WEAKLY-SUPERVISED GROUP DISENTANGLEMENT USING TOTAL CORRELATION

Linh Tran[1,2,*], **Saeid Asgari Taghanaki** [1], **Amir Hosein Khasahmadi**[1], **Aditya Sanghi**[1]

[1]*Autodesk AI Lab*
[2]*Imperial College London*

ABSTRACT

Learning disentangled representations that uncover factors of variation in data remains an ongoing key challenge in representation learning. Recent concerns about the feasibility of learning disentangled representations in an unsupervised fashion have motivated a shift toward weak supervision. One way to incorporate weak supervision is through *match pairing*, i.e., using observations as pairs that share at least one factor of variation. Existing match pairing approaches only consider the structural constraints with an average approximate posterior over observations of a shared group. We show the limitations of these approaches and propose a novel formulation to enforce disentangled representations of groups through total correlation, which improves overall disentanglement on various image datasets.

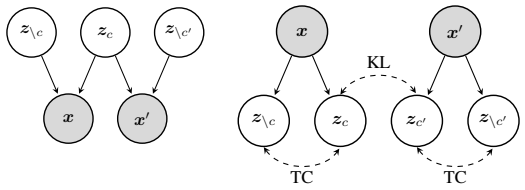
1 INTRODUCTION

Decomposing data into disjoint independent factors of variations and thus learning disentangled representations is essential for interpretable and controllable machine learning Shu et al. (2019). Recent works have shown the usefulness of disentangled representation with respect to abstract reasoning (van Steenkiste et al. (2019)), fairness (Locatello et al. (2019a); Creager et al. (2019)), reinforcement learning (Higgins et al. (2017b)) and general predictive performance (Locatello et al. (2019b)). Even though unsupervised disentanglement methods (Higgins et al. (2017a); Kim & Mnih (2018); Chen et al. (2018)) have shown promising results to learn disentangled representations, Locatello et al. (2019b) showed in a rigorous study that it is impossible to disentangle variations of data without any supervision or inductive bias. Since then, there has been a shift toward weakly supervised disentanglement learning Locatello et al. (2019b), Shu et al. (2019) such as *match pairing* (Shu et al. (2019)) which uses paired observations during optimization. In this work, we present a framework to learn group-disentangled representations using total correlation in a weakly-supervised setting. Our work can be considered learning different levels of weakly-supervised group disentanglement with total correlation. Closely related work is the one of Creager et al. (2019) which proposed to minimize the mutual information between the sensitive latent variable and sensitive labels. Similarly, Klys et al. (2018) proposed to minimize mutual information between the latent variable and a conditional subspace. Both works require either supervised labels or conditions to estimate the mutual information, whereas we only use *weak* supervision for learning disentangled group representations. Locatello et al. (2020) proposed to disentangle groups of variations with only knowing the number of common groups which can be considered as a complementary component to our method. We show that our approach can flexibly disentangle between and within groups of factors of variation. Further, we demonstrate that we improve on disentanglement for various image datasets.

In summary, we make the following contributions:

1. We show limitations of existing group disentanglement approaches (Bouchacourt et al. (2018) and Hosoya (2019)) in terms of latent variable collapse and batch size sensitivity and propose a weakly-supervised way for addressing these weaknesses.
2. We propose a new way of learning disentangled representations from paired observations using total correlation. We also show how to enforce different levels of inter-group and intra-group disentanglement through total correlation.

[*]Corresponding author: linh.tran@autodesk.com



(a) Generative (b) Inference

Figure 1: **The proposed generative and inference model.** Shaded nodes denote observed quantities, and unshaded nodes represent unobserved (latent) variables. Dotted arrows represent either minimizing the TC or the KL divergence between variables.

2 Background

VAEs are latent variable models and aim to learn latent variables $\boldsymbol{z}$ which should capture information about the observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. They are trained to maximize the evidence lower bound (ELBO) given as $\log p(\boldsymbol{x}) \geq \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))$. To be consistent with the works of Bouchacourt et al. (2018) and Hosoya (2019), let us assume that the observations $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are collected in $\mathcal{G}$ distinct groups. Within a group, all observations share some *fixed* factors of variations. Each group $g \in \mathcal{G}$ splits $\mathcal{X}$ into disjoint partitions with arbitrary sizes. For simplicity, we define two groups g_C and $g_{\setminus C}$; where g_C represents information about the actual content whereas $g_{\setminus C}$ represents any variation not contained in C. Given a pair of observations $(\boldsymbol{x}, \boldsymbol{x}')$ which share group factors c, we define two variables $\boldsymbol{z} = (\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c})$ and $\boldsymbol{z}' = (\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c})$ to capture content $(\boldsymbol{z}_c, \boldsymbol{z}_{c'})$ and non-content information $(\boldsymbol{z}_{\setminus c}, \boldsymbol{z}_{\setminus c'})$, e.g. style or background. (Bouchacourt et al. (2018); Hosoya (2019)) learn a group-specific latent variable $\bar{\boldsymbol{z}}_{c,c'}$ by averaging over the corresponding content latent variable $\boldsymbol{z}_c, \boldsymbol{z}_{c'}$ during inference. The modified objective considers the ELBO of paired observations $(\boldsymbol{x}, \boldsymbol{x}')$ i.i.d. sampled from group g_C

$$\begin{aligned}\mathcal{L}_{\text{WS-ELBO}}(\boldsymbol{x}, \boldsymbol{x}'; \theta, \phi) = \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}|\bar{\boldsymbol{z}}_{c,c'}, \boldsymbol{z}_{\setminus c})] + \mathbb{E}_{q_\phi}[\log p_\theta(\boldsymbol{x}'|\bar{\boldsymbol{z}}_{c,c'}, \boldsymbol{z}_{\setminus c'})] \\ - \beta \mathrm{D}_{\mathrm{KL}}(q_\phi(\bar{\boldsymbol{z}}_{c,c'}, \boldsymbol{z}_{\setminus c}|\boldsymbol{x}) \parallel p(\boldsymbol{z})) - \beta \mathrm{D}_{\mathrm{KL}}(q_\phi(\bar{\boldsymbol{z}}_{c,c'}, \boldsymbol{z}_{\setminus c'}|\boldsymbol{x}') \parallel p(\boldsymbol{z})),\end{aligned} \tag{1}$$

where $\bar{\boldsymbol{z}}_{c,c'}$ is sampled from either a Normal distribution over the average of learned means and covariances (Hosoya (2019)) or a product of Normal distributions (Bouchacourt et al. (2018)).

3 Learning Group Similarities using Total Correlation

Considering the setting in Section 2, we would like 1) z_c to be highly correlated with group C and $z_{\setminus c}$ to be highly correlated with group $\setminus C$ and 2) $\boldsymbol{z}_c \approx \boldsymbol{z}_{c'}$ if the paired observations share the same content c or $\boldsymbol{z}_{\setminus c} \approx \boldsymbol{z}_{\setminus c'}$ if the paired observations share the same non-content $\setminus c$. In what follows, we describe our approach with the generative and inference model visualized in Figure 1. Although existing works showed promising results, in practice, minimizing the objective in (1) will not necessarily fulfill the first requirement, i.e., the model will learn representations z_c and $z_{\setminus c}$ that are uncorrelated with each other. Therefore, along with maximizing the variational lower bound, we propose to minimize the total correlation between latent variables z_c and $z_{\setminus c}$.

The total correlation of $\boldsymbol{z}_c$ and $\boldsymbol{z}_{\setminus c}$ is defined as

$$\mathrm{TC}(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c}) = \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c}) \parallel \bar{q}(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c})) \tag{2}$$

where $\bar{q}(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c})$ denotes the desired factorization of the aggregated posterior $q(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c})$. With different factorization, we can enforce different levels of disentanglement:

1. Inter-group disentanglement: $[\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c}]$ is said to be disentangled if its aggregate posterior factorizes as $\bar{q}(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c}) = q(\boldsymbol{z}_c) \cdot q(\boldsymbol{z}_{\setminus c})$. Note that under this disentanglement criteria, each $\boldsymbol{z}_c$ and $\boldsymbol{z}_{\setminus c}$ can be correlated among themselves. However, they must be independent of each other.

(a) Inter-group disentanglement ($q(z_c) \cdot q(z_{\setminus c})$) (b) Inter- and intra-group disentanglement ($q(z_c) \cdot \prod_j q(z_{\setminus c,i})$) (c) Total disentanglement ($\prod_i q(z_{c,j}) \cdot \prod_j q(z_{\setminus c,i})$)

Figure 2: **Different factorizations encourage different levels of disentanglement.** The latent variable $\boldsymbol{z}_c$ representing content information is visualized as red tiles whereas $\boldsymbol{z}_{\setminus c}$ representing non-content information is visualized as blue tiles. The different shades of colors represents correlation to different factors of variation.

2. Inter-group disentanglement and intra-group disentanglement of one group: $[z_c, \boldsymbol{z}_{\setminus c}]$ and $\boldsymbol{z}_{\setminus c}$ are disentangled if the aggregate posterior factorizes as $q(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c}) = q(\boldsymbol{z}_c) \cdot \prod_j q(\boldsymbol{z}_{\setminus c,i})$. With this criteria, $\boldsymbol{z}_c$ is still free to co-vary together, but must be independent from all $\boldsymbol{z}_{\setminus c,i}$. Further each dimension of $\boldsymbol{z}_{\setminus c}$ is disentangled. This kind of disentanglement was also used by Creager et al. (2019).
3. Inter-group disentanglement and intra-group disentanglement of all groups: We can enforce total disentanglement if the aggregate posterior factorizes as $q(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c}) = \prod_i q(\boldsymbol{z}_{c,j}) \cdot \prod_j q(\boldsymbol{z}_{s,i})$. This is equivalent to disentanglement achieved in the FactorVAE objective Kim & Mnih (2018).

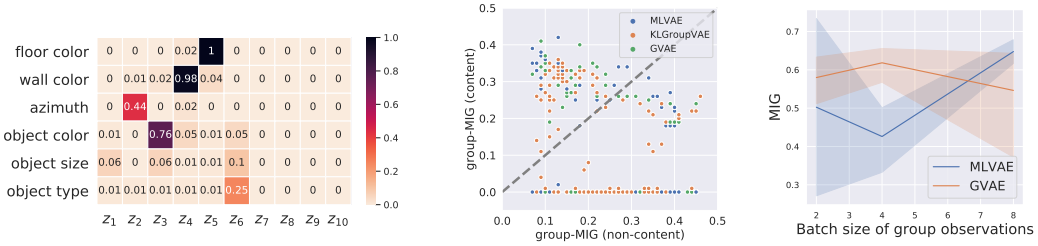
A visualization of these levels can be found in Figure 2. These types of disentanglement are useful for cases where only high-level labels are available, e.g., content vs. style, and only some groups can be further disentangled. Existing works have addressed the second requirement for group disentanglement ("shared observations have approx. same corresponding group latent variable") by using some average over the content latent variable. However, this estimate is biased and requires a certain amount of observations that share the same factors. This might be difficult with sparse, incomplete, and small datasets. We propose a KL-based regularization between the group latent variables of paired observations. This has an analytical form when the latent variables are Normal distributed and does not have any batch-size dependency. Using the total correlation (TC) and a KL term on the group latent variable, the objective becomes

$$\begin{aligned}\mathcal{L} = \sum_{\tilde{\boldsymbol{x}}=\boldsymbol{x},\boldsymbol{x}'} \Bigg(\underbrace{\mathbb{E}_{q_\phi}[\log p_\theta(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{z}}_c, \tilde{\boldsymbol{z}}_{\setminus c})]}_{\text{reconstruction}} - \underbrace{\mathrm{D}_{\mathrm{KL}}(q_\phi(\tilde{\boldsymbol{z}}_c, \tilde{\boldsymbol{z}}_{\setminus c}|\tilde{\boldsymbol{x}}) \parallel p(\boldsymbol{z}))}_{\text{KL between approx. posterior and prior}} \Bigg) \\ - \sum_{\tilde{\boldsymbol{z}}=\boldsymbol{z},\boldsymbol{z}'} \underbrace{\beta \cdot \mathrm{TC}(\tilde{\boldsymbol{z}}_c, \tilde{\boldsymbol{z}}_{\setminus c})}_{\text{total correlation}} - \underbrace{\gamma \cdot \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{z}_g) \parallel q(\boldsymbol{z}_{g'}))}_{\text{KL between shared group latent variables}} ,\end{aligned} \tag{3}$$

where $g \in \{c, \setminus c\}$ is the group which is being shared by the paired observations $(\boldsymbol{x}, \boldsymbol{x}')$. For evaluation, we used a binary adversary which approximates the log density ratio (Kim & Mnih (2018)) to estimate the total correlation loss. We use an adversarial network which attempts to classify between "true" samples from the aggregate posterior $q(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c})$ and "fake" samples from the product of the marginals $\bar{q}(\boldsymbol{z}_c, \boldsymbol{z}_{\setminus c})$. The latent variables are independent from each other if the samples are indistinguishable and the adversary cannot do it better than random chance.

4 Evaluation

Following the experimental setup in (Locatello et al. (2019b); Chen et al. (2018)), we treated learning disentangled representations as a statistical problem instead of empirical risk minimization and hence, did not use the separate train and test sets. For evaluation, we used two datasets, namely, 3DShapes (Burgess & Kim (2018)) and dSprites (Matthey et al. (2017)). We compare our model, *group-tcVAE*, with MLVAE, Bouchacourt et al. (2018), and GVAE, Hosoya (2019). Locatello et al. (2020) have already shown that both works by Bouchacourt et al. (2018) and Hosoya (2019) are superior to the unsupervised disentanglement approaches, hence, we do not compare with them. We quantitatively compare the strength of disentanglement with the Mutual Information Gap (MIG) (Chen et al. (2018)). Further, we introduce *group-MIG*, a metric based on MIG, which quantitatively estimates the mutual information between groups and corresponding latent variables. Formally, we define group-MIG as $\frac{1}{K} \sum_{k=1}^{K} \frac{1}{H(v_k)} (\max I(z_{i=f_g(v_k)}; v_k) - \max I(z_{i \neq f_g(v_k)}; v_k))$ where K is the

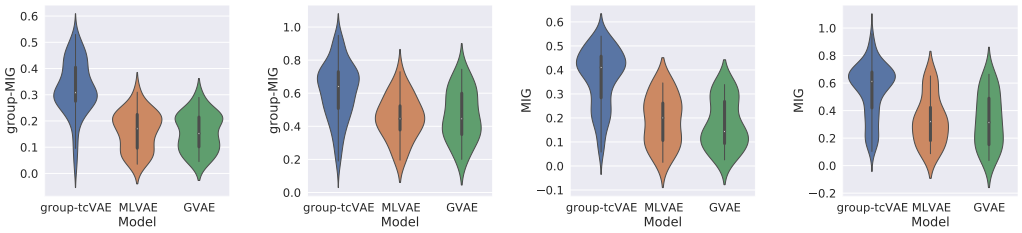


(a) 3DShapes: MI between latent dimensions and factors of variation of trained GVAE model with MIG $= 0.55$ and group-MIG $= 0.44$.

(b) dSprites: group-MIG of content and non-content information for all hyperparameter runs for MLVAE and GVAE.

(c) dSprites: MIG w.r.t. different number of shared observations for MLVAE and GVAE.

Figure 3: **Collapse and sensitity of existing weakly-supervised disentanglement models.** In all the sub-figures higher is better.



(a) dSprites (b) 3DShapes (c) dSprites (d) 3DShapes

Figure 4: **Comparisons between group-tcVAE and comparisons.** Density plots of group-MIG and MIG for group-tcVAE, MLVAE and GVAE) over all runs (*higher is better*).

number of known factors, v_k is the ground truth factor, $f_g(v_k) \in \{c, \setminus c\}$ returns the group that the factor belongs to and $I(z; v_k)$ is an empirical estimate of mutual information between continuous variable z and v_k.

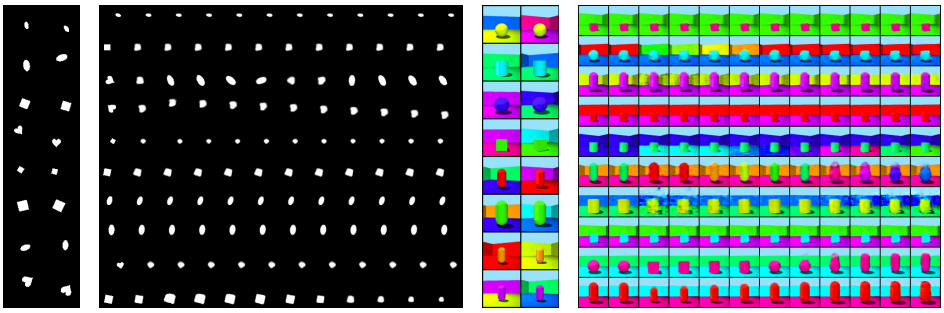
4.1 Empirical analysis of existing weakly-supervised methods

Collapse of content latent variable. The latent variable $\boldsymbol{z}_c$ can either collapse to a single factor of variation or it might even contain almost no information to any content. We visualize such behavior in Figure 3 (a) on a GVAE model trained on 3DShapes with two groups of variations $c =$\{object color, object size and object type\} and $\setminus c =$\{floor color, wall color, azimuth\}. Ideally, $\boldsymbol{z}_1 - \boldsymbol{z}_5$ contains high mutual information with group factors $\setminus c$ and $\boldsymbol{z}_6 - \boldsymbol{z}_{10}$ contains high mutual information with group factors c. However, most information is captured in $\boldsymbol{z}_1 - \boldsymbol{z}_5$, whereas only a little information about object type is contained in $\boldsymbol{z}_6$. As shown in Figure 3 (b), we make similar observations with dataset dSprites in which both MLVAE and GVAE fail to capture content-specific information in the corresponding latent variable.

Sensitivity to group batch size. In practice, always having a certain number of observations that share the same group variations might be difficult, which is a requirement for MLVAE and GVAE with dSprites. This results in performance degradation and high variance (Figure 3(c)).

4.2 Weakly-supervised disentanglement

We perform an extensive evaluation on group-tcVAE to assess its performance in comparison to MLVAE and GVAE. For MLVAE and GVAE, we experimented with the hyperparameters as in Locatello et al. (2020). For group-tcVAE, we used the hyperparameter ranges $\beta = [10, 20, 30, 40, 50, 100]$, $\lambda = [1, 8, 16, 32, 64]$ and a batch size of 32 paired observations ($= 32 \times 2$). For all models, we performed five runs with different random seeds. We plotted all results in Figure 4. For both group-MIG



(a) (b) (c) (d)

Figure 5: **Qualitative results of group-tcVAE.** Training samples (a, c) and reconstruction as well as interpolations (b, d) for dSprites (a, b) and 3DShapes (c, d). Each row represents a pair of observations in (a, c). For the interpolations, each i-th row represents interpolating over only the i-th dimension of $\boldsymbol{z}$.

and MIG, group-tcVAE outperforms MLVAE and GVAE w.r.t. average and best MIG and group-MIG. With dSprites, group-tcVAE almost doubles the average and best performance, whereas with 3DShapes we observe an increase of at least 10% w.r.t. group-MIG and MIG. For the best performing models, we also plotted training samples and qualitative results in Figure 5.

5 Conclusion

We have analyzed existing weakly-supervised disentanglement models and identified challenges w.r.t. latent variable collapse and batch size sensitivity. We proposed a new framework based on total correlation for weakly-supervised disentanglement and showed through empirical evaluations on image datasets that our model improves learning disentangled representations. For future work, we plan to apply our proposed framework to challenging real-world data sets and non-image domains and extend it to semi weakly-supervised and active learning settings.

References

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in neural information processing systems*, pp. 2610–2620, 2018.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017a.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loïc Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: improving zero-shot transfer in reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1480–1490. PMLR, 2017b.

Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, pp. 2506–2513, 2019.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Jack Klys, Jake Snell, and Richard S. Zemel. Learning latent subspaces in variational autoencoders. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 6445–6455, 2018.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pp. 14611–14624, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019b.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14222–14235, 2019.