Evaluating Disentanglement in Data Representation

Ziyuan (Roger) Xu
Department of Computer Science
Vanderbilt University
ziyuan.xu@vanderbilt.edu

Yuanhe Li
Department of Computer Science
Vanderbilt University
yuanhe.li@vanderbilt.edu

Abstract

Autoencoders (AEs) are widely used for tasks like image reconstruction and compression. Our project investigates the disentanglement and reconstruction quality of different AE variants, including AE, VAE, β-VAE, and FactorVAE, using the 3D Shapes Dataset, which provides ground-truth factors. We employ a unified set of metrics, such as Mutual Information Gap (MIG) and FID, to evaluate axis alignment and reconstruction fidelity. Additionally, we analyze the trade-offs between disentanglement and reconstruction, leveraging insights from prior research. Our findings aim to deepen understanding of disentangled representations and demonstrate their potential in real-world applications, such as video frame prediction and domain adaptation.

1 Introduction

The project's motivation stems from Google's technology of RAISR Sharp Images with Machine Learning to compress images using AE techniques to greatly reduce bandwidth by 75% [Google, 2016], and the decoded image can be even sharper (upsampled) - software engineers and hardware engineers could never dream of this.

Variational Autoencoders (VAEs) have profound impact on various fields in Machine Learning. In [Bengio et al., 2013], it is claimed that representation learning has a strong impact in various areas, such as speech recognition, object recognition, natural language processing, transfer learning, and domain adaptation. Different explanatory factors of the data tend to change independently of each other in the input distribution, and only a few at a time tend to change when one considers a sequence of consecutive real-world inputs. The most robust approach to feature learning is to disentangle as many factors as possible, discarding as little information about the data as is practical.

In recent advancements, VAEs have been utilized in diverse real-world applications, including enhancing the capability of agents in reinforcement learning. One notable example is DARLA [Higgins et al., 2017], which leverages disentangled visual representations learned through a β-VAE to enable robust domain adaptation. By separating high-level generative factors such as object properties and environmental attributes, DARLA allows agents to effectively distinguish between objects, floors, and other elements within complex environments. Another useful application using disentangled representation is to learn disentangled image representations from video such that each frame is factorized into a stationary part, i.e. content, and a temporally varying component, i.e. pose [Denton et al., 2017]. This enables the stable and coherent long-range prediction of future frames by applying a standard LSTM to the pose latent features, conditioned on the content latent features from the last observed frame.

In this research project, we used 3DShapes [Burgess and Kim, 2018] as our training dataset. It is clearly labeled with 6 ground-truth independent latent factors: floor color, wall color, object color, scale, shape and orientation. The images are $64 \times 64 \times 3$ RGB images. It is needed to provide the reconstructed images an actual reference for calculating the MIG and FID scores. Compared to

Preprint. Under review.

ground-truth labeling datasets such as dSprites, the contrast in color and orientation of 3DShapes could provide a better visual contrast, also larger dimensions make factor alignments more challenging. Also, due to the size of the dataset, we only selected 50,000 out of 480,000 images, uniformly at random, to be split into training and validation sets.

Now, to investigate the disentanglement in data representation, or latent space, we aim to evaluate the disentanglement on the learned representation using four models: AE, VAE, β-VAE, and FactorVAE.

2 Hypothesis

For the four models, We hypothesize the following:

1. AE will have the best reconstruction quality, but the worst disentanglement, because of lack of the KL-divergence term in its objective.
2. VAE (β-VAE with β=1) will have worse image quality than AE because of its probabilistic nature in sampling during decoding, but the disentanglement is slightly better due to the regularization for the posterior distribution $p(z|x)$ to match the unit Gaussian.
3. β-VAE with β>1 will have a better disentanglement than VAE due to a larger penalty on the KL term in order to match the unit Gaussian prior. As β increases, the posterior distribution is more enforced to match the unit Gaussian distribution, but at the same time, the reconstruction quality will be sacrificed.
4. FactorVAE claims [Kim and Mnih, 2018] to be able to further disentangle the latent space via a Total Correlation term to encourage independence in the latent distribution. Hence, we expect FactorVAE to outperform β-VAE in terms of disentanglement.

3 Methodology

3.1 Model

For methods with regularization terms, we make sure to search for the best parameter. Due to the time consumed to train one model, we tuned the hyper-parameters by hand, while try to make the hyper-parameters consistent (by taking $\beta = 1$ across models). Assume q is a function or distribution characterized by ϕ, we characterize the following loss functions:

1. Autoencoder:
$$\mathcal{L} = \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - q(f_\theta(\mathbf{x}^{(i)}))\|^2$$
2. VAE (β-VAE with β=1):
$$\mathcal{L} = \sum_{i=1}^{N} \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p(\mathbf{x}^{(i)} \mid \mathbf{z})\right] - D_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x}^{(i)}) \| p(\mathbf{z})\right)\right]$$
3. β-VAE:
$$\mathcal{L} = \sum_{i=1}^{N} \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p(\mathbf{x}^{(i)} \mid \mathbf{z})\right] - \beta D_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x}^{(i)}) \| p(\mathbf{z})\right)\right]$$
4. FactorVAE:
$$\mathcal{L} = \sum_{i=1}^{N} \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p(\mathbf{x}^{(i)} \mid \mathbf{z})\right] - D_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x}^{(i)}) \| p(\mathbf{z})\right) - \gamma\, D_{\mathrm{KL}}\left(q(\mathbf{z}) \| \bar{q}(\mathbf{z})\right)\right]$$
where each element in $\bar{q}(\mathbf{z})$ is a d-dimensional sample, each dimension sampled $q(\mathbf{z})$.

3.2 Evaluation metric

Below are used metrics in evaluating our models.

3.2.1 Disentanglement

Disentanglement is a measurement of how the latent dimensions identified by the models are clearly separable (indicated by heatmap and data), and also as close to being correct as possible. According to [Carbonneau et al., 2022], we pick one of each dimensions for measurement. We decide on the following:

1. **Intervention-based:** latent space traversal. It is a deterministic approach for evaluating latent factor identification. By randomly selecting each latent dimension and keeping others fixed, we observe how the generated images change, which gives insight into whether the model captures interpretable and disentangled representations. This is a qualitative measurement of disentanglement. We traverse across the 6 dimensions in the dataset using 10 steps within the range of the mean ± 1 std.
2. **Information-based:** Alignment analysis using Mutual Information Gap (MIG) proposed in [Chen et al., 2018].

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H\left(v_k\right)}\left(I_n\left(z_{j^{(k)}}, v_k\right) - \max_{j\neq j^{(k)}} I_n\left(z_j, v_k\right)\right)$$

where $j^{(k)} = \arg\max_j I_n(z_j, v_k)$ and K is the number of known factors.
Note that a single factor can have high mutual information with multiple latent variables. To achieve a good axis-alignment, we want each latent dimension zi only captures one ground-truth factor vj. Without axis alignment, each latent variable can contain a decent amount of information regarding two or more factors, making it difficult to interpret the latent space. Axis-alignment can be assessed by measuring the difference between the top two latent variables with highest mutual information with a given factor. We will use MIP to compare the disentanglement quality and alignment across models. Higher MIG values indicate better separation and identification of latent factors. A higher MIG score indicates that one single latent dimension is highly informative about a particular ground-truth factor, suggesting strong axis alignment. A low MIG score implies that there is not a prominent latent dimension that is informative about a particular ground-truth factor, indicating poor axis alignment.

3.2.2 Reconstructed Image Analysis

Image reconstruction quality is a generalizable problem. We choose to use the averaged FID score, a metrics that compares the distribution of generated images with the distribution of a set of ground truths. Since we are modeling the VAE as multivariate Gaussians, its FID score is defined as:

$$\|\mu - \mu'\|_2^2 + \mathrm{tr}\left(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{\frac{1}{2}}\right)$$

assuming two points drawn from $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$.

4 Implementation Details

Because we output the pixel values within range $[0, 1]$, we use binary cross entropy as our reconstruction loss instead of MSE loss, to provide a better and stable gradient. Also, in our VAE training, we calculate reconstruction loss and KL loss both per-image to ensure they are on the same scale.

We use the following architectures to build the aforementioned 4 models. To ensure fair comparisons, we trained all models with 15 epochs, used Adam optimizer with an initial learning rate of 3×10^{-4}, and used batch size 64. We chose to encode into 6 latent dimensions (to match the number of ground-truth factors) for each model using the specified settings. Details of the implementation are as follows:

1. **Autoencoder**: we used a convolutional encoder and decoder. The encoder consists of 6 convolutional layers with output channel numbers $32, 64, 128, 256, 256$, and 6 (latent dimension), with ReLU activation. The decoder mirrors the encoder with 1 convolution layer and 5 transposed convolution layers for reconstruction.

2. **VAE (β-VAE with $\beta = 1$)**: the encoder first uses 5 convolution layers (with output channel numbers: $32, 64, 128, 256, 256$), followed by two convolution layers to output the mean and log variance of dimension 6 respectively. The decoder mirrors this structure.
3. **β-VAE**: we employed an encoder and decoder structure the same as the standard β-VAE.
4. **FactorVAE**: The architecture was adapted from Kim and Mnih [2018]. The VAE component is the same as the standard β-VAE. In addition, we implement a discriminator to distinguish joint vs. factorized samples in order to compute the total correlation term in the objective. This discriminator consists of two fully connected neural networks with hidden dimension 1000, and output logits for 2 classes, joint or factorized.

5 Results and Evaluation

Below are our experiment results. We analyze reconstruction and disentanglement separately. The code for reproducing the results are available here.

5.1 Reconstruction Analysis

Here are the reconstructed sample images, from all four models given the same training epochs, batch size, and learning rate for a fair comparison (see implementation details above). From FID scores shown in Table 1, it is observed that β-VAE and FactorVAE's reconstruction performance are worse than AE and VAE. This aligns with our hypothesis that a greater constraint on the latent space will sacrifice some reconstruction quality. Visually inspecting our reconstructed images below roughly agrees with the FID scores.

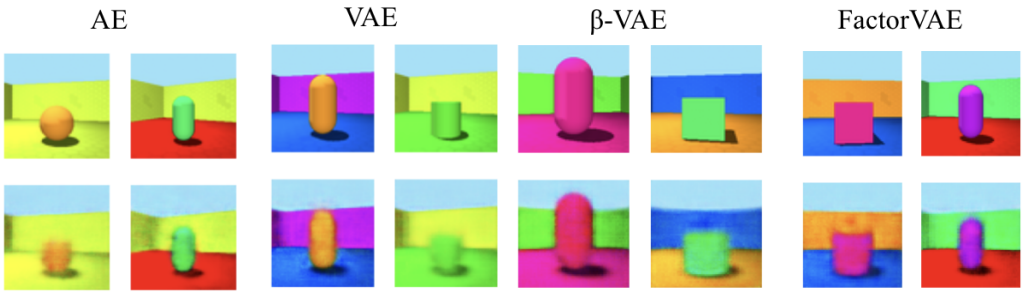


Figure 1: Comparison between Ground-Truth and Reconstructed Images

5.2 Alignment Analysis

5.2.1 Latent Traversal

For all four models, we fix all other latent dimensions but varying only one of the latent dimensions to inspect how the reconstructed outputs change in response to the variation of this specific latent dimension. In this example for autoencoder, multiple factors are simultaneously changing, including floor hue, orientation, object hue, and scale. This indicates entanglement in the latent space, as latent dimension 1 encodes multiple factors rather than only one factor.

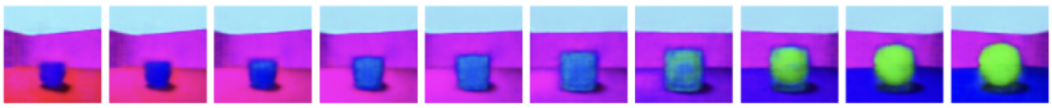


Figure 2: Autoencoder: Traversal Dimension 1

In this example for VAE, multiple factors are simultaneously changing, including floor hue and wall hue. This indicates entanglement in the latent space, as latent dimension 2 encodes multiple factors

rather than only one factor. $\beta = 1$ may not be sufficient enough to enforce a unit Gaussian posterior and achieve disentanglement.

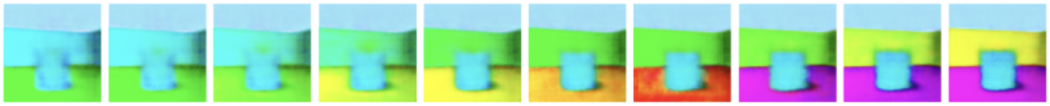


Figure 3: VAE: Traversal Dimension 2

For β-VAE ($\beta = 15$), we could observe disentanglement in the latent space, indicated by a clear trend of changes of features in the reconstructed images.

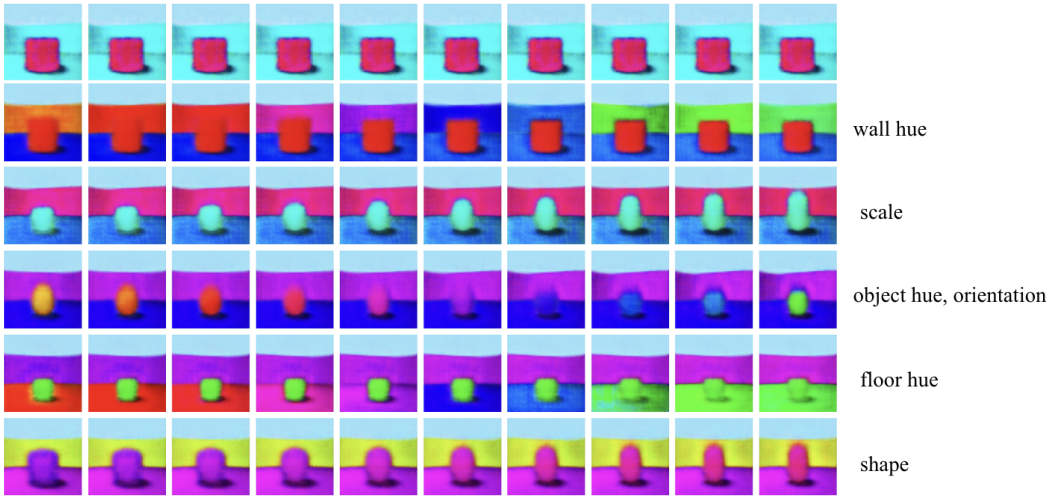


Figure 4: β-VAE: All-Dimension Traversal

Each row represents a latent traversal along one of the latent dimensions. For instance, in the first row, we show the results of the latent traversal of latent dimension 0, and the second row shows the latent traversal of latent dimension 1, and so on. Consider the second row, we can see that the variation in latent dimension 1 changes the wall hue in the reconstructed images but keeps all other image features the same, showing that this latent dimension encodes wall hue. Notably, latent dimension 0 suffered from posterior collapse, where the variation in latent dimension 0 does not change the reconstructed output significantly. This dimension is unused by the model to encode information of the images. The fourth latent dimension encodes two features, both object hue and orientation, as seen in that the wall behind switches orientation while the color of the object is also changing.

For Factor-VAE (γ = 500), we also observe disentanglement in the latent space.

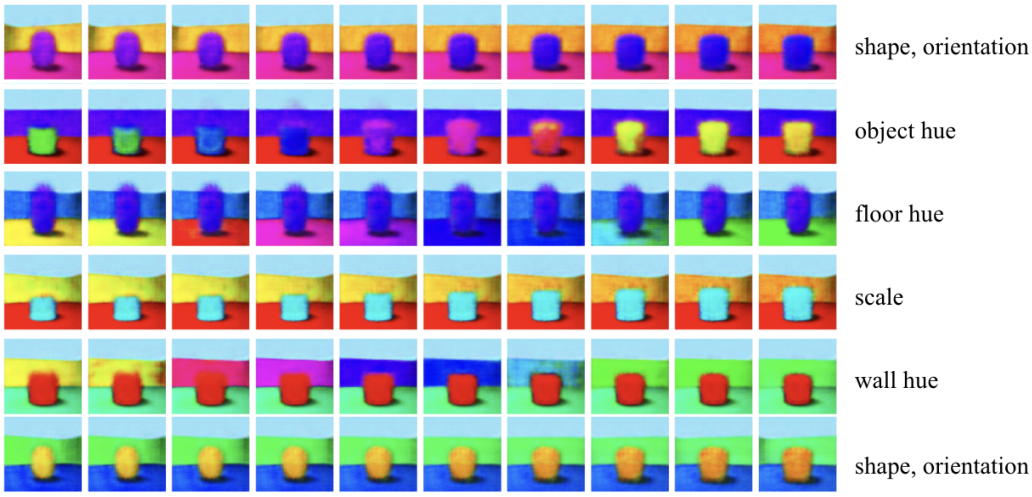


Figure 5: Factor-VAE: All-Dimension Traversal

Note that latent dimensions 0 and 5 both encode shape and orientation information, indicating some amount of entanglement and dependencies between the latent dimensions. The other latent dimensions show a clear encoding of one of the image features.

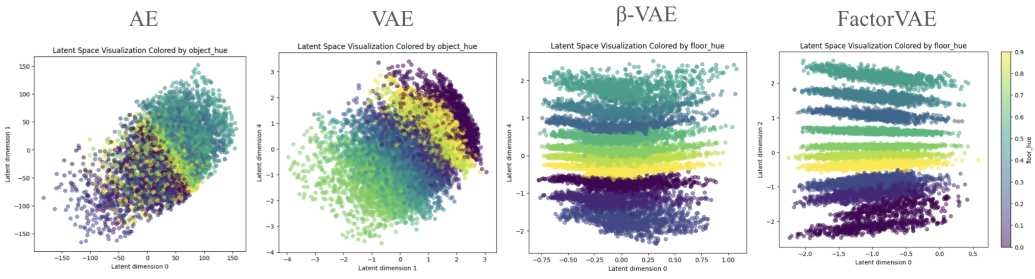
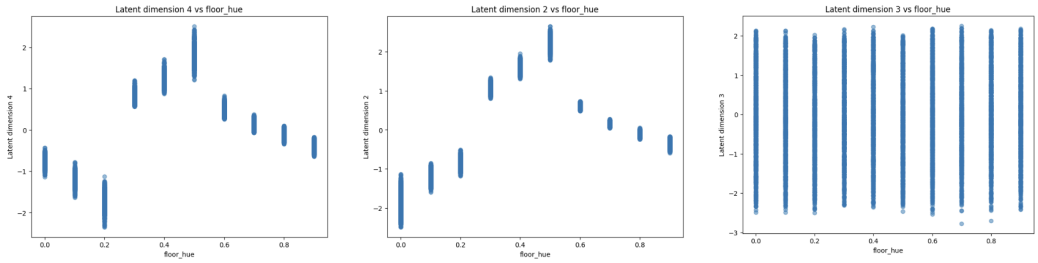


Figure 6: All Model Comparison: Visualization of Latent Space

Here we visualize the latent space directly, comparing all the four models. Autoencoder, as expected, did the worst in separating our latent variables. Notice that both latent dimensions 0 and 1 encode the object hue information, indicating dependencies between latent dimensions and hence entanglement. VAE's graph is tilted, showing clusters formed in the 2D latent space. The two latent dimensions together encode a single ground-truth factor, demonstrating entanglement. Out of expectation, FactorVAE and β-VAE both show a much better disentanglement, indicated by the horizontal strip-like clusters. This means that only one latent dimension encodes a given latent factor, as images with the same values of this latent factor will be encoded into similar values for this latent dimension. FactorVAE shows a slightly better disentanglement compared to β-VAE, despite having a lower MIG score (see Table 1). We suspect that the β-VAE's posterior collapse in latent dimension 0 greatly increases its MIG. Since MIG is a global statistic of disentanglement considering all latent dimensions, the result that great separation with one dimension yet having a lower MIG score is possible.



β-VAE FactorVAE β-VAE, failed

Figure 7: Latent factor vs. a single latent dimension

Interestingly, from Figure 7, we observed that models, while learning a particular latent factor like floor hue, exhibit the following common patterns.

1. Similar graph pattern across models, i.e., different models encode similar values in one latent dimension for the same latent factor value. For instance, FactorVAE and β-VAE encode values both near 1 for latent dimension 2 and 4 respectively, for floor hue $= 0.3$. This might be because the similarity between the formulae - what we changed is merely some regularization terms.
2. Similar pattern of confidence for each hue category across models. For example, β-VAE shows large confidence (indicated by the small variance) in predicting hue from 0.6 to 0.8, which is the same case (but more confident) in FactorVAE. There are slight exceptions: for example, $0 - 0.2$.

We also included a latent factor-dimension mismatch in the last graph. If traversing a different latent dimension against a given ground-truth, this latent dimension is unable to tell the difference.

	FID	**MIG**
Autoencoder	0.0641	0.152
VAE (β = 1)	0.0641	0.0797
VAE (β = 15)	0.1135	0.409
FactorVAE (β = 1, γ = 500)	0.0963	0.368

Table 1: FID and MIG Scores for Different Models

Lower FID score represents lower distance to the ground-truth image, hence better image reconstruction quality; higher MIG score represents better disentanglement. Here, we see the trade-off between reconstruction quality and disentanglement: if a model has a low FID score, it also has a low MIG score. β-VAE with $\beta = 15$ has the highest MIG score, indicating the best disentanglement, but it also has the highest FID score, indicating the worst reconstruction quality. Compare VAE ($\beta = 1$) with FactorVAE, due to the total correlation term in FactorVAE's loss, FactorVAE has a higher MIG score.

6 Limitations

MIP metric limitations. In [Carbonneau et al., 2022], MIG can measure compactness well but has limitations particularly when evaluating disentanglement properties such as modularity. MIG measures the difference in mutual information between the two most informative latent dimensions for each factor of variation, capturing how uniquely a single factor is encoded by a specific dimension. While this makes it a useful measure of compactness, it fails to capture modularity, which requires that each factor only influences a distinct subset of the representation space without overlap. MIG

does not penalize situations where multiple factors are encoded within the same latent subspace, nor does it assess whether factors are properly grouped across different dimensions. This narrow focus on distinctiveness between top candidates overlooks how the information is distributed across the entire latent space, making MIG an inadequate measure of modularity. Despite these drawbacks, MIG can still be effective for simpler datasets like 3D Shapes, where factors are easily disentangled and relationships are straightforward, allowing the metric to highlight distinct latent features. However, in more complex real-world scenarios, compactness is less useful because many factors (e.g., speaker identity, color, illumination) are inherently multi-dimensional and cannot be neatly represented by a single latent dimension, emphasizing the greater importance of modularity in practical applications. In the future, we can try some other disentanglement metrics such as the SAP score proposed in [Kumar et al., 2017].

Data Selection Bias. We are only selecting 50,000 out of 480,000 possible images. This selection is done randomly, so there could be bias in having one latent factor more than the other in the dataset. Also, when splitting into training and validation sets, the same bias could also happen. This could affect the learning efficiency over one factor - for example, the model could overfit to the more frequently observed factors, leading to poor generalization. This imbalance can skew the disentanglement evaluation, as the latent representations may disproportionately favor factors with higher representation in the dataset. In the future, we will consider a well-balanced subset with equal representation on each feature, ensuring a fairer evaluation of the disentanglement capabilities of VAEs.

Training limitations. Currently, we only train 15 epochs due to time and hardware resource constraints, and the hyperparameters are chosen empirically and based on prior literature. In the future, we can increase the training epochs to achieve better result and systematically choose hyperparameters.

7 Miscellaneous

The main basis of our paper are [Burgess et al., 2018] and [Kim and Mnih, 2018], which describes β-VAE and FactorVAE, respectively. These are famous and general VAEs (meaning not used for specific domains or goals), and they use some regularized versions of ELBO for loss function. Upon a quick review at [Carbonneau et al., 2022], we realized that CCA and MIG are most used for alignment measurement, but MIG is selected, because it captures the similarities of the mutual traits across multiple models, which is what we intend to measure.

Reproducing the Result

Our code is stored in a public GitHub repository, available here. To run the project, install the necessary dependencies and simply hit run for each .ipynb file!

References

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.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. `https://github.com/google-deepmind/3d-shapes/`, 2018.

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.

Marc-André Carbonneau, Julian Zaidi, Jonathan Boilard, and Ghyslain Gagnon. Measuring disentanglement: A review of metrics. *IEEE transactions on neural networks and learning systems*, 2022.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems*, 30, 2017.

Google. Enhance! raisr sharp images with machine learning, 2016. URL `https://research.google/blog/enhance-raisr-sharp-images-with-machine-learning/`.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: improving zero-shot transfer in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1480–1490. JMLR.org, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018.

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