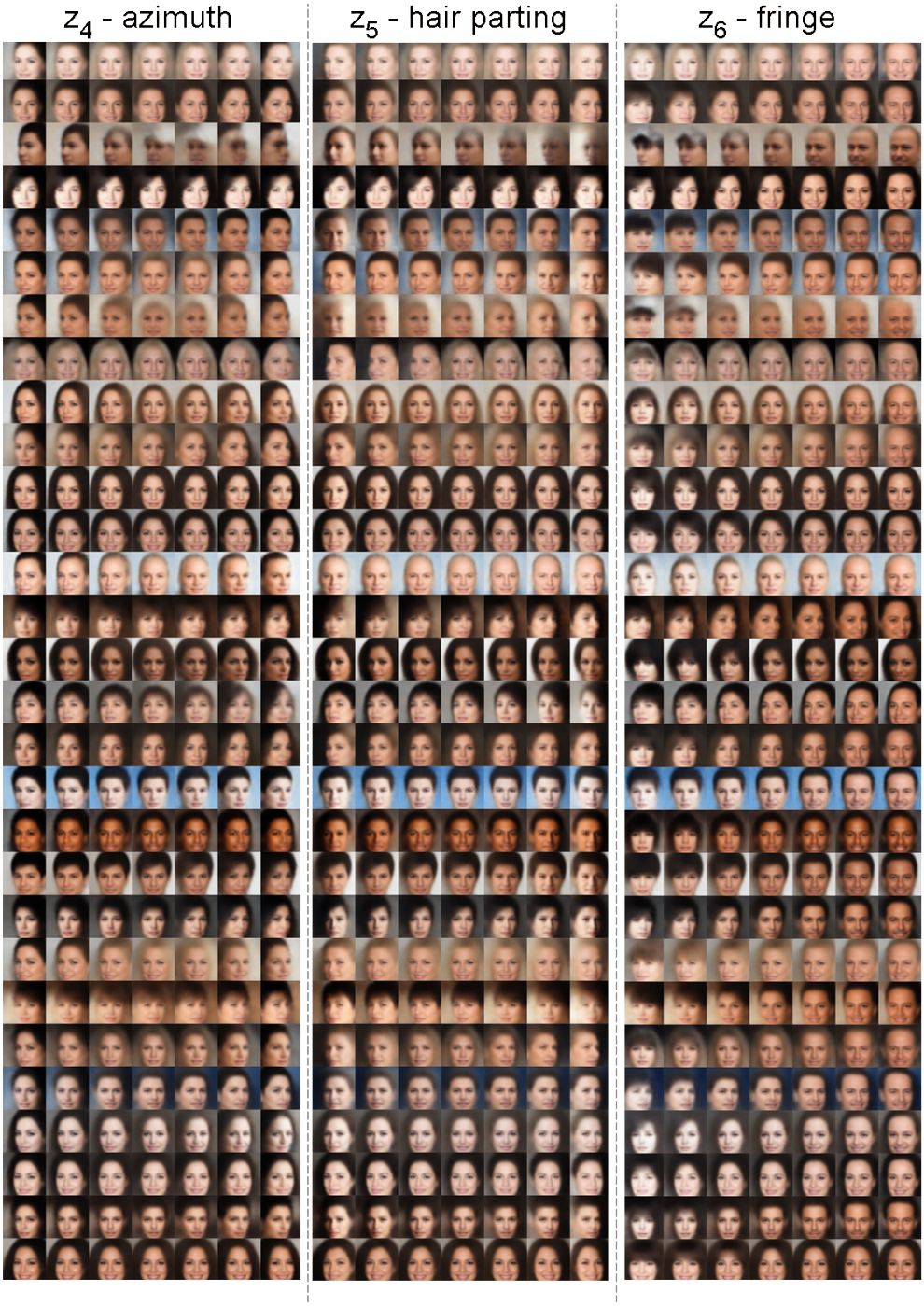

Figure 13: Latent traversal plots from β-VAE that learnt disentangled representations on the CelebA dataset.

Figure 14: Latent traversal plots from β-VAE that learnt disentangled representations on the CelebA dataset.