Overlooked Implications of the Reconstruction Loss for VAE Disentanglement

Nathan Michlo, Richard Klein, Steven James
University of the Witwatersrand, Johannesburg, South Africa
nathan.michlo1@students.wits.ac.za, {richard.klein, steven.james}@wits.ac.za

Abstract

Learning disentangled representations with variational autoencoders (VAEs) is often attributed to the regularisation component of the loss. In this work, we highlight the interaction between data and the reconstruction term of the loss as the main contributor to disentanglement in VAEs. We show that standard benchmark datasets have unintended correlations between their subjective ground-truth factors and perceived axes in the data according to typical VAE reconstruction losses. Our work exploits this relationship to provide a theory for what constitutes an adversarial dataset under a given reconstruction loss. We verify this by constructing an example dataset that prevents disentanglement in state-of-the-art frameworks while maintaining human-intuitive ground-truth factors. Finally, we re-enable disentanglement by designing an example reconstruction loss that is once again able to perceive the ground-truth factors. Our findings demonstrate the subjective nature of disentanglement and the importance of considering the interaction between the ground-truth factors, data and notably, the reconstruction loss, which is under-recognised in the literature.

1 Introduction

A fundamental challenge in machine learning is discovering useful representations from high-dimensional data that can be used to solve subsequent tasks effectively. Recently, deep learning approaches have showcased the ability of neural networks to extract meaningful features from high-dimensional inputs for tasks such as classification [Krizhevsky et al., 2012] and reinforcement learning [Mnih et al., 2015]. However, these learned representations are often not semantically meaningful, which can negatively impact interpretability, fairness [Locatello et al., 2019a], and downstream task performance [Locatello et al., 2019b].

Prior work has therefore argued that it is desirable to learn a representation that is disentangled [Bengio et al., 2013]. While there is no consensus on what constitutes a disentangled representation, it is generally agreed that it should be factorised so that each latent variable corresponds to a single explanatory variable responsible for generating the data [Burgess et al., 2017]. For example, a single image from a video game may be represented by continuous latent variables governing the $x$ and $y$ positions of the player or enemies, as well as categorical variables governing their clothing or appearance.

A common approach to discovering these representations is variational autoencoders (VAEs) [Kingma and Welling, 2014], which are trained on unlabelled data to learn a low-dimensional representation capable of reconstructing the input. However, it has been shown that unsupervised methods cannot reliably learn representations without the introduction of supervision or inductive biases [Locatello et al., 2019b]. The recently introduced Ada-GVAE framework partially overcame this problem by using a weakly supervised signal to discover underlying factors [Locatello et al., 2020], but there remains room for improvement.

Interestingly, VAEs do not have an explicit mechanism that encourages the learning of disentangled representations, but it is theorised that this behaviour is related to the regularisation term and the information bottleneck principle [Burgess et al., 2017; Mathieu et al., 2019; Rolinek et al., 2019]. However, despite this hypothesis, there is still no explicit reason for why the representations learnt by these frameworks should align with generative factors in the data. Nonetheless, these frameworks have been shown to produce disentangled representations when trained on synthetically generated data, as measured by appropriate metrics [Eastwood and Williams, 2018; Chen et al., 2018; Zaidi et al., 2020].

In this paper, we aim to understand why VAEs implicitly learn disentangled representations by investigating the interaction between the reconstruction loss of the VAE and the input data. We provide compelling evidence that disentanglement occurs not because of special algorithmic choices or the regularisation term, but because of how VAEs perceive distances between observations in the datasets themselves according to the reconstruction loss, and the fact that these distances accidentally correlate to the chosen ground-truth factors generating the data. In particular, we find that standardised benchmarks are constructed in such a way that they unintentionally encourage models to learn what appear to be disentangled representations.

The main, summarised contributions of this paper\(^1\) are:

---

\(^1\)Our extended paper with factor importance results and full experiment details is available at: https://arxiv.org/abs/2202.13341
(i) We introduce the concept of perceived distance, in terms of the VAE reconstruction loss, to measure overlap or similarity between dataset pairs. We demonstrate that perceived distances in existing datasets unintentionally correspond to the distances between ground-truth factors, and that VAEs learn these distances, explaining why learnt representations may appear disentangled.

(ii) We provide a technique to visualise the correlation between perceived distances in the data and ground-truth factors generating the data. We use this understanding to provide a theory for what constitutes an adversarial dataset under a given reconstruction loss.

(iii) We reveal the ineffectiveness of state-of-the-art models by using our theory to design a simple, example adversarial dataset with constant perceived distance between elements, over which VAE-based frameworks fail to learn disentangled representations.

(iv) We provide an example solution to the adversarial dataset that modifies the reconstruction loss, and thus perceived distances across the dataset, so that VAE frameworks are again able to capture the ground-truth factors.

(v) We contribute Disent, a general PyTorch [Paszke et al., 2017] disentanglement framework, with common models, metrics, and datasets.\(^2\)

2 Background

Assume a dataset \(\mathcal{X} = \{x^{(0)}, \ldots, x^{(n)}\}\) is a set of independent and identically distributed (i.i.d) observations \(x \in \mathbb{R}^N\), generated by some random process involving an unobserved random variable \(z \in \mathbb{R}^D\) of lower dimensionality \(D \ll N\). Additionally, the true prior distribution \(z \sim p_\star(z)\) and true conditional distribution \(x \sim p_\star(x|z)\) are unknown. Variational autoencoders (VAEs) aim to learn this generative process. Unlike autoencoders (AEs), which consist of an encoder \(f_\phi(x) = z\) and decoder \(g_\theta(z) = \hat{x}\) with weights \(\phi\) and \(\theta\), VAEs instead construct a probabilistic encoder by outputing the output from the encoder or inference model parameterise approximate posterior distributions \(z \sim q_\phi(z|x)\). The approximate posterior is then sampled during training to obtain representations \(z\), which are then decoded using the generative model to obtain reconstructions \(\hat{x} \sim p_\theta(x|z)\).

A factorised Gaussian encoder [Kingma and Welling, 2014] is commonly used. The posterior is modelled using a multivariate Gaussian distribution with diagonal covariance \(z \sim \mathcal{N}(\mu_\phi(x), \sigma_\phi(x))\), and the prior is given by the multivariate normal distribution \(p_\star(z) \sim \mathcal{N}(0, I)\), with a mean of 0 and diagonal covariance \(I\). To enable backpropagation, the reparameterisation trick in Equation (1) is used to sample from the posterior distribution by offsetting the distribution means by scaled noise values.\(^3\)

\[
    z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, I) \tag{1}
\]

VAEs maximise the evidence lower bound (ELBO) by minimising the loss given by Equation (4). VAE-based approaches often make slight modifications to this loss [Higgins et al., 2016; Zhao et al., 2017; Hou et al., 2017; Kumar et al., 2018; Chen et al., 2018; Kim and Mnih, 2018; Locatello et al., 2020], but the terms of these modified loss functions can usually still be grouped into reconstruction and regularisation components, given by Equations 2 and 3 respectively. The regularisation term affects the representations learnt by the encoder, while the reconstruction term improves the outputs from the decoder. These terms usually contradict in practice, with strong regularisation leading to worse reconstructions but often better disentanglement [Higgins et al., 2016; Burgess et al., 2017].

\[
    \mathcal{L}_{\text{rec}}(x, \hat{x}) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \tag{2}
\]

\[
    \mathcal{L}_{\text{reg}}(x) = -D_{\text{KL}}(q_\phi(z|x) \parallel p_\star(z)) \tag{3}
\]

\[
    \mathcal{L}_{\text{VAE}}(x, \hat{x}) = \mathcal{L}_{\text{rec}}(x, \hat{x}) + \mathcal{L}_{\text{reg}}(x) \tag{4}
\]

2.1 Random Sampling Reorganises VAE Embeddings

Disentanglement in VAEs is generally attributed to the regularisation term in Equation (3); however, we highlight that regularisation only enables the underlying disentanglement mechanism. Disentanglement arises rather as a result of VAEs reorganising the latent space to minimise reconstruction mistakes due to random sampling from the probabilistic encoder during training. Through this mechanism, a VAE will place similar observations according to the reconstruction loss in Equation (2) close together in the latent space [Burgess et al., 2017; Mathieu et al., 2019; Zietlow et al., 2021], as this action minimises sampling errors.

The regularisation term enables this interaction by controlling the overlap between latent distributions corresponding to different inputs. If these distributions overlap sufficiently, the decoder will often attribute a random sample to an incorrect input, see Figure 1. Thus, a mistake will be made during the decoding process, which encourages reorganisation to minimise the reconstruction error.

3 Related Work

The following works are the most applicable to our research, falling under three general categories: (i) explanations for disentanglement, (ii) the role of the reconstruction loss in disentanglement, and (iii) problems with disentanglement.

Firstly, Burgess et al. [2017] relate VAEs to the information bottleneck principle. Which explains that random sampling...
leads to a local minimisation of the reconstruction loss which reorganises the latent space so that points close in pixel space are close in the latent space. Mathieu et al. [2019] argue that VAEs do not explicitly encourage disentanglement through their design. Rather, they provide the explanation that the diagonal prior typically used in VAEs when combined with random sampling produces a similar effect to PCA. Our work takes inspiration from these ideas to develop the theory of perceived overlap in VAEs, which we use to analyse datasets and improve or hinder disentanglement.

Secondly, inspired by Burgess et al. [2017] most modern frameworks offer some way to balance the regularisation and reconstruction components of the loss, the ControlVAE automates this process [Shao et al., 2020]. Hou et al. [2017] instead swap out the reconstruction loss of VAEs for a perceptual loss function, which can improve the representations learnt by the model. Zietlow et al. [2021] extend the analysis of Mathieu et al. [2019]; however, emphasis is placed on constructing adversarial datasets that hinder disentanglement performance using a mild transformation, obtained from trained models which achieve poor disentanglement scores. Our work provides intuition by constructing an example adversarial dataset that targets a specific reconstruction loss, and then remedies this problem by adjusting the loss.

Finally, Locatello et al. [2019b] show that useful representations cannot be reliably learnt with unsupervised methods, unless inductive biases are introduced, and Gondal et al. [2019] show that representations learnt on synthetic data often do not transfer well to real-world data. Our work investigates the interplay between the reconstruction loss and data as the main bias in VAEs, standard choices accidentally disentangle synthetic data.

4 Existing Disentanglement Datasets

Consider the 3D Shapes dataset [Burgess and Kim, 2018] in Figure 2a, which contains observations of shapes fixed in the centre of the image with progressively changing attributes or factors such as size and colour. If, as humans, we are given unordered observations from a traversal along the size factor of 3D Shapes, it would be easy to order these observations using a perceived increase or decrease in the size of the shape. We might even say that the shapes in the images overlap by

\[ d_{gt}(x^{(a)}, x^{(b)}) = \|y^{(a)} - y^{(b)}\|_1. \]

4.2 VAE Perceived Distance

With the idea of ground-truth distances between observations, we need a distance measure between observations as perceived by VAE frameworks. We derive the perceived distance between dataset elements from the noisy sampling procedure and the chosen reconstruction loss in a VAE framework.

Let \( z^{(b)} \sim q_{\theta}(z|x^{(a)}) \) be a (possibly incorrect) sample from the posterior distribution corresponding to some input element \( x^{(a)} \in \mathcal{X} \). Since the regularisation term encourages latent distributions to overlap, this sample \( z^{(b)} \) may be incorrectly attributed by the decoder to a distribution corresponding to some other element from the dataset \( x^{(b)} \in \mathcal{X} \), with reconstruction \( \hat{x}^{(b)} \approx x^{(b)} \). As the VAE objective consisting of the regularisation and reconstruction losses is jointly optimised, the decoder becomes better at reconstructing the inputs. In an ideal scenario, the inputs map to outputs \( \hat{x} \rightarrow x \), and reconstructions are samples from our dataset: \( \hat{x} \in \mathcal{X} \). While this is not the case in practice due to the regularisation term, we derive the perceived distance in Equation (7) from this assumption that \( \hat{x} \rightarrow x \). This allows us to directly compare the elements \( x^{(a)}, x^{(b)} \in \mathcal{X} \) within a dataset using the reconstruction loss as a distance function:

\[ d_{pcv}(x^{(a)}, x^{(b)}) = \lim_{\hat{x} \rightarrow x} L_{rec}(x^{(a)}, \hat{x}^{(b)}) \]

The perceived distance depends on the choice of reconstruction loss, which in literature is usually the pixel-wise Mean Squared Error (MSE) for data that is assumed to be normally distributed. We assume MSE is used throughout the rest of this work, unless specified. Note that analyses are similar for other pixel-wise losses, such as Binary Cross-Entropy (BCE).

4.3 Perceived Distances Correspond to Ground-Truth

In Section 4.1, all the ground-truth factors of a dataset are defined as the set \( \mathcal{Y} = [f_1] \times \ldots \times [f_p] \). In Equation (9), we now define a factor traversal \( \mathcal{Y}^{(a,b)} \in \mathcal{Y} \) as the ordered set of all the coordinates along a factor \( i \in [F] \) such that the set passes through a point \( y^{(a)} \in \mathcal{Y} \). The number of

\[ f_i \in \mathbb{N}^+ \]. The set of all factors used for generating the dataset is written as \( \mathcal{Y} = [f_1] \times \ldots \times [f_p] \). The full dataset is generated from this set of factors using some ground-truth generative process \( \mathcal{X} = \{y, (y) \mid y \in \mathcal{Y}\} \). Examples of this generative process are given in Figure 2.

With this construction of synthetic datasets, it is fitting to describe the ground-truth distances between observations \( x^{(a)}, x^{(b)} \in \mathcal{X} \) using the Manhattan or \( l_1 \) distance between their corresponding ground-truth factors \( y^{(a)}, y^{(b)} \in \mathcal{Y} \), as in Equation (5). It is important to note that this choice may not be optimal for single factors; rather, \( l_1 \) distance naturally aligns with how the datasets are constructed.

\[ d_{gt}(x^{(a)}, x^{(b)}) = \|y^{(a)} - y^{(b)}\|_1. \]
We compute distance matrices over a trained VAE with diagonal priors are rotationally invariant [Mathieu et al., 2016], thus the same distances between the means \( \mu \) of latent factors. Columns: represent a traversal along a single factor.

\[ D^{(a,i)} = \ldots \times \left\{ y_{i-1}^{(a)} \right\} \times [f_i] \times \left\{ y_{i+1}^{(a)} \right\} \times \ldots \]  
\[ = \left\{ (\ldots, y_{i-1}^{(a)}, j, y_{i+1}^{(a)}, \ldots) \mid \forall j \in [f_i] \right\} \]

We compute the distance matrix \( \tilde{D}^{(a,i)} \in \mathbb{R}^{f_i \times f_i} \), for some distance function \( d \), between pairwise elements along a factor traversal \( \mathcal{Y}^{(a,i)} \), written in Equation (10) using matrix notation.

\[ \tilde{D}^{(a,i)} = (d(x^{(a)}, x^{(v)})) \in \mathbb{R}^{f_i \times f_i}, \forall u, v \in \mathcal{Y}^{(a,i)} \]  

To examine the ground-truth factors within our dataset, we compute the average distance matrix \( D^{(i)} = \mathbb{E}_{a \in \mathcal{Y}[D^{(a,i)}]} \) for each factor \( i \in [F] \). We plot these results in Figure 3 for both the ground-truth distance \( d_{gt} \) and perceived distance \( d_{pcv} \).

It is immediately obvious from these plots that the ground-truth factors naturally correspond with distances in the dataset. If these perceived distances were to change such that they do not correspond to the ground-truth distances, VAEs might not be able to learn meaningful representations. This is highlighted by the fact that VAEs are already known to perform poorly on real-world data [Gondal et al., 2019].

### 5 Adversarial Datasets

In the previous section, we highlighted the striking similarity between the ground-truth distances and the perceived distances between observations in synthetic ground-truth datasets. This suggests that disentanglement occurs because latent distances accidentally correspond to ground-truth distances, when the latent space is reorganised to minimise reconstruction errors and perceived distances from the data space are captured.

Consider the example of a single chess piece moving across a chess board; there are no smooth transitions between grid points, since the piece is only valid when placed in the middle of squares. We describe such a dataset as having constant perceived distance. This property is adversarial in nature as it is impossible for a VAE to order these observations using pixel-wise perceived distance. It is tempting to think that a harder case is if the perceived distances do not correspond to ground-truth distances; however, an (incorrect) ordering can then still be found. Existing datasets such as Cars3D already satisfy this incorrect ordering, which may explain the generally worse disentanglement performance compared to other datasets, see Figure 3.

Formally, we say that a dataset has constant overlap when the pairwise distances over factor traversals are all equal. Let \( i \in [F] \) be a factor and \( y^{(a)} \in \mathcal{Y} \) be ground-truth coordinate vector. Then, for all elements over the factor traversal \( \forall y^{(b)} \in \mathcal{Y}^{(a,i)} / \{ y^{(a)} \} \), the corresponding perceived distance is constant such that \( d_{pcv}(x^{(a)}, x^{(b)}) = C_f \) with \( C_f \in \mathbb{R} \) and \( C_f > 0 \). Along factor traversals in such a dataset, no distinct ordering of elements can be found when a VAE tries to minimise the sampling error over the reconstruction loss. Going forward, we only consider the case where \( \forall f \in [F] \), \( C_f = \bar{C} \) for some \( \bar{C} > 0 \).

### 5.1 Example XYSquares Adversarial Dataset

Taking inspiration from the chess piece example, we design a synthetic adversarial dataset called XYSquares (See Figure 5)
that specifically targets VAEs that use a pixel-wise reconstruction loss such as MSE, resulting in constant perceived distances. The dataset consists of three $8 \times 8$ pixel squares in a world of size of $64 \times 64$. This leaves 8 grid positions along each axis without any pixel-wise overlap. The three squares are each assigned a colour according to $R(1,0,0), G(0,1,0)$ and $B(0,0,1)$ to avoid any channel-wise overlap. With 6 ground-truth factors (three squares moving along two axes), each with 8 possible values, this gives a total dataset size of $8^6 = 262144$ observations. In the rightmost column of Figure 8, we validate that this leads to constant perceived distances between observation pairs in factor traversals.

5.2 Experimental Setup

We now investigate the performance of VAEs on our new dataset. In particular, we use the unsupervised $\beta$-VAE [Higgins et al., 2016] and the state-of-the-art weakly supervised Ada-GVAE [Locatello et al., 2020]. The $\beta$-VAE scales the VAE regularisation term with a coefficient $\beta > 0$, while the Ada-GVAE breaks symmetry and encourages shared latent variables between pairs of observations. This is achieved by averaging together latent distributions between observation pairs that are estimated to remain unchanged when the KL divergence is below some threshold. We note that if the weakly supervised Ada-GVAE performs poorly, then it is highly likely that another unsupervised method will also perform poorly.

We use the same Adam [Kingma and Ba, 2015] optimiser and convolutional neural architecture as Burgess et al. [2017]. To evaluate disentangled representations, we use the MIG [Chen et al., 2018] (Mutual Information Gap) and DCI Disentanglement [Eastwood and Williams, 2018] scores. MIG measures the mutual information between the highest and second highest latent units for each factor, and DCI Disentanglement measures how much each latent unit captures a ground-truth factor using a predictive model.

Finally, we perform an extensive hyper-parameter grid search for existing frameworks and datasets before running our own experiments. Hyperparameters include the learning rate, size of the latent dimension, training steps, batch size and $\beta$ values. See our extended paper in Footnote 1 for further details on all experiments conducted throughout the remainder of the paper.

5.3 Example Adversarial Dataset Results

Figure 6 shows that the disentanglement performance over XYsquares is extremely poor compared to existing datasets, even with the state-of-the-art Ada-GVAE. We are concerned only with the maximum score obtained for each model and dataset, as the graph is plotted over the hyper-parameter sweeps. This validates our adversarial dataset hypothesis in Section 5. Not only is the disentanglement performance poor, but much smaller values for $\beta$ are needed when tuning the regularisation loss. Example latent traversals from a VAE trained over the adversarial dataset are given in Figure 7, results are far from disentangled and do not correspond in any way to the ground-truth factors in Figure 5.

5.4 Example of Varying Levels of Overlap

We have examined the effect of training on existing datasets with significant amounts of overlap, as well as our own advers-
sarial dataset with constant perceived distances according to pixel-wise losses. However, we have not investigated increasing levels of overlap in datasets, or rather reducing perceived between observations that are also close in ground-truth factor space. To do so, we modify XYSquares by decreasing the spacing between grid points while keeping the number of grid points constant along each factor, ensuring the dataset size remains fixed at $8^6 = 262144$ observations.

The original adversarial dataset, with a spacing of 8, has a constant distance value of $d_{pcv}(x^{(a)}, x^{(b)}) = C_8$. As the spacing $s$ decreases from $8 \rightarrow 1$ over the datasets, the probability increases that any two observations re-sampled along a single factor traversal overlap $p(d_{pcv}(x^{(a)}, x^{(b)}) < C_8)$ and should thus be placed closer together in the latent space. More overlap leads to more unique distance values which in turn allows for easier ordering of data points. We visualise this concept using ground-truth and perceived distance matrices in Figure 8.

We verify our statements through the experimental results in Figure 9, where the $\beta$-VAE and Ada-GVAE are trained on these datasets. As the spacing decreases and overlap is introduced, the disentanglement performance improves, since it is easier for a VAE to introduce an ordering over representations. Even for the XYSquares dataset with 1 pixel of overlap between grid points, an ordering of elements along factor traversals can be induced. However, the probability of a VAE encountering these scenarios in the latent space due to random sampling is low, and thus it is still not always easy for the model to learn disentangled representations over such a dataset.

6 Example of Introducing Overlap

The previous section focused on increasing overlap by changing the underlying dataset; however, this still does not solve the case for the original XYSquares dataset with constant pixel-wise perceived distance. Throughout this paper, we have provided evidence that VAEs disentangle based on their reconstruction loss, which happens to align with ground-truth factors of variation in common benchmark datasets. This correspondence is not optimal for all tasks and we propose that this leads to the poor disentanglement performance in these settings. Our solution is to choose a loss function that modifies perceived distances such that they also correspond to ground-truth distances.

The new loss function we choose cannot be a pixel-wise approach, as this does not capture the distances due to the spatial nature of the XYSquares dataset. For the sake of simplicity in this example, we convert the existing pixel-wise loss function into a spatially aware loss function by introducing a differentiable augmentation to its inputs. An appropriate augmentation for our dataset is a channel-wise box blur. The problem, however, is that the decoder needs to be able to reconstruct the data, and so purely replacing the pixel-wise loss with the augmented loss may not succeed. Rather, in Equation (11), we append the augmented term to the existing loss and scale it by a constant $\alpha > 0$.

$$L_{\text{Overlap}}(x, ˆx) = L_{\text{rec}}(x, ˆx) + \alpha L_{\text{rec}}(\text{blur}(x), \text{blur}(ˆx))$$  

(11)
We choose a channel-wise box blur for our loss function with work. \( V \) AE frameworks to learn disentangled representations. We spatially-aware loss function to the original pixel-wise loss. the MSE loss and our modified loss function. Introducing a spa- sorial XYSquares dataset. between observations and allows the models to disentangle the adver- tially aware loss function allows us to capture ground-truth distances

\[
\beta
\]
for different datasets, to improve disentanglement, as future leave learning or identification of optimal reconstruction losses function changes perceived distances and affects the ability of

\[
tant
\]
because it provides the intuition that changing the loss dataset, disentanglement results are impressive. This is impor- glement of these specific

\[
x
\]
tions that align with the ground-truth factors.

Finally, in Figure 11, we compare the performance of the spatially-aware loss function to the original pixel-wise loss. Our new loss significantly improves the disentanglement performance over the adversarial dataset. This is because it allows our models to capture perceived distances between observations that align with the ground-truth factors.

While our choice of loss may not be optimal for disentan- glement of these specific \( x \) and \( y \) factors from our adversarial dataset, disentanglement results are impressive. This is impor- tant because it provides the intuition that changing the loss function changes perceived distances and affects the ability of VAE frameworks to learn disentangled representations. We leave learning or identification of optimal reconstruction losses for different datasets, to improve disentanglement, as future work.

### 7 Considerations for Disentanglement

Research

We highlight the similarity between introducing overlap in Section 6 through the reconstruction loss function and varying levels of overlap in Section 5.4 through modifications to the construction of the dataset itself. Both methods aim to improve disentanglement by changing perceived distances to better correspond to the ground-truth factors, while keeping ground-truth factors fixed.

The problem is that ground-truth factors can indeed change, and this choice, while at the discretion of the researcher, is largely ignored in literature. For example, a researcher may choose RGB, HSV or categorical representations for colours, they may choose binary or continuous encodings for positions, or they may split or merge various factors together.

As our work shows, disentanglement is largely dependent on the chosen reconstruction loss and not special algorithmic choices. Obtaining improved disentanglement results under current VAE disentanglement frameworks will ultimately require supervision from the researcher to adjust perceived dis- tances of the model to the task at hand. This contradicts the current notion that unsupervised and weakly supervised disen- tanglement methods can automatically uncover these human interpretable ground-truth factors [Higgins et al., 2016].

Ultimately, benchmarking against synthetic datasets with already subjective ground-truth factors will thus always remain problematic. There are infinitely many datasets with infinitely many choices as to what constitutes their ground-truth factors. Accurate disentanglement through future methods may need general world knowledge so that the methods can adapt to the task at hand.

### 8 Conclusion

In this paper, we demonstrated that there are fundamental char- acteristics of existing datasets that encourage VAEs to learn disentangled representations. Our work provides a theory for how VAEs perceive distances between pairs of observations in datasets. We used this theory to provide intuition by constructing an adversarial dataset for pixel-wise losses over-which state-of-the-art VAEs fail to learn disentangled representations. Finally, we re-enabled disentanglement over the example adversarial dataset by again adjusting perceived distances, instead through a change of the VAE reconstruction loss to capture the ground-truth factors of the dataset.

Our results highlight issues in current representation learn- ing approaches. We find that the focus on regularisation for dis- entanglement is misplaced, rather, disentanglement is largely accidental, and careful choice of the reconstruction loss or data is needed to capture the ultimately subjective ground-truth fac- tors. This is impractical in the real world, since perceived distances cannot be a prerequisite for true disentanglement. More advanced methods are therefore required that can un- cover true meaning within the data.

### Acknowledgements

Computations were performed using the High Performance Computing infrastructure provided by the Mathematical Sciences Support unit at the University of the Witwatersrand.
References

[Bengio et al., 2013] 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.

[Burgess and Kim, 2018] Christopher Burgess and Hyunjik Kim. 3D shapes dataset. https://github.com/deepmind/3dshapes-dataset, 2018. Accessed: 2023-01-01.

[Burgess et al., 2017] Christopher Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. In *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems*, 2017.

[Chen et al., 2018] Ricky Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2615–2625, 2018.

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

[Gondal et al., 2019] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32:15740–15751, 2019.

[Higgins et al., 2016] 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. 2016.

[Hou et al., 2017] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision*, pages 1133–1141. IEEE, 2017.

[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 Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Kingma and Welling, 2014] Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[Krizhevsky et al., 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[Kumar et al., 2018] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

[LeCun et al., 2004] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[Locatello et al., 2019a] Francesco Locatello, Gabriele Abbat, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pages 14584–14597, 2019.

[Locatello et al., 2019b] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

[Locatello et al., 2020] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

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

[Mathieu et al., 2017] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset, 2017. Accessed: 2023-01-01.

[Mnih et al., 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Paszke et al., 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.

[Reed et al., 2015] Scott Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in Neural Information Processing Systems*, 28:12406–12415, 2015.

[Rolinek et al., 2019] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions by accident. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

[Shao et al., 2020] Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. Controlvae: Controllable variational autoencoder. In *International Conference on Machine Learning*, pages 8655–8664. PMLR, 2020.
[Zaidi et al., 2020] Julian Zaidi, Jonathan Boilard, Ghyslain Gagnon, and Marc-André Carbonneau. Measuring disentanglement: A review of metrics. *arXiv preprint arXiv:2012.09276*, 2020.

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

[Zietlow et al., 2021] Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying inductive biases for $\beta$-VAE based architectures. *arXiv preprint arXiv:2102.06822*, 2021.