On Disentangled Representations Learned from Correlated Data

Frederik Träuble [1] **Elliot Creager** [2] **Niki Kilbertus** [3] **Francesco Locatello** [4] **Andrea Dittadi** [5] **Anirudh Goyal** [6] **Bernhard Schölkopf** [1] **Stefan Bauer** [1 7]

Abstract

The focus of disentanglement approaches has been on identifying independent factors of variation in data. However, the causal variables underlying real-world observations are often not statistically independent. In this work, we bridge the gap to real-world scenarios by analyzing the behavior of the most prominent disentanglement approaches on correlated data in a large-scale empirical study (including 4260 models). We show and quantify that systematically induced correlations in the dataset are being learned and reflected in the latent representations, which has implications for downstream applications of disentanglement such as fairness. We also demonstrate how to resolve these latent correlations, either using weak supervision during training or by post-hoc correcting a pre-trained model with a small number of labels.

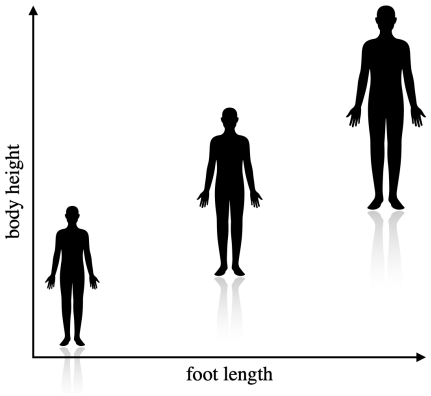

Figure 1. While in principle we would like to consider foot length and body height to be independent features of our model, they exhibit strong positive correlation in observed data.

1. Introduction

Disentangled representations promise generalization to unseen scenarios (Higgins et al., 2017b; Locatello et al., 2020b), increased interpretability (Adel et al., 2018; Higgins et al., 2018), faster learning on downstream tasks (van Steenkiste et al., 2019) or better sim2real transfer (Dittadi et al., 2021) as well as a connection to robustness and causal inference (Suter et al., 2019).

While the advantages of disentangled representations have been well established, they generally assume the existence of natural factors that vary independently within the given dataset, which is rarely the case in real-world settings. As an example, consider a dataset containing images of various persons (see Fig. 1). Higher-level factors of this representation, such as foot length and body height are in fact found to be statistically correlated (Agnihotri et al., 2007; Grivas et al., 2008). Humans can still conceive varying both factors independently across their entire range. However, only given a correlated dataset, learning systems may be tempted to encode both factors simultaneously in a single *size* factor. It is thus argued that what we actually want to infer are independent (causal) mechanisms (Peters et al., 2017; Suter et al., 2019; Goyal et al., 2021; Leeb et al., 2020; Schölkopf et al., 2021).

A complex generative model can be thought of as the composition of independent mechanisms, which generate high-dimensional observations (such as images or videos). In the causality community, this is often considered a prerequisite to achieve representations that are robust to interventions upon variables determined by such models (Peters et al., 2017). The notion of *disentangled* representation is one particular instantiation of this idea (Bengio et al., 2013). The goal of disentanglement learning is to find a representation of the data which captures all the ground-truth factors of variation (FoV) independently.

Despite the recent growth of the field, state-of-the-art disentanglement learners have mostly been trained and evaluated

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Toronto and Vector Institute [3]Helmholtz AI, Munich [4]Amazon (work partly done when FL was at ETH Zurich and MPI-IS) [5]Technical University of Denmark [6]Mila and Université de Montréal [7]CIFAR Azrieli Global Scholar. Correspondence to: Frederik Träuble <frederik.traeuble@tuebingen.mpg.de>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

on datasets in which all FoV indeed vary independently from each other. Their performance remains unknown for more realistic settings where FoV are correlated during training. Given the potential societal impact in the medical domain (Chartsias et al., 2018) or fair decision making (Locatello et al., 2019a; Madras et al., 2018; Creager et al., 2019), the evaluation of the usefulness of disentangled representations trained on correlated data is of high importance.

The idealized settings employed for validating disentanglement learning so far would typically build on a dataset that has as many persons with small feet and small body height as small feet and large body height. Realistic unlabeled datasets are rather unlikely to be of this type. It is thus an open question to what degree existing inductive biases from the encoder/decoder architecture, but most importantly these dataset biases, affect the learned representation. In this work, we introduce dataset correlations in a controlled manner to understand in detail to what degree state-of-the-art approaches can cope with such correlations.

To this end, we report a large-scale empirical study to systematically assess the effect of induced correlations between pairs of factors of variation in training data on the learned representations. To provide a qualitative and quantitative evaluation, we investigate multiple datasets with access to ground-truth labels. Moreover, we study the generalization abilities of the representations learned on correlated data as well as their performance in particular for the downstream task of fair decision making.

Contributions. We summarize our contributions:

- We present the first large-scale empirical study (4260 models)[1] that examines how modern disentanglement learners perform when ground truth factors of the observational data are *correlated*.
- We find that, in this case, factorization-based inductive biases are insufficient to learn disentangled representations. Existing methods fail to disentangle correlated factors of variation as the latent space dimensions become statistically entangled, which has implications for downstream applications of disentanglement such as fairness. We corroborate this theoretically by showing that, in the correlated setting, all generative models that match the marginal likelihood of the ground truth model are not disentangled.
- We propose a simple post-hoc procedure that can align entangled latents with only very few labels and investigate the usefulness of weakly-supervised approaches to resolve latent entanglement already at train time when only correlated data is available.

[1]Each model was trained for 300,000 iterations on Tesla V100 GPUs. Reproducing these experiments requires approximately 0.79 GPU years.

2. Background

Current state-of-the-art disentanglement approaches use the framework of variational autoencoders (VAEs) (Kingma & Welling, 2014; Rezende et al., 2014). The (high-dimensional) observations $\mathbf{x}$ are modeled as being generated from latent features $\mathbf{z}$ according to the probabilistic model $p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$. Typically, the prior $p(\mathbf{z})$ is fixed, while the generative model $p_\theta(\mathbf{x}|\mathbf{z})$ and the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ are parameterized by neural networks with parameters θ and ϕ respectively. These are optimized by maximizing the variational lower bound (ELBO) of the log likelihood $\log p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ of data $D = \{\mathbf{x}^{(i)}\}_{i=1}^N$:

$$\mathcal{L}_{ELBO} = \sum_{i=1}^{N} \Big(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z})] - D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)}) \| p(\mathbf{z})\big) \Big) \tag{1}$$

This objective does not enforce any structure on the latent space, except for similarity (in KL divergence) to the prior $p(\mathbf{z})$ (typically chosen as an isotropic Gaussian). Thus, no specific structure and semantic meaning of latent representations is encouraged by this objective. Consequently, various works propose new evaluation metrics to quantify different notions of disentanglement of the learned representations, as well as new disentanglement learning methods, such as β-VAE, AnnealedVAE, FactorVAE, β-TCVAE, and DIP-VAE, that incorporate suitable structure-imposing regularizers (Higgins et al., 2017a; Burgess et al., 2018; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2018; Eastwood & Williams, 2018; Mathieu et al., 2019).

Since unsupervised disentanglement by optimizing the marginal likelihood in a generative model is impossible (Locatello et al., 2019b, Theorem 1), inductive biases like grouping information (Bouchacourt et al., 2018) or access to labels (Locatello et al., 2020c) is required. To address this theoretical limitation, methods have been proposed that only require weak label information (Locatello et al., 2020b; Shu et al., 2020). Changes in natural environments, which typically correspond to changes of only a few underlying FoV, can provide a weak supervision signal for representation learning algorithms (Goyal et al., 2021; Földiák, 1991; Schmidt et al., 2007; Bengio et al., 2020; Ke et al., 2019; Klindt et al., 2021). In the absence of correlations it has been shown on common disentanglement datasets that this weak supervision facilitates learning disentangled representations (Locatello et al., 2020b; Shu et al., 2020).

Most popular datasets in the disentanglement literature exhibit perfect independence in their FoV such as *dSprites* (Higgins et al., 2017a), *Cars3D* (Reed et al., 2015), *SmallNORB* (LeCun et al., 2004), *Shapes3D* (Kim & Mnih, 2018) or *MPI3D* variants (Gondal et al., 2019). At some level this is sensible as it reflects the underlying assumption

in the inductive bias being studied. However, this assumption is unlikely to hold in practice as is also shown by Li et al. (2019).

3. The Problem with Correlated Data

Most work on learning disentangled representations assumes that there is an underlying set of independent ground truth variables that govern the generative process of observable data. These methods are hence predominantly trained and evaluated on datasets that obey independence in the true factors of variation by design, which we then consider to be the correct factorization. Formally, disentanglement methods typically assume $\mathbf{x} \sim \int_{\mathbf{c}} p^*(\mathbf{x}|\mathbf{c})p^*(\mathbf{c})d\mathbf{c}$ where the prior over ground truth factors $\mathbf{c}$ factorizes as

$$p^*(c_1, c_2, \ldots, c_n) = \prod_{i=1}^{n} p^*(c_i). \tag{2}$$

Note that this is the dataset-generating prior and distinct from the latent prior $p(\mathbf{z})$ of the model.

In the real world, however, we generally expect correlations in the collected datasets, i.e., the joint distribution of the ground truth factors $\{c_i\}_{i=1}^n$ does not factorize:

$$p^*(c_1, c_2, \ldots, c_n) \neq \prod_{i=1}^{n} p^*(c_i). \tag{3}$$

In this case, we speak of dependence between the random variables, also commonly referred to as correlation. Even though there might be correlation between multiple variables simultaneously, for a principled systematic analysis we will mostly consider pairwise correlations. Correlation between two variables can stem from various sources, e.g., from a direct causal relationship (one variable affects the other), or from unobserved circumstances (confounders) affecting both. Real-world datasets display many of these "spurious" and often a priori unknown correlations (Geirhos et al., 2020; Li et al., 2019). In the introductory example, various (potentially hidden) confounders such as age or sex may indeed affect both foot length and body height. Even though we expect such a dataset to exhibit strong correlations in these factors, we would like to model them independently of each other.

To establish some intuition for why such correlations pose problems for generative models attempting to learn a disentangled representation, assume two FoV c_1 and c_2 that are correlated in the dataset. In a perfectly disentangled representation, a single latent dimension z_{c1} would model factor c_1 and z_{c2} would model factor c_2. However, because our model has an independent latent prior, the resulting generative model would generate all combinations of the two correlated factors, i.e., put probability mass also outside of the training distribution. Hence, perfect disentanglement would lead to a sub-optimal log likelihood. We can show this more formally as follows (see Appendix A for a proof).

Proposition 1 *Consider the latent variable model* $p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})d\mathbf{z}$ *and let the true data distribution be* $p^*(\mathbf{x}) = \int_{\mathbf{c}} p^*(\mathbf{x} \mid \mathbf{c})p^*(\mathbf{c})d\mathbf{c}$ *where* $\mathbf{z} \in \mathcal{Z}$ *and* $\mathbf{c} \in \mathcal{C}$. *Assume that the fixed prior* $p(\mathbf{z}) = \prod_i p(z_i)$ *factorizes, the true FoV prior* $p^*(\mathbf{c}) \neq \prod_i p^*(c_i)$ *does not, and that there exists a smooth bijection* $f^* : \mathcal{C} \to \mathcal{Z}$ *with smooth inverse (i.e., a diffeomorphism) that transforms* $p^*(\mathbf{c})$ *into* $p(\mathbf{z})$. *Then, the likelihood* $p_\theta(\mathbf{x})$ *can be equal to the optimal likelihood* $p^*(\mathbf{x})$ *if and only if the representations* $\mathbf{z}$ *are entangled w.r.t. the true factors of variation* $\mathbf{c}$.

Note that we define a representation to be *disentangled* if the Jacobian of f^* is diagonal up to a permutation, in line with Locatello et al. (2019b) & Locatello et al. (2020b). Under the assumptions of this proposition, we established that a generative model that matches the true marginal likelihood $p^*(\mathbf{x})$ cannot be disentangled. This suggests that, if some FoV are correlated in the dataset and the prior $p(\mathbf{z})$ factorizes, methods that optimize a lower bound to the log likelihood might have a bias *against* disentanglement.[2] The primary goal of this work is thus to empirically assess to what extent the additional inductive biases of state-of-the-art disentanglement methods still suffice to learn disentangled representations when the training data exhibits correlations.

Why not encode train-time correlations? The prior discussion raises the question why we would like to learn representations that do not encode data correlations. Why not encode foot size and body height in a single latent dimension when we are only shown data in which they are strongly correlated? Maximizing the log likelihood, which amounts to matching the data distribution at train time, would embrace such correlations, and it would later allow us to sample novel examples from this distribution. We now provide three major reasons why this may not always be the goal in real-world applications.

First, if we want to build models that generalize well across multiple tasks and dataset distributions the standard approach is myopic. A key promise of disentanglement is to learn meaningful and compact representations, in which relevant semantic aspects of the data are structurally disentangled and can be analyzed or controlled independently from each other during inference or generation. From this perspective, resolving instead of exploiting correlations in the train data is desirable at test time, where such representa-

[2] While the impossibility result of Locatello et al. (2019b) states that there may be many generative models achieving the optimal likelihood for a ground-truth model with independent FoVs, Proposition 1 states that if the ground-truth prior is correlated, a disentangled representation will never achieve this optimal likelihood and therefore entangled representations are preferred.

tions would remain robust and meaningful under distribution shifts. Returning to our running example, assume training data is constrained to a fixed sex and age bracket, both being confounders for foot length and body height (Grivas et al., 2008). If we could disentangle these factors despite the correlation, the model would arguably be more robust to distribution shifts, for example when testing on a different sex or different ages.

Second, we would like to be able to sample out-of-distribution (OOD) examples or intervene on individual factors independently of the others. Should a good model not be able to meaningfully reason about and imagine someone's foot length changing independently of their body height?

Lastly, disentangled representation are relevant in fairness settings, where representations are used in downstream tasks for consequential decisions that ought to be fair with respect to sensitive (protected) variables (Locatello et al., 2019a; Creager et al., 2019). Undesirable data correlations with such protected variables are a major problem in fairness applications. Therefore, it is crucial to evaluate the validity of factorization-based inductive biases to learn disentangled representations from correlated observational data.

How can we resolve these latent correlations? To resolve potential latent correlations, we will investigate two approaches. First, we explore the scenario in which few labels are given after training with the specific purpose of resolving these correlations. For this, we propose a simple approach in Section 5, which we find to be effective. Second, we employ the performance of the weakly-supervised approach of Locatello et al. (2020b). They showed (on uncorrelated data) that access to pairs of observations which display differences in a known number of FoV (without knowing which ones specifically) suffices to learn disentangled representations. These additional weak assumptions render the generative model identifiable in contrast to unsupervised disentanglement and may arguably be indeed available in certain practical settings, e.g., in temporally close frames from a video of a moving robot arm where some factors remain unchanged.[3] We test the Ada-GVAE algorithm from Locatello et al. (2020b), which requires a pair of observations that differ in an unknown number of factors. To impose this structure in the latent space, the latent factors that are shared by the two observations are estimated from the k largest coordinate-wise KL divergences $D_{KL}(q_\phi(z_i|\mathbf{x}^{(1)})\|q_\phi(z_i|\mathbf{x}^{(2)}))$. They then maximize the following modified β-VAE (Higgins et al., 2017a) objective

[3]On the other hand, in applications with fairness concerns it may be impossible to intervene on FoV representing sensitive attributes of individuals (sex, race, etc.); we refer to (Kilbertus et al., 2017; Madras et al., 2019) for a more complete discussion.

corr. dataset	data source	1st correlated FoV	2nd correlated FoV
A	Shapes3D	object size	azimuth
B	dSprites	orientation	x-position
C	MPI3D	1st DoF	2nd DoF
D	Shapes3D	object color	object size
E	Shapes3D	object color	azimuth

Table 1. Dataset variants with introduced correlations under investigation within this empirical study.

for the pair of observations

$$\sum_{i\in\{1,2\}} \mathbb{E}_{(\mathbf{x}^{(1)},\mathbf{x}^{(2)})}\Big[\mathbb{E}_{\tilde{q}_\phi^{(i)}(\hat{\mathbf{z}}|\mathbf{x}^{(1)},\mathbf{x}^{(2)})}\log(p_\theta(\mathbf{x}^{(i)}|\hat{\mathbf{z}})) - \beta D_{KL}\left(\tilde{q}_\phi^{(i)}(\hat{\mathbf{z}}|\mathbf{x}^{(1)},\mathbf{x}^{(2)})\|p(\hat{\mathbf{z}})\right)\Big], \quad (4)$$

with $\tilde{q}_\phi^{(i)}(\hat{z}_j|\mathbf{x}^{(1)},\mathbf{x}^{(2)}) = q_\phi(\hat{z}_j|\mathbf{x}^{(i)})$ for the latent dimensions z_j that are inferred to be changing and $\tilde{q}_\phi^{(i)}(\hat{z}_j|\mathbf{x}^{(1)},\mathbf{x}^{(2)}) = a(q_\phi(\hat{z}_j|\mathbf{x}^{(1)}), q_\phi(\hat{z}_j|\mathbf{x}^{(2)}))$ for those inferred to be shared. The averaging function a forces the approximate posterior of the shared latent variable to be the same in the two observations.

4. Unsupervised Disentanglement under Correlated Factors

In this section, we present the key findings from our empirical study of unsupervised disentanglement learning approaches on a particular variant of correlated datasets. We start by outlining the experimental design of our studies in Section 4.1. Based on this, in Section 4.2 we present a latent space analysis of unsupervised disentanglement learners and find that factorization-based inductive biases are insufficient to learn disentangled representations from correlated observational data. Although these methods fail to disentangle correlated factors, in Section 4.3 we qualitatively show that they still generalize to FoV combinations that are out-of-distribution with respect to the training data.

4.1. Experimental Design

To begin to systematically understand the unknown behavior of SOTA disentanglement approaches on correlated data, we focus on linear correlations with Gaussian noise between pairs of FoV, denoted by c_1 and c_2.[4] For our experiments we introduce five correlated dataset variants comprising dSprites, Shapes3D and MPI3D with correlations between

[4]We emphasize that this seemingly narrow class of correlations already captures most relevant effects of more general correlations between FoV at train time on the learned representation. This is not to say that one can draw general conclusions about highly nonlinear correlated settings. However, rigorously understanding the linear case in a wide range of controlled settings already represents a considerable set of experiments and is a vital step in bridging the gap between highly idealized settings and real-world applications.

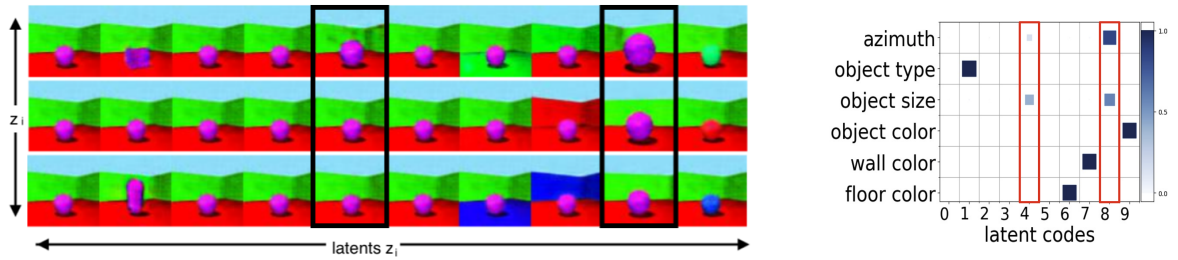

Figure 2. **Left:** Latent traversals for the model with best DCI score (commonly used metric for measuring disentanglement (Eastwood & Williams, 2018)) among all 180 models trained on strongly correlated ($\sigma = 0.2$) Shapes3D data (A). The traversals in latent dimensions 5 and 9 (highlighted in black) encode a mixture of azimuth and object size, reflecting the main correlation line of the joint distribution and a smaller, locally orthogonal one. **Right:** A heat map of the GBT feature importance matrix of this model indicates an entanglement of azimuth and object size encoded into both latent codes.

single pairs of FoV. Table 1 shows the correlated FoV in each dataset variant and the names we refer to them by. We then parameterize correlations by sampling the training dataset from the joint distribution

$$p(c_1, c_2) \propto \exp\left(-\frac{(c_1 - \alpha c_2)^2}{2\sigma^2}\right)$$

with $c_1 \in \{0, \dots, c_1^{\max}\}$ and $c_2 \in \{0, \dots, c_2^{\max}\}$ and $\alpha = c_2^{\max}/c_1^{\max}$. The correlation strength can be tuned by σ, for which we choose 0.2, 0.4, 0.7 in normalized units with respect to the range of values in $\{c_1, c_2\}$. Lower σ indicates stronger correlation. See Fig. 4 for an example of $p(c_1, c_2)$ for correlating azimuth and object size in Shapes3D. Additionally, we study the uncorrelated limit ($\sigma = \infty$), which amounts to the case typically considered in the literature. All remaining factors are sampled uniformly at random.

We train the same six VAE methods as Locatello et al. (2019b), i.e., β-VAE, FactorVAE, AnnealedVAE, DIP-VAE-I, DIP-VAE-II and β-TC-VAE, each with 6 hyperparameter settings and 5 random seeds. These 180 models are trained on datasets A, B, and C with $\sigma \in \{0.2, 0.4, 0.7, \infty\}$ and datasets D and E with $\sigma \in \{0.2, 0.4\}$, totaling 2880 models. Appendix B describes additional implementation details.[5]

4.2. Unsupervised Methods for Correlated Data

Here we assess the applicability of SOTA unsupervised learning approaches to correlated data. Are the factorization-based inductive biases introduced by these methods enough to disentangle correlated FoV although it is sub-optimal from a standard VAE perspective?

Latent structure and pairwise entanglement. We start by visually inspecting latent traversals of some trained models on Shapes3D (A). For strong correlations ($\sigma = 0.2$ and $\sigma = 0.4$), we typically observe trained models with two latent codes encoding the two correlated variables simultaneously. In these cases, one of the latent codes corresponds to data along the major axis of the correlation line whereas the other latent code dimension manifests in an orthogonal change of the two variables along the minor axis. Perhaps unsurprisingly, a full traversal of the code corresponding to the minor axis often seems to cover only observations within the variance of the correlation line, i.e., in the training distribution. Fig. 2 (left) shows this effect for the latent space of a model trained on Shapes3D (A) with strongest correlation ($\sigma = 0.2$). To quantify this observation, we analyze the importance of individual latent codes in predicting the value of a given ground truth FoV. An importance weight for each pair of {FoV, latent dimension} is computed by training a gradient boosting tree (GBT) classifier to predict the ground truth labels from the latents (10,000 examples). In the right panel of Fig. 2, we show these importance weights for the model used to generate traversals in the left panel. The corresponding evaluation for a model trained on the same dataset with a much weaker correlation of $\sigma = 0.7$ does not reveal this feature visually (see Fig. 7 in Appendix C.1).

Corr. strength		$\sigma = 0.2$	$\sigma = 0.4$	$\sigma = 0.7$	$\sigma = \infty$ (uc)
Shapes3D (A)	object size - azimuth	0.38	0.26	0.13	0.08
	median uncorrelated pairs	0.09	0.09	0.09	0.08
dSprites (B)	orientation - position x	0.17	0.16	0.14	0.11
	median uncorrelated pairs	0.13	0.13	0.13	0.13
MPI3D (C)	First DOF - Second DOF	0.2	0.19	0.17	0.16
	median uncorrelated pairs	0.16	0.16	0.15	0.15

Table 2. The means of the pairwise entanglement scores for the correlated pair (red) and the median of the uncorrelated pairs. Stronger correlation leads to statistically more entangled latents compared to the baseline score without correlation (blue), thus uncovering still existent correlations in the latent representation.

To support this claim empirically across the full study with all datasets, we calculate a pairwise entanglement score that allows us to measure how difficult it is to separate two factors of variation from their latent codes. This computation involves grouping FoV into pairs based on an ordering of their pairwise mutual information or GBT feature importance between latents and FoV; we refer to Appendix B

[5] Code for reproducing experiments is available under https://github.com/ftraeuble/disentanglement_lib

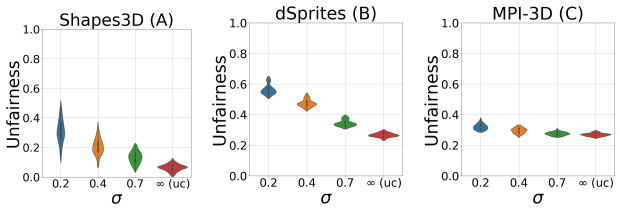

Figure 3. Stronger correlations in the train data lead to substantially elevated unfairness scores for the correlated pair of factors.

for a detailed description of this procedure. Table 2 shows that across all datasets the pair of correlated FoV has a substantially higher score than the median of all other pairs, indicating that they are harder to disentangle. This score decreases with weaker correlation, i.e., the pair becomes easier to disentangle for $\sigma \geq 0.7$. We conclude that models still disentangle weakly correlated factors, but inductive biases become insufficient to avoid latent entanglements for stronger correlations. In Appendix C.1 we further consolidate our claims by showing that the correlated pair is more entangled in the latent representation across all unsupervised experiments and datasets. We also further corroborate our finding that standard disentanglement metrics when evaluated on all FoV are insufficient to reveal this entanglement.

Does this matter for downstream tasks? Correlations between variables are of crucial importance in fairness applications, motivating an additional investigation on ramifications of these entangled latent spaces. In this setting we are interested in the unfairness of predicting one of the two correlated variables when the other represents a protected or sensitive attribute. In the following, we use a variant of demographic parity (Dwork et al., 2012) that computes pairwise mutual information between latents and FoV (Locatello et al., 2019a). In Fig. 3 we evaluate this score when correlations are present across all unsupervised experiments and datasets. Unfairness tracks correlation strength in this scenario. Locatello et al. (2019a) show that representations learned without supervision may exhibit unfairness even without correlations. Our results suggest that the problem is substantially aggravated in the presence of correlations between a sensitive attribute and another ground truth FoV.

Summary. Factorization-based inductive biases are insufficient to learn disentangled representations from correlated observational data in the unsupervised case. We observed persisting pairwise entanglement in the latent space, both visually and quantitatively using appropriate metrics, which might be particularly problematic for fairness applications.

4.3. Generalization Under Changing Correlations

We now aim to understand the robustness and generalization capabilities of these models on data that is sampled not iid from the correlated train distribution but instead out-of-

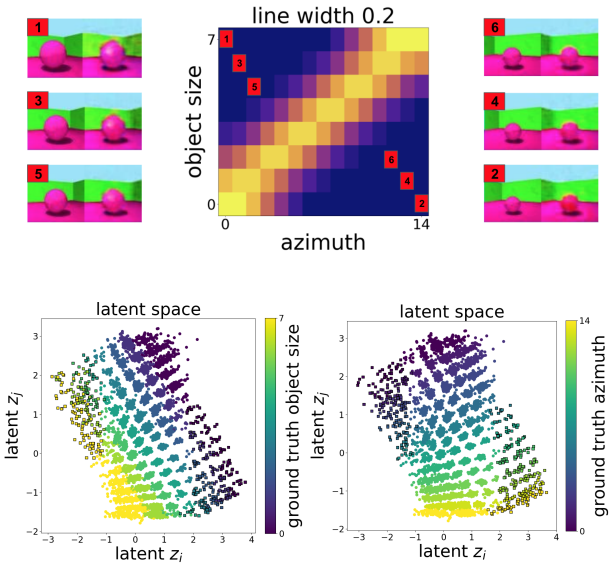

Figure 4. Generalization to out-of-distribution (OOD) test data. **Top:** Reconstructions of observations the model has never seen during training. **Bottom:** Latent space distribution of the two entangled dimensions. Circles without edges represents latent encodings of (correlated) training examples. Circles with edges are OOD examples that break the correlation pattern.

distribution (OOD) from a low-probability region given the correlation structure, e.g., large feet but short body height. In our Shapes3D experiments, this amounts to large object size and small azimuth which has *zero* probability under the correlated train distribution. We focus on the model from Fig. 2, which has disentangled all non-correlated factors well, allowing us to only focus on the two latent dimensions encoding the entangled variables.

First, we analyze examples with object size and azimuth values that have numerically zero probability under the correlated data distribution as shown as numbers 1 to 6 in Fig. 4 (top panel) in the FoV space. The model can reconstruct these examples despite never having seen such configurations during training. This suggests that OOD examples are meaningfully encoded into the existing structure of the latent space *and* that the decoder is equally capable of generating observations from such unseen representations. Note that in contrast to Zhao et al. (2018) who test this combinatorial generalization in the interpolation regime from uncorrelated, unbiased, but sparsely populated train data, to the best of our knowledge this extrapolation OOD generalization has not been studied on strongly correlated data so far. To test this hypothesis further, we analyzed latent traversals originating from these OOD points and observe that changes in the remaining factors reliably yield the expected reconstructions. Traversals with respect to the two entangled latent codes continue to encode object size and azimuth.

To fully understand these models' generalization properties we visualize the latent space encoding of the training distribution projected onto the two identified entangled dimensions (i.e., marginalizing over all others) in Fig. 4 (bottom panel). We depict the ground truth value of each correlated variable via a color gradient. The two sets of depicted points are (1) latent codes for samples from the correlated training data (no edges) and (2) latent codes for samples with (object size, azimuth) configurations that have zero probability under the correlated training distribution (edges). As expected, contours of equal color (ground truth) are not aligned with the latent axes, indicating that the two latent dimensions encode both FoV at the same time. The generalization capabilities of this model away from the training data at least for the decoder can be described as follows: If we imagine the contour lines of equal color in both plots for the training data, i.e., only circles without edges, and linearly extend those lines, they form a grid in the latent space beyond where the training data lies. Latent encodings of unseen combinations of ground truth FoV happen to meaningfully extend and complete this grid structure in the latent space. Note that in this example, there is a natural ordering on both correlated FoV (numerical values for azimuth and size). Even though for categorical factors such as object color in datasets D and E we cannot expect the ad-hoc ordering of colors to be preserved in the latent space, the latent encodings still extrapolate meaningfully for examples with unseen combinations of ground truth FoV. In Appendix C.2, we show some of these characteristic latent space visualizations with similar extrapolation and generalization capabilities.

Summary. We conclude from these results that even though the models fail to disentangle the correlated FoV, they are still incorporating enough structure in the latent space to generalize towards unseen FoV configurations that are OOD w.r.t. the correlated train distribution.

5. Recovering Factorized Latent Codes

We now investigate the usefulness of semi- and weakly-supervised approaches to resolve latent entanglement.

Post-hoc alignment correction with few labels. When a limited number of FoV labels $\{\mathbf{c}_i\}_{i=1}^{M}$ can be accessed, a reasonable option for resolving entangled dimensions of the latent code is by *fast adaptation*. To identify the two entangled dimensions $\mathbf{z}_{fa} := (z_i, z_j)$ we look at the maximum feature importance for a given FoV from a GBT trained using these labels only. We then train a *substitution function* $f_\theta : \mathbb{R}^2 \to \{0, \dots, c_1^{\max}\} \times \{0, \dots, c_2^{\max}\}$ via supervised learning to infer the two ground truth FoV c_1, c_2 from the entangled latent codes z_i, z_j, $f_\theta(\mathbf{z}_{fa}) = (c_1, c_2)$. We then use this prediction to replace these two dimension of the latent codes. Crucially, both steps of this procedure rely on the same M FoV labels, of which we assume only very few being available in practice.

# Labels		0	100	1000	10000
Shapes3D (A) $\sigma = 0.2$	object size - azimuth	0.38	0.17	0.15	0.15
	median uncorrelated pairs	0.09	0.08	0.07	0.07
Shapes3D (A) $\sigma = 0.4$	object size - azimuth	0.26	0.1	0.1	0.1
	median uncorrelated pairs	0.09	0.08	0.08	0.08

Table 3. **Fast adaption with few labels:** Pairwise entanglement scores for correlated FoV pair in Shapes3D (A). The correlated pair is highlighted (red). Zero labels reflects the unsupervised baseline without any fast adaptation. Growing number of labels show that fast adaption using linear regression reduces these correlations with as little as 100 labels (blue). Reported pairwise scores are averaged over 180 models per correlation strength.

Table 3 shows the pairwise entanglement score of the correlated FoV under this fast adaptation with a linear regression as the substitution function, which succeeds with as few as 100 labels (we only need the correlated factor labels), corresponding to less than 0.02% of all ground truth labels in Shapes3D. However, fast adaptation with linear regression substitution fails in some settings: when no two latent dimensions encode the applied correlation isolated from the other latent codes, or when the correlated variables do not have a unique natural ordering (e.g., categorical variables). Accordingly, a nonlinear substitution function such as an MLP can further reduce pairwise entanglement in these cases (see additional results in Appendix D.1).

Weak supervision mitigates learning latent entanglement. We now return to the weakly-supervised method from Section 2 and evaluate its applicability when training data is correlated. Specifically, we study the variant where the difference in the observation pair is present in one random ground truth factor. Each time we construct such a pair, all underlying ground truth factors share the same values except for one factor. Whenever this happens to be one of the correlated factors, the values of this factor within each pair are drawn from the marginal probability distribution conditioned on the other correlated factor. In these cases the difference in this factor is typically very small and depends on the correlation strength. Note that this procedure assures that constructed pairs are consistent with the observational data such that the correlation is never broken. In this part of our study, we consider datasets A, D, and E, and train β-VAEs with the same 6 hyperparameters and 5 random seeds as in the unsupervised study, yielding 360 models. Fig. 5 summarizes the weak supervision results on Shapes3D (A) when imposing correlations in object size and azimuth. We consistently observe much better disentangled models, often achieving perfect DCI score irrespective of correlations in the dataset. The latent spaces tend to strongly align their coordinates with the ground truth label axis. Finally, weak-supervision reduces unfairness relative to the unsupervised baseline from Fig. 3, and occasionally even

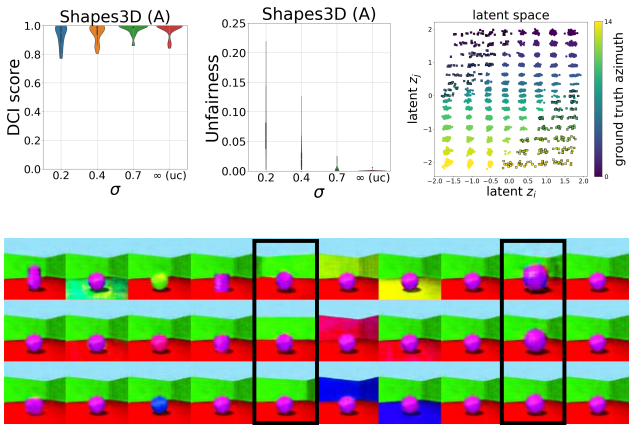

Figure 5. **Top left:** We show the disentanglement (DCI score) for models trained on Shapes3D *with weak supervision* when object size and azimuth are correlated with different strength. Weak supervision helps to recover improved, often perfect, disentanglement. **Top middle:** Unfairness scores between correlated FoV are drastically reduced when using weak supervision (see scale) across all correlation strengths. **Top right:** Latent dimensions of the best DCI model with strong correlation ($\sigma = 0.2$) when using weak supervision. Representations are axis-aligned with respect to both of the correlated ground truth factors. **Lower:** Latent traversals of this model trained on strong correlation ($\sigma = 0.2$).

achieves zero unfairness score. Additional results on the other datasets can be found in Appendix D.2 including two additional training regimes in which the weak-supervision pairs are sampled from an interventional (but still correlated) distribution, representing an additional 1020 models. We consistently observe the same strong trends regarding disentangled correlations in all of the above studies using weak supervision.

These results suggest that weak supervision can provide a strong inductive bias capable of finding the right factorization and resolving spurious correlations for datasets of unknown degree of correlation. As a prominent example, this is an issue in the fairness context where real-world datasets often display unknown discriminatory correlations.

Summary. When ground truth information about the correlated FoV is available, our fast adaptation method can resolve latent correlations using only few labels of correlated FoV. When no labels are available, recently proposed approaches in weakly supervised disentanglement learning applied on correlated data offer an alternative to overcoming pairwise correlations at train time. We have thus shown how to leverage even weak supervision signals to learn disentangled representations from correlated data.

6. Other Related Work

ICA and Disentanglement. Ideas related to disentangling the factors of variation date back to the non-linear ICA literature (Comon, 1994; Jutten & Karhunen, 2003; Hyvärinen & Pajunen, 1999; Hyvarinen et al., 2019; Hyvarinen & Morioka, 2016; Gresele et al., 2019). Recent work combines non-linear ICA with disentanglement (Khemakhem et al., 2020; Sorrenson et al., 2020).

Entangled and rotated latent spaces. Zhou et al. (2021) measure disentanglement based on topological similarity of latent submanifolds, arguing that this may help uncover latent entanglements. While having orthogonal goals to ours, it complements our empirical findings regarding the problematic structure of entangled latent spaces. Rolinek et al. (2019) contributed a related theoretical result showing that VAEs promote latent spaces pursued by (locally) orthogonal PCA embeddings due to the role of the diagonal covariance of the latent prior. They confirm this experimentally using the uncorrelated dSprites dataset. Stühmer et al. (2020) use a family of rotationally asymmetric distributions as the latent prior, which can help learning disentangled subspaces. In contrast to the modeling perspective, we studied the effect of dependencies from the data perspective with strong correlations in the data generating prior $p^*(\mathbf{c})$.

Studies on correlated data. The literature so far is missing a systematic large-scale empirical study how popular inductive biases such as factorized priors behave when actually learning from correlated datasets, although several smaller experiments along these lines can be acknowledged. Chen et al. (2018) studied correlated 3DFaces (Paysan et al., 2009) by fixing all except three factors in which the authors conclude that the β-TC-VAE regularizer can help to disentangle imposed correlations based on the MIG metric. However, the latent structure was not studied in detail; our findings suggest that global disentanglement metrics are insufficient to fully diagnose models learned from correlated data. Based on the observation that the assumption of independent factors is unlikely to hold in practice and certain factors might be correlated in data Li et al. (2019) propose methods based on a pairwise independence assumption instead. Brekelmans et al. (2019) show that Echo noise results in superior disentanglement compared to standard β-VAE in a small experiment on a downsampled dSprites variant. Creager et al. (2019) based some of the evaluations of a proposed new autoencoder architecture in the fairness context on a biased dSprites variant and Yang et al. (2020) study a linear SCM in a VAE architecture on datasets with dependent variables. However, their studies focused on representation learners that require strong supervision via FoV labels at train time.

7. Conclusion

We have taken first steps to understand the gap between idealized datasets and realistic applications of modern disentanglement learners by presenting the first large-scale empirical study examining how such models cope with correlated observational data. We find that existing methods fail to learn disentangled representations for strongly correlated factors of variation. We discuss and quantify practical implications for downstream tasks like fair decision making. Despite these shortcomings, the learned latent space structure of some models naturally accommodates unseen examples via meaningful extrapolation, leading to out-of-distribution generalization capabilities. Based on these findings, we ultimately demonstrate how to correct for latent entanglement via fast adaptation and other weakly supervised training methods. Future work is needed to address open question surrounding whether these results extend to broader classes of correlations in particular exhibited by time series (Miladinović et al., 2020) from high resolution inputs (Miladinović et al., 2021) and their impact on downstream tasks in the real-world (Ahmed et al., 2021). Finally, our findings draw additional attention towards how inductive biases and weak supervision can be combined for successful disentanglement and under what circumstances this leads to strong out-of-distribution generalization.

Acknowledgments

Frederik Träuble would like to thank Felix Leeb, Georgios Arvanitidis, Dominik Zietlow and Julius von Kügelgen for fruitful discussions. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Frederik Träuble and CIFAR.

References

Adel, T., Ghahramani, Z., and Weller, A. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning (ICML)*, pp. 50–59, 2018.

Agnihotri, A. K., Purwar, B., Googoolye, K., Agnihotri, S., and Jeebun, N. Estimation of stature by foot length. *Journal of forensic and legal medicine*, 14(5):279–283, 2007.

Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, Y., Schölkopf, B., Wüthrich, M., and Bauer, S. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *International Conference on Learning Representations (ICLR)*, 2021.

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.

Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., and Pal, C. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Brekelmans, R., Moyer, D., Galstyan, A., and Ver Steeg, G. Exact rate-distortion in autoencoders via echo noise. In *Advances in Neural Information Processing Systems*, pp. 3889–3900, 2019.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in beta-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D., Dharmakumar, R., and Tsaftaris, S. A. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 490–498. Springer, 2018.

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

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

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning (ICML)*, 2019.

Dittadi, A., Träuble, F., Locatello, F., Wüthrich, M., Agrawal, V., Winther, O., Bauer, S., and Schölkopf, B. On the transfer of disentangled representations in realistic settings. In *International Conference on Learning Representations (ICLR)*, 2021.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations.

In *International Conference on Learning Representations (ICLR)*, 2018.

Földiák, P. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gondal, M. W., Wüthrich, M., Miladinović, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Advances in Neural Information Processing Systems*, 2019.

Goyal, A., Lamb, A., Hoffmann, J., Sodhani, S., Levine, S., Bengio, Y., and Schölkopf, B. Recurrent independent mechanisms. In *International Conference on Learning Representations (ICLR)*, 2021.

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

Grivas, T. B., Mihas, C., Arapaki, A., and Vasiliadis, E. Correlation of foot length with height and weight in school age children. *Journal of forensic and legal medicine*, 15 (2):89–95, 2008.

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 (ICLR)*, 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 (ICML)*, 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 (ICLR)*, 2018.

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

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 1999.

Hyvarinen, A., Sasaki, H., and Turner, R. E. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Jutten, C. and Karhunen, J. Advances in nonlinear blind source separation. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 245–256, 2003.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pp. 656–666, 2017.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018.

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

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations (ICLR)*, 2021.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations (ICLR)*, 2018.

LeCun, Y., Huang, F. J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

Leeb, F., Annadani, Y., Bauer, S., and Schölkopf, B. Structured representation learning using structural autoencoders and hybridization. *arXiv preprint arXiv:2006.07796*, 2020.

Li, Z., Tang, Y., Li, W., and He, Y. Learning disentangled representation with pairwise independence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4245–4252, 2019.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., and Bachem, O. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pp. 14584–14597, 2019a.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019b.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. A sober look at the unsupervised learning of disentangled representations and their evaluation. *Journal of Machine Learning Research*, 21(209):1–62, 2020a.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning (ICML)*, pp. 7753–7764, July 2020b.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variations using few labels. In *International Conference on Learning Representations (ICLR)*, April 2020c.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning (ICML)*, 2018.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 349–358, 2019.

Mathieu, E., Rainforth, T., Siddharth, N., and Whye Teh, Y. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning (ICML)*, 2019.

Miladinović, Đ., Gondal, M. W., Schölkopf, B., Buhmann, J. M., and Bauer, S. Disentangled state space representations. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2020.

Miladinović, Đ., Stanić, A., Bauer, S., Schmidhuber, J., and Buhmann, J. M. Spatial dependency networks: Neural layers for improved generative image modeling. *International Conference on Learning Representations (ICLR)*, 2021.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 296–301. Ieee, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

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

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Rolinek, M., Zietlow, D., and Martius, G. Variational autoencoders recover pca directions (by accident). In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2019.

Schmidt, M., Niculescu-Mizil, A., Murphy, K., et al. Learning graphical model structure using l1-regularization paths. In *AAAI*, volume 7, pp. 1278–1283, 2007.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations (ICLR)*, 2020.

Sorrenson, P., Rother, C., and Köthe, U. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations (ICLR)*, 2020.

Stühmer, J., Turner, R., and Nowozin, S. Independent subspace analysis for unsupervised learning of disentangled representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1200–1210. PMLR, 2020.

Suter, R., Miladinović, D., Bauer, S., and Schölkopf, B. Interventional robustness of deep latent variable models. In *International Conference on Machine Learning (ICML)*, 2019.

van Steenkiste, S., Locatello, F., Schmidhuber, J., and Bachem, O. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, 2019.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

Zhao, S., Ren, H., Yuan, A., Song, J., Goodman, N., and Ermon, S. Bias and generalization in deep generative models: An empirical study. In *Advances in Neural Information Processing Systems*, 2018.

Zhou, S., Zelikman, E., Lu, F., Ng, A. Y., and Ermon, S. Evaluating the disentanglement of deep generative models through manifold topology. In *International Conference on Learning Representations (ICLR)*, 2021.