Weakly-Supervised Disentanglement Without Compromises

Francesco Locatello [1 2] **Ben Poole** [3] **Gunnar Rätsch** [1]
Bernhard Schölkopf [2] **Olivier Bachem** [3] **Michael Tschannen** [3]

Abstract

Intelligent agents should be able to learn useful representations by observing changes in their environment. We model such observations as pairs of non-i.i.d. images sharing at least one of the underlying factors of variation. First, we theoretically show that only knowing *how many* factors have changed, but not which ones, is sufficient to learn disentangled representations. Second, we provide practical algorithms that learn disentangled representations from pairs of images without requiring annotation of groups, individual factors, or the number of factors that have changed. Third, we perform a large-scale empirical study and show that such pairs of observations are sufficient to reliably learn disentangled representations on several benchmark data sets. Finally, we evaluate our learned representations and find that they are *simultaneously* useful on a diverse suite of tasks, including generalization under covariate shifts, fairness, and abstract reasoning. Overall, our results demonstrate that weak supervision enables learning of useful disentangled representations in realistic scenarios.

1. Introduction

A recent line of work argued that representations which are *disentangled* offer useful properties such as interpretability (Adel et al., 2018; Bengio et al., 2013; Higgins et al., 2017a), predictive performance (Locatello et al., 2019b; 2020), reduced sample complexity on abstract reasoning tasks (van Steenkiste et al., 2019), and fairness (Locatello et al., 2019a; Creager et al., 2019). The key underlying assumption is that high-dimensional observations $\mathbf{x}$ (such as images or videos) are in fact a manifestation of a low-dimensional set of independent ground-truth factors of variation $\mathbf{z}$ (Locatello et al., 2019b; Bengio et al., 2013; Tschannen et al., 2018). The goal of disentangled representation learning is to learn a function $r(\mathbf{x})$ mapping the observations to a low-dimensional vector that contains all the information about each factor of variation, with each coordinate (or a subset of coordinates) containing information about only one factor. Unfortunately, Locatello et al. (2019b) showed that the unsupervised learning of disentangled representations is theoretically impossible from i.i.d. observations without inductive biases. In practice, they observed that unsupervised models exhibit significant variance depending on hyperparameters and random seed, making their training somewhat unreliable.

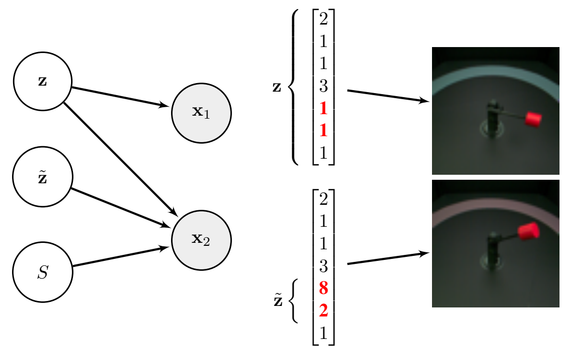

Figure 1. (**left**) The proposed generative model. We observe pairs of observations $(\mathbf{x}_1, \mathbf{x}_2)$ sharing a random subset S of latent factors: $\mathbf{x}_1$ is generated by $\mathbf{z}$; $\mathbf{x}_2$ is generated by combining the subset S of $\mathbf{z}$ and resampling the remaining entries (modeled by $\tilde{\mathbf{z}}$). (**right**) Real-world example of the model: A pair of images from MPI3D (Gondal et al., 2019) where all factors are shared except the first degree of freedom and the background color (red values). This corresponds to a setting where few factors in a causal generative model change, which, by the *independent causal mechanisms* principle, leaves the others invariant (Schölkopf et al., 2012).

On the other hand, many data modalities are *not* observed as i.i.d. samples from a distribution (Dayan, 1993; Storck et al., 1995; Hochreiter & Schmidhuber, 1999; Bengio et al., 2013; Peters et al., 2017; Thomas et al., 2017; Schölkopf, 2019). Changes in natural environments, which typically correspond to changes of only a few underlying factors of variation, provide a weak supervision signal for representation learning algorithms (Földiák, 1991; Schmidt et al., 2007; Bengio, 2017; Bengio et al.,

[1]Department of Computer Science, ETH Zurich [2]Max Planck Institute for Intelligent Systems [3]Google Research, Brain Team. Correspondence to: <francesco.locatello@inf.ethz.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

2019). State-of-the-art weakly-supervised disentanglement methods (Bouchacourt et al., 2018; Hosoya, 2019; Shu et al., 2020) assume that observations belong to annotated groups where two things are known at training time: (i) the relation between images in the same group, and (ii) the group each image belongs to. Bouchacourt et al. (2018); Hosoya (2019) consider groups of observations differing in precisely one of the underlying factors. An example of such a group are images of a given object with a fixed orientation, in a fixed scene, but of varying color. Shu et al. (2020) generalized this notion to other relations (e.g., single shared factor, ranking information). In general, precise knowledge of the groups and their structure may require either explicit human labeling or at least strongly controlled acquisition of the observations. As a motivating example, consider the video feedback of a robotic arm. In two temporally close frames, both the manipulated objects and the arm may have changed their position, the objects themselves may be different, or the lighting conditions may have changed due to failures.

In this paper, we consider learning disentangled representations from pairs of observations which differ by a few factors of variation (Bengio, 2017; Schmidt et al., 2007; Bengio et al., 2019) as in Figure 1. Unlike previous work on weakly-supervised disentanglement, we consider the realistic and broadly applicable setting where we observe pairs of images and have no additional annotations: It is unknown *which* and *how many* factors of variation have changed. In other words, we do not know which group each pair belongs to, and what is the precise relation between the two images. The only condition we require is that the two observations are different and that the change in the factors is not dense. The key contributions of this paper are:

- We present simple adaptive group-based disentanglement methods which do not require annotations of the groups, as opposed to (Bouchacourt et al., 2018; Hosoya, 2019; Shu et al., 2020). Our approach is readily applicable to a variety of settings where groups of non-i.i.d. observations are available with no additional annotations.
- We theoretically show that identifiability is possible from non-i.i.d. pairs of observations under weak assumptions. Our proof motivates the setup we consider, which is identifiable as opposed to the standard one, which was proven to be non-identifiable (Locatello et al., 2019b). Further, we use theoretical arguments to inform the design of our algorithms, recover existing group-based VAE methods (Bouchacourt et al., 2018; Hosoya, 2019) as special cases, and relax their impractical assumptions.
- We perform a large-scale reproducible experimental study training over $15\,000$ disentanglement models and over one million downstream classifiers[1] on five different data sets, one of which consisting of real images of a robotic platform (Gondal et al., 2019).
- We demonstrate that one can reliably learn disentangled representations with weak supervision only, *without relying on supervised disentanglement metrics for model selection*, as done in previous works. Further, we show that these representations are useful on a diverse suite of downstream tasks, including a novel experiment targeting strong generalization under covariate shifts, fairness (Locatello et al., 2019a) and abstract visual reasoning (van Steenkiste et al., 2019).

[1]Our experiments required ~ 5.85 GPU years (NVIDIA P100).

2. Related work

Recovering independent components of the data generating process is a well-studied problem in machine learning. It has roots in the independent component analysis (ICA) literature, where the goal is to unmix independent non-Gaussian sources of a d-dimensional signal (Comon, 1994). Crucially, identifiability is not possible in the nonlinear case from i.i.d. observations (Hyvärinen & Pajunen, 1999). Recently, the ICA community has considered weak forms of supervision such as temporal consistency (Hyvarinen & Morioka, 2016; 2017), auxiliary supervised information (Hyvarinen et al., 2019; Khemakhem et al., 2019), and multiple views (Gresele et al., 2019). A parallel thread of work has studied distribution shifts by identifying changes in causal generative factors (Zhang et al., 2015; 2017; Huang et al., 2017), which is linked to a causal view of disentanglement (Suter et al., 2019; Schölkopf, 2019).

On the other hand, more applied machine learning approaches have experienced the opposite shift. Initially, the community focused on more or less explicit and task dependent supervision (Reed et al., 2014; Yang et al., 2015; Kulkarni et al., 2015; Cheung et al., 2014; Mathieu et al., 2016; Narayanaswamy et al., 2017). For example, a number of works rely on known relations between the factors of variation (Karaletsos et al., 2015; Whitney et al., 2016; Fraccaro et al., 2017; Denton & Birodkar, 2017; Hsu et al., 2017; Yingzhen & Mandt, 2018; Locatello et al., 2018; Ridgeway & Mozer, 2018; Chen & Batmanghelich, 2020) and disentangling motion and pose from content (Hsieh et al., 2018; Fortuin et al., 2019; Deng et al., 2017; Goroshin et al., 2015).

Recently, there has been a renewed interest in the unsupervised learning of disentangled representations (Higgins et al., 2017a; Burgess et al., 2018; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2018) along with quantitative evaluation (Kim & Mnih, 2018; Eastwood & Williams, 2018; Kumar et al., 2018; Ridgeway & Mozer, 2018; Duan et al., 2019). After the theoretical impossibility result of Locatello et al. (2019b), the focus shifted back to semi-supervised (Locatello et al., 2020; Sorrenson et al., 2020; Khemakhem et al., 2019) and weakly-supervised approaches (Bouchacourt et al., 2018; Hosoya, 2019; Shu et al., 2020).

3. Generative models

We first describe the generative model commonly used in the disentanglement literature, and then turn to the weakly-supervised model used in this paper.

Unsupervised generative model First, a $\mathbf{z}$ is drawn from a set of independent ground-truth factors of variation $p(\mathbf{z}) = \prod_i p(\mathbf{z}_i)$. Second, the observations are obtained as draws from $p(\mathbf{x}|\mathbf{z})$. The factors of variation $\mathbf{z}_i$ do not need to be one-dimensional but we assume so to simplify the notation.

Disentangled representations The goal of disentanglement learning is to learn a mapping $r(\mathbf{x})$ where the effect of the different factors of variation is axis-aligned with different coordinates. More precisely, each factor of variation z_i is associated with exactly one coordinate (or group of coordinates) of $r(\mathbf{x})$ and vice-versa (and the groups are non-overlapping). As a result, varying one factor of variation and keeping the others fixed results in a variation of exactly one coordinate (group of coordinates) of $r(\mathbf{x})$. Locatello et al. (2019b) theoretically showed that learning such a mapping r is theoretically impossible without inductive biases or some other, possibly weak, form of supervision.

Weakly-supervised generative model We study learning of disentangled image representations from paired observations, for which some (but not all) factors of variation have the same value. This can be modeled as sampling two images from the causal generative model with an intervention (Peters et al., 2017) on a random subset of the factors of variation. Our goal is to use the additional information given by the pair (as opposed to a single image) to learn a disentangled image representations. We generally do not assume knowledge of which or how many factors are shared, i.e., we do not require controlled acquisition of the observations. This observation model applies to many practical scenarios. For example, we may want to learn a disentangled representation of a robot arm observed through a camera: In two temporally close frames some joint angles will likely have changed, but others will have remained constant. Other factors of variation may also change independently of the actions of the robot. An example can be seen in Figure 1 (right) where the first degree of freedom of the arm and the color of the background changed. More generally this observation model applies to many natural scenes with moving objects (Földiák, 1991). More formally, we consider the following generative model. For simplicity of exposition, we assume that the number of factors k in which the two observations differ is constant (we present a strategy to deal with varying k in Section 4.1). The generative model is given by

$$p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i), \quad p(\tilde{\mathbf{z}}) = \prod_{i=1}^{k} p(\tilde{z}_i), \quad S \sim p(S) \tag{1}$$

$$\mathbf{x}_1 = g^\star(\mathbf{z}), \qquad \mathbf{x}_2 = g^\star(f(\mathbf{z}, \tilde{\mathbf{z}}, S)), \tag{2}$$

where S is the subset of shared indices of size $d - k$ sampled from a distribution $p(S)$ over the set $\mathcal{S} = \{S \subset [d] \colon |S| = d - k\}$, and the $p(z_i)$ and $p(\tilde{z}_j)$ are all identical. The generative mechanism is modeled using a function $g^\star \colon \mathcal{Z} \to \mathcal{X}$, with $\mathcal{Z} = \mathrm{supp}(\mathbf{z}) \subseteq \mathbb{R}^d$ and $\mathcal{X} \subset \mathbb{R}^m$, which maps the latent variable to observations of dimension m, typically $m \gg d$. To make the relation between $\mathbf{x}_1$ and $\mathbf{x}_2$ explicit, we use a function f obeying

$$f(\mathbf{z}, \tilde{\mathbf{z}}, S)_S = \mathbf{z}_S \qquad \text{and} \qquad f(\mathbf{z}, \tilde{\mathbf{z}}, S)_{\bar{S}} = \tilde{\mathbf{z}}$$

with $\bar{S} = [d] \backslash S$. Intuitively, to generate $\mathbf{x}_2$, f selects entries from $\mathbf{z}$ with index in S and substitutes the remaining factors with $\tilde{\mathbf{z}}$, thus ensuring that the factors indexed by S are shared in the two observations. The generative model (1)–(2) does not model additive noise; we assume that noise is explicitly modeled as a latent variable and its effect is manifested through $g^\star$ as done by (Bengio et al., 2013; Locatello et al., 2019b; Higgins et al., 2018; 2017a; Suter et al., 2019; Reed et al., 2015; LeCun et al., 2004; Kim & Mnih, 2018; Gondal et al., 2019). For simplicity, we consider the case where groups consisting of two observations (pairs), but extensions to more than two observations are possible (Gresele et al., 2019).

4. Identifiability and algorithms

First, we show that, as opposed to the unsupervised case (Locatello et al., 2019b), the generative model (1)–(2) is identifiable under weak additional assumptions. Note that the joint distribution of all random variables factorizes as

$$p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}, \tilde{\mathbf{z}}, S) = p(\mathbf{x}_1|\mathbf{z})p(\mathbf{x}_2|f(\mathbf{z}, \tilde{\mathbf{z}}, S))p(\mathbf{z})p(\tilde{\mathbf{z}})p(S) \tag{3}$$

where the likelihood terms have the same distribution, i.e., $p(\mathbf{x}_1|\bar{\mathbf{z}}) = p(\mathbf{x}_2|\bar{\mathbf{z}}), \forall \bar{\mathbf{z}} \in \mathrm{supp}(p(\mathbf{z}))$. We show that to learn a disentangled generative model of the data $p(\mathbf{x}_1, \mathbf{x}_2)$ it is therefore sufficient to recover a factorized latent distribution with factors $p(\hat{z}_i) = p(\hat{\tilde{z}}_j)$, a corresponding likelihood $q(\mathbf{x}_1|\cdot) = q(\mathbf{x}_2|\cdot)$, as well as a distribution $p(\hat{S})$ over $\mathcal{S}$, which together satisfy the constraints of the true generative model (1)–(2) and match the true $p(\mathbf{x}_1, \mathbf{x}_2)$ after marginalization over $\hat{\mathbf{z}}, \hat{\tilde{\mathbf{z}}}, \hat{S}$ when substituted into (3).

Theorem 1. *Consider the generative model (1)–(2). Further assume that $p(z_i) = p(\tilde{z}_i)$ are continuous distributions, $p(S)$ is a distribution over $\mathcal{S}$ s.t. for $S, S' \sim p(S)$ we have $P(S \cap S' = \{i\}) > 0, \forall i \in [d]$. Let $g^\star \colon \mathcal{Z} \to \mathcal{X}$ in (2) be smooth and invertible on $\mathcal{X}$ with smooth inverse (i.e., a diffeomorphism). Given unlimited data from $p(\mathbf{x}_1, \mathbf{x}_2)$ and the true (fixed) k, consider all tuples $(p(\hat{z}_i), q(\mathbf{x}_1|\hat{\mathbf{z}}), p(\hat{S}))$ obeying these assumptions and matching $p(\mathbf{x}_1, \mathbf{x}_2)$ after marginalization over $\hat{\mathbf{z}}, \hat{\tilde{\mathbf{z}}}, \hat{S}$ when substituted in (3). Then, the posteriors $q(\hat{\mathbf{z}}|\mathbf{x}_1) = q(\mathbf{x}_1|\hat{\mathbf{z}})p(\hat{\mathbf{z}})/p(\mathbf{x}_1)$ are disentangled in the sense that the aggregate posteriors $q(\hat{\mathbf{z}}) = \int q(\hat{\mathbf{z}}|\mathbf{x}_1)p(\mathbf{x}_1)\mathrm{d}\mathbf{x}_1 = \iint q(\hat{\mathbf{z}}|\mathbf{x}_1)p(\mathbf{x}_1|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x}_1$ are* *coordinate-wise reparameterizations of the ground-truth prior $p(\mathbf{z})$ up to a permutation of the indices of $\mathbf{z}$.*

Discussion Under the assumptions of this theorem, we established that all generative models that match the true marginal over the observations $p(\mathbf{x}_1, \mathbf{x}_2)$ must be disentangled. Therefore, constrained distribution matching is sufficient to learn disentangled representations. Formally, the aggregate posterior $q(\hat{\mathbf{z}})$ is a coordinate-wise reparameterization of the true distribution of the factors of variation (up to index permutations). In other words, there exists a one-to-one mapping between every entry of $\mathbf{z}$ and a unique matching entry of $\hat{\mathbf{z}}$, and thus a change in a single coordinate of $\mathbf{z}$ implies a change in a single matching coordinate of $\hat{\mathbf{z}}$ (Bengio et al., 2013). Changing the observation model from single i.i.d. observations to non-i.i.d. pairs of observations generated according to the generative model (1)–(2) allows us to bypass the non-identifiability result of (Locatello et al., 2019b). Our result requires strictly weaker assumptions than the result of Shu et al. (2020) as we do not require group annotations, but only knowledge of k. As we shall see in Section 4.1, k can be cheaply and reliably estimated from data at run-time. Although the weak assumptions of Theorem 1 may not be satisfied in practice, we will show that the proof can inform practical algorithm design.

4.1. Practical adaptive algorithms

We conceive two β-VAE (Higgins et al., 2017a) variants tailored to the weakly-supervised generative model (1)–(2) and a selection heuristic to deal with unknown and random k. We will see that these simple models can very reliably learn disentangled representations.

The key differences between theory and practice are that: (i) we use the ELBO and amortized variational inference for distribution matching (the true and learned distributions will not exactly match after training), (ii) we have access to a finite number of data only, and (iii) the theory assumes known, fixed k, but k might be unknown and random.

Enforcing the structural constraints Here we present a simple structure for the variational family that allows us to tractably perform approximate inference on the weakly-supervised generative model. First note that the alignment constraints imposed by the generative model (see (7) and (8) evaluated for $g = g^\star$ in Appendix A) imply for the true posterior

$$p(z_i|\mathbf{x}_1) = p(z_i|\mathbf{x}_2) \quad \forall i \in S, \tag{4}$$
$$p(z_i|\mathbf{x}_1) \neq p(z_i|\mathbf{x}_2) \quad \forall i \in \bar{S}, \tag{5}$$

(with probability 1) and we want to enforce these constraints on the approximate posterior $q_\phi(\hat{\mathbf{z}}|\mathbf{x})$ of our learned model. However, the set S is unknown. To obtain an estimate $\hat{S}$ of S we therefore choose for every pair $(\mathbf{x}_1, \mathbf{x}_2)$ the $d-k$ coordinates with the smallest $D_{\mathrm{KL}}(q_\phi(\hat{z}_i|\mathbf{x}_1)||q_\phi(\hat{z}_i|\mathbf{x}_2))$. To impose the constraint (4) we then replace each shared coordinate with some average a of the two posteriors

$$\begin{aligned}
\tilde{q}_\phi(\hat{z}_i|\mathbf{x}_1) &= a(q_\phi(\hat{z}_i|\mathbf{x}_1), q_\phi(\hat{z}_i|\mathbf{x}_2)) \qquad && \forall i \in \hat{S}, \\
\tilde{q}_\phi(\hat{z}_i|\mathbf{x}_1) &= q_\phi(\hat{z}_i|\mathbf{x}_1) && \text{else},
\end{aligned}$$

and obtain $\tilde{q}_\phi(\mathbf{z}_i|\mathbf{x}_2)$ in analogous manner. As we later simply use the averaging strategies of the Group-VAE (GVAE) (Hosoya, 2019) and the Multi Level-VAE (ML-VAE) (Bouchacourt et al., 2018), we term variants of our approach which infers the groups and their properties adaptively *Adaptive-Group-VAE* (Ada-GVAE) and *Adaptive-ML-VAE* (Ada-ML-VAE), depending on the choice of the averaging function a. We then optimize the following variant of the β-VAE objective

$$\begin{aligned}
\max_{\phi,\theta} \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)} \mathbb{E}_{\tilde{q}_\phi(\hat{\mathbf{z}}|\mathbf{x}_1)} & \log(p_\theta(\mathbf{x}_1|\hat{\mathbf{z}})) \\
&+ \mathbb{E}_{\tilde{q}_\phi(\hat{\mathbf{z}}|\mathbf{x}_2)} \log(p_\theta(\mathbf{x}_2|\hat{\mathbf{z}})) \\
&- \beta D_{KL}\left(\tilde{q}_\phi(\hat{\mathbf{z}}||\mathbf{x}_1)|p(\hat{\mathbf{z}})\right) \\
&- \beta D_{KL}\left(\tilde{q}_\phi(\hat{\mathbf{z}}||\mathbf{x}_2)|p(\hat{\mathbf{z}})\right),
\end{aligned} \tag{6}$$

where $\beta \geq 1$ (Higgins et al., 2017a). The advantage of this averaging-based implementation of (4), over implementing it, for instance, via a D_{KL}-term that encourages the distributions of the shared coordinates $\hat{S}$ to be similar, is that averaging imposes a hard constraint in the sense that $q_\phi(\hat{\mathbf{z}}|\mathbf{x}_1)$ and $q_\phi(\hat{\mathbf{z}}|\mathbf{x}_2)$ can jointly encode only one value per shared coordinate. This in turn implicitly enforces the constraint (5) as the non-shared dimensions need to be efficiently used to encode the non-shared factors of $\mathbf{x}_1$ and $\mathbf{x}_2$.

We emphasize that the objective (6) is a simple modification of the β-VAE objective and is very easy to implement. Finally, we remark that invoking Theorem 4 of (Khemakhem et al., 2019), we achieve consistency under maximum likelihood estimation up to the equivalence class in our Theorem 1, for $\beta = 1$ and in the limit of infinite data and capacity.

Inferring k In the (practical) scenario where k is unknown, we use the threshold

$$\tau = \frac{1}{2}(\max_i \delta_i + \min_i \delta_i),$$

where $\delta_i = D_{\mathrm{KL}}(q_\phi(\hat{z}_i|\mathbf{x}_1)||q_\phi(\hat{z}_i|\mathbf{x}_2))$, and average the coordinates with $\delta_i < \tau$. This heuristic is inspired by the "elbow method" (Ketchen & Shook, 1996) for model selection in k-means clustering and k-singular value decomposition and we found it to work surprisingly well in practice (see the experiments in Section 5). This estimate relies on the assumption that not all factors have changed. All our adaptive methods use this heuristic. Although a formal recovery argument cannot be made for arbitrary data sets, inductive biases may limit the impact of an approximate

k in practice. We further remark that this heuristic always yields the correct k if the encoder is disentangled.

Relation to prior work Closely related to the proposed objective (6) the GVAE of Hosoya (2019) and the ML-VAE of Bouchacourt et al. (2018) assume S is known and implement a using different averaging choices. Both assume Gaussian approximate posteriors where μ_j, Σ_j are the mean and variance of $q(\hat{\mathbf{z}}_S|\mathbf{x}_j)$ and μ, Σ are the mean and variance, of $\tilde{q}(\hat{\mathbf{z}}_S|\mathbf{x}_j)$. For the coordinates in S, the GVAE uses a simple arithmetic mean ($\mu = \frac{1}{2}(\mu_1 + \mu_2)$ and $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$) and the ML-VAE takes the product of the encoder distributions, with μ, Σ taking the form:

$$\Sigma^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1}, \quad \mu^T = (\mu_1^T \Sigma_1^{-1} + \mu_2^T \Sigma_2^{-1})\Sigma.$$

Our approach critically differs in the sense that S is not known and needs to be estimated for every pair of images.

Recent work combines non-linear ICA with disentanglement (Khemakhem et al., 2019; Sorrenson et al., 2020). Critically, these approaches are based on the setup of Hyvarinen et al. (2019) which requires access to label information $\mathbf{u}$ such that $p(\mathbf{z}|\mathbf{u})$ factorizes as $\prod_i p(z_i|\mathbf{u})$. In contrast, we base our work on the setup of Gresele et al. (2019), which only assumes access to two *sufficiently distinct views* of the latent variable. Shu et al. (2020) train the same type of generative models over paired data but use a GAN objective where inference is not required. However, they require known and fixed k as well as annotations of which factors change in each pair.

5. Experimental results

Experimental setup We consider the setup of Locatello et al. (2019b). We use the five data sets where the observations are generated as deterministic functions of the factors of variation: *dSprites* (Higgins et al., 2017a), *Cars3D* (Reed et al., 2015), *SmallNORB* (LeCun et al., 2004), *Shapes3D* (Kim & Mnih, 2018), and the real-world robotics data set *MPI3D* (Gondal et al., 2019). Our unsupervised baselines correspond to a cohort of 9000 unsupervised models (β-VAE (Higgins et al., 2017a), AnnealedVAE (Burgess et al., 2018), Factor-VAE (Kim & Mnih, 2018), β-TCVAE (Chen et al., 2018), DIP-VAE-I and II (Kumar et al., 2018)), each with the same six hyperparameters from Locatello et al. (2019b) and 50 random seeds.

To create data sets with weak supervision from the existing disentanglement data sets, we first sample from the discrete $\mathbf{z}$ according to the ground-truth generative model (1)–(2). Then, we sample k factors of variation that should not be shared by the two images and re-sample those coordinates to obtain $\tilde{\mathbf{z}}$. This ensures that each image pair differs in at most k factors of variation. For k we consider the range from 1 to $d-1$. This last setting corresponds to the case where all but one factor of variation are re-sampled. We study both the case where k is constant across all pairs in the data set and where k is sampled uniformly in the range $[d-1]$ for every training pair ($k = \texttt{Rnd}$ in the following). Unless specified otherwise, we aggregate the results for all values of k.

For each data set, we train four weakly-supervised methods: Our adaptive and vanilla (group-supervision) variants of GVAE (Hosoya, 2019) and ML-VAE (Bouchacourt et al., 2018). For each approach we consider six values for the regularization strength and 10 random seeds, training a total of 6000 weakly-supervised models. We perform model selection using the weakly-supervised reconstruction loss (i.e., the sum of the first two terms in (6))[2]. We stress that we *do not require labels for model selection*.

To evaluate the representations, we consider the disentanglement metrics in Locatello et al. (2019b): *BetaVAE* score (Higgins et al., 2017a), *FactorVAE* score (Kim & Mnih, 2018), *Mutual Information Gap (MIG)* (Chen et al., 2018), *Modularity* (Ridgeway & Mozer, 2018), *DCI Disentanglement* (Eastwood & Williams, 2018) and *SAP score* (Kumar et al., 2018). To directly compare the disentanglement produced by different methods, we report the DCI Disentanglement (Eastwood & Williams, 2018) in the main text and defer the plots with the other scores to the appendix as the same conclusions can be drawn based on these metrics. Appendix B contains full implementation details.

5.1. Is weak supervision enough for disentanglement?

In Figure 2, we compare the performance of the weakly-supervised methods with $k = \texttt{Rnd}$ against the unsupervised methods. Unlike in unsupervised disentanglement with β-VAEs where $\beta \gg 1$ is common, we find $\beta = 1$ (the ELBO) performs best in most cases. We clearly observe that weakly-supervised models outperform the unsupervised ones. In Figure 6 in the appendix, we further observe that they are competitive even if we allow fully supervised model selection on the unsupervised models. The Ada-GVAE performs similarly to the Ada-ML-VAE. For this reason, we focus the following analysis on the Ada-GVAE, and include Ada-ML-VAE results in Appendix C.

Summary With weak supervision, we reliably learn disentangled representations that outperform unsupervised ones. Our representations are competitive even if we perform fully supervised model selection on the unsupervised models.

5.2. Are our methods adaptive to different values of k?

In Figure 3 (left), we report the performance of Ada-GVAE without model selection for different values of k on MPI3D

[2] In Figure 9 in the appendix, we show that the training loss and the ELBO correlate similarly with disentanglement.

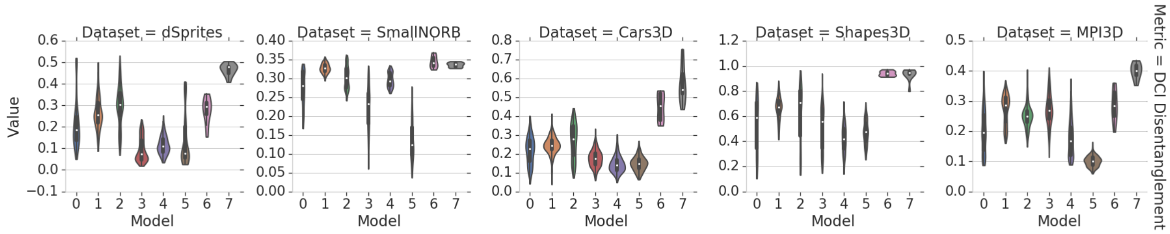

Figure 2. Our adaptive variants of the group-based disentanglement methods (models 6 and 7) significantly and consistently outperform unsupervised methods. In particular, the Ada-GVAE consistently yields the same or better performance than the Ada-ML-VAE. In this experiment, we consider the case where the number of shared factors of variation is random and different for every pair with high probability ($k = \texttt{Rnd}$). Legend: 0=β-VAE, 1=FactorVAE, 2=β-TCVAE, 3=DIP-VAE-I, 4=DIP-VAE-II, 5=AnnealedVAE, 6=Ada-ML-VAE, 7=Ada-GVAE

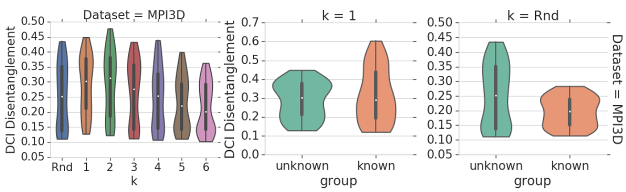

Figure 3. (**left**) Performance of the Ada-GVAE with different k on MPI3D. The algorithm adapts well to the unknown k and benefits from sparser changes. (**center** and **right**) Comparison of Ada-ML-VAE with the vanilla ML-VAE which assumes group knowledge. We note that group knowledge may improve performance (**center**) but can also hurt when it is incomplete (**right**).

(see Figure 10 in the appendix for the other data sets). We observe that Ada-GVAE is indeed adaptive to different values of k and it achieves better performance when the change between the factors of variation is sparser. Note that our method is agnostic to the sharing pattern between the image pairs. In applications where the number of shared factors is known to be constant, the performance may thus be further improved by injecting this knowledge into the inference procedure.

Summary Our approach makes no assumption of which and how many factors are shared and successfully adapts to different values of k. The sparser the difference on the factors of variation, the more effective our method is in using weak supervision and learning disentangled representations.

5.3. Supervision-performance trade-offs

The case $k = 1$ where we actually know which factor of variation is not shared was previously considered in (Bouchacourt et al., 2018; Hosoya, 2019; Shu et al., 2020). Clearly, this additional knowledge should lead to improvements over our method. On the other hand, this information may be correct but incomplete in practice: For every pair of images, we know about one factor of variation that has changed but it may not be the only one. We therefore also consider the setup where $k = \texttt{Rnd}$ but the algorithm is only informed about one factor. Note that the original GVAE assumes group knowledge, so we directly compare its performance with our Ada-GVAE. We defer the comparison with ML-VAE (Bouchacourt et al., 2018) and with the GAN-based approaches of (Shu et al., 2020) to Appendix C.3.

In Figure 3 (center and right), we observe that when $k = 1$, the knowledge of which factor was changed generally improves the performance of weakly-supervised methods on MPI3D. On the other hand, the GVAE is not robust to incomplete knowledge as its performance degrades when the factor that is labeled as non-shared is not the only one. The performance degradation is stronger on the data sets with more factors of variation (dSprites/Shapes3D/MPI3D) as can be seen in Figure 12 in the appendix. This may not come as a surprise as group-based disentanglement methods all assume that the group knowledge is precise.

Summary Whenever the groups are fully and precisely known, this information can be used to improve disentanglement. Even though our adaptive method does not use group annotations, its performance is often comparable to the methods of (Bouchacourt et al., 2018; Hosoya, 2019; Shu et al., 2020). On the other hand, in practical applications there may not be precise control of which factors have changed. In this scenario, relying on incomplete group knowledge significantly harms the performance of GVAE and ML-VAE as they assume exact group knowledge. A blend between our adaptive variant and the vanilla GVAE may further improve performance when only partial group knowledge is available.

5.4. Are weakly-supervised representations useful?

In this section, we investigate whether the representations learned by our Ada-GVAE are useful on a variety of tasks. We show that representations with small weakly-supervised reconstruction loss (the sum of the first two terms in (6)) achieve improved downstream performance (Locatello et al., 2019b; 2020), improved downstream generalization (Peters et al., 2017) under covariate shifts (Shimodaira, 2000; Quionero-Candela et al., 2009; Ben-David et al., 2010), fairer downstream predictions (Locatello et al., 2019a), and improved sample complexity on an abstract reasoning task (van Steenkiste et al., 2019). To the best of our

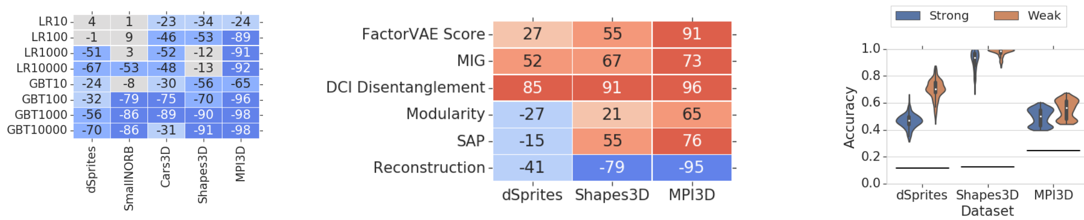

Figure 4. (**left**) Rank correlation between our weakly-supervised reconstruction loss and performance of downstream prediction tasks with logistic regression (LR) and gradient boosted decision-trees (GBT) at different sample sizes for Ada-GVAE. We observe a general negative correlation that indicates that models with a low weakly-supervised reconstruction loss may also be more accurate. (**center**) Rank correlation between the strong generalization accuracy under covariate shifts and disentanglement scores as well as weakly-supervised reconstruction loss, for Ada-GVAE. (**right**) Distribution of vanilla (weak) generalization and under covariate shifts (strong generalization) for Ada-GVAE. The horizontal line corresponds to the accuracy of a naive classifier based on the prior only.

knowledge, strong generalization under covariate shift has not been tested on disentangled representations before.

Key insight We remark that the usefulness insights of Locatello et al. (2019b; 2020; 2019a); van Steenkiste et al. (2019) are based on the assumption that disentangled representations can be learned without observing the factors of variation. They consider models trained without supervision and argue that *some* of the *supervised disentanglement scores* (which require explicit labeling of the factors of variation) correlate well with desirable properties. *In stark contrast, we here show that all these properties can be achieved simultaneously using only weakly-supervised data.*

5.4.1. Downstream performance

In this section, we consider the prediction task of Locatello et al. (2019b) that predicts the values of the factors of variation from the representation. We also evaluate whether our weakly-supervised reconstruction loss is a good proxy for downstream performance. We use a setup identical to Locatello et al. (2019b) and train the same logistic regression and gradient boosted decision trees (GBT) on the learned representations using different sample sizes (10/100/1000/10 000). All test sets contain 5000 examples.

In Figure 4 (left), we observe that the weakly-supervised reconstruction loss of Ada-GVAE is generally anti-correlated with downstream performance. The best weakly-supervised disentanglement methods thus learn representations that are useful for training accurate classifiers downstream.

Summary The weakly-supervised reconstruction loss of our Ada-GVAE is a useful proxy for downstream accuracy.

5.4.2. Generalization under covariate shift

Assume we have access to a large pool of unlabeled paired data and our goal is to solve a prediction task for which we have a smaller labeled training set. Both the labeled training set and test set are biased, but with different biases. For example, we want to predict object shape but our training set contains *only* red objects, whereas the test set does not contain *any* red objects. We create a biased training set by performing an intervention on a random factor of variation (other than the target variable), so that its value is constant in the whole training set. We perform another intervention on the test set, so that the same factor can take all other values. We train a GBT classifier on 10000 examples from the representations learned by Ada-GVAE. For each target factor of variation, we repeat the training of the classifier 10 times for different random interventions. For this experiment, we consider only dSprites, Shapes3D and MPI3D since Cars3D and SmallNORB are too small (after an intervention on their most fine grained factor of variation, they only contain 96 and 270 images respectively).

In Figure 4 (center) we plot the rank correlation between disentanglement scores and weakly-supervised reconstruction, and the results for generalization under covariate shifts for Ada-GVAE. We note that both the disentanglement scores and our weakly-supervised reconstruction loss are correlated with strong generalization. In Figure 4 (right), we highlight the gap between the performance of a classifier trained on a normal train/test split (which we refer to as *weak* generalization) as opposed to this covariate shift setting. We do not perform model selection, so we can show the performance of the whole range of representations. We observe that there is a gap between weak and strong generalization but the distributions of accuracies significantly overlap and are significantly better than a naive classifier based on the prior distribution of the classes.

Summary Our results provide compelling evidence that disentanglement is useful for strong generalization under covariate shifts. The best Ada-GVAE models in terms of weakly-supervised reconstruction loss are useful for training classifiers that generalize under covariate shifts.

5.4.3. Fairness

Recently, Locatello et al. (2019a) showed that disentangled representations may be useful to train robust classifiers that are fairer to unobserved sensitive variables independent of the target variable. While they observed a strong correlation

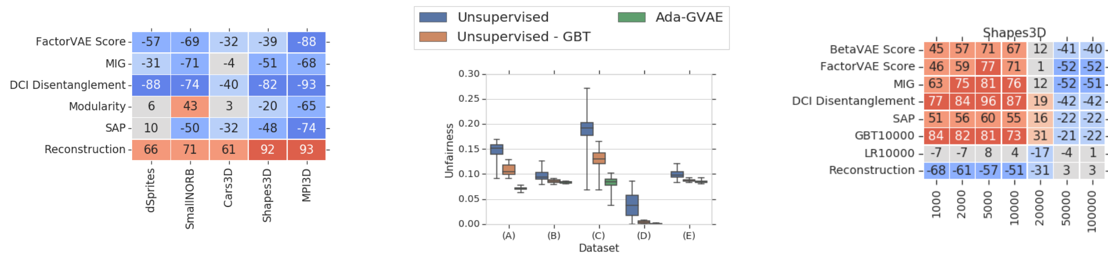

Figure 5. (**left**) Rank correlation between both disentanglement scores and our weakly-supervised reconstruction loss with the unfairness of GBT10000 on all the data sets for Ada-GVAE. (**center**) Unfairness of the unsupervised methods with the semi-supervised model selection heuristic of (Locatello et al., 2019a) and our weakly-supervised Ada-GVAE with $k = 1$. (**right**) Rank correlation with down-stream accuracy of the abstract visual reasoning models of (van Steenkiste et al., 2019) throughout training (i.e., for different sample sizes).

between demographic parity (Calders et al., 2009; Zliobaite, 2015) and disentanglement, the applicability of their approach is limited by the fact that disentangled representations are difficult to identify without access to explicit observations of the factors of variation (Locatello et al., 2019b).

Our experimental setup is identical to the one of Locatello et al. (2019a) and we measure *unfairness* of a classifier as in Locatello et al. (2019a, Section 4). In Figure 5 (left), we show that the weakly-supervised reconstruction loss of our Ada-GVAE correlates with unfairness as strongly as the disentanglement scores, even though the former can be computed without observing the factors of variation. In particular, we can perform model selection without observing the sensitive variable. In Figure 5 (center), we show that our Ada-GVAE with $k = 1$ and model selection allows us to train and identify fairer models compared to the unsupervised models of Locatello et al. (2019a). Furthermore, their model selection heuristic is based on downstream performance which requires knowledge of the sensitive variable. From both plots we conclude that our weakly-supervised reconstruction loss is a good proxy for unfairness and allows us to train fairer classifiers in the setup of Locatello et al. (2019a) even if the sensitive variable is not observed.

Summary We showed that using weak supervision, we can train and identify fairer classifiers in the sense of demographic parity (Calders et al., 2009; Zliobaite, 2015). As opposed to Locatello et al. (2019a), we do not need to observe the target variable and yet, our principled weakly-supervised approach outperforms their semi-supervised heuristic.

5.4.4. Abstract Visual Reasoning

Finally, we consider the abstract visual reasoning task of van Steenkiste et al. (2019). This task is based on Raven's progressive matrices (Raven, 1941) and requires completing the bottom right missing panel of a sequence of context panels arranged in a 3×3 grid (see Figure 18 (left) in the appendix). The algorithm is presented with six potential answers and needs to choose the correct one. To solve this task, the model has to infer the abstract relationships between the panels. We replicate the experiment of van Steenkiste et al. (2019) on Shapes3D under the same exact experimental conditions (see Appendix B for more details).

In Figure 5 (right), one can see that at low sample sizes, the weakly-supervised reconstruction loss is strongly anti-correlated with performance on the abstract visual reasoning task. As previously observed by van Steenkiste et al. (2019), this benefit only occurs at low sample sizes.

Summary We demonstrated that training a relational network on the representations learned by our Ada-GVAE improves its sample efficiency. This result is in line with the findings of van Steenkiste et al. (2019) where disentanglement was found to correlate positively with improved sample complexity.

6. Conclusion

In this paper, we considered the problem of learning disentangled representations from pairs of non-i.i.d. observations sharing an unknown, random subset of factors of variation. We demonstrated that, under certain technical assumptions, the associated disentangled generative model is identifiable. We extensively discussed the impact of the different supervision modalities, such as the degree of group-level supervision, and studied the impact of the (unknown) number of shared factors. These insights will be particularly useful to practitioners having access to specific domain knowledge. Importantly, we show how to select models with strong performance on a diverse suite of downstream tasks *without using supervised disentanglement metrics*, relying exclusively on weak supervision. This result is of great importance as the community is becoming increasingly interested in the practical benefits of disentangled representations (van Steenkiste et al., 2019; Locatello et al., 2019a; Creager et al., 2019; Chao et al., 2019; Iten et al., 2020; Chartsias et al., 2019; Higgins et al., 2017b). Future work should apply the proposed framework to challenging real-world data sets where the factors of variation are not observed and extend it to an interactive setup involving reinforcement learning.

Acknowledgments: The authors thank Stefan Bauer, Ilya Tolstikhin, Sarah Strauss and Josip Djolonga for helpful discussions and comments. Francesco Locatello is supported by the Max Planck ETH Center for Learning Systems, by an ETH core grant (to Gunnar Rätsch), and by a Google Ph.D. Fellowship. This work was partially done while Francesco Locatello was at Google Research, Brain Team, Zurich.

References

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

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010.

Bengio, Y. The consciousness prior. *arXiv:1709.08568*, 2017.

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. *arXiv:1901.10912*, 2019.

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

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

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops*, pp. 13–18, 2009.

Chao, M. A., Kulkarni, C., Goebel, K., and Fink, O. Hybrid deep fault detection and isolation: Combining deep neural networks and system performance models. *arXiv:1908.01529*, 2019.

Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D. E., Dharmakumar, R., and Tsaftaris, S. A. Disentangled representation learning in cardiac image analysis. *Medical Image Analysis*, 58:101535, 2019.

Chen, J. and Batmanghelich, K. Weakly supervised disentanglement by pairwise similarities. In *AAAI Conference on Artificial Intelligence*, 2020.

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.

Cheung, B., Livezey, J. A., Bansal, A. K., and Olshausen, B. A. Discovering hidden factors of variation in deep networks. *arXiv:1412.6583*, 2014.

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

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pp. 1436–1445, 2019.

Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993.

Deng, Z., Navarathna, R., Carr, P., Mandt, S., Yue, Y., Matthews, I., and Mori, G. Factorized variational autoencoders for modeling audience reactions to movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Denton, E. L. and Birodkar, V. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, 2017.

Duan, S., Watters, N., Matthey, L., Burgess, C. P., Lerchner, A., and Higgins, I. A heuristic for unsupervised model selection for variational disentangled representation learning. *arXiv:1905.12614*, 2019.

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

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

Fortuin, V., Hüser, M., Locatello, F., Strathmann, H., and Rätsch, G. Deep self-organization: Interpretable discrete representation learning on time series. In *International Conference on Learning Representations*, 2019.

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*, 2017.

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.

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

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*, 2019.

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*, 2017b.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv:1812.02230*, 2018.

Hochreiter, S. and Schmidhuber, J. Feature extraction through lococode. *Neural Computation*, 11(3):679–714, 1999.

Hosoya, H. Group-based learning of disentangled representations with generalizability for novel contents. In *International Joint Conference on Artificial Intelligence*, pp. 2506–2513, 2019.

Hsieh, J.-T., Liu, B., Huang, D.-A., Fei-Fei, L. F., and Niebles, J. C. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, 2018.

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*, 2017.

Huang, B., Zhang, K., Zhang, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Behind distribution shift: Mining driving forces of changes and causal arrows. In *IEEE International Conference on Data Mining*, pp. 913–918, 2017.

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

Hyvarinen, A. and Morioka, H. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469, 2017.

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.

Iten, R., Metger, T., Wilming, H., Del Rio, L., and Renner, R. Discovering physical concepts with neural networks. *Physical Review Letters*, 124(1):010508, 2020.

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

Ketchen, D. J. and Shook, C. L. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17(6):441–458, 1996.

Khemakhem, I., Kingma, D. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. *arXiv:1907.04809*, 2019.

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

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, 2015.

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

Locatello, F., Vincent, D., Tolstikhin, I., Rätsch, G., Gelly, S., and Schölkopf, B. Competitive training of mixtures of independent deep generative models. In *Workshop at the 6th International Conference on Learning Representations (ICLR)*, 2018.

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*, 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*, 2019b.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variation using few labels. *International Conference on Learning Representations*, 2020.

Mathieu, M. F., Zhao, J. 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*, 2016.

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*, 2017.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. MIT Press, 2017.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Raven, J. C. Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, 19(1):137–150, 1941.

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

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

Ridgeway, K. A survey of inductive biases for factorial representation-learning. *arXiv:1612.05299*, 2016.

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

Santoro, A., Hill, F., Barrett, D., Morcos, A., and Lillicrap, T. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pp. 4477–4486, 2018.

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. Causality for machine learning, 2019. arXiv:1911.10500.

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

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

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

Sorrenson, P., Rother, C., and Köthe, U. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN). *arXiv:2001.04872*, 2020.

Storck, J., Hochreiter, S., and Schmidhuber, J. Reinforcement driven information acquisition in non-deterministic environments. In *International Conference on Artificial Neural Networks*, pp. 159–164, 1995.

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

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. *Learning Disentangled Representations Workshop at NeurIPS*, 2017.

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

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.

Whitney, W. F., Chang, M., Kulkarni, T., and Tenenbaum, J. B. Understanding visual concepts with continuation learning. *arXiv:1602.06822*, 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*, 2015.

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

Zhang, K., Gong, M., and Schölkopf, B. Multi-source domain adaptation: A causal view. In *AAAI Conference on Artificial Intelligence*, pp. 3150–3157, 2015.

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *International Joint Conference on Artificial Intelligence*, pp. 1347–1353, 2017.

Zliobaite, I. On the relation between accuracy and fairness in binary classification. *arXiv:1505.05723*, 2015.