Disentangling Autoencoders (DAE)

Jaehoon Cha
Scientific Machine Learning group
Rutherford Appleton Laboratory
Science and Technology Facilities Council
United Kingdom
jaehoon.cha@stfc.ac.uk

Jeyan Thiyagalingam
Scientific Machine Learning group
Rutherford Appleton Laboratory
Science and Technology Facilities Council
United Kingdom
t.jeyan@stfc.ac.uk

***Abstract*—Noting the importance of factorizing (or disentangling) the latent space, we propose a novel, non-probabilistic disentangling framework for autoencoders, based on the principles of symmetry transformations in group-theory. To the best of our knowledge, this is the first deterministic model that is aiming to achieve disentanglement based on autoencoders without regularizers. The proposed model is compared to seven state-of-the-art generative models based on autoencoders and evaluated based on five supervised disentanglement metrics. The experimental results show that the proposed model can have better disentanglement when variances of each features are different. We believe that this model leads to a new field for disentanglement learning based on autoencoders without regularizers.**

***Index Terms*—disentanglement, generative models, unsupervised learning, latent representation**

I. Introduction

Learning generalizable representations of data is one of the fundamental aspects of modern machine learning [Rudin et al.(2022)Rudin, Chen, Chen, Huang, Semenova, and Zhong]. In fact, better representations are more than a luxury now, and is a key to achieve generalization, interpretability, and robustness of machine learning models [Bengio et al.(2013)Bengio, Courville, and Vincent], [Brakel and Bengio(2017)], [Spurek et al.(2020)Spurek, Nowak, Tabor, Maziarka, and Jastrzebski]. One of the primary and desired characteristics of the learned representation is factorizability or disentanglement so that latent representation is composed of multiple, independent generative factors of variations. The disentanglement process renders the latent space features to become independent of one another, and thus provides the basis for novel applications, such as scene rendering, interpretability, and unsupervised deep learning [Eslami et al.(2018)Eslami, Rezende, Besse, Viola, Morcos, Garnelo, Ruderman, Rusu, Danihelka, Gregor, et al.], [Iten et al.(2020)Iten, Metger, Wilming, Del Rio, and Renner], [Higgins et al.(2021)Higgins, Chang, Langston, Hassabis, Summerfield, Tsao, and Botvinick]. Deep generative models, particularly that build on autoencoders, from the vanilla variational autoencoder (VAE) model [Kingma and Welling(2013)] to various derivatives of VAE, including, β-VAE [Higgins et al.(2017)Higgins, Matthey, Pal, Burgess, Glorot, Botvinick, Mohamed, and Lerchner], [Burgess et al.(2018)Burgess, Higgins, Pal, Matthey, Watters, Desjardins, and Lerchner], β-Total Correlation Variational Autoencoder (TCVAE) [Chen et al.(2018)Chen, Li, Grosse, and Duvenaud], Controlled Capacity Increase-VAE(CCI-VAE) [Burgess et al.(2018)Burgess, Higgins, Pal, Matthey, Watters, Desjardins, and Lerchner], Factor-VAE (FVAE) [Kim and Mnih(2018)], Information Maximizing Variational Autoencoders (InfoVAE) [Zhao et al.(2019)Zhao, Song, and Ermon], and Wasserstein-AE (WAE) [Tolstikhin et al.(2018)Tolstikhin, Bousquet, Gelly, and Schölkopf], have shown to be effective in learning factored representations. The disentangling mechanism, and hence the underpinning functionality of these generative models, rely on two forms of losses: regularization and reconstruction losses [Higgins et al.(2017)Higgins, Matthey, Pal, Burgess, Glorot, Botvinick, Mohamed, and Lerchner], [Chen et al.(2018)Chen, Li, Grosse, and Duvenaud], [Burgess et al.(2018)Burgess, Higgins, Pal, Matthey, Watters, Desjardins, and Lerchner], [Kim and Mnih(2018)], [Tolstikhin et al.(2018)Tolstikhin, Bousquet, Gelly, and Schölkopf].

Although these approaches have advanced the disentangled representation learning, there are a number of issues that limit their full potential. Among these, two of the salient issues that directly conflict with the process of deriving disentangled representations are:

- The tension of balancing two loss components in VAE (and their derivatives) is a delicate and a well-known issue [Asperti and Trentin(2020)]. While the KL-divergence acts as a regularizer by normalizing the smoothness of the latent space (with potential overlapping of latent variables), the reconstruction loss focuses on improving the visual quality of the resulting images. However, the process of improving reconstruction loss (and hence the visual quality of the output) is oblivious to the shape of the latent space. These contrasting effects render the balancing process more delicate, and when not done correctly, the visual quality of the generated images degrade.
- The notion of known prior distribution is the cornerstone of VAEs and often assumed to be simple isotropic Gaussian distribution. Even with approaches that relax the expressive constrains around the prior exists, such as [Tomczak and Welling(2018)], [Takahashi et al.(2019)Takahashi, Iwata, Yamanaka, Yamada, and Yagi], [Zhang et al.(2020)Zhang, Zhang, Li, Bengio, and Paull], [Aneja et al.(2021)Aneja, Schwing, Kautz, and

Under review.

Vahdat], the presence of a prior (even if optimal) can easily create a tension between the true distribution and the prior. Hence, this can exert an additional pressure on latent space regularization, particularly if the distribution of the real data does not match the prior.

In this paper, we propose a novel autoencoder (AE)-based non-probabilistic approach for deriving disentangled representations while addressing the concerns highlighted above. More specifically, the proposed approach, which we name as Disentangling Auto-Encoder (DAE), relies on the concept symmetry transformation [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner], which is often formalized using group theory. By carefully deriving a set of symmetry transformations on the latent space for each latent variables, we achieve a powerful method for obtaining disentangled representations. The proposed model has the following advantages over conventional VAE-based approaches:

1 It is a non-probabilistic, group theory-based approach. As such, neither there is any assumption of any priors nor the process of learning any posteriors from the input data; and
2 As a consequence of (1), the proposed approach fully eliminates the need for any distribution regularization mechanism (such as KL-divergence) in the latent space, and thus the approach renders a model that improves a reconstruction loss whilst maintaining disentangled representations.

Our evaluation, covering seven state-of-the-art VAE-based models across five different supervised disentanglement metrics, shows that the proposed model has a powerful disentangling ability without regularizers. This is particularly proven to be true across our evaluation when the variances of each feature are different. This provides an additional advantage where the method has potential to analyse real datasets which have a combination of categorical and continuous factors.

The rest of this paper is organized as follows. In Section II we review the related work, particularly focusing on VAE-based approaches due to its nature of strongly principled yet simplistic approach to disentanglement. This is then followed by a derivation of AE-based non-probabilistic approach for deriving disentangled representations in Section III. In Section IV, we evaluate the proposed method against a number of relevant models with a toy example and three benchmark datasets, and discuss our findings. We then conclude the paper in Section V with directions for further research.

II. Related Work

A. Disentanglement

Disentangled representation learning [Bengio et al.(2013)Bengio, Courville, and Vincent], [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner] focuses on learning independent factors that have useful but minimal information for a given task, such that their variations are orthogonal with each other and can account for the entire dataset. Decoupling any correlations between latent variables matches single underlying factor with one feature of latent variables and can serve a number of downstream applications including the improvement of predictive performance [Locatello et al.(2019)Locatello, Tschannen, Bauer, Rätsch, Schölkopf, and Bachem], effective learning with a small number of samples [Van Steenkiste et al.(2019)Van Steenkiste, Locatello, Schmidhuber, and Bachem], [Yue et al.(2021)Yue, Wang, Sun, Hua, and Zhang], discovery of physical concepts [Iten et al.(2020)Iten, Metger, Wilming, Del Rio, and Renner] and enabling 3D shape reconstruction from 2D images [Pan et al.(2020)Pan, Dai, Liu, Loy, and Luo].

A large body of work can be found around disentanglement, and ideal properties of a disentangled representation can be found in [Ridgeway(2016)], [Eastwood and Williams(2018)], [Ridgeway and Mozer(2018)], [Zaidi et al.(2020)Zaidi, Boilard, Gagnon, and Carbonneau]. Among a number of desirable properties of disentanglement, modularity, compactness and explicitness are three critically important properties. The modularity property focuses on the effect of one feature of learnt representation on others, or in other words, independence. The compactness property measures how effectively one feature of the learnt representation covers one of the ground truth factor. The explicitness property measures the relationship between the learned factors and true factors of data. A number of metrics have been proposed in the literature to quantify these properties [Higgins et al.(2017)Higgins, Matthey, Pal, Burgess, Glorot, Botvinick, Mohamed, and Lerchner], [Kim and Mnih(2018)], [Eastwood and Williams(2018)], [Chen et al.(2018)Chen, Li, Grosse, and Duvenaud], [Do and Tran(2019)], [Sepliarskaia et al.(2019)Sepliarskaia, Kiseleva, de Rijke, et al.]. In our work, we use the notions outlined in [Zaidi et al.(2020)Zaidi, Boilard, Gagnon, and Carbonneau], where the metrics are divided into three classes, namely, Intervention-based metrics, Predictor-based metrics, and Information-based metrics. These metrics are all used in a supervised manner and can be of indicators to quantify modularity, compactness, explicitness robustness to noise, nonlinear relationships between learnt representations and ground truth factors.

B. Probabilistic Generative Models based on Autoencoder Model

Autoencoder (AE), which consists of an encoder E_ϕ that maps an observation space to a lower-dimensional latent space, and a decoder D_θ that re-maps the latent space to the observation space, effectively learn meaningful representations in the latent space by minimizing the reconstruction loss, $\mathcal{L}_{recon}$ (cross-entropy or L_2).

Probabilistic generative models based on AE are achieved by replacing the conventional encoder and decoder with probabilistic variants of them [Kingma and Welling(2013)], [Rezende and Mohamed(2015)], [Higgins et al.(2017)Higgins, Matthey, Pal, Burgess, Glorot, Botvinick, Mohamed, and Lerchner], [Tolstikhin et al.(2018)Tolstikhin, Bousquet, Gelly, and Schölkopf], respectively. Given an observation $\mathbf{x} \in \mathbb{R}^n$, the VAE [Kingma and Welling(2013)]-based approaches rely on the

TABLE I
COMPARISON OF DIFFERENT VAE-BASED MODELS W.R.T THE REGULARIZERS THEY EMPLOY.

Model	$L_{reg}(\phi)$	**Notes**
VAE	$KL(q_\phi(\mathbf{z}\|\mathbf{x}), p(\mathbf{z}))$	–
β-VAE	$\beta KL(q_\phi(\mathbf{z}\|\mathbf{x}), p(\mathbf{z}))$	Usually, β is greater than 1
β-TCVAE	$I(\mathbf{z}, \mathbf{x}) + \beta KL(q(\mathbf{z}), \prod_j q(\mathbf{z}_j)) + \sum_j KL(q(\mathbf{z}_j), p(\mathbf{z}_j))$	$I(\cdot, \cdot)$ is a mutual information
CCI-VAE	$\beta\|\|KL(q_\phi(\mathbf{z}\|\mathbf{x}), p(\mathbf{z})) - C\|\|$	C is a capacity
FVAE	$KL(q_\phi(\mathbf{z}\|\mathbf{x}), p(\mathbf{z})) + \gamma KL(q(\mathbf{z}), \prod_j q(\mathbf{z}_j)))$	The second term is minimised using density-ratio trick
InfoVAE	$KL(q_\phi(\mathbf{z}\|\mathbf{x}), p(\mathbf{z}))$ + $\lambda MMD(q_\phi(\mathbf{z}\|\mathbf{x}), p(\mathbf{z}))$	$MMD(\cdot, \cdot)$ is Maximum Mean Discrepancy
WAE	$\lambda MMD(q_\phi(\mathbf{z}\|\mathbf{x}), p(\mathbf{z}))$	λ is a regularization coefficient

variational theory. They use the probabilistic encoder, denoted by $q_\phi(\mathbf{z}|\mathbf{x})$, to approximate the intractable true posterior and the probabilistic decoder, denoted by $p_\theta(\mathbf{x}|\mathbf{z})$ that reconstructs the $\mathbf{x}$ from $\mathbf{z}$. In an ideal world, the resulting posterior $q_\phi(\mathbf{z}|\mathbf{x})$ should match well with the prior distribution $p(\mathbf{z})$. However, this is rarely the case, and weights in the encoder and decoder are trained accounting this fact by relying on a loss function that measures not only the reconstruction loss, but also the similarity of the posterior and prior distributions. The similarity between two different distributions is, usually, computed using the KL-divergence, but alternative techniques can be used [Tolstikhin et al.(2018)Tolstikhin, Bousquet, Gelly, and Schölkopf]. The combined loss is referred to as the Evidence Lower Bound (ELBO) [Kingma and Welling(2013)], and defined as follows,

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

The first term in (1) can be estimated from samples $\mathbf{z}$ drawn from the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ using reparameterization trick [Kingma and Welling(2013)]. The second term plays a crucial role as a regularizer to minimize the difference between $q_\phi(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$.

Majority of the previous work on disentangled representation learning are predominantly based on probabilistic models, particularly building on VAE. They enforce regularization in the latent space that either regularizes the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ or the aggregate posterior $q(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^{N} q_\phi(\mathbf{z}|\mathbf{x}^{(i)})$, as summarized in [Tschannen et al.(2018)Tschannen, Bachem, and Lucic]. The overall objective of majority of the VAE-based methods can be expressed as:

$$\mathcal{L}_{recon}(\phi, \theta) + L_{reg}(\phi) \tag{2}$$

where $L_{reg}(\phi)$ is a regularizer of a generative model, which often includes one or more hyperparameters, to strike a balance between the two losses. A carefully designed regularizer should enable the model achieving better disentanglement either by controlling the capacity of the latent space, or by measuring the total correlation between latent variables. In our evaluation, we compare the proposed model against seven other VAE-based derivatives, namely, vanilla VAE, β-VAE, β-TCVAE, CCI-VAE, FVAE, InfoVAE and WAE. All these models vary based on the underlying regularizer $L_{reg}(\phi)$. For example, the β-VAE model constraints on the latent space using β to limit the capacity of the latent space, which encourages the model to learn the most efficient representation of the data. The regularization term of these different models (Column 2) are summarized in Table I along with relevant notes (Column 3).

Depending on the selection of the regularizer, each model provides different disentangling capabilities. In contrast, the method we propose here is not a probabilistic model, and thus, does not rely on variational inference or any approximation of posteriors or assumption of priors, totally eliminating the need for any regularizers. Instead, the proposed model relies on a deterministic AE model for deriving the latent space, which is then manipulated very carefully to derive the disentangled latent representations.

III. FRAMEWORK FOR DAE

The deterministic, non-probabilistic approach we propose here in this paper, builds on the autoencoder (rather than variational autoencoders). As such, we first provide a relevant background in Section III-A based on [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner]. We then establish the relationship between the autoencoder model and disentangled representation in Section III-B. We then define the relevant mathematical framework and a corresponding neural network architecture implementing the proposed disentangling autoencoder.

A. Disentangled representation

The notion of disentangled representation is mathematically defined using the concept of symmetry in [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner]. For example, horizontal and vertical translations are symmetry transformations in two-dimensional grid, and, hence, such transformations change the location of an object in this two-dimensional grid. From the definitions of symmetry group in [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner], a symmetry group can be decomposed as a product of multiple subgroups, if suitable subgroups can be identified. This can render an intuitive method to disentangle the latent space, if subgroups that independently act on subspaces of a latent space, can be found. If actions by transformations of each subgroup only affect the corresponding subspace, the actions are called *disentangled group actions*. In other words, disentangled group actions only change a specific property of the state of an object, and leaves the other properties invariant. If

there is a transformation in a vector space of representations, corresponding to a disentangled group action, the representation is called a *disentangled representation*. We reproduce the formal definitions of disentangled group action and disentangled representation from [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner], as Definitions III.1 and III.2, respectively.

Definition III.1. *Suppose that we have a group action* $\cdot : G \times X \to X$*, and the group* G *decomposes as a direct product* $G = G_1 \times \cdots \times G_n$*. Let the action of the full group, and the actions of each subgroups be referred to as* $\cdot$ *and* $\cdot_i$*, respectively. Then, the action is disentangled if there is a decomposition* $X = X_1 \times \cdots \times X_n$*, and actions* $\cdot_i : G_i \times X_i \to X_i$*,* $i \in \{1, \cdots, n\}$ *such that:*

$$(g_1, \cdots, g_n) \cdot (\mathbf{x_1}, \cdots, \mathbf{x_n}) = (g_1 \cdot \mathbf{x_1}, \cdots, g_n \cdot \mathbf{x_n}) \quad (3)$$

for all $g_i \in G_i$ *and* $\mathbf{x_i} \in X_i$*.*

Now, to derive the definition of disentangled representation from the definition of disentangled group action, consider a set of world-states, denoted by W. Furthermore, assume that: (a) there is a generative process $b : W \to O$ leading from world-states to observations, O, (b) and an inference process $h : O \to Z$ leading from observations to an agent's representations, Z. With these, consider the composition $f : W \to Z$, $f = h \circ b$. In terms of transformation, assume that these transformations are represented by a group G of symmetries acting on W via an action $\cdot : G \times W \to W$.

The overarching goal of disentangling the latent space now relies on finding a corresponding action $\cdot : G \times Z \to Z$ so that the symmetry structure of W is reflected in Z. In other words, an action on Z corresponding to the action on W is desirable. This can be achieved if the following condition is satisfied:

$$g \cdot f(\mathbf{w}) = f(g \cdot \mathbf{w}) \quad \forall g \in G, \mathbf{w} \in W. \quad (4)$$

In other words, the action, $\cdot$, should commute with f, which adheres to the definition of the equivariant map, and thus, f is an equivariant map, as shown below.

$$\begin{array}{ccc} G \times W & \xrightarrow{\cdot_{\mathbf{W}}} & W \\ {\scriptstyle id_G \times f}\downarrow & & \downarrow{\scriptstyle f} \\ G \times Z & \xrightarrow{\cdot_{\mathbf{Z}}} & Z \end{array}$$

A very good example of an equivariant map from [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner] is,

$$f(\mathbf{z}) = (e^{i\mathbf{z}_1}, \cdots, e^{i\mathbf{z}_n}). \quad (5)$$

From [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner], a disentangled representation can be defined as follows:

Definition III.2. *The representation* Z *is disentangled with respect to* $G = G_1 \times \cdots \times G_n$ *if*

1. *There is an action* $\cdot : G \times Z \to Z$*,*
2. *The map* $f : W \to Z$ *is equivariant between the actions on* W *and* Z*, and*
3. *There is a decomposition* $Z = Z_1 \times \cdots \times Z_n$ *or* $Z = Z_1 \oplus \cdots \oplus Z_n$ *such that each* Z_i *is fixed by the action of all* G_j*,* $j \neq i$ *and affected only by* G_i*.*

B. Association between the Disentangled Representation and Autoencoder

With the definition of equivariant map in place, the overarching goal of finding a disentangled representation is equivalent to finding f that satisfies (4). However, in general, one cannot control the nature of the generative process. In addition, without loss of generality, we can easily assume the generative process b is an equivariant map. In other words, action on the set of world-states commute with b,

$$g \cdot b(\mathbf{w}) = b(g \cdot \mathbf{w}) \quad \forall g \in G, \mathbf{w} \in W. \quad (6)$$

Now, consider the inference process h, defined above.

Theorem III.3. *Suppose a generative process* b *is an equivariant map satisfying (6). Then, there exists a function* f *that satisfies (4) if an inference process* $h : O \to Z$ *is an equivariant map satisfying,*

$$g \cdot h(\mathbf{o}) = h(g \cdot \mathbf{o}) \quad \forall g \in G, \mathbf{o} \in O. \quad (7)$$

Proof. Suppose that there b satisfies (6) and h is an equivariant map. Then

$$\begin{aligned} g \cdot f(\mathbf{w}) &= g \cdot h(b(\mathbf{w})) && (8)\\ &= h(g \cdot b(\mathbf{w})) && (9)\\ &= h(b(g \cdot \mathbf{w})) && (10)\\ &= f(g \cdot \mathbf{w})) && (11)\\ & && (12) \end{aligned}$$

$\forall g \in G, \mathbf{w} \in W.$ □

Following the Theorem III.3, the goal of disentangling is same as finding an inference process $h : O \to Z$. Although there is no guarantee that one can find a compatible action $\cdot : G \times Z \to Z$ satisfying (7), if h is bijective then (7) can be expressed as follows,

$$g \cdot \mathbf{z} = h(g \cdot h^{-1}(\mathbf{z})) \quad (13)$$

However, as h is a bijective function, simple neural network-based models cannot learn the overall equivariant map. However, the equivariant map, such as one outlined in (5) can be learned by the autoencoders, which is the central contribution of this paper. To show this mapping, let h and h^{-1} be an encoder, E_ϕ, and a decoder, D_θ, of an autoencoder. Then, the group action $\cdot : G \times Z \to Z$ can be defined as follows:

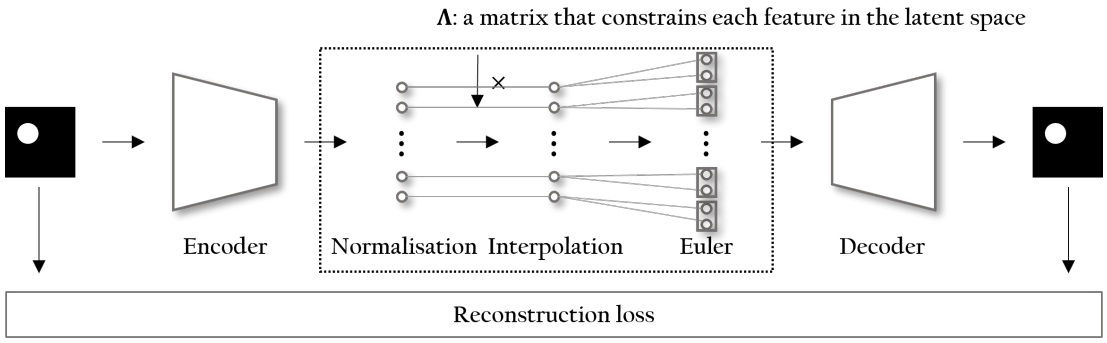


Fig. 1. Illustration of the DAE architecture.

$$G \times Z \xrightarrow{id_G \times D_\theta} G \times O \xrightarrow{\cdot \mathbf{o}} O \xrightarrow{E_\phi} Z$$

This shows that the equivariant map can indeed be learned by an autoencoder. However, this is not without a number of challenges, which we discuss in Section III-C below.

C. Towards Disentangling Autoencoder: Challenges

Consider the generic equivariant map f stated in (5), now applied to an n dimensional latent space vector $\mathbf{z}$. One way this mapping can be made more specific to our case is from [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner], which can be expressed as:

$$f(\mathbf{x_1}, \ldots, \mathbf{x_n}) = (e^{2\pi i \mathbf{x_1}/N_1}, \ldots, e^{2\pi i \mathbf{x_n}/N_n}). \tag{14}$$

where N_j (for $j = 1, \ldots, n$) is the number of elements in subgroup j. Given that $e^{2\pi i\theta} = \cos(2\pi\theta) + i\sin(2\pi\theta)$, (14) provides an excellent route for disentangling groups. However, there are still a number of challenges in realising the overall idea to be of practical utility, particularly in the AE setting. These are:

- **Number of Elements in a Subgroup**: The number of possible elements in the subgroups N_j ($j = 1, \ldots, n$), or at least the relative ratio of the number of elements between the subgroups are not known a priori. Without access to this information, learning (14) becomes impossible.
- **Robustness to Small Perturbations**: Although mapping like (14) renders an approach for disentanglement, the model is not resilient to small perturbations (such as due to noise), which is essential for the model to behave in robust manner when presented with unseen examples.
- **Spatial Distribution of Features**: An ideal factorized latent space must have the features spatially distributed in an equally likely manner. However, the equivariant map we discussed above alone may not take care of this.

Although it is possible to address some of these concerns from the theoretical stand point, nearly all of these are addressable by carefully designing the architecture that exploits both the AE and the equivariant map principle discussed above to achieve the best disentanglement process. We discuss this in the next sub section.

D. Architecture of the DAE

In mapping our theory to an architecture, we build on the AE model, which constitutes an encoder, that maps the observation space O to a factorized latent space Z, followed by the disentangling process that factorizes/disentangles the latent space Z to Z', and finally the decoding layer, that maps the factorized latent Z' to regenerated observation space O'. Each of the concerns that were discussed in Section III-C are handled by a network layer in our architecture, as shown in Figure 1. We describe how each of these layers addresses the concerns in the following sub sections.

1) Number of Elements in a Subgroup: Although the number of elements in a subgroup is not known a priori, these numbers or the relative ratio of the possible number of elements across subgroups can be estimated using techniques that can extract the variance information from compressed information, such as principal component analysis (PCA) [Jolliffe(2002)], independent component analysis (ICA) [Hyvärinen and Oja(2000)], or even a variational encoder (VAE). In this paper, for the reasons of simplification, we will be using the PCA technique. Assume that Λ denotes the relative ratio of the possible number of elements across subgroups.

2) Uniform Spatial Distribution of Features using Batch min-max Normalisation: To ensure that each feature is equally/likely distributed across the latent space, we introduce a normalisation layer, where we apply batch min-max normalisation to the outputs of the encoder. This layer uses the batch minimum and the maximum of each feature of the encoder output during training. As minimum and maximum values vary from batch (mini-batch) to batch (mini-batch), we update the moving minimum and maximum values during the training process, and use them during the test phase, akin to a batch normalization layer [Ioffe and Szegedy(2015)]. In order to slowly learn the moving minimum and maximum values, they are initialized close to the middle point of $[0, 1)$. After batch min-max normalisation, we need to multiply Λ (obtained using PCA method in our case) to the output of the batch min-max normalisation layer to consider the different number of possible elements at different features.

Since the singular values from PCA are proportional to the variances of the principal components of compressed data, these values are used to obtain relative ratio of the number of possible element in the subgroups [Wall et al.(2003)Wall, Rechtsteiner, and Rocha]. Then, all singular values are divided by the maximum values and are rounded to the nearest one decimal place. The values smaller than unity are replaced with hyperparameter α. The relevant algorithm is shown in Algorithm 1 in the supplementary material.

3) Adding Robustness to Small Perturbations using Interpolation Layer: We achieve this by introducing a layer (Interpolation layer) that performs Gaussian interpolation on the output of the normalized latent space. In [Vincent et al.(2010)Vincent, Larochelle, Lajoie, Bengio, and Manzagol], [Berthelot et al.(2018)Berthelot, Raffel, Roy, and Goodfellow] show that interpolation by Gaussian noise helps mapping unseen examples to known examples, and also makes the latent space locally smooth. Since the proposed model is deterministic, it is important to map a number of unseen examples to the learned representations. This is achieved by adding weight-sensitive Gaussian noise to the outputs of the previous layer during training. Weight-sensitive Gaussian is obtained based on the closest proximal distance of each dimension of the representations. This approach enables unseen examples to fall into the closet representations in the latent space. The relevant algorithm is shown in Algorithm 2 in the supplementary material. It is worth noting that this layer will not be used during the inference / test phase.

4) Mapping using Euler Layer: The final stage of the disentangling process is to perform the mapping outlined in (5). We define a dedicated layer, referred to as the Euler layer, by mapping each latent variable to its cosine and sine values by

$$\mathbf{z_j} \rightarrow (cos(2\pi \mathbf{z_j}), sin(2\pi \mathbf{z_j})) \tag{15}$$

for all $j \in \{1, \ldots, n\}$ as discussed in Section III-C. We illustrate this in Figure 1, where the outputs from the interpolation layer are mapped to cosine and sine values as discussed above.

IV. Evaluation and Results

A. Evaluation Method

We perform our evaluation using five different supervised disentanglement metrics to show the disentanglement ability of the proposed model. Therefore, we use datasets which have ground truth factors for disentanglement analysis.

1) Datasets: One of the critical challenges around evaluating disentanglement is identifying suitable datasets. It is difficult to identify a common dataset that can be used to study this problem. In the literature, different datasets have been used for different purposes. For example, dSprite [Matthey et al.(2017)Matthey, Higgins, Hassabis, and Lerchner] dataset has been used in β-VAE, β-TCVAE, CCI-VAE and FVAE. Although this dataset is useful to understand the traversal order of the latent space, the lack of possibility to fully disentangle the feature space of this dataset prevents us from using this for our study. Similarly, majority of the datasets, such as 3D Chair [Burgess and Kim(2018)] and CelebA [Liu et al.(2015)Liu, Luo, Wang, and Tang] despite having the ground truth, they lack the coherent labelling needed for quantifying the disentanglement. Therefore, in this paper, we utilise the datasets that have been first utilised in [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner], with relevant enhancements, which we describe in the supplementary material (See E). In addition to this toy dataset, we also use three benchmark datasets to evaluate our model, namely, 3D Shape Dataset [Burgess and Kim(2018)], 3D Teapots Dataset [Eastwood and Williams(2018)] and 3D Face Model Dataset [Paysan et al.(2009)Paysan, Knothe, Amberg, Romdhani, and Vetter].

2) Baseline Models: We considered seven different baselines for evaluation, namely, VAE, β-VAE, β-TCVAE, CCI-VAE, FVAE, InfoVAE and WAE. As the proposed technique is purely an AE-based method, we have not included any GAN-specific baselines. To render a fair evaluation mechanism, we used the same encoder and decoder architectures, and same latent space dimensions (for each baseline) throughout the evaluation. We provide a detailed description of these, including the details of the system on which these evaluations were carried out as part of the supplementary material.

3) Performance Metrics: As outlined in Section II-A, there are a number of metrics that can be used to study the performance of disentanglement, depending on the nature of the dataset, access to ground truth, availability of latent factors, and the number of dimensions in the latent space. We use two metrics: **(a) Visualization of the latent space**, and (b) **(b) Numerical disentanglement score**. The former metric permits one to visualize the orthogonality between features and can be used to demonstrate that the model handles combination of categorical and continuous factors in the latent space. The second metric provides a quantifiable method of the disentanglement. We have used five supervised disentanglement scores each of the disentanglement metric classes (see Section II-A), namely, **z-diff** and **z-min** from the intervention-based, **dci-rf** from the predictor-based, and **jemmig** and **dcimig** from the information-based metric classes.

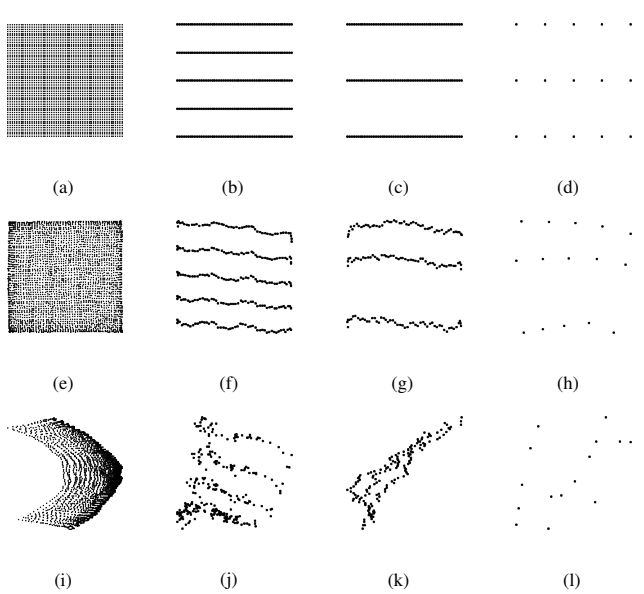
4) Hyperparameter Setting: The proposed model relies on an easily determinable hyperparameter, namely, Λ, that captures the relative ratio of the number of elements for each feature. As shall we discussed later, although the proposed model is not heavily sensitive to this hyperparameter, providing sensible value can lead to best outcomes. Each of the dataset used for the evaluation has varying number of features, and hence the variance between these features. As stated before, we used Algorithm 1 to obtain the values for this hyperparameter, which utilises the PCA method (in our case), and the typical values for $\bar{S}$ in Algorithm 1 are shown in Table IX.

B. Results and Discussions

Our exhaustive evaluation has produced a considerable volume of results, and accounting the limitations of space here, we make the following measures: **(a)** we present the results only for the 2D Toy dataset as main part of the paper, **(b)** we present the remaining results (for the 3D Shape, 3D Face Model and 3D

(a) DAE (b) β-VAE (c) β-TCVAE (d) CCI-VAE (e) FVAE (f) InfoVAE (g) VAE (g) WAE

Fig. 2. Positional relationships (X-Y) in the latent space learned by different models when a dataset has only x and y positional features. In this case, most models are able to disentangle x and y positional features in the latent space.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 3. Relationships between X-Y, X-C, X-S and C-S features in each column, respectively. The first row shows the ideal relationships in the latent space. The second and third rows show the learned latent variables for the DAE and CCI-VAE models.

Teapots datasets) as part of the supplementary material. Please see additional notes provided in the supplementary material, and **(c)** we list additional details, such as hyperparameters of each of the baseline models that yields the best possible outcomes for corresponding baseline model, as part of the supplementary material.

1) Results for the 2D Toy Dataset: Visualization Metric: We show the disentangled (two-dimensional) latent space for the XY dataset in Figure 2 (please see Table X in Appendix A for details of relevant hyperparameters). As x and y positions collectively have 53 possible elements, the ratio of the number of elements in each sub-group is simply one, and hence the notion of hyperparameter for the proposed model under this setting is irrelevant. As can be seen in the figure, the proposed model, in general, provides the ideal grid-shape outlined in [Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner]. The plain vanilla VAE model offers the worst performance. Other models, such as β-VAE, β-TCVAE and CCI-VAE models also come closer to the ideal pattern, and thus most models are able to disentangle x and y positions. However, when colour or shape feature is added to this XY dataset (i.e., for XYC, XYS and $XYCS$ datasets), the disentanglement can become a significant challenge, other than for the proposed model. We show this in Figure 3. For the reasons of brevity, we present the ideal X-Y, X-C, X-S, and C-S relationships and the relationships learned by top two models based on disentanglement scores. These are from the proposed and CCI-VAE models. As we can see, the learned latent space using the proposed model is almost same as the ideal case. However, CCI-VAE fails to ideally disentangle the X-Y positions when colours and shape features are added. In addition to these pairs of latent space, reconstructions of latent traversals across each latent dimension of all datasets are shown in the supplementary materials along with the latent spaces for the other models.

2) Results for the 2D Toy Dataset: Disentanglement Scores: We present the disentanglement scores for XY and $XYCS$ datasets in Tables II and III, respectively, with best performing results bolded.

TABLE II
DISENTANGLEMENT SCORES FOR THE XY DATASET

Models/Metrics	z-diff	z-var	dci-rf	jemmig	dcimig
DAE	**1.00**	**1.00**	**0.99**	**0.85**	**0.84**
VAE	**1.00**	0.84	0.23	0.38	0.25
β-VAE	**1.00**	**1.00**	0.91	0.63	0.60
β-TCVAE	**1.00**	**1.00**	0.93	0.69	0.68
CCI-VAE	**1.00**	**1.00**	0.97	0.82	0.81
FVAE	**1.00**	**1.00**	0.94	0.68	0.65
InfoVAE	**1.00**	**1.00**	0.20	0.34	0.20
WAE	**1.00**	**1.00**	0.58	0.51	0.43

TABLE III
DISENTANGLEMENT SCORES FOR THE $XYCS$ DATASET

Models/Metrics	z-diff	z-var	dci-rf	jemmig	dcimig
DAE	**1.00**	**1.00**	**0.95**	**0.83**	**0.84**
VAE	0.82	0.24	0.08	0.27	0.09
β-VAE	0.97	0.89	0.51	0.40	0.37
β-TCVAE	**1.00**	0.73	0.55	0.50	0.52
CCI-VAE	**1.00**	0.99	0.62	0.48	0.41
FVAE	0.99	0.92	0.19	0.27	0.15
InfoVAE	0.90	0.50	0.21	0.31	0.13
WAE	0.83	0.58	0.20	0.27	0.13

From the results presented in this paper (including the ones includes as part of the supplementary), we can draw the following key observations. First, the proposed model outperforms all models across all metrics for the 2D Toy dataset

(covering XY, XYC, XYS, XYCS datasets), and 3D Teapots datasets (See Table VII). Second, the proposed model is the only model that can successfully disentangle the 3D Teapots dataset (See Table VII and Figure 20). Third, for the 2D Toy dataset, the proposed model maintains the reconstruction loss as small as possible whilst offering improved disentanglement scores (i.e. scores increase) (See Figures 15-18). On the other hand, the reconstruction losses for the β-VAE, β-TCVAE and CCI-VAE models increase along with their disentanglement scores. Finally, β-VAE, β-TCVAE, CCI-VAE and FVAE show relatively better performance than the other models. However, their **dic-rf**, **jemmig** and **dcimig** scores decrease when colour and shape factors, which have much smaller number of elements, are added to the dataset.

V. Conclusions

In the context of representation learning, being able to factorize or disentangle the latent space dimensions is crucial for obtaining latent representations that is composed of multiple, independent factors of variations. On this aspect, deep generative models, particularly that build on autoencoders, play an important role. AE-, particularly, VAE-based models employ two forms of losses to balance the two conflicting goals representation learning: reconstruction loss and factorizability. To favour one over the other, many factorizing models rely on one or more hyperparameters which increases disentanglement ability while reducing reconstruction ability.

In this paper, we presented a non-probabilistic, disentangling autoencoder model, namely, DAE, to address this problem. By exploiting the principles of symmetry transformations in group-theory, we presented a model that only has a reconstruction loss. Although the model relies on a hyperparameter, the model is not overly sensitive to this, and the value can easily be obtained using a number of techniques, such as PCA or ICA or VAE. Our evaluations, performed against a number of VAE-based models, using a number of metrics show that our model can offer the best performance on a number of datasets.

Although the results are encouraging, a number of aspects remain to be investigated. We intend to investigate a number of issues, including evaluation against other metrics and public datasets, and automatic determination of an optimal value for the hyperparameter.

References

[Aneja et al.(2021)Aneja, Schwing, Kautz, and Vahdat] Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. *Advances in Neural Information Processing Systems*, 34, 2021.

[Asperti and Trentin(2020)] Andrea Asperti and Matteo Trentin. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access*, 8:199440–199448, 2020.

[Bengio et al.(2013)Bengio, Courville, and Vincent] 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.

[Berthelot et al.(2018)Berthelot, Raffel, Roy, and Goodfellow] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

[Brakel and Bengio(2017)] Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.

[Burgess and Kim(2018)] Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

[Burgess et al.(2018)Burgess, Higgins, Pal, Matthey, Watters, Desjardins, and Lerchner] 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.

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

[Do and Tran(2019)] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*, 2019.

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

[Eslami et al.(2018)Eslami, Rezende, Besse, Viola, Morcos, Garnelo, Ruderman, Rusu, Dar] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[Higgins et al.(2017)Higgins, Matthey, Pal, Burgess, Glorot, Botvinick, Mohamed, and Lerc] 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.

[Higgins et al.(2018)Higgins, Amos, Pfau, Racaniere, Matthey, Rezende, and Lerchner] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

[Higgins et al.(2021)Higgins, Chang, Langston, Hassabis, Summerfield, Tsao, and Botvinic] Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):1–14, 2021.

[Hyvärinen and Oja(2000)] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13 (4-5):411–430, 2000.

[Ioffe and Szegedy(2015)] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[Iten et al.(2020)Iten, Metger, Wilming, Del Rio, and Renner] Raban Iten, Tony Metger, Henrik Wilming, Lídia Del Rio, and Renato Renner. Discovering physical concepts with neural networks. *Physical review letters*, 124(1):010508, 2020.

[Jolliffe(2002)] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[Kim and Mnih(2018)] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

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

[Liu et al.(2015)Liu, Luo, Wang, and Tang] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[Locatello et al.(2019)Locatello, Tschannen, Bauer, Rätsch, Schölkopf, and Bachem] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

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

[Pan et al.(2020)Pan, Dai, Liu, Loy, and Luo] Xingang Pan, Bo Dai, Ziwei Liu, Chen Change Loy, and Ping Luo. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844*, 2020.

[Paysan et al.(2009)Paysan, Knothe, Amberg, Romdhani, and Vetter] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas

Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.

[Rezende and Mohamed(2015)] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

[Ridgeway(2016)] Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

[Ridgeway and Mozer(2018)] Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. *arXiv preprint arXiv:1802.05312*, 2018.

[Rudin et al.(2022)Rudin, Chen, Chen, Huang, Semenova, and Zhong] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16: 1–85, 2022.

[Sepliarskaia et al.(2019)Sepliarskaia, Kiseleva, de Rijke, et al.] Anna Sepliarskaia, Julia Kiseleva, Maarten de Rijke, et al. Evaluating disentangled representations. *arXiv preprint arXiv:1910.05587*, 2019.

[Spurek et al.(2020)Spurek, Nowak, Tabor, Maziarka, and Jastrzebski] Przemyslaw Spurek, Aleksandra Nowak, Jacek Tabor, Lukasz Maziarka, and Stanislaw Jastrzebski. Non-linear ica based on cramer-wold metric. In *International Conference on Neural Information Processing*, pages 294–305. Springer, 2020.

[Takahashi et al.(2019)Takahashi, Iwata, Yamanaka, Yamada, and Yagi] Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Variational autoencoder with implicit optimal priors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 5066–5073, 2019.

[Tolstikhin et al.(2018)Tolstikhin, Bousquet, Gelly, and Schölkopf] I Tolstikhin, O Bousquet, S Gelly, and B Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR 2018)*. OpenReview. net, 2018.

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

[Tschannen et al.(2018)Tschannen, Bachem, and Lucic] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

[Van Steenkiste et al.(2019)Van Steenkiste, Locatello, Schmidhuber, and Bachem] Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv:1905.12506*, 2019.

[Vincent et al.(2010)Vincent, Larochelle, Lajoie, Bengio, and Manzagol] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[Wall et al.(2003)Wall, Rechtsteiner, and Rocha] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[Yue et al.(2021)Yue, Wang, Sun, Hua, and Zhang] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021.

[Zaidi et al.(2020)Zaidi, Boilard, Gagnon, and Carbonneau] Julian Zaidi, Jonathan Boilard, Ghyslain Gagnon, and Marc-André Carbonneau. Measuring disentanglement: A review of metrics. *arXiv preprint arXiv:2012.09276*, 2020.

[Zhang et al.(2020)Zhang, Zhang, Li, Bengio, and Paull] 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.

[Zhao et al.(2019)Zhao, Song, and Ermon] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5885–5892, 2019.

APPENDIX

A. Algorithms

Algorithm 1: Obtaining Λ using PCA

Input: X: the entire dataset and α: hyperparameter less than 1
Output: $\Lambda = [w_1, w_2, \cdots, w_n]$
$S = [s_1, s_2, \cdots, s_n]$: singular values from PCA(X)
$\bar{S} = [\bar{s_1}, \bar{s_2}, \cdots, \bar{s_n}] = S/max(S)$
$\Lambda = [w_1, w_2, \cdots, w_n]$: round to 1 decimal place of $\bar{S}$
If there exists i such that $w_i < 1$, then $w_i = \alpha$

Algorithm 2: Interpolation layer

Input: x over a mini-batch: $B = \{x_1, \cdots, x_m\}$.
Output: $\{y_i = I(x_i)\}$
$w_i^k = min_{j \in \{1,\cdots,m\}} d(x_i^k, x_j^k)$ where
$x_i = (x_i^k)_{k=1,\cdots,n}$
$y_i^k = x_i^k + w_i^k * \varepsilon \equiv S(x)$ where ε $\mathcal{N}(0, 1)$

B. Disentanglement scores

TABLE IV
DISENTANGLEMENT SCORES FOR THE XYC DATASET

Models / Metrics	z-diff	z-var	dci-rf	jemmig	dcimig
DAE	**1.00**	**1.00**	**0.99**	**0.91**	**0.91**
VAE	**1.00**	0.70	0.14	0.24	0.16
β-VAE	**1.00**	**1.00**	0.83	0.58	0.52
β-TCVAE	**1.00**	**1.00**	0.94	0.78	0.77
CCI-VAE	**1.00**	**1.00**	0.91	0.66	0.62
FVAE	**1.00**	**1.00**	0.27	0.34	0.20
InfoVAE	**1.00**	0.67	0.25	0.29	0.21
WAE	**1.00**	0.77	0.21	0.28	0.14

TABLE V
DISENTANGLEMENT SCORES FOR THE XYS DATASET

Models / Metrics	z-diff	z-var	dci-rf	jemmig	dcimig
DAE	**1.00**	**1.00**	**0.98**	**0.85**	**0.86**
VAE	**1.00**	0.78	0.37	0.37	0.33
β-VAE	**1.00**	**1.00**	0.96	0.69	0.65
β-TCVAE	**1.00**	**1.00**	0.97	0.83	0.80
CCI-VAE	**1.00**	**1.00**	0.91	0.68	0.65
FVAE	**1.00**	0.85	0.48	0.46	0.39
InfoVAE	**1.00**	0.68	0.28	0.31	0.28
WAE	0.99	0.46	0.10	0.24	0.15

TABLE VI
DISENTANGLEMENT SCORES FOR 3D SHAPE DATASET

Models / Metrics	z-diff	z-var	dci-rf	jemmig	dcimig
DAE	**1.00**	0.96	0.90	0.74	0.73
VAE	0.96	0.59	0.36	0.26	0.22
β-VAE	**1.00**	**1.00**	0.91	0.70	0.68
β-TCVAE	**1.00**	0.85	0.78	0.61	0.63
CCI-VAE	0.98	0.89	0.74	0.59	0.59
FVAE	**1.00**	**1.00**	**0.97**	**0.82**	**0.81**
InfoVAE	0.99	0.73	0.33	0.23	0.20
WAE	0.93	0.43	0.13	0.12	0.06

TABLE VII
DISENTANGLEMENT SCORES FOR 3D TEAPOTS DATASET

Models / Metrics	z-diff	z-var	dci-rf	jemmig	dcimig
DAE	**1.00**	**1.00**	**0.80**	**0.54**	**0.53**
VAE	0.99	0.77	0.44	0.38	0.22
β-VAE	0.90	0.73	0.46	0.36	0.23
β-TCVAE	**1.00**	0.85	0.68	0.47	0.37
CCI-VAE	0.89	0.62	0.41	0.35	0.14
FVAE	0.99	0.79	0.50	0.39	0.26
InfoVAE	0.99	0.70	0.46	0.37	0.24
WAE	0.77	0.52	0.16	0.22	0.05

TABLE VIII
DISENTANGLEMENT SCORES FOR 3D FACE MODEL DATASET

Models / Metrics	z-diff	z-var	dci-rf	jemmig	dcimig
DAE	**1.00**	0.82	0.57	0.46	**0.44**
VAE	**1.00**	0.66	0.48	0.38	0.23
β-VAE	**1.00**	0.74	**0.68**	0.48	0.36
β-TCVAE	**1.00**	**0.83**	0.65	**0.54**	**0.44**
CCI-VAE	**1.00**	**0.83**	0.61	0.47	0.34
FVAE	**1.00**	0.65	0.48	0.37	0.21
InfoVAE	0.99	0.68	0.46	0.39	0.21
WAE	**1.00**	0.75	0.21	0.26	0.16

C. Hyperparameters

TABLE IX
$\bar{S}$ VALUES FOR DIFFERENT DATASETS

Dataset	$\bar{S}$
XY	[1.0, 1.0]
XYC	[1.0, 1.0, 0.8]
XYS	[1.0, 1.0, 0.8]
XYCS	[1.0, 1.0, 0.8, 0.8]
3D Shape	[1.0, 1.0, 1.0, 1.0, 0.5, 0.5]
3D Teapots	[1.0, 0.8, 0.8, 0.4, 0.3, 0.3]
3D Face Model	[1.0, 0.4, 0.4, 0.3]

TABLE X
BEST HYPERPARAMETERS FOR MODELS FOR DIFFERENT DATASETS.

Model / Dataset	XY	XYC	XYS	XYCS
DAE (α)	–	0.005	0.001	0.0005
β-VAE (β)	16	64	64	32
β-TCVAE (β)	32	64	128	128
CCI-VAE (C)	500	100	100	100
FVAE (γ)	200	100	100	500
InfoVAE (λ)	100	100	100	500
WAE (λ)	1	50	30	50

TABLE XI
BEST HYPERPARAMETERS FOR MODELS FOR DIFFERENT DATASETS.

Model / Dataset	3D Shape	3D Teapots	3D Face Model
DAE (α)	0.01	0.1	0.1
β-VAE (β)	64	6	16
β-TCVAE (β)	32	6	32
CCI-VAE (C)	100	50	100
FVAE (γ)	5	1	1
InfoVAE (λ)	100	50	2000
WAE (λ)	50	10	50

D. Encoder and Decoder architectures

TABLE XII
ARCHITECTURE FOR 2D TOY DATASET

Encoder	Decoder
Input $84\times84\times1$ image	3×3 1 Conv $\downarrow$, Sigmoid
10×10 8 Conv $\downarrow$, BN, LReLU	10×10 1 Conv $\uparrow$, BN, LReLU
10×10 16 Conv $\downarrow$, BN, LReLU	10×10 8 Conv $\uparrow$, BN, LReLU
FC 64	FC 256, LReLU
FC The number of features	FC 64, LReLU

TABLE XIII
ARCHITECTURE FOR 3D SHAPE DATASET

Encoder	Decoder
Input $64\times64\times3$ image	3×3 1 Conv $\downarrow$, Sigmoid
4×4 32 Conv $\downarrow$, BN, LReLU	4×4 3 Conv $\uparrow$, BN, LReLU
4×4 32 Conv $\downarrow$, BN, LReLU	4×4 32 Conv $\uparrow$, BN, LReLU
4×4 64 Conv $\downarrow$, BN, LReLU	4×4 32 Conv $\uparrow$, BN, LReLU
4×4 64 Conv $\downarrow$, BN, LReLU	4×4 64 Conv $\uparrow$, BN, LReLU
FC 256	FC 1024, LReLU
FC 6	FC 256, LReLU

TABLE XIV
ARCHITECTURE FOR 3D TEAPOTS DATASET

Encoder	Decoder
Input $64\times64\times3$ image	3×3 1 Conv $\downarrow$
4×4 32 Conv $\downarrow$, BN, ReLU	4×4 3 Conv $\uparrow$, BN, ReLU
4×4 32 Conv $\downarrow$, BN, ReLU	4×4 32 Conv $\uparrow$, BN, ReLU
4×4 64 Conv $\downarrow$, BN, ReLU	4×4 32 Conv $\uparrow$, BN, ReLU
4×4 64 Conv $\downarrow$, BN, ReLU	4×4 64 Conv $\uparrow$, BN, ReLU
FC 128	FC 1024, LReLU
FC 6	FC 128, LReLU

TABLE XV
ARCHITECTURE FOR 3D FACE MODEL DATASET

Encoder	Decoder
Input $64\times64\times1$ image	3×3 1 Conv $\downarrow$, Sigmoid
4×4 32 Conv $\downarrow$, BN, LReLU	4×4 1 Conv $\uparrow$, BN, LReLU
4×4 32 Conv $\downarrow$, BN, LReLU	4×4 32 Conv $\uparrow$, BN, LReLU
4×4 64 Conv $\downarrow$, BN, LReLU	4×4 32 Conv $\uparrow$, BN, LReLU
4×4 64 Conv $\downarrow$, BN, LReLU	4×4 64 Conv $\uparrow$, BN, LReLU
FC 128	FC 1024, LReLU
FC 4	FC 128, LReLU

E. Dataset

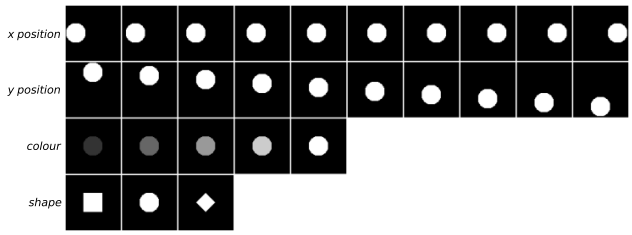
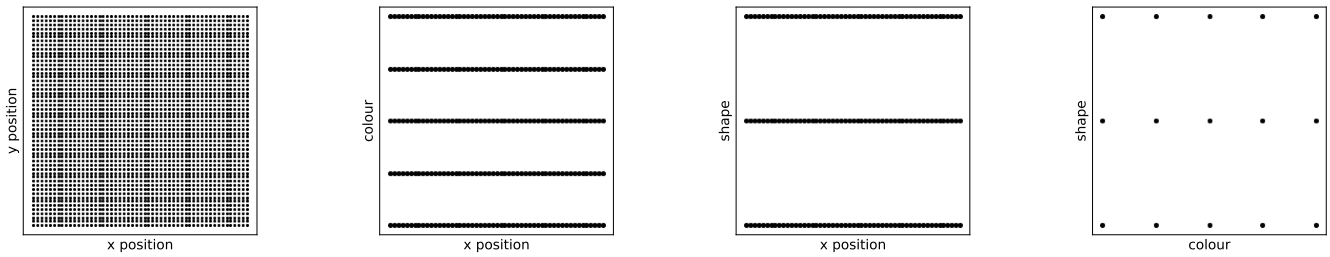


Fig. 4. Four factors in datasets. x and y positions have 53 elements, colour has 5 elements and shape has 3 elements.

1) 2D Toy Dataset: This dataset has objects with three shapes (S) (a circles, a rectangles and a diamonds), and variations to their x and y positions and colour information (more specifically, the brightness). This is a rather small, but very effective, dataset. There are 53 unique x positions (X), 53 unique y positions (Y) and 5 colours (C). We create XY, XYC, XYS and $XYCS$ sub-datasets to show the differences of the latent space when the combination of categorical and continuous factors are presented.
2) 3D Shape Dataset [Burgess and Kim(2018)]: This dataset has $480,000$, three-channel RGB, $64 \times 64 \times 3$ images of 3D objects with ground truth factors of four shapes, eight scales, 15 orientations, 10 floor colour, 10 wall colours, and 10 object colours.
3) 3D Teapots Dataset [Eastwood and Williams(2018)]: This dataset has two million, three-channel RGB, $64 \times 64 \times 3$ images of a 3D object (teapot) with ground truth factors of independently sampled from its respective distribution: azimuth $\sim U[0, 2\pi]$, elevation $\sim U[0, \pi/2]$, and three colours, namely, red (R), green (G) and blue (B), sampled with $R \sim U[0, 1]$, $G \sim U[0, 1]$, and $B \sim U[0, 1]$. This dataset is very effective to evaluate model when all factors are independently from the uniform distributions.
4) 3D Face Model Dataset [Paysan et al.(2009)Paysan, Knothe, Amberg, Romdhani, and Vetter]: This dataset has $127,050$, greyscale, 64×64 images of 3D faces with ground truth factors of 50 different face ids, 21 azimuth, 11 elevation and 11 lighting conditions.

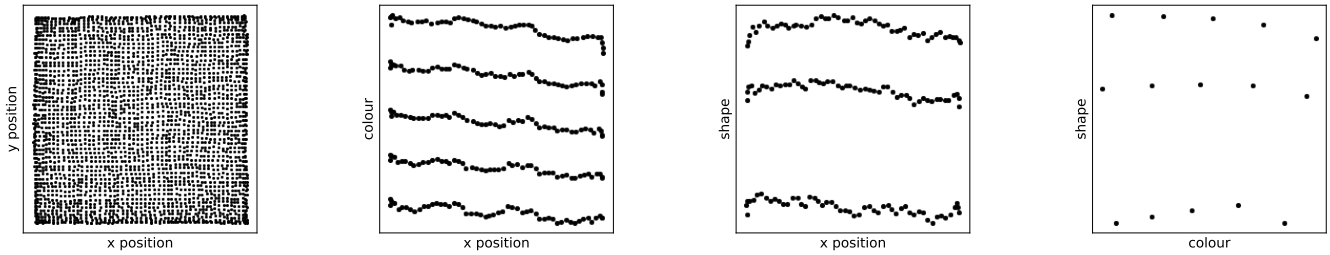
F. System and Model Configurations

All of our experiments were run on a single hardware consisting two DGX2 nodes, collectively consisting of 32-V100 GPUs, 1.5GB GPU RAM, and 3TB System RAM. Encoder and decoder architecture are the same in all experiments. Encoder has two convolutional layers followed by Batch Normalization layer and LeakyReLU activation. After convolutional layers, there is one fully-connected layer with 64 nodes and another layer which maps to the latent space. The decode part is symmetric to the encoder part. C for CCI-VAE is set as 25 for all experiments.



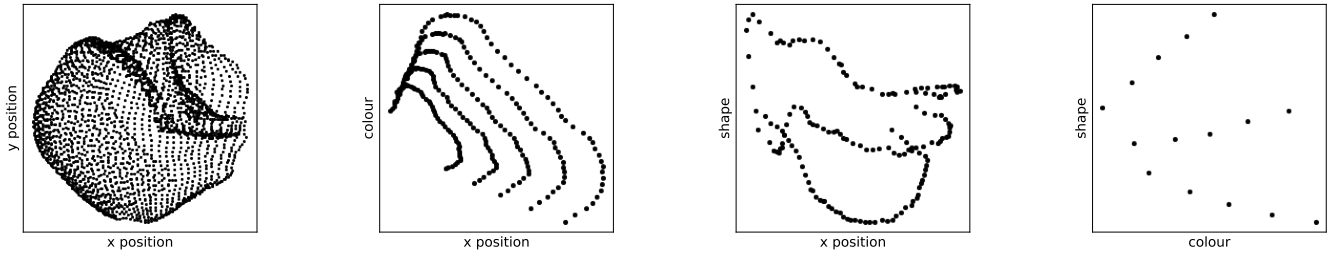
(a) Ideal X-Y relationship (b) Ideal X-C relationship (c) Ideal X-S relationship (d) Ideal C-S relationship

Fig. 5. Ideal relationships between X-Y, X-C, X-S and C-S features.



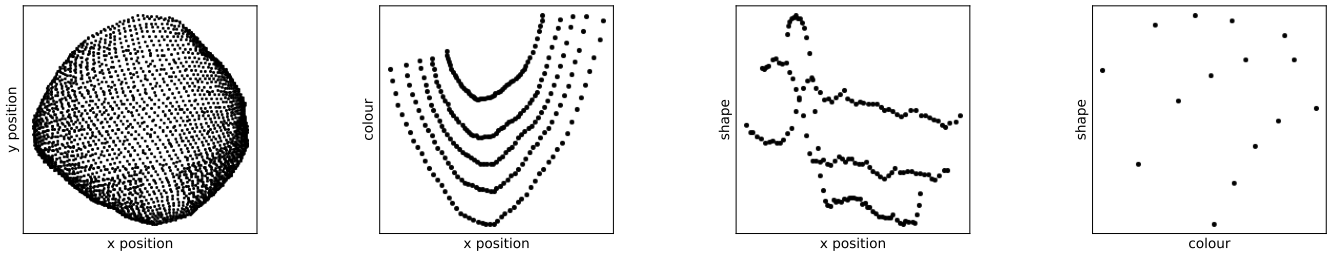
(a) X-Y relationship in DAE (b) X-C relationship in DAE (c) X-S relationship in DAE (d) C-S relationship in DAE

Fig. 6. Relationships between X-Y, X-C, X-S and C-S features in DAE.



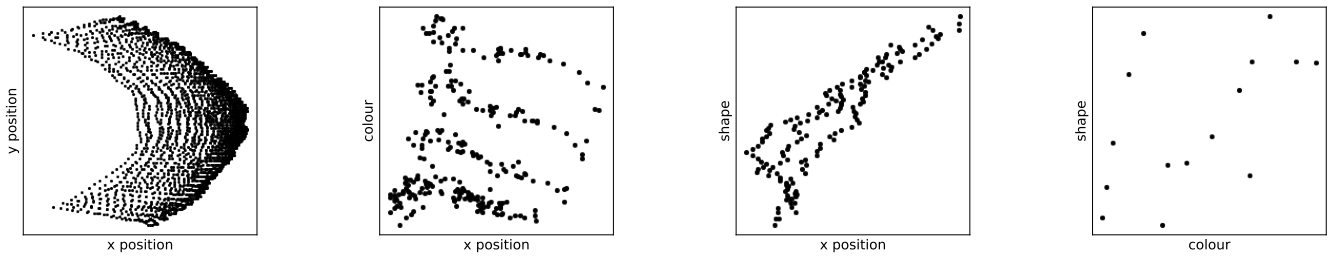
(a) X-Y relationship in β-VAE (b) X-C relationship in β-VAE (c) X-S relationship in β-VAE (d) C-S relationship in β-VAE

Fig. 7. Relationships between X-Y, X-C, X-S and C-S features in β-VAE.



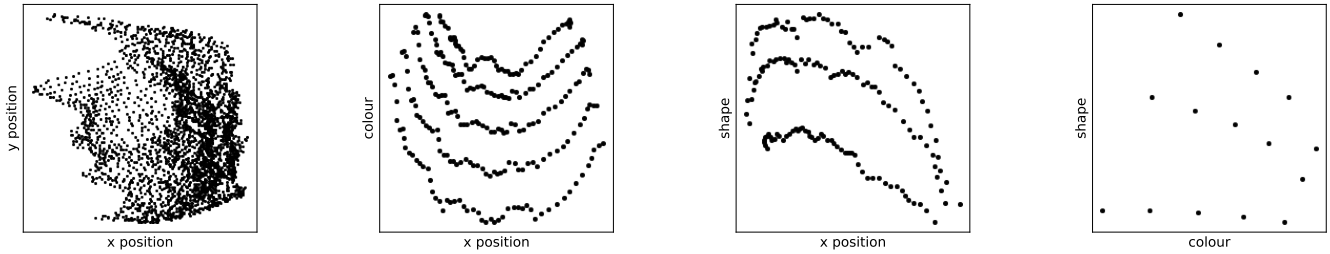
(a) X-Y relationship in β-TCVAE (b) X-C relationship in β-TCVAE (c) X-S relationship in β-TCVAE (d) C-S relationship in β-TCVAE

Fig. 8. Relationships between X-Y, X-C, X-S and C-S features in β-TCVAE.



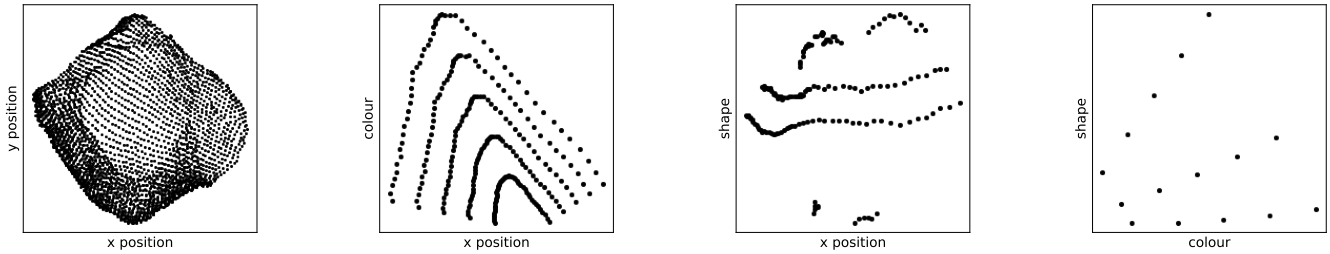
(a) X-Y relationship in CCIVAE (b) X-C relationship in CCI-VAE (c) X-S relationship in CCI-VAE (d) C-S relationship in CCI-VAE

Fig. 9. Relationships between X-Y, X-C, X-S and C-S features in CCI-VAE.



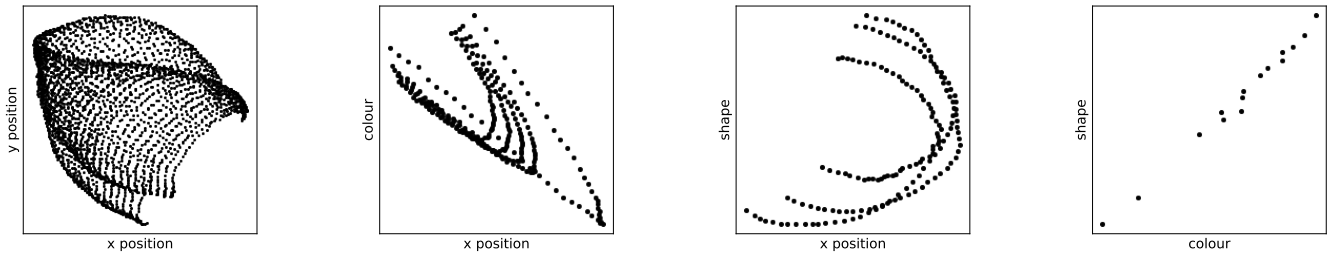
(a) X-Y relationship in FVAE (b) X-C relationship in FVAE (c) X-S relationship in FVAE (d) C-S relationship in FVAE

Fig. 10. Relationships between X-Y, X-C, X-S and C-S features in FVAE.



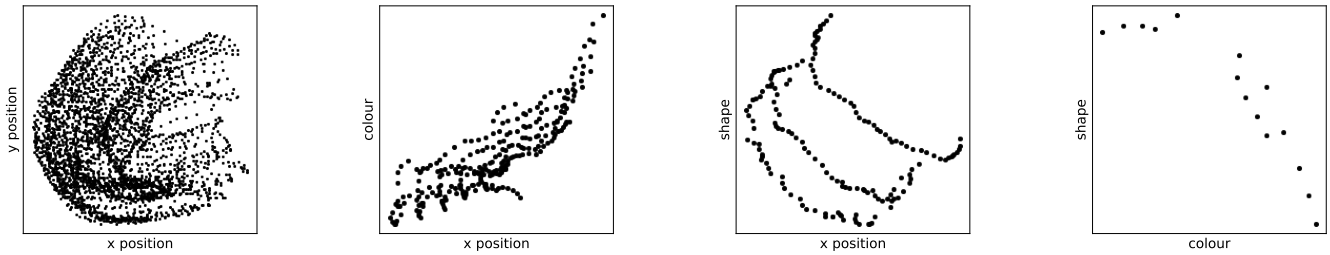
(a) X-Y relationship in InfoVAE (b) X-C relationship in InfoVAE (c) X-S relationship in InfoVAE (d) C-S relationship in InfoVAE

Fig. 11. Relationships between X-Y, X-C, X-S and C-S features in InfoVAE.



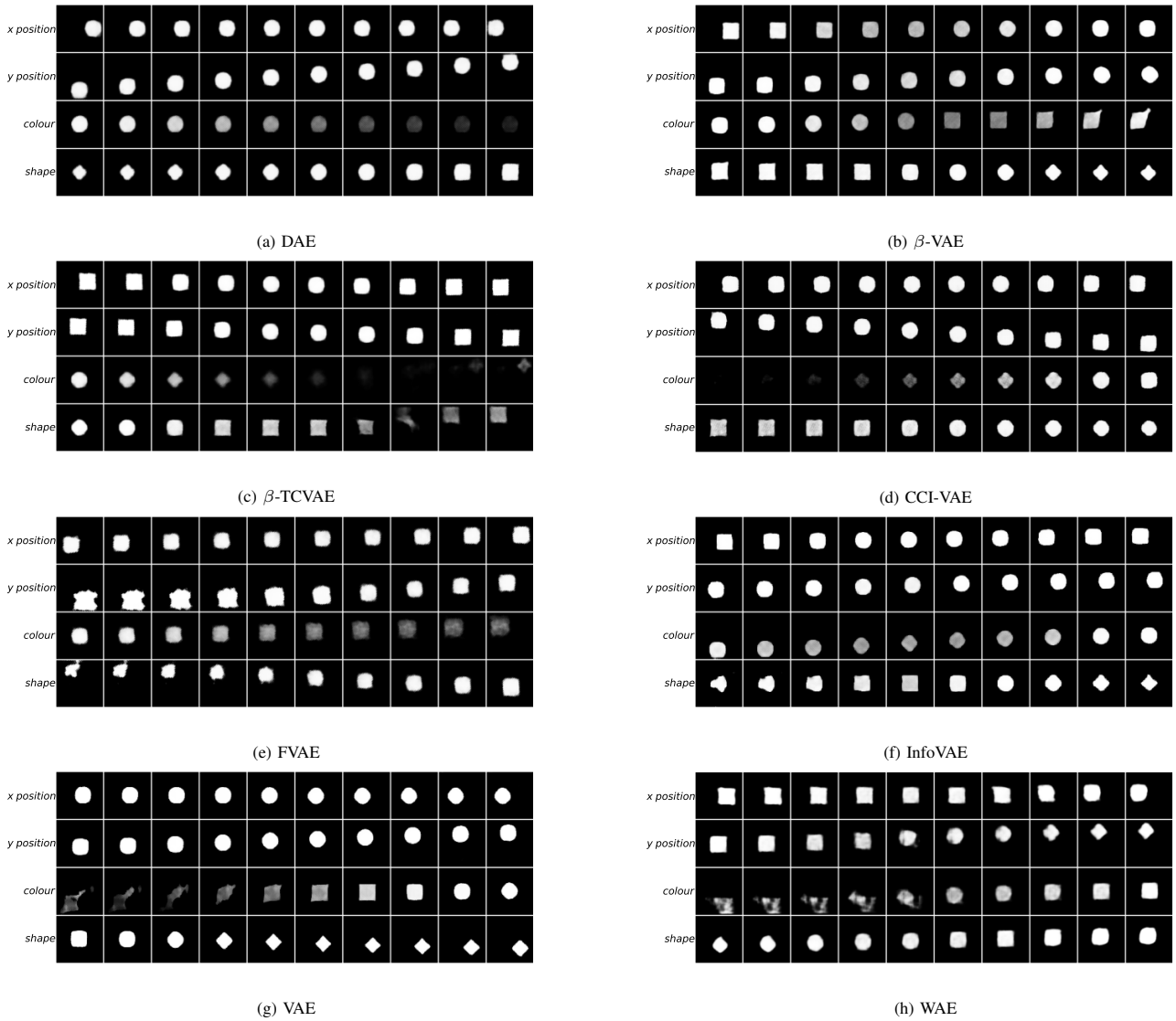
(a) X-Y relationship in VAE (b) X-C relationship in VAE (c) X-S relationship in VAE (d) C-S relationship in VAE

Fig. 12. Relationships between X-Y, X-C, X-S and C-S features in VAE.



(a) X-Y relationship in WAE (b) X-C relationship in WAE (c) X-S relationship in WAE (d) C-S relationship in WAE

Fig. 13. Relationships between X-Y, X-C, X-S and C-S features in WAE.



(a) DAE (b) β-VAE

(c) β-TCVAE (d) CCI-VAE

(e) FVAE (f) InfoVAE

(g) VAE (h) WAE

Fig. 14. Reconstructions of latent traversals across each latent dimension in the XYCS dataset.

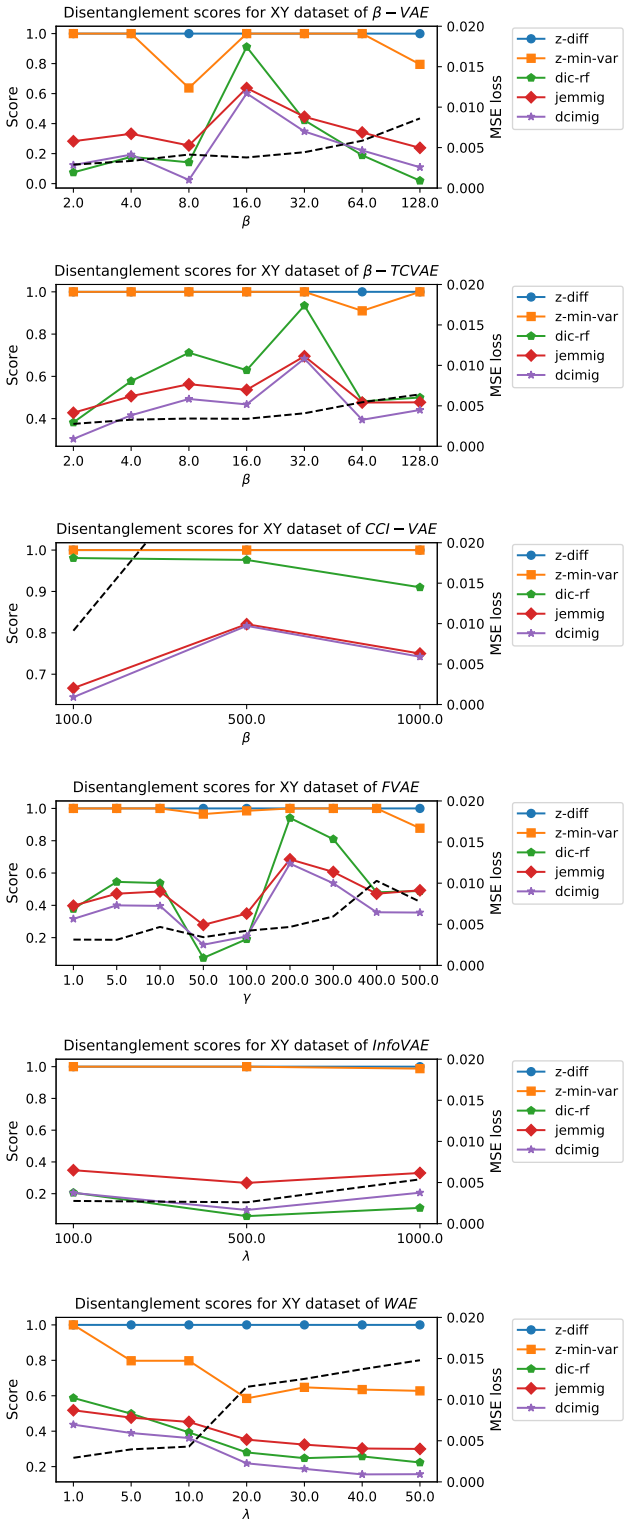


Fig. 15. Disentanglement scores with XY dataset with respect to hyperparameters.

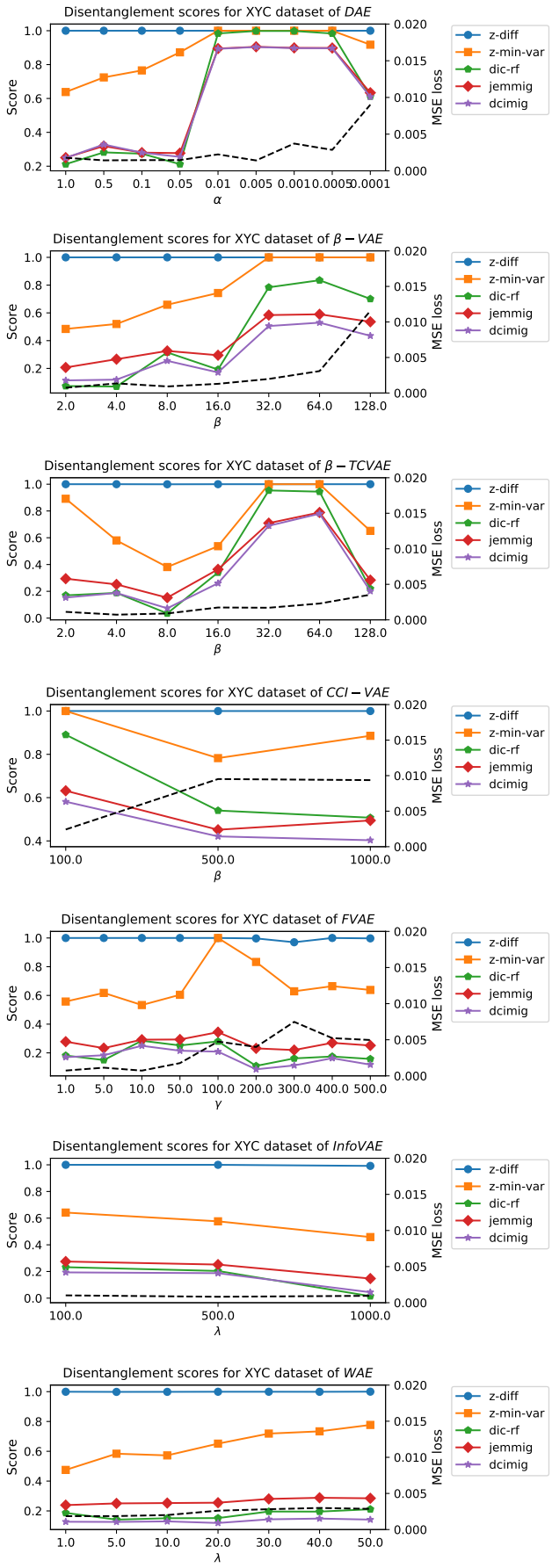


Fig. 16. Disentanglement scores with XYC dataset with respect to hyperparameters.

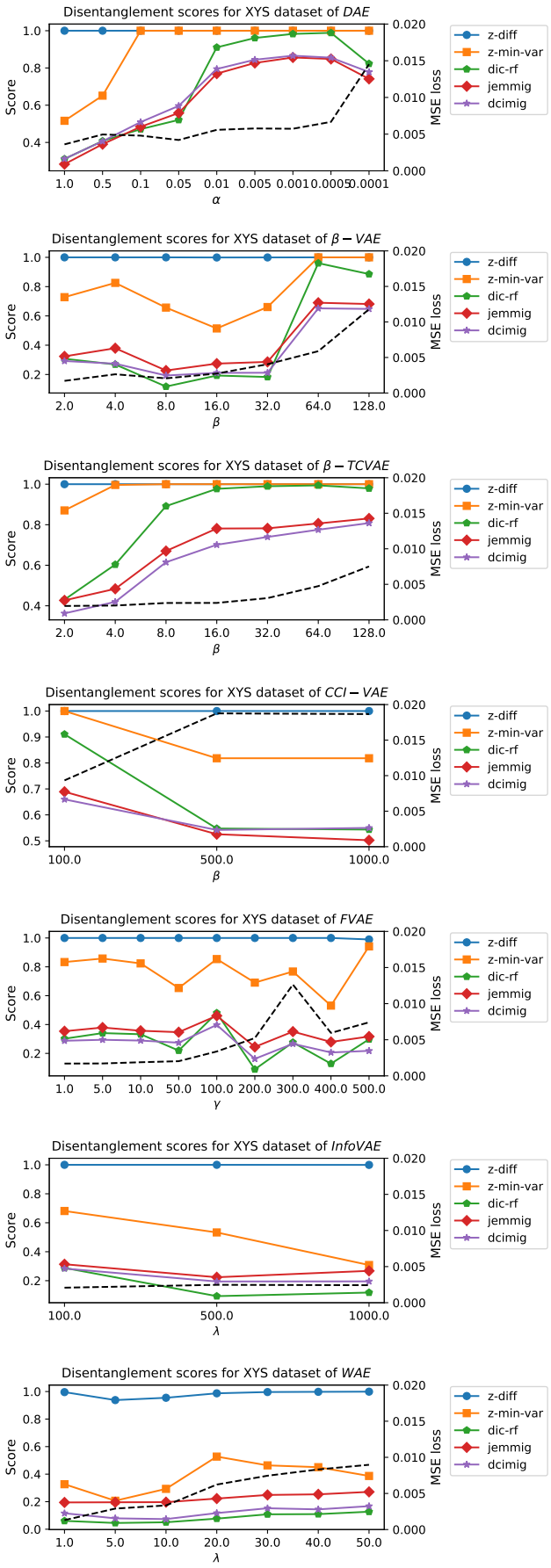


Fig. 17. Disentanglement scores with XYS dataset with respect to hyperparameters.

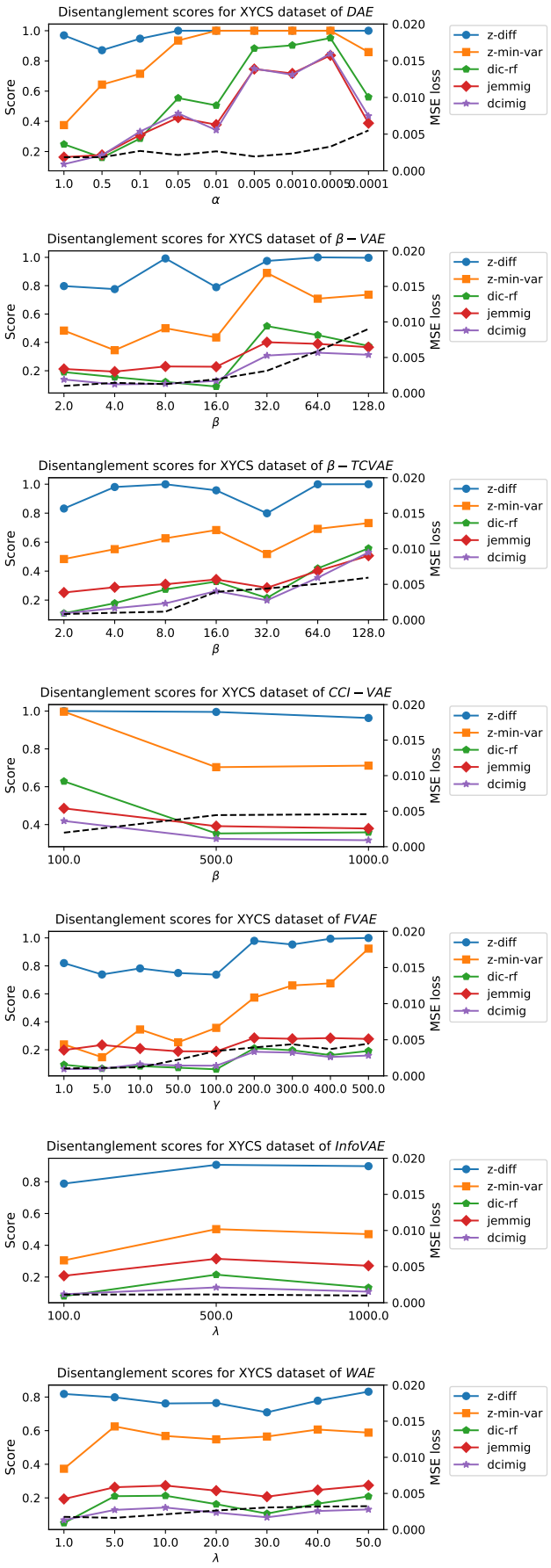
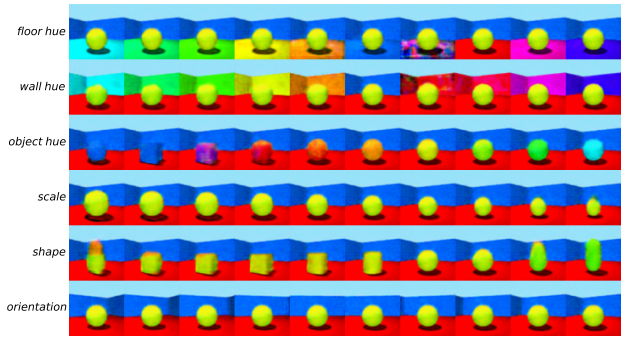
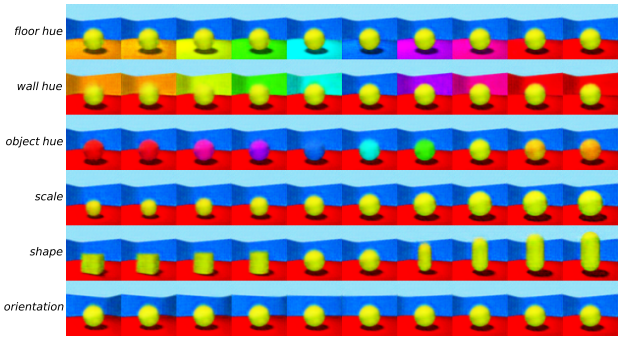


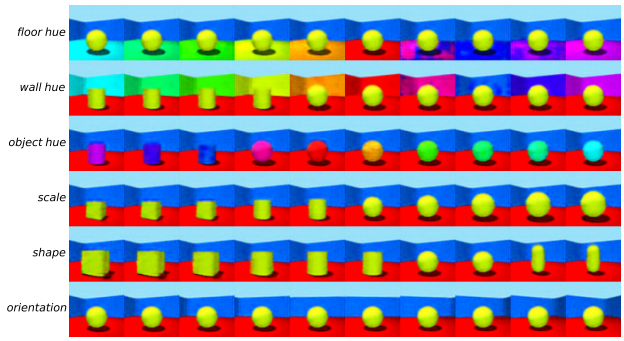
Fig. 18. Disentanglement scores with XYCS dataset with respect to hyperparameters.



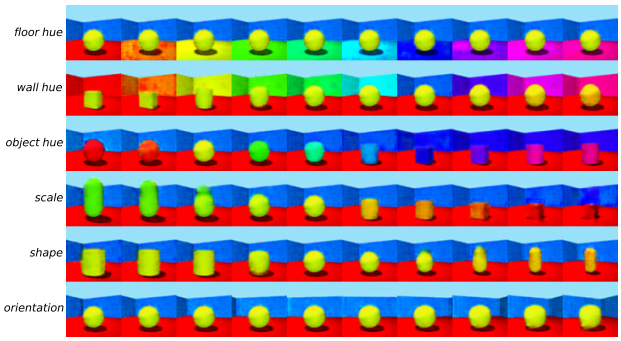
(a) DAE



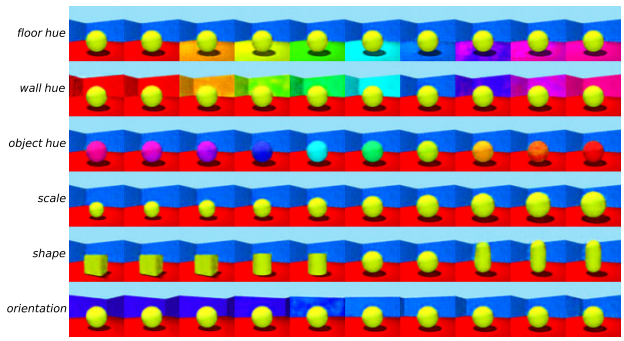
(b) β-VAE



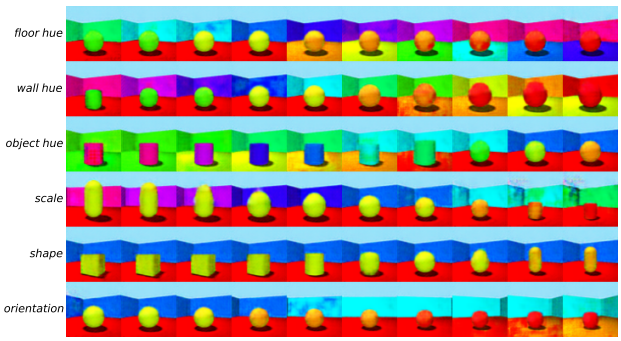
(c) β-TCVAE



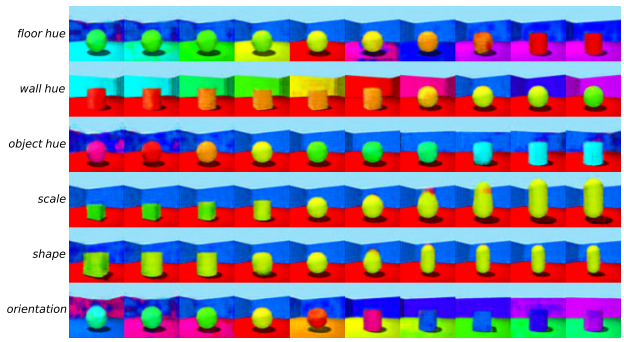
(d) CCI-VAE



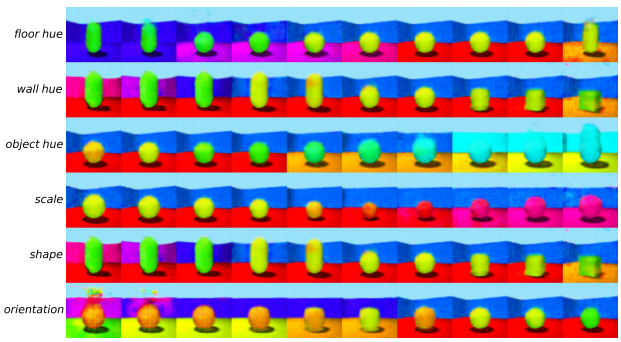
(e) FVAE



(f) InfoVAE

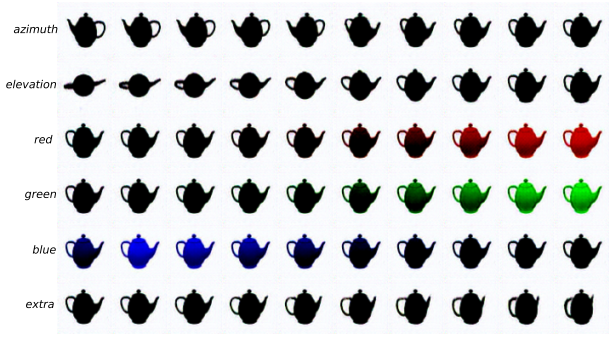


(g) VAE

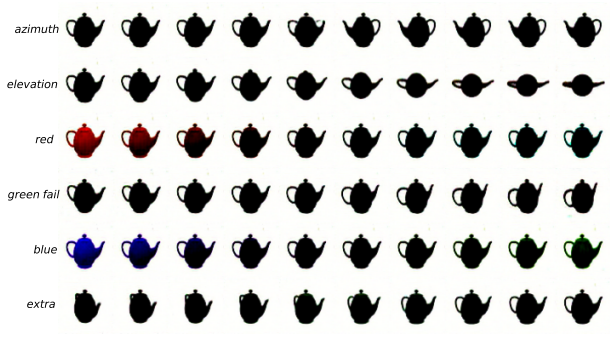


(h) WAE

Fig. 19. Reconstructions of latent traversals across each latent dimension in the 3D Shape dataset.

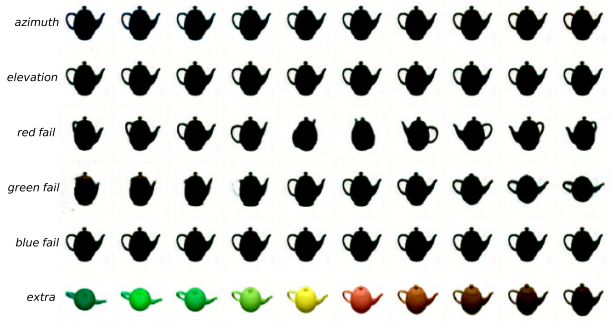


(a) DAE

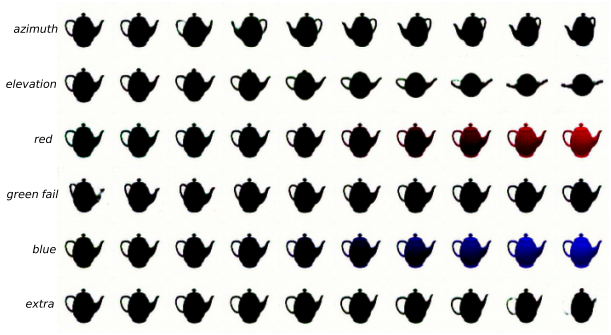


(b) β-VAE

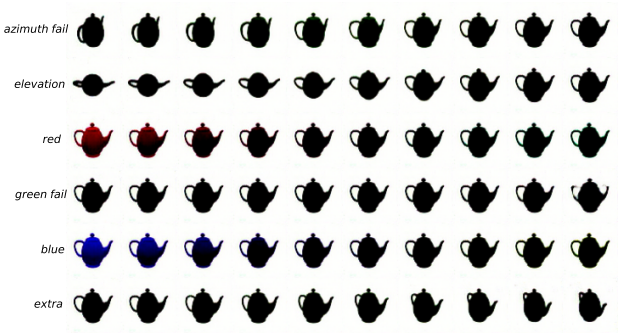
(c) β-TCVAE



(d) CCI-VAE

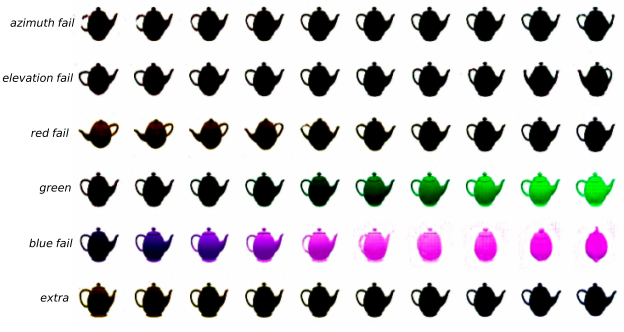


(e) FVAE



(f) InfoVAE

(g) VAE



(h) WAE

Fig. 20. Reconstructions of latent traversals across each latent dimension in the 3D Teapots dataset.

(a) DAE

(b) β-VAE

(c) β-TCVAE

(d) CCI-VAE

(e) FVAE

(g) VAE

Fig. 21. Reconstructions of latent traversals across each latent dimension in the 3D Face Model dataset. We do not visualize results of InfoVAE and WAE since both models fail to disentangle the data.