# Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

**Francesco Locatello** [1 2] **Stefan Bauer** [2] **Mario Lucic** [3] **Gunnar Rätsch** [1] **Sylvain Gelly** [3] **Bernhard Schölkopf** [2] **Olivier Bachem** [3]

## Abstract

The key idea behind the *unsupervised* learning of *disentangled* representations is that real-world data is generated by a few explanatory factors of variation which can be recovered by unsupervised learning algorithms. In this paper, we provide a sober look at recent progress in the field and challenge some common assumptions. We first theoretically show that the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases on both the models and the data. Then, we train more than 12 000 models covering most prominent methods and evaluation metrics in a reproducible large-scale experimental study on seven different data sets. We observe that while the different methods successfully enforce properties "encouraged" by the corresponding losses, well-disentangled models seemingly cannot be identified without supervision. Furthermore, increased disentanglement does not seem to lead to a decreased sample complexity of learning for downstream tasks. Our results suggest that future work on disentanglement learning should be explicit about the role of inductive biases and (implicit) supervision, investigate concrete benefits of enforcing disentanglement of the learned representations, and consider a reproducible experimental setup covering several data sets.

## 1. Introduction

In representation learning it is often assumed that real-world observations $\mathbf{x}$ (e.g., images or videos) are generated by a two-step generative process. First, a multivariate latent random variable $\mathbf{z}$ is sampled from a distribution $P(\mathbf{z})$. Intuitively, $\mathbf{z}$ corresponds to semantically meaningful factors of variation of the observations (e.g., content + position of objects in an image). Then, in a second step, the observation $\mathbf{x}$ is sampled from the conditional distribution $P(\mathbf{x}|\mathbf{z})$. The key idea behind this model is that the high-dimensional data $\mathbf{x}$ can be explained by the substantially lower dimensional and semantically meaningful latent variable $\mathbf{z}$ which is mapped to the higher-dimensional space of observations $\mathbf{x}$. Informally, the goal of representation learning is to find useful transformations $r(\mathbf{x})$ of $\mathbf{x}$ that "make it easier to extract useful information when building classifiers or other predictors" (Bengio et al., 2013).

A recent line of work has argued that representations that are *disentangled* are an important step towards a better representation learning (Bengio et al., 2013; Peters et al., 2017; LeCun et al., 2015; Bengio et al., 2007; Schmidhuber, 1992; Lake et al., 2017; Tschannen et al., 2018). They should contain all the information present in $\mathbf{x}$ in a compact and interpretable structure (Bengio et al., 2013; Kulkarni et al., 2015; Chen et al., 2016) while being independent from the task at hand (Goodfellow et al., 2009; Lenc & Vedaldi, 2015). They should be useful for (semi-)supervised learning of downstream tasks, transfer and few shot learning (Bengio et al., 2013; Schölkopf et al., 2012; Peters et al., 2017). They should enable to integrate out nuisance factors (Kumar et al., 2017), to perform interventions, and to answer counterfactual questions (Pearl, 2009; Spirtes et al., 1993; Peters et al., 2017).

While there is no single formalized notion of disentanglement (yet) which is widely accepted, the key intuition is that a disentangled representation should separate the distinct, informative *factors of variations* in the data (Bengio et al., 2013). A change in a single underlying factor of variation $z_i$ should lead to a change in a single factor in the learned representation $r(\mathbf{x})$. This assumption can be extended to groups of factors as, for instance, in Bouchacourt et al. (2018) or Suter et al. (2018). Based on this idea, a variety of disentanglement evaluation protocols have been proposed leveraging the statistical relations between the learned

[1]ETH Zurich, Department for Computer Science [2]Max-Planck Institute for Intelligent Systems [3]Google Research, Brain Team. Correspondence to: Francesco Locatello <francesco.locatello@tuebingen.mpg.de>, Olivier Bachem <bachem@google.com>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

representation and the ground-truth factor of variations. Disentanglement is then measured as a particular structural property of these relations (Higgins et al., 2017a; Kim & Mnih, 2018; Eastwood & Williams, 2018; Kumar et al., 2017; Chen et al., 2018; Ridgeway & Mozer, 2018).

State-of-the-art approaches for unsupervised disentanglement learning are largely based on *Variational Autoencoders (VAEs)* (Kingma & Welling, 2014): One assumes a specific prior $P(\mathbf{z})$ on the latent space and then uses a deep neural network to parameterize the conditional probability $P(\mathbf{x}|\mathbf{z})$. Similarly, the distribution $P(\mathbf{z}|\mathbf{x})$ is approximated using a variational distribution $Q(\mathbf{z}|\mathbf{x})$, again parametrized using a deep neural network. The model is then trained by minimizing a suitable approximation to the negative log-likelihood. The representation for $r(\mathbf{x})$ is usually taken to be the mean of the approximate posterior distribution $Q(\mathbf{z}|\mathbf{x})$. Several variations of VAEs were proposed with the motivation that they lead to better disentanglement (Higgins et al., 2017a; Burgess et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2017; Rubenstein et al., 2018). The common theme behind all these approaches is that they try to enforce a factorized aggregated posterior $\int_{\mathbf{x}} Q(\mathbf{z}|\mathbf{x})P(\mathbf{x})d\mathbf{x}$, which should encourage disentanglement.

**Our contributions.** In this paper, we challenge commonly held assumptions in this field in both theory and practice. Our key contributions can be summarized as follows:

- We theoretically prove that (perhaps unsurprisingly) the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases both on the considered learning approaches and the data sets.
- We investigate current approaches and their inductive biases in a reproducible large-scale experimental study[1] with a sound experimental protocol for unsupervised disentanglement learning. We implement six recent unsupervised disentanglement learning methods as well as six disentanglement measures from scratch and train more than 12 000 models on seven data sets.
- We release `disentanglement_lib`[2], a new library to train and evaluate disentangled representations. As reproducing our results requires substantial computational effort, we also release more than 10 000 trained models which can be used as baselines for future research.
- We analyze our experimental results and challenge common beliefs in unsupervised disentanglement learning: (i) While all considered methods prove effective at ensuring that the individual dimensions of the aggregated posterior (which is sampled) are not correlated, we observe that the dimensions of the representation (which is taken to be the mean) are correlated. (ii) We do not find any evidence that the considered models can be used to reliably learn disentangled representations in an *unsupervised* manner as random seeds and hyperparameters seem to matter more than the model choice. Furthermore, good trained models seemingly cannot be identified without access to ground-truth labels even if we are allowed to transfer good hyperparameter values across data sets. (iii) For the considered models and data sets, we cannot validate the assumption that disentanglement is useful for downstream tasks, for example through a decreased sample complexity of learning.
- Based on these empirical evidence, we suggest three critical areas of further research: (i) The role of inductive biases and implicit and explicit supervision should be made explicit: unsupervised model selection persists as a key question. (ii) The concrete practical benefits of enforcing a specific notion of disentanglement of the learned representations should be demonstrated. (iii) Experiments should be conducted in a reproducible experimental setup on data sets of varying degrees of difficulty.

[1] Reproducing these experiments requires approximately 2.52 GPU years (NVIDIA P100).

[2] https://github.com/google-research/disentanglement_lib

## 2. Other related work

In a similar spirit to disentanglement, (non-)linear independent component analysis (Comon, 1994; Bach & Jordan, 2002; Jutten & Karhunen, 2003; Hyvarinen & Morioka, 2016) studies the problem of recovering independent components of a signal. The underlying assumption is that there is a generative model for the signal composed of the combination of statistically independent non-Gaussian components. While the identifiability result for linear ICA (Comon, 1994) proved to be a milestone for the classical theory of factor analysis, similar results are in general not obtainable for the nonlinear case and the underlying sources generating the data cannot be identified (Hyvarinen & Pajunen, 1999). The lack of almost any identifiability result in nonlinear ICA has been a main bottleneck for the utility of the approach (Hyvarinen et al., 2018) and partially motivated alternative machine learning approaches (Desjardins et al., 2012; Schmidhuber, 1992; Cohen & Welling, 2015). Given that unsupervised algorithms did not initially perform well on realistic settings most of the other works have considered some more or less explicit form of supervision (Reed et al., 2014; Zhu et al., 2014; Yang et al., 2015; Kulkarni et al., 2015; Cheung et al., 2015; Mathieu et al., 2016; Narayanaswamy et al., 2017; Suter et al., 2018). (Hinton et al., 2011; Cohen & Welling, 2014) assume some knowledge of the effect of the factors of variations even though they are not observed. One can also exploit known relations between factors in different samples (Karaletsos et al., 2015; Goroshin et al., 2015; Whitney et al., 2016; Fraccaro et al.,

2017; Denton & Birodkar, 2017; Hsu et al., 2017; Yingzhen & Mandt, 2018) or explicit inductive biases (Locatello et al., 2018). This is not a limiting assumption especially in sequential data, i.e., for videos. We focus our study on the setting where factors of variations are not observable at all, i.e. we only observe samples from $P(\mathbf{x})$.

## 3. Impossibility result

The first question that we investigate is whether unsupervised disentanglement learning is even possible for arbitrary generative models. Theorem 1 essentially shows that without inductive biases both on models and data sets the task is fundamentally impossible. The proof is provided in Appendix A.

**Theorem 1.** *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\boldsymbol{u})}{\partial u_j} \neq 0$ almost everywhere for all $i$ and $j$ (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \boldsymbol{u}) = P(f(\mathbf{z}) \leq \boldsymbol{u})$ for all $\boldsymbol{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

Consider the commonly used "intuitive" notion of disentanglement which advocates that a change in a single ground-truth factor should lead to a single change in the representation. In that setting, Theorem 1 implies that unsupervised disentanglement learning is *impossible* for arbitrary generative models with a factorized prior[3] in the following sense: Assume we have $p(\mathbf{z})$ and some $P(\mathbf{x}|\mathbf{z})$ defining a generative model. Consider any unsupervised disentanglement method and assume that it finds a representation $r(\mathbf{x})$ that is perfectly disentangled with respect to $\mathbf{z}$ in the generative model. Then, Theorem 1 implies that there is an equivalent generative model with the latent variable $\hat{\mathbf{z}} = f(\mathbf{z})$ where $\hat{\mathbf{z}}$ is completely *entangled* with respect to $\mathbf{z}$ and thus also $r(\mathbf{x})$: as all the entries in the Jacobian of $f$ are non-zero, a change in a single dimension of $\mathbf{z}$ implies that all dimensions of $\hat{\mathbf{z}}$ change. Furthermore, since $f$ is deterministic and $p(\mathbf{z}) = p(\hat{\mathbf{z}})$ almost everywhere, both generative models have the same marginal distribution of the observations $\mathbf{x}$ by construction, i.e., $P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}$. Since the (unsupervised) disentanglement method only has access to observations $\mathbf{x}$, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them.

This may not be surprising to readers familiar with the causality and ICA literature as it is consistent with the following argument: After observing $\mathbf{x}$, we can construct infinitely many generative models which have the same marginal distribution of $\mathbf{x}$. Any one of these models could be the true causal generative model for the data, and the right model cannot be identified given only the distribution of $\mathbf{x}$ (Peters et al., 2017). Similar results have been obtained in the context of non-linear ICA (Hyvarinen & Pajunen, 1999). The main novelty of Theorem 1 is that it allows the explicit construction of latent spaces $\mathbf{z}$ and $\hat{\mathbf{z}}$ that are completely *entangled* with each other in the sense of (Bengio et al., 2013). We note that while this result is very intuitive for multivariate Gaussians it also holds for distributions which are not invariant to rotation, for example multivariate uniform distributions.

While Theorem 1 shows that unsupervised disentanglement learning is fundamentally impossible for arbitrary generative models, this does not necessarily mean it is an impossible endeavour in practice. After all, real world generative models may have a certain structure that could be exploited through suitably chosen inductive biases. However, Theorem 1 clearly shows that inductive biases are required both for the models (so that we find a specific set of solutions) and for the data sets (such that these solutions match the true generative model). We hence argue that the role of inductive biases should be made explicit and investigated further as done in the following experimental study.

## 4. Experimental design

**Considered methods.** All the considered methods augment the VAE loss with a regularizer: The $\beta$-VAE (Higgins et al., 2017a), introduces a hyperparameter in front of the KL regularizer of vanilla VAEs to constrain the capacity of the VAE bottleneck. The AnnealedVAE (Burgess et al., 2017) progressively increase the bottleneck capacity so that the encoder can focus on learning one factor of variation at the time (the one that most contribute to a small reconstruction error). The FactorVAE (Kim & Mnih, 2018) and the $\beta$-TCVAE (Chen et al., 2018) penalize the total correlation (Watanabe, 1960) with adversarial training (Nguyen et al., 2010; Sugiyama et al., 2012) or with a tractable but biased Monte-Carlo estimator respectively. The DIP-VAE-I and the DIP-VAE-II (Kumar et al., 2017) both penalize the mismatch between the aggregated posterior and a factorized prior. Implementation details and further discussion on the methods can be found in Appendix B and G.

**Considered metrics.** The *BetaVAE* metric (Higgins et al., 2017a) measures disentanglement as the accuracy of a linear classifier that predicts the index of a fixed factor of variation. Kim & Mnih (2018) address several issues with this metric in their *FactorVAE* metric by using a majority vote classifier on a different feature vector which accounts for a corner case in the BetaVAE metric. The *Mutual Information Gap (MIG)* (Chen et al., 2018) measures for each factor of vari-

[3]Theorem 1 only applies to factorized priors; however, we expect that a similar result can be extended to non-factorizing priors.

ation the normalized gap in mutual information between the highest and second highest coordinate in $r(\mathbf{x})$. Instead, the *Modularity* (Ridgeway & Mozer, 2018) measures if each dimension of $r(\mathbf{x})$ depends on at most a factor of variation using their mutual information. The Disentanglement metric of Eastwood & Williams (2018) (which we call *DCI Disentanglement* for clarity) computes the entropy of the distribution obtained by normalizing the importance of each dimension of the learned representation for predicting the value of a factor of variation. The *SAP score* (Kumar et al., 2017) is the average difference of the prediction error of the two most predictive latent dimensions for each factor. Implementation details and further descriptions can be found in Appendix C.

**Data sets.** We consider four data sets in which $\mathbf{x}$ is obtained as a deterministic function of $\mathbf{z}$: *dSprites* (Higgins et al., 2017a), *Cars3D* (Reed et al., 2015), *SmallNORB* (LeCun et al., 2004), *Shapes3D* (Kim & Mnih, 2018). We also introduce three data sets where the observations $\mathbf{x}$ are stochastic given the factor of variations $\mathbf{z}$: *Color-dSprites*, *Noisy-dSprites* and *Scream-dSprites*. In *Color-dSprites*, the shapes are colored with a random color. In *Noisy-dSprites*, we consider white-colored shapes on a noisy background. Finally, in *Scream-dSprites* the background is replaced with a random patch in a random color shade extracted from the famous *The Scream* painting (Munch, 1893). The dSprites shape is embedded into the image by inverting the color of its pixels. Further details on the preprocessing of the data can be found in Appendix H.

**Inductive biases.** To fairly evaluate the different approaches, we separate the effect of regularization (in the form of model choice and regularization strength) from the other inductive biases (e.g., the choice of the neural architecture). Each method uses the same convolutional architecture, optimizer, hyperparameters of the optimizer and batch size. All methods use a Gaussian encoder where the mean and the log variance of each latent factor is parametrized by the deep neural network, a Bernoulli decoder and latent dimension fixed to 10. We note that these are all standard choices in prior work (Higgins et al., 2017a; Kim & Mnih, 2018).

We choose six different regularization strengths, i.e., hyperparameter values, for each of the considered methods. The key idea was to take a wide enough set to ensure that there are useful hyperparameters for different settings for each method and not to focus on specific values known to work for specific data sets. However, the values are partially based on the ranges that are prescribed in the literature (including the hyperparameters suggested by the authors).

We fix our experimental setup in advance and we run all the considered methods on each data set for 50 different random seeds and evaluate them on the considered metrics. The full details on the experimental setup are provided in the Appendix G. Our experimental setup, the limitations of this study, and the differences with previous implementations are extensively discussed in Appendices D-F.

## 5. Key experimental results

In this section, we highlight our key findings with plots specifically picked to be representative of our main results. In Appendix I, we provide the full experimental results with a complete set of plots for different methods, data sets and disentanglement metrics.

### 5.1. Can current methods enforce a uncorrelated aggregated posterior and representation?

While many of the considered methods aim to enforce a factorizing and thus uncorrelated aggregated posterior (e.g., regularizing the total correlation of the sampled representation), they use the mean vector of the Gaussian encoder as the representation and not a sample from the Gaussian encoder. This may seem like a minor, irrelevant modification; however, it is not clear whether a factorizing aggregated posterior also ensures that the dimensions of the mean representation are uncorrelated. To test the impact of this, we compute the total correlation of both the mean and the sampled representation based on fitting Gaussian distributions for each data set, model and hyperparameter value (see Appendix C and I.2 for details).

Figure 1 (left) shows the total correlation based on a fitted Gaussian of the *sampled* representation plotted against the regularization strength for each method except AnnealedVAE on Color-dSprites. We observe that the total correlation of the sampled representation generally decreases with the regularization strength. One the other hand, Figure 1 (right) shows the total correlation of the *mean* representation plotted against the regularization strength. It is evident that the total correlation of the mean representation generally increases with the regularization strength. The only exception is DIP-VAE-I for which we observe that the total correlation

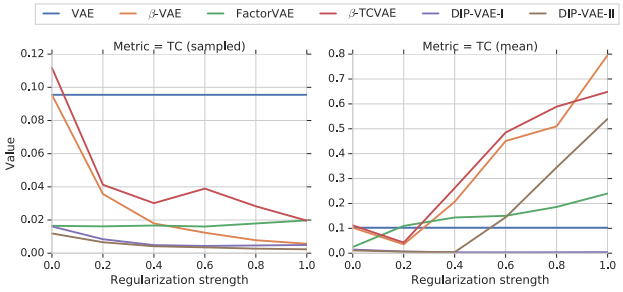


*Figure 1.* Total correlation based on a fitted Gaussian of the sampled (left) and the mean representation (right) plotted against regularization strength for Color-dSprites and approaches (except AnnealedVAE). The total correlation of the sampled representation decreases while the total correlation of the mean representation increases as the regularization strength is increased.

of the mean representation is consistently low. This is not surprising as the DIP-VAE-I objective directly optimizes the covariance matrix of the mean representation to be diagonal which implies that the corresponding total correlation (as we measure it) is low. These findings are confirmed by our detailed experimental results in Appendix I.2 (in particular Figures 8-9) which considers all different data sets. Furthermore, we observe largely the same pattern if we consider the average mutual information between different dimension of the representation instead of the total correlation (see Figures 27-28 in Appendix J).

**Implications.** Overall, these results lead us to conclude with minor exceptions that the considered methods are effective at enforcing an aggregated posterior whose individual dimensions are not correlated but that this does not seem to imply that the dimensions of the mean representation (usually used for representation) are uncorrelated.

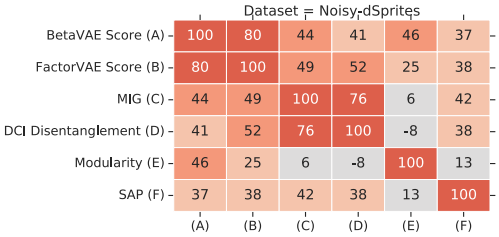


*Figure 2.* Rank correlation of different metrics on Noisy-dSprites. Overall, we observe that all metrics except Modularity seem mildly correlated with the pairs BetaVAE and FactorVAE, and MIG and DCI Disentanglement strongly correlated with each other.

### 5.2. How much do the disentanglement metrics agree?

As there exists no single, common definition of disentanglement, an interesting question is to see how much different proposed metrics agree. Figure 2 shows the Spearman rank correlation between different disentanglement metrics on Noisy-dSprites whereas Figure 12 in Appendix I.3 shows the correlation for all the different data sets. We observe that all metrics except Modularity seem to be correlated strongly on the data sets dSprites, Color-dSprites and Scream-dSprites and mildly on the other data sets. There appear to be two pairs among these metrics that capture particularly similar notions: the BetaVAE and the FactorVAE score as well as the MIG and DCI Disentanglement.

**Implication.** All disentanglement metrics except Modularity appear to be correlated. However, the level of correlation changes between different data sets.

### 5.3. How important are different models and hyperparameters for disentanglement?

The primary motivation behind the considered methods is that they should lead to improved disentanglement. This raises the question how disentanglement is affected by the model choice, the hyperparameter selection and randomness (in the form of different random seeds). To investigate this, we compute all the considered disentanglement metrics for each of our trained models.

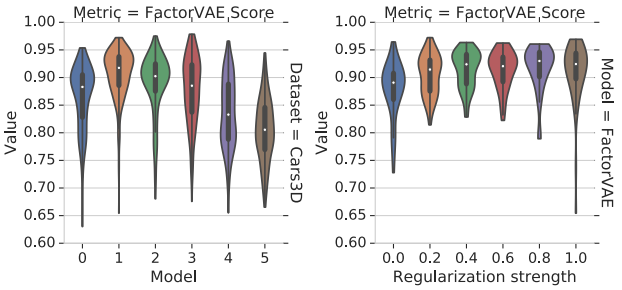


*Figure 3.* (left) FactorVAE score for each method on Cars3D. Models are abbreviated (0=$\beta$-VAE, 1=FactorVAE, 2=$\beta$-TCVAE, 3=DIP-VAE-I, 4=DIP-VAE-II, 5=AnnealedVAE). The variance is due to different hyperparameters and random seeds. The scores are heavily overlapping. (right) Distribution of FactorVAE scores for FactorVAE model for different regularization strengths on Cars3D. In this case, the variance is only due to the different random seeds. We observe that randomness (in the form of different random seeds) has a substantial impact on the attained result and that a good run with a bad hyperparameter can beat a bad run with a good hyperparameter.

In Figure 3 (left), we show the range of attainable FactorVAE scores for each method on Cars3D. We observe that these ranges are heavily overlapping for different models leading us to (qualitatively) conclude that the choice of hyperparameters and the random seed seems to be substantially more important than the choice of objective function. These results are confirmed by the full experimental results on all the data sets presented in Figure 13 of Appendix I.4: While certain models seem to attain better maximum scores on specific data sets and disentanglement metrics, we do not observe any consistent pattern that one model is consistently better than the other. At this point, we note that in our study we have fixed the range of hyperparameters *a priori* to six different values for each model and did not explore additional hyperparameters based on the results (as that would bias our study). However, this also means that specific models may have performed better than in Figure 13 (left) if we had chosen a different set of hyperparameters.

In Figure 3 (right), we further show the impact of randomness in the form of random seeds on the disentanglement scores. Each violin plot shows the distribution of the FactorVAE metric across all 50 trained FactorVAE models for each hyperparameter setting on Cars3D. We clearly see that randomness (in the form of different random seeds) has a substantial impact on the attained result and that a good run with a bad hyperparameter can beat a bad run with a good hyperparameter in many cases. Again, these findings

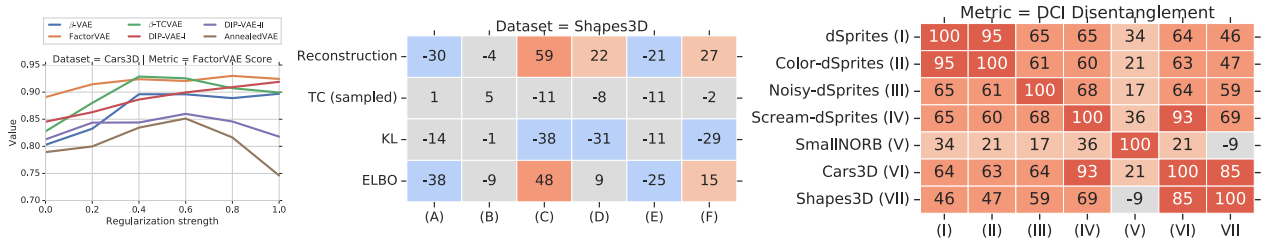


*Figure 4.* (left) FactorVAE score vs hyperparameters for each score on Cars3d. There seems to be no model dominating all the others and for each model there does not seem to be a consistent strategy in choosing the regularization strength. (center) Unsupervised scores vs disentanglement metrics on Shapes3D. Metrics are abbreviated ((A)=BetaVAE Score, (B)=FactorVAE Score, (C)=MIG , (D)=DCI Disentanglement, (E)=Modularity, (F)=SAP). The unsupervised scores we consider do not seem to be useful for model selection. (right) Rank-correlation of DCI disentanglement metric across different data sets. Good hyperparameters only seem to transfer between dSprites and Color-dSprites but not in between the other data sets.

are consistent with the complete set of plots provided in Figure 14 of Appendix I.4.

Finally, we perform a variance analysis by trying to predict the different disentanglement scores using ordinary least squares for each data set: If we allow the score to depend only on the objective function (treated as a categorical variable), we are only able to explain $37\%$ of the variance of the scores on average (see Table 5 in Appendix I.4 for further details). Similarly, if the score depends on the Cartesian product of objective function and regularization strength (again categorical), we are able to explain $59\%$ of the variance while the rest is due to the random seed.

**Implication.** The disentanglement scores of unsupervised models are heavily influenced by randomness (in the form of the random seed) and the choice of the hyperparameter (in the form of the regularization strength). The objective function appears to have less impact.

### 5.4. Are there reliable recipes for model selection?

In this section, we investigate how good hyperparameters can be chosen and how we can distinguish between good and bad training runs. In this paper, we advocate that that model selection *should not* depend on the considered disentanglement score for the following reasons: The point of unsupervised learning of disentangled representation is that there is no access to the labels as otherwise we could incorporate them and would have to compare to semi-supervised and fully supervised methods. All the disentanglement metrics considered in this paper require a substantial amount of ground-truth labels or the full generative model (for example for the BetaVAE and the FactorVAE metric). Hence, one may substantially bias the results of a study by tuning hyperparameters based on (supervised) disentanglement metrics. Furthermore, we argue that it is not sufficient to fix a set of hyperparameters *a priori* and then show that one of those hyperparameters and a specific random seed achieves a good disentanglement score as it amounts to showing the existence of a good model, but does not guide the practitioner in finding it. Finally, in many practical settings, we might not even have access to adequate labels as it may be hard to identify the true underlying factor of variations, in particular, if we consider data modalities that are less suitable to human interpretation than images.

In the remainder of this section, we hence investigate and assess different ways how hyperparameters and good model runs could be chosen. In this study, we focus on choosing the learning model and the regularization strength corresponding to that loss function. However, we note that in practice this problem is likely even harder as a practitioner might also want to tune other modeling choices such architecture or optimizer.

**General recipes for hyperparameter selection.** We first investigate whether we may find generally applicable "rules of thumb" for choosing the hyperparameters. For this, we plot in Figure 4 (left) the FactorVAE score against different regularization strengths for each model on the Cars3D data set whereas Figure 16 in Appendix I.5 shows the same plot for all data sets and disentanglement metrics. The values correspond to the median obtained values across 50 random seeds for each model, hyperparameter and data set. Overall, there seems to be no model consistently dominating all the others and for each model there does not seem to be a consistent strategy in choosing the regularization strength to maximize disentanglement scores. Furthermore, even if we could identify a good objective function and corresponding hyperparameter value, we still could not distinguish between a good and a bad training run.

**Model selection based on unsupervised scores.** Another approach could be to select hyperparameters based on unsupervised scores such as the reconstruction error, the KL divergence between the prior and the approximate posterior, the Evidence Lower BOund or the estimated total correlation of the sampled representation (mean representation gives similar results). This would have the advantage that

*Table 1.* Probability of outperforming random model selection on a different random seed. A random disentanglement metric and data set is sampled and used for model selection. That model is then compared to a randomly selected model: (i) on the same metric and data set, (ii) on the same metric and a random different data set, (iii) on a random different metric and the same data set, and (iv) on a random different metric and a random different data set. The results are averaged across 10 000 random draws.

| | Random data set | Same data set |
|---|---|---|
| Random metric | 54.9% | 62.6% |
| Same metric | 59.3% | 80.7% |

we could select specific trained models and not just good hyperparameter settings whose median trained model would perform well. To test whether such an approach is fruitful, we compute the rank correlation between these unsupervised metrics and the disentanglement metrics and present it in Figure 4 (center) for Shapes3D and in Figure 16 of Appendix I.5 for all the different data sets. While we do observe some correlations, no clear pattern emerges which leads us to conclude that this approach is unlikely to be successful in practice.

**Hyperparameter selection based on transfer.** The final strategy for hyperparameter selection that we consider is based on transferring good settings across data sets. The key idea is that good hyperparameter settings may be inferred on data sets where we have labels available (such as dSprites) and then applied to novel data sets. Figure 4 (right) shows the rank correlations obtained between different data sets for the DCI disentanglement (whereas Figure 17 in Appendix I.5 shows it for all data sets). We find a strong and consistent correlation between dSprites and Color-dSprites. While these results suggest that some transfer of hyperparameters is possible, it does not allow us to distinguish between good and bad random seeds on the target data set.

To illustrate this, we compare such a transfer based approach to hyperparameter selection to random model selection as follows: First, we sample one of our 50 random seeds, a random disentanglement metric and a data set and use them to select the hyperparameter setting with the highest attained score. Then, we compare that selected hyperparameter setting to a randomly selected model on either the same or a random different data set, based on either the same or a random different metric and for a randomly sampled seed. Finally, we report the percentage of trials in which this transfer strategy outperforms or performs equally well as random model selection across 10 000 trials in Table 1. If we choose the same metric and the same data set (but a different random seed), we obtain a score of 80.7%. If we aim to transfer for the same metric across data sets, we achieve around 59.3%. Finally, if we transfer both across metrics and data sets, our performance drops to 54.9%.

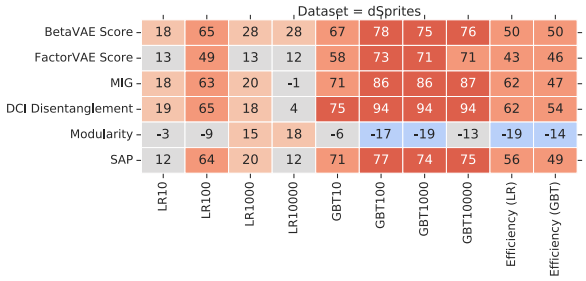


*Figure 5.* Rank correlations between disentanglement metrics and downstream performance (accuracy and efficiency) on dSprites.

**Implications.** Unsupervised model selection remains an unsolved problem. Transfer of good hyperparameters between metrics and data sets does not seem to work as there appears to be no unsupervised way to distinguish between good and bad random seeds on the target task.

### 5.5. Are these disentangled representations useful for downstream tasks in terms of the sample complexity of learning?

One of the key motivations behind disentangled representations is that they are assumed to be useful for later downstream tasks. In particular, it is argued that disentanglement should lead to a better sample complexity of learning (Bengio et al., 2013; Schölkopf et al., 2012; Peters et al., 2017). In this section, we consider the simplest downstream classification task where the goal is to recover the true factors of variations from the learned representation using either multi-class logistic regression (LR) or gradient boosted trees (GBT).

Figure 5 shows the rank correlations between the disentanglement metrics and the downstream performance on dSprites. We observe that all metrics except Modularity seem to be correlated with increased downstream performance on the different variations of dSprites and to some degree on Shapes3D but not on the other data sets. However, it is not clear whether this is due to the fact that disentangled representations perform better or whether some of these scores actually also (partially) capture the informativeness of the evaluated representation. Furthermore, the full results in Figure 19 of Appendix I.6 indicate that the correlation is weaker or inexistent on other data sets (e.g. Cars3D).

To assess the sample complexity argument we compute for each trained model a statistical efficiency score which we define as the average accuracy based on 100 samples divided by the average accuracy based on 10 000 samples. Figure 6 show the sample efficiency of learning (based on GBT) versus the FactorVAE Score on dSprites. We do not observe that higher disentanglement scores reliably lead to a higher sample efficiency. This finding which appears to be consistent with the results in Figures 20-23 of Appendix I.6.

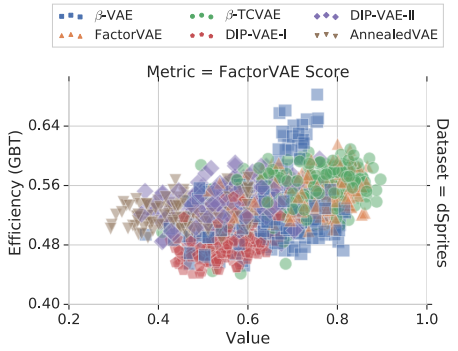


*Figure 6.* Statistical efficiency of the FactorVAE Score for learning a GBT downstream task on dSprites.

**Implications.** While the empirical results in this section are negative, they should also be interpreted with care. After all, we have seen in previous sections that the models considered in this study fail to reliably produce disentangled representations. Hence, the results in this section might change if one were to consider a different set of models, for example semi-supervised or fully supervised one. Furthermore, there are many more potential notions of usefulness such as interpretability and fairness that we have not considered in our experimental evaluation. Nevertheless, we argue that the lack of concrete examples of useful disentangled representations necessitates that future work on disentanglement methods should make this point more explicit. While prior work (Steenbrugge et al., 2018; Laversanne-Finot et al., 2018; Nair et al., 2018; Higgins et al., 2017b; 2018) successfully applied disentanglement methods such as $\beta$-VAE on a variety of downstream tasks, it is not clear to us that these approaches and trained models performed well *because of disentanglement*.

## 6. Conclusions

In this work we first theoretically show that the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases. We then performed a large-scale empirical study with six state-of-the-art disentanglement methods, six disentanglement metrics on seven data sets and conclude the following: (i) A factorizing aggregated posterior (which is sampled) does not seem to necessarily imply that the dimensions in the representation (which is taken to be the mean) are uncorrelated. (ii) Random seeds and hyperparameters seem to matter more than the model but tuning seem to require supervision. (iii) We did not observe that increased disentanglement implies a decreased sample complexity of learning downstream tasks. Based on these findings, we suggest three main directions for future research:

**Inductive biases and implicit and explicit supervision.** Our theoretical impossibility result in Section 3 highlights the need of inductive biases while our experimental results indicate that the role of supervision is crucial. As currently there does not seem to exist a reliable strategy to choose hyperparameters in the unsupervised learning of disentangled representations, we argue that future work should make the role of inductive biases and implicit and explicit supervision more explicit. We would encourage and motivate future work on disentangled representation learning that deviates from the static, purely unsupervised setting considered in this work. Promising settings (that have been explored to some degree) seem to be for example (i) disentanglement learning with interactions (Thomas et al., 2017), (ii) when weak forms of supervision e.g. through grouping information are available (Bouchacourt et al., 2018), or (iii) when temporal structure is available for the learning problem. The last setting seems to be particularly interesting given recent identifiability results in non-linear ICA (Hyvarinen & Morioka, 2016).

**Concrete practical benefits of disentangled representations.** In our experiments we investigated whether higher disentanglement scores lead to increased sample efficiency for downstream tasks and did not find evidence that this is the case. While these results only apply to the setting and downstream task used in our study, we are also not aware of other prior work that compellingly shows the usefulness of disentangled representations. Hence, we argue that future work should aim to show concrete benefits of disentangled representations. Interpretability and fairness as well as interactive settings seem to be particularly promising candidates to evaluate usefulness. One potential approach to include inductive biases, offer interpretability, and generalization is the concept of independent causal mechanisms and the framework of causal inference (Pearl, 2009; Peters et al., 2017).

**Experimental setup and diversity of data sets.** Our study also highlights the need for a sound, robust, and reproducible experimental setup on a diverse set of data sets in order to draw valid conclusions. We have observed that it is easy to draw spurious conclusions from experimental results if one only considers a subset of methods, metrics and data sets. Hence, we argue that it is crucial for future work to perform experiments on a wide variety of data sets to see whether conclusions and insights are generally applicable. This is particularly important in the setting of disentanglement learning as experiments are largely performed on toy-like data sets. For this reason, we released `disentanglement_lib`, the library we created to train and evaluate the different disentanglement methods on multiple data sets. We also released more than 10 000 trained models to provide a solid baseline for future methods and metrics.

## Acknowledgements

The authors thank Ilya Tolstikhin, Paul Rubenstein and Josip Djolonga for helpful discussions and comments. This research was partially supported by the Max Planck ETH Center for Learning Systems and by an ETH core grant (to Gunnar Rätsch). This work was partially done while Francesco Locatello was at Google Research Zurich.

## References

Arcones, M. A. and Gine, E. On the bootstrap of u and v statistics. *The Annals of Statistics*, pp. 655–674, 1992.

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul): 1–48, 2002.

Bengio, Y., LeCun, Y., et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI*, 2018.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in beta-vae. In *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems*, 2017.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2615–2625, 2018.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2172–2180, 2016.

Cheung, B., Livezey, J. A., Bansal, A. K., and Olshausen, B. A. Discovering hidden factors of variation in deep networks. In *Workshop at International Conference on Learning Representations*, 2015.

Cohen, T. and Welling, M. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, pp. 1755–1763, 2014.

Cohen, T. S. and Welling, M. Transformation properties of learned visual representations. In *International Conference on Learning Representations*, 2015.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Denton, E. L. and Birodkar, v. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pp. 4414–4423, 2017.

Desjardins, G., Courville, A., and Bengio, Y. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012.

Eastwood, C. and Williams, C. K. I. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Fraccaro, M., Kamronn, S., Paquet, U., and Winther, O. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, pp. 3601–3610, 2017.

Goodfellow, I., Lee, H., Le, Q. V., Saxe, A., and Ng, A. Y. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pp. 646–654, 2009.

Goroshin, R., Mathieu, M. F., and LeCun, Y. Learning to linearize under uncertainty. In *Advances in Neural Information Processing Systems*, pp. 1234–1242, 2015.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017a.

Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., and Lerchner, A. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pp. 1480–1490, 2017b.

Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C. P., Bošnjak, M., Shanahan, M., Botvinick, M., Hassabis, D., and Lerchner, A. Scan: Learning hierarchical compositional visual concepts. In *International Conference on Learning Representations*, 2018.

Hinton, G. E., Krizhevsky, A., and Wang, S. D. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pp. 44–51. Springer, 2011.

Hsu, W.-N., Zhang, Y., and Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*, pp. 1878–1889, 2017.

Hyvarinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pp. 3765–3773, 2016.

Hyvarinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Hyvarinen, A., Sasaki, H., and Turner, R. E. Nonlinear ica using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

Jutten, C. and Karhunen, J. Advances in nonlinear blind source separation. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 245–256, 2003.

Karaletsos, T., Belongie, S., and Rätsch, G. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.

Kim, H. and Mnih, A. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2649–2658, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pp. 2539–2547, 2015.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2017.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Laversanne-Finot, A., Pere, A., and Oudeyer, P.-Y. Curiosity driven exploration of learned disentangled goal spaces. In *Conference on Robot Learning*, pp. 487–504, 2018.

LeCun, Y., Huang, F. J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pp. II–104. IEEE, 2004.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Lenc, K. and Vedaldi, A. Understanding image representations by measuring their equivariance and equivalence. In *IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Locatello, F., Vincent, D., Tolstikhin, I., Rätsch, G., Gelly, S., and Schölkopf, B. Competitive training of mixtures of independent deep generative models. *arXiv preprint arXiv:1804.11130*, 2018.

Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., and LeCun, Y. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pp. 5040–5048, 2016.

Munch, E. The scream, 1893.

Nair, A. V., Pong, V., Dalal, M., Bahl, S., Lin, S., and Levine, S. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, pp. 9209–9220, 2018.

Narayanaswamy, S., Paige, T. B., Van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., and Torr, P. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pp. 5925–5935, 2017.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Reed, S., Sohn, K., Zhang, Y., and Lee, H. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pp. 1431–1439, 2014.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, pp. 1252–1260, 2015.

Ridgeway, K. and Mozer, M. C. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pp. 185–194, 2018.

Rubenstein, P. K., Schoelkopf, B., and Tolstikhin, I. Learning disentangled representations with wasserstein auto-encoders. In *Workshop at International Conference on Learning Representations*, 2018.

Schmidhuber, J. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *International Conference on Machine Learning*, pp. 1255–1262, 2012.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, prediction, and search*. Springer-Verlag. (2nd edition MIT Press 2000), 1993.

Steenbrugge, X., Leroux, S., Verbelen, T., and Dhoedt, B. Improving generalization for abstract reasoning tasks using disentangled feature representations. In *Workshop on Relational Representation Learning at Conference on Neural Information Processing Systems*, 2018.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Suter, R., Miladinović, Đ., Bauer, S., and Schölkopf, B. Interventional robustness of deep latent variable models. *arXiv preprint arXiv:1811.00007*, 2018.

Thomas, V., Bengio, E., Fedus, W., Pondard, J., Beaudoin, P., Larochelle, H., Pineau, J., Precup, D., and Bengio, Y. Disentangling the independently controllable factors of variation by interacting with the world. In *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems*, 2017.

Tschannen, M., Bachem, O., and Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

Whitney, W. F., Chang, M., Kulkarni, T., and Tenenbaum, J. B. Understanding visual concepts with continuation learning. In *Workshop at International Conference on Learning Representations*, 2016.

Yang, J., Reed, S. E., Yang, M.-H., and Lee, H. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pp. 1099–1107, 2015.

Yingzhen, L. and Mandt, S. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pp. 5656–5665, 2018.

Zhu, Z., Luo, P., Wang, X., and Tang, X. Multi-view perceptron: a deep model for learning face identity and view representations. In *Advances in Neural Information Processing Systems*, pp. 217–225, 2014.