



Learning Disentangled Discrete Representations

David Friede¹(✉), Christian Reimers², Heiner Stuckenschmidt¹,
and Mathias Niepert^{3,4}

¹ University of Mannheim, Mannheim, Germany
{david,heiner}@informatik.uni-mannheim.de

² Max Planck Institute for Biogeochemistry, Jena, Germany
creimers@bgc-jena.mpg.de

³ University of Stuttgart, Stuttgart, Germany
mathias.niepert@simtech.uni-stuttgart.de

⁴ NEC Laboratories Europe, Heidelberg, Germany

Abstract. Recent successes in image generation, model-based reinforcement learning, and text-to-image generation have demonstrated the empirical advantages of discrete latent representations, although the reasons behind their benefits remain unclear. We explore the relationship between discrete latent spaces and disentangled representations by replacing the standard Gaussian variational autoencoder (VAE) with a tailored categorical variational autoencoder. We show that the underlying grid structure of categorical distributions mitigates the problem of rotational invariance associated with multivariate Gaussian distributions, acting as an efficient inductive prior for disentangled representations. We provide both analytical and empirical findings that demonstrate the advantages of discrete VAEs for learning disentangled representations. Furthermore, we introduce the first unsupervised model selection strategy that favors disentangled representations.

Keywords: Categorical VAE · Disentanglement

1 Introduction

Discrete variational autoencoders based on categorical distributions [17, 28] or vector quantization [45] have enabled recent success in large-scale image generation [34, 45], model-based reinforcement learning [13, 14, 31], and perhaps most notably, in text-to-image generation models like Dall-E [33] and Stable Diffusion [37]. Prior work has argued that discrete representations are a natural fit for complex reasoning or planning [17, 31, 33] and has shown empirically that a discrete latent space yields better generalization behavior [10, 13, 37]. Hafner et al. [13] hypothesize that the sparsity enforced by a vector of discrete latent variables could encourage generalization behavior. However, they admit that “we do not know the reason why the categorical variables are beneficial.”

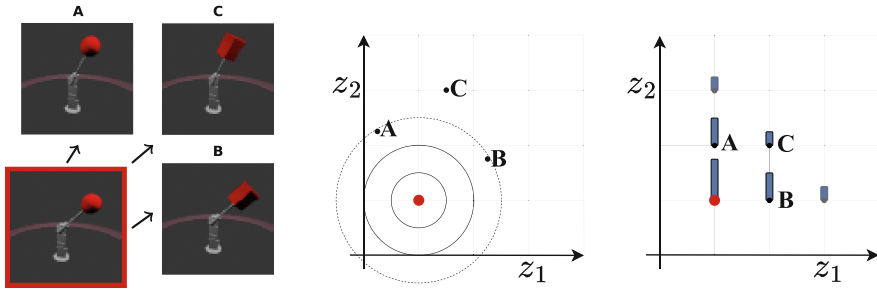


Fig. 1. Four observations and their latent representation with a Gaussian and discrete VAE. Both VAEs encourage similar inputs to be placed close to each other in latent space. **Left:** Four examples from the MPI3D dataset [11]. The horizontal axis depicts the object’s shape, and the vertical axis depicts the angle of the arm. **Middle:** A 2-dimensional latent space of a Gaussian VAE representing the four examples. Distances in the Gaussian latent space are related to the Euclidean distance. **Right:** A categorical latent space augmented with an order of the categories representing the same examples. The grid structure of the discrete latent space makes it more robust against rotations constituting a stronger inductive prior for disentanglement.

We focus on an extensive study of the *structural impact* of discrete representations on the latent space. The disentanglement literature [3, 15, 25] provides a common approach to analyzing the structure of latent spaces. Disentangled representations [3] recover the low-dimensional and independent ground-truth factors of variation of high-dimensional observations. Such representations promise interpretability [1, 15], fairness [7, 24, 42], and better sample complexity for learning [3, 32, 38, 46]. State-of-the-art unsupervised disentanglement methods enrich *Gaussian* variational autoencoders [20] with regularizers encouraging disentangling properties [5, 6, 16, 19, 22]. Locatello et al. [25] showed that unsupervised disentanglement without inductive priors is theoretically impossible. Thus, a recent line of work has shifted to weakly-supervised disentanglement [21, 26, 27, 40].

We focus on the impact on disentanglement of replacing the standard variational autoencoder with a slightly tailored *categorical* variational autoencoder [17, 28]. Most disentanglement metrics assume an ordered latent space, which can be traversed and visualized by fixing all but one latent variable [6, 9, 16]. Conventional categorical variational autoencoders lack sortability since there is generally no order between the categories. For direct comparison via established disentanglement metrics, we modify the categorical variational autoencoder to represent each category with a *one-dimensional* representation. While regularization and supervision have been discussed extensively in the disentanglement literature, the variational autoencoder is a component that has mainly remained constant. At the same time, Watters et al. [50] have observed that Gaussian VAEs might suffer from rotations in the latent space, which can harm disentangling properties. We analyze the rotational invariance of multivariate Gaussian distributions in more detail and show that the underlying grid structure of categorical distributions mitigates this problem and acts as an efficient inductive prior for disentangled representations. We first show that the observation from

[5] still holds in the discrete case, in that neighboring points in the data space are encouraged to be also represented close together in the latent space. Second, the categorical latent space is less rotation-prone than its Gaussian counterpart and thus, constitutes a stronger inductive prior for disentanglement as illustrated in Fig. 1. Third, the categorical variational autoencoder admits an unsupervised disentangling score that is correlated with several disentanglement metrics. Hence, to the best of our knowledge, we present the first disentangling model selection based on unsupervised scores.

2 Disentangled Representations

The disentanglement literature is usually premised on the assumption that a high-dimensional observation \mathbf{x} from the data space \mathcal{X} is generated from a low-dimensional latent variable \mathbf{z} whose entries correspond to the dataset’s ground-truth factors of variation such as position, color, or shape [3, 43]. First, the *independent* ground-truth factors are sampled from some distribution $\mathbf{z} \sim p(\mathbf{z}) = \prod p(z_i)$. The observation is then a sample from the conditional probability $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$. The goal of disentanglement learning is to find a representation $r(\mathbf{x})$ such that each ground-truth factor z_i is recovered in one and only one dimension of the representation. The formalism of variational autoencoders [20] enables an estimation of these distributions. Assuming a known prior $p(\mathbf{z})$, we can depict the conditional probability $p_\theta(\mathbf{x}|\mathbf{z})$ as a parameterized probabilistic decoder. In general, the posterior $p_\theta(\mathbf{z}|\mathbf{x})$ is intractable. Thus, we turn to variational inference and approximate the posterior by a parameterized probabilistic encoder $q_\phi(\mathbf{z}|\mathbf{x})$ and minimize the Kullback-Leibler (KL) divergence $D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z}|\mathbf{x}))$. This term, too, is intractable but can be minimized by maximizing the evidence lower bound (ELBO)

$$\mathcal{L}_{\theta, \phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})). \quad (1)$$

State-of-the-art unsupervised disentanglement methods assume a Normal prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ as well as an amortized diagonal Gaussian for the approximated posterior distribution $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}_\phi(\mathbf{x})\mathbf{I})$. They enrich the ELBO with regularizers encouraging disentangling [5, 6, 16, 19, 22] and choose the representation as the mean of the approximated posterior $r(\mathbf{x}) = \boldsymbol{\mu}_\phi(\mathbf{x})$ [25].

Discrete VAE. We propose a variant of the categorical VAE modeling a joint distribution of n *Gumbel-Softmax* random variables [17, 28]. Let n be the dimension of \mathbf{z} , m be the number of categories, $\alpha_i^j \in (0, \infty)$ be the unnormalized probabilities of the categories and $g_i^j \sim \text{Gumbel}(0, 1)$ be i.i.d. samples drawn from the Gumbel distribution for $i \in [n], j \in [m]$. For each dimension $i \in [n]$, we sample a Gumbel-softmax random variable $z_i \sim \text{GS}(\boldsymbol{\alpha}_i)$ over the simplex $\Delta^{m-1} = \{\mathbf{y} \in \mathbb{R}^n \mid y^j \in [0, 1], \sum_{j=1}^m y^j = 1\}$ by setting

$$z_i^j = \frac{\exp(\log \alpha_i^j + g_i^j)}{\sum_{k=1}^m \exp(\log \alpha_i^k + g_i^k)} \quad (2)$$

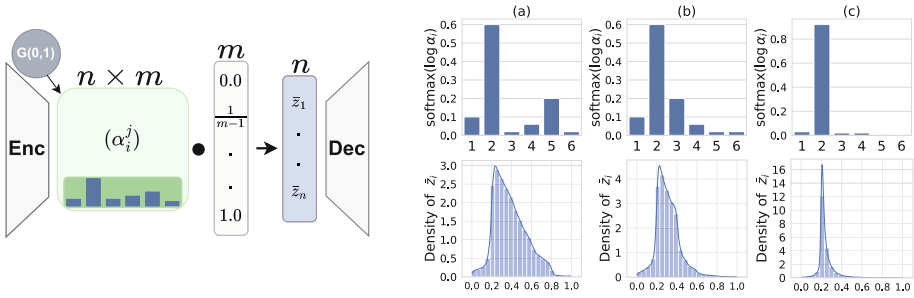


Fig. 2. We utilize n Gumbel-softmax distributions (GS) to approximate the posterior distribution. **Left:** An encoder learns nm parameters α_i^j for the n joint distributions. Each m -dimensional sample is mapped into the one-dimensional unit interval as described in Sect. 3.1. **Right:** Three examples of (normalized) parameters of a single Gumbel-softmax distribution and the corresponding one-dimensional distribution of \bar{z}_i .

for $j \in [m]$. We set the approximated posterior distribution to be a joint distribution of n Gumbel-softmax distributions, i.e., $q_\phi(\mathbf{z}|\mathbf{x}) = \text{GS}^n(\mathbf{z} | \boldsymbol{\alpha}_\phi(\mathbf{x}))$ and assume a joint discrete uniform prior distribution $p(\mathbf{z}) = \mathcal{U}^n\{1, m\}$. Note that \mathbf{z} is of dimension $n \times m$. To obtain the final n -dimensional latent variable $\bar{\mathbf{z}}$, we define a function $f: \Delta^{m-1} \rightarrow [0, 1]$ as the dot product of \mathbf{z}_i with the vector $\mathbf{v}_m = (v_m^1, \dots, v_m^m)$ of m equidistant entries $v_m^j = \frac{j-1}{m-1}$ of the interval¹ $[0, 1]$, i.e.,

$$\bar{z}_i = f(\mathbf{z}_i) = \mathbf{z}_i \cdot \mathbf{v}_m = \frac{1}{m-1} \sum_{j=1}^m j z_i^j \tag{3}$$

as illustrated in Fig. 2. We will show in Sect. 3.2 that this choice of the latent variable $\bar{\mathbf{z}}$ has favorable disentangling properties. The representation is obtained by the standard softmax function $r(\mathbf{x})_i = f(\text{softmax}(\log \boldsymbol{\alpha}_\phi(\mathbf{x})_i))$.

3 Learning Disentangled Discrete Representations

Using a discrete distribution in the latent space is a strong inductive bias for disentanglement. In this section, we introduce some properties of the discrete latent space and compare it to the latent space of a Gaussian VAE. First, we show that mapping the discrete categories into a shared unit interval as in Eq. 3 causes an ordering of the discrete categories and, in turn, enable a definition of neighborhoods in the latent space. Second, we derive that, in the discrete case, neighboring points in the data space are encouraged to be represented close together in the latent space. Third, we show that the categorical latent space is less rotation-prone than its Gaussian counterpart and thus, constituting a stronger inductive prior for disentanglement. Finally, we describe how to select models with better disentanglement using the straight-through gap.

¹ The choice of the unit interval is arbitrary.

3.1 Neighborhoods in the Latent Space

In the Gaussian case, neighboring points in the observable space correspond to neighboring points in the latent space. The ELBO Loss Eq. 1, more precisely the reconstruction loss as part of the ELBO, implies a topology of the observable space. For more details on this topology, see Appendix 2. In the case, where the approximated posterior distribution, $q_\phi(\mathbf{z}|\mathbf{x})$, is Gaussian and the covariance matrix, $\Sigma(\mathbf{x})$, is diagonal, the topology of the latent space can be defined in a similar way: The negative log-probability is the weighted Euclidean distance to the mean $\boldsymbol{\mu}(\mathbf{x})$ of the distribution

$$C - \log q_\phi(\mathbf{z}|\mathbf{x}) = \frac{1}{2} [(\mathbf{z} - \boldsymbol{\mu}(\mathbf{x}))^\top \Sigma(\mathbf{x})(\mathbf{z} - \boldsymbol{\mu}(\mathbf{x}))]^2 = \sum_{i=1}^n \frac{(z_i - \mu_i(\mathbf{x}))^2}{2\sigma_i(\mathbf{x})} \quad (4)$$

where C denotes the logarithm of the normalization factor in the Gaussian density function. Neighboring points in the observable space will be mapped to neighboring points in the latent space to reduce the log-likelihood cost of sampling in the latent space [5].

In the case of categorical latent distributions, the induced topology is not related to the euclidean distance and, hence, it does not encourage that points that are close in the observable space will be mapped to points that are close in the latent space. The problem becomes explicit if we consider a single categorical distribution. In the latent space, neighbourhoods entirely depend on the shared representation of the m classes. The canonical representation maps a class j into the one-hot vector $\mathbf{e}^j = (e_1, e_2, \dots, e_m)$ with $e_k = 1$ for $k = j$ and $e_k = 0$ otherwise. The representation space consists of the m -dimensional units vectors, and all classes have the same pairwise distance between each other.

To overcome this problem, we inherit the canonical order of \mathbb{R} by depicting a 1-dimensional representation space. We consider the representation $\bar{z}_i = f(\mathbf{z}_i)$ from Eq. 3 that maps a class j on the value $\frac{j-1}{m-1}$ inside the unit interval. In this way, we create an ordering on the classes $1 < 2 < \dots < m$ and define the distance between two classes by $d(j, k) = \frac{1}{m-1}|j - k|$. In the following, we discuss properties of a VAE using this representation space.

3.2 Disentangling Properties of the Discrete VAE

In this section, we show that neighboring points in the observable space are represented close together in the latent space and that each data point is represented discretely by a single category j for each dimension $i \in \{1, \dots, n\}$. First, we show that reconstructing under the latent variable $\bar{z}_i = f(\mathbf{z}_i)$ encourages each data point to utilize neighboring categories rather than categories with a larger distance. Second, we discuss how the Gumbel-softmax distribution is encouraged to approximate the discrete categorical distribution. For the Gaussian case, this property was shown by [5]. Here, the ELBO (Eq. 1) depicts an inductive prior that encourages disentanglement by encouraging neighboring points in the data space to be represented close together in the latent space [5]. To show these properties for the D-VAE, we use the following proposition. The proof can be found in Appendix 1.

Proposition 1. Let $\alpha_i \in [0, \infty)^m$, $\mathbf{z}_i \sim \text{GS}(\alpha_i)$ be as in Eq. 2 and $\bar{z}_i = f(\mathbf{z}_i)$ be as in Eq. 3. Define $j_{\min} = \operatorname{argmin}_j \{\alpha_i^j > 0\}$ and $j_{\max} = \operatorname{argmax}_j \{\alpha_i^j > 0\}$. Then it holds that

$$(a) \operatorname{supp}(f) = \left(\frac{j_{\min}}{m-1}, \frac{j_{\max}}{m-1} \right)$$

$$(b) \frac{\alpha_i^j}{\sum_{k=1}^m \alpha_i^k} \rightarrow 1 \Rightarrow \mathbb{P}(z_i^j = 1) = 1 \wedge f(\mathbf{z}_i) = \mathbb{1}_{\left\{ \frac{j}{m-1} \right\}}.$$

Proposition 1 has multiple consequences. First, a class j might have a high density regarding $\bar{z}_i = f(\mathbf{z}_i)$ although $\alpha_i^j \approx 0$. For example, if j is positioned between two other classes with large α_i^k (e.g. $j = 3$ in Fig. 2(a)). Second, if there is a class j such that $\alpha_i^k \approx 0$ for all $k \geq j$ or $k \leq j$, then the density of these classes is also almost zero (Figure 2(a-c)). Note that a small support benefits a small reconstruction loss since it reduces the probability of sampling a wrong class. The probabilities of Fig. 2 (a) and (b) are the same with the only exception that $\alpha_i^3 \leftrightarrow \alpha_i^5$ are swapped. Since the probability distribution in (b) yields a smaller support and consequently a smaller reconstruction loss while the KL divergence is the same for both probabilities,² the model is encouraged to utilize probability (b) over (a). This encourages the representation of similar inputs in neighboring classes rather than classes with a larger distance.

Consequently, we can apply the same argument as in [5] Sect. 4.2 about the connection of the posterior overlap with minimizing the ELBO. Since the posterior overlap is highest between neighboring classes, confusions caused by sampling are more likely in neighboring classes than those with a larger distance. To minimize the penalization of the reconstruction loss caused by these confusions, neighboring points in the data space are encouraged to be represented close together in the latent space. Similar to the Gaussian case [5], we observe an increase in the KL divergence loss during training while the reconstruction loss continually decreases. The probability of sampling confusion and, therefore, the posterior overlap must be reduced as much as possible to reduce the reconstruction loss. Thus, later in training, data points are encouraged to utilize exactly one category while accepting some penalization in the form of KL loss, meaning that $\alpha_i^j / (\sum_{k=1}^m \alpha_i^k) \rightarrow 1$. Consequently, the Gumbel-softmax distribution approximates the discrete categorical distribution, see Proposition 1 (b). An example is shown in Fig. 2(c). This training behavior results in the unique situation in which the latent space approximates a discrete representation while its classes maintain the discussed order and the property of having neighborhoods.

3.3 Structural Advantages of the Discrete VAE

In this section, we demonstrate that the properties discussed in Sect. 3.2 aid disentanglement. So far, we have only considered a single factor \mathbf{z}_i of the approximated posterior $q_\phi(\mathbf{z}|\mathbf{x})$. To understand the disentangling properties regarding the full latent variable \mathbf{z} , we first highlight the differences between the continuous and the discrete approach.

² The KL divergence is invariant under permutation.

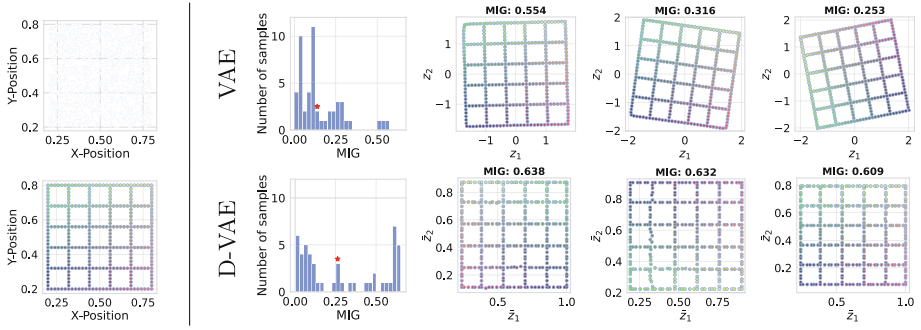


Fig. 3. Geometry analysis of the latent space of the circles experiment [50]. **Col 1, top:** The generative factor distribution of the circles dataset. **Bottom:** A selective grid of points in generative factor space spanning the data distribution. **Col 2:** The Mutual Information Gap (MIG) [6] for 50 Gaussian VAE (top) and a categorical VAE (bottom), respectively. The red star denotes the median value. **Col 3 - 5:** The latent space visualized by the representations of the selective grid of points. We show the best, 5th best, and 10th model determined by the MIG score of the Gaussian VAE (top) and the categorical VAE (bottom), respectively.

In the continuous case, neighboring points in the observable space are represented close together in the latent space. However, this does not imply disentanglement, since the first property is invariant under rotations over \mathbb{R}^n while disentanglement is not. Even when utilizing a diagonal covariance matrix for the approximated posterior $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\sigma}(\mathbf{x})\mathbf{I})$, which, in general, is not invariant under rotation, there are cases where rotations are problematic, as the following proposition shows. We provide the proof in Appendix 1.

Proposition 2 (Rotational Equivariance). *Let $\alpha \in [0, 2\pi)$ and let $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\Sigma = \boldsymbol{\sigma}\mathbf{I}$, $\boldsymbol{\sigma} = (\sigma_0, \dots, \sigma_n)$. If $\sigma_i = \sigma_j$ for some $i \neq j \in [n]$, then \mathbf{z} is equivariant under any i, j -rotation, i.e., $R_{ij}^\alpha \mathbf{z} \stackrel{d}{=} \mathbf{y}$ with $\mathbf{y} \sim \mathcal{N}(R_{ij}^\alpha \boldsymbol{\mu}, \Sigma)$.*

Since, in the Gaussian VAE, the KL-divergence term in Eq. 1 is invariant under rotations, Proposition 2 implies that its latent space can be arbitrarily rotated in dimensions i, j that hold equal variances $\sigma_i = \sigma_j$. Equal variances can occur, for example, when different factors exert a similar influence on the data space, e.g., X-position and Y-position or for factors where high log-likelihood costs of potential confusion causes lead to variances close to zero. In contrast, the discrete latent space is invariant only under rotations that are axially aligned.

We illustrate this with an example in Fig. 3. Here we illustrate the 2-dimensional latent space of a Gaussian VAE model trained on a dataset generated from the two ground-truth factors, X-position and Y-position. We train 50 copies of the model and depicted the best, the 5th best, and the 10th best latent space regarding the Mutual Information Gap (MIG) [6]. All three latent spaces exhibit rotation, while the disentanglement score is strongly correlated

with the angle of the rotation. In the discrete case, the latent space is, according to Proposition 1 (b), a subset of the regular grid \mathbb{G}^n with $\mathbb{G} = \{\frac{j}{m-1}\}_{j=0}^{m-1}$ as illustrated in Fig. 1 (right). Distances and rotations exhibit different geometric properties on \mathbb{G}^n than on \mathbb{R}^n . First, the closest neighbors are axially aligned. Non-aligned points have a distance at least $\sqrt{2}$ times larger. Consequently, representing neighboring points in the data space close together in the latent space encourages disentanglement. Secondly, \mathbb{G}^n is invariant only under exactly those rotations that are axially aligned. Figure 3 (bottom right) illustrates the 2-dimensional latent space of a D-VAE model trained on the same dataset and with the same random seeds as the Gaussian VAE model. Contrary to the Gaussian latent spaces, the discrete latent spaces are sensible of the axes and generally yield better disentanglement scores. The set of all 100 latent spaces is available in Figs. 10 and 11 in Appendix 7.

3.4 The Straight-Through Gap

We have observed that sometimes the models approach local minima, for which \mathbf{z} is not entirely discrete. As per the previous discussion, those models have inferior disentangling properties. We leverage this property by selecting models that yield discrete latent spaces. Similar to the Straight-Through Estimator [4], we round \mathbf{z} off using argmax and measure the difference between the rounded and original ELBO, i.e., $\text{Gap}_{ST}(\mathbf{x}) = |\mathcal{L}_{\theta, \phi}^{ST}(\mathbf{x}) - \mathcal{L}_{\theta, \phi}(\mathbf{x})|$, which equals zero if \mathbf{z} is discrete. Figure 4 (left) illustrates the Spearman rank correlation between Gap_{ST} and various disentangling metrics on different datasets. A smaller Gap_{ST} value indicates high disentangling scores for most datasets and metrics.

4 Related Work

Previous studies have proposed various methods for utilizing discrete latent spaces. The REINFORCE algorithm [51] utilizes the log derivative trick. The Straight-Through estimator [4] back-propagates through hard samples by replacing the threshold function with the identity in the backward pass. Additional prior work employed the nearest neighbor look-up called vector quantization [45] to discretize the latent space. Other approaches use reparameterization tricks [20] that enable the gradient computation by removing the dependence of the density on the input parameters. Maddison et al. [28] and Jang et al. [17] propose the Gumbel-Softmax trick, a continuous reparameterization trick for categorical distributions. Extensions of the Gumbel-Softmax trick discussed control variates [12, 44], the local reparameterization trick [39], or the behavior of multiple sequential discrete components [10]. In this work, we focus on the structural impact of discrete representations on the latent space from the viewpoint of disentanglement.

State-of-the-art unsupervised disentanglement methods enhance Gaussian VAEs with various regularizers that encourage disentangling properties. The β -VAE model [16] introduces a hyperparameter to control the trade-off between the

Table 1. The median MIG scores in % for state-of-the-art unsupervised methods compared to the discrete methods. Results taken from [25] are marked with an asterisk (*). We have re-implemented all other results with the same architecture as in [25] for the sake of fairness. The last row depicts the scores of the models selected by the smallest Gap_{ST} . The 25% and the 75% quantiles can be found in Table 5 in Appendix 7.

Model	dSprites	C-dSprites	SmallNORB	Cars3D	Shapes3D	MPI3D
β -VAE [16]	11.3*	12.5*	20.2*	9.5*	n.a.	n.a.
β -TCVAE [6]	17.6*	14.6*	21.5*	12.0*	n.a.	n.a.
DIP-VAE-I [22]	3.6*	4.7*	16.7*	5.3*	n.a.	n.a.
DIP-VAE-II [22]	6.2*	4.9*	24.1*	4.2*	n.a.	n.a.
AnnealedVAE [5]	7.8*	10.7*	4.6*	6.7*	n.a.	n.a.
FactorVAE [19]	17.4	14.3	25.3	9.0	34.7	11.1
D-VAE	17.4	9.4	19.0	8.5	28.8	12.8
FactorDVAE	21.7	15.5	23.2	14.9	42.4	30.5
Selection	39.5	20.0	22.7	19.1	40.1	32.3

reconstruction loss and the KL-divergence term, promoting disentangled latent representations. The annealedVAE [5] adapts to the β -VAE by annealing the β hyperparameter during training. FactorVAE [19] and β -TCVAE [6] promote independence among latent variables by controlling the total correlation between them. DIP-VAE-I and DIP-VAE-II [22] are two variants that enforce disentangled latent factors by matching the covariance of the aggregated posterior to that of the prior. Previous research has focused on augmenting the standard variational autoencoder with discrete factors [8, 18, 29] to improve disentangling properties. In contrast, our goal is to replace the variational autoencoder with a categorical one, treating every ground-truth factor as a discrete representation.

5 Experimental Setup

Methods. The experiments aim to compare the Gaussian VAE with the discrete VAE. We consider the unregularized version and the total correlation penalizing method, VAE, D-VAE, FactorVAE [19] and FactorDVAE a version of FactorVAE for the D-VAE. We provide a detailed discussion of FactorDVAE in Appendix 3. For the semi-supervised experiments, we augment each loss function with the supervised regularizer R_s as in Appendix 3. For the Gaussian VAE, we choose the BCE and the L_2 loss for R_s , respectively. For the discrete VAE, we select the cross-entropy loss, once without and once with masked attention where we incorporate the knowledge about the number of unique variations. We discuss the corresponding learning objectives in more detail in Appendix 3.

BetaVAE	-13	17	-13	-2	-30	-36	BetaVAE	-20	-17	-38	-36	-53	-67
FactorVAE	-21	17	-3	-11	-25	-24	FactorVAE	-42	-33	-39	-30	-54	-70
MIG	-29	-8	46	-25	-26	-8	MIG	21	51	23	62	32	58
DCI	-19	3	-49	-49	-52	-35	DCI	32	59	39	19	-39	-19
Modularity	-35	-8	-20	-22	-22	-14	Modularity	-62	-76	28	-27	-37	-68
SAP	-4	-23	7	-15	-14	4	SAP	2	59	7	27	-37	33
	(A)	(B)	(C)	(D)	(E)	(F)		(A)	(B)	(C)	(D)	(E)	(F)

Fig. 4. The Spearman rank correlation between various disentanglement metrics and Gap_{ST} (**left**) and the statistical sample efficiency, i.e., the downstream task accuracy based on 100 samples divided by the one on 10 000 samples (**right**) on different datasets: dSprites (A), C-dSprites (B), SmallNORB (C), Cars3D (D), Shapes3D (E), MPI3D (F). **Left:** Correlation to Gap_{ST} indicates the disentanglement skill. **Right:** Only a high MIG score reliably leads to a higher sample efficiency over all six datasets.

Datasets. We consider six commonly used disentanglement datasets which offer explicit access to the ground-truth factors of variation: *dSprites* [16], *C-dSprites* [25], *SmallNORB* [23], *Cars3D* [35], *Shapes3D* [19] and *MPI3D* [11]. We provide a more detailed description of the datasets in Table 8 in Appendix 6.

Metrics. We consider the commonly used disentanglement metrics that have been discussed in detail in [25] to evaluate the representations: *BetaVAE* metric [16], *FactorVAE* metric [19], *Mutual Information Gap* (MIG) [6], *DCI Disentanglement* (DCI) [9], *Modularity* [36] and *SAP score* (SAP) [22]. As illustrated on the right side of Fig. 4, the MIG score seems to be the most reliable indicator of sample efficiency across different datasets. Therefore, we primarily focus on the MIG disentanglement score. We discuss this in more detail in Appendix 4.

Experimental Protocol. We adopt the experimental setup of prior work ([25, 27]) for the unsupervised and for the semi-supervised experiments, respectively. Specifically, we utilize the same neural architecture for all methods so that all differences solely emerge from the distribution of the type of VAE. For the unsupervised case, we run each considered method on each dataset for 50 different random seeds. Since the two unregularized methods do not have any extra hyperparameters, we run them for 300 different random seeds instead. For the semi-supervised case, we consider two numbers (100/1000) of perfectly labeled examples and split the labeled examples (90%/10%) into a training and validation set. We choose 6 values for the correlation penalizing hyperparameter γ and for the semi-supervising hyperparameter ω from Eq. 6 and 7 in Appendix 3, respectively. We present the full implementation details in Appendix 5.

6 Experimental Results

First, we investigate whether a discrete VAE offers advantages over Gaussian VAEs in terms of disentanglement properties, finding that the discrete model generally outperforms its Gaussian counterpart and showing that the FactorD-VAE achieves new state-of-the-art MIG scores on most datasets. Additionally, we

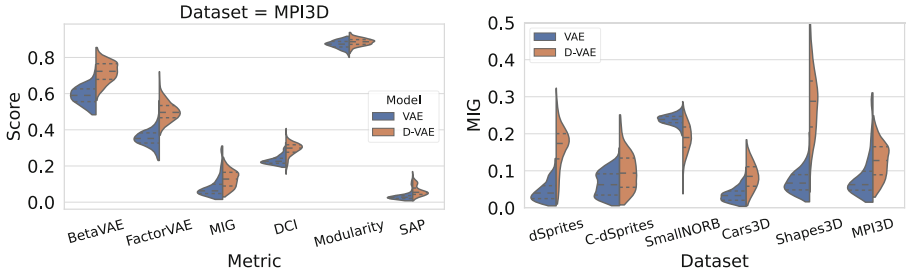


Fig. 5. Comparison between the unregularized Gaussian VAE and the discrete VAE by kernel density estimates of 300 runs, respectively. **Left:** Comparison on the MPI3D dataset w.r.t. the six disentanglement metrics. The discrete model yields a better score for each metric, with median improvements ranging from 2% for Modularity to 104% for MIG. **Right:** Comparison on all six datasets w.r.t. the MIG metric. With the exception of SmallNORB, the discrete VAE yields a better score for all datasets with improvements of the median score ranging from 50% on C-dSprites to 336% on dSprites.

propose a model selection criterion based on Gap_{ST} to find good discrete models solely using unsupervised scores. Lastly, we examine how incorporating label information can further enhance discrete representations. The implementations are in JAX and Haiku and were run on a RTX A6000 GPU.³

6.1 Improvement in Unsupervised Disentanglement Properties

Comparison of the Unregularized Models. In the first experiment, we aim to answer our main research question of whether discrete latent spaces yield structural advantages over their Gaussian counterparts. Figure 5 depicts the comparison regarding the disentanglement scores (left) and the datasets (right). The discrete model achieves a better score on the MPI3D dataset for each metric with median improvements ranging from 2% for Modularity to 104% for MIG. Furthermore, the discrete model yields a better score for all datasets but SmallNORB with median improvements ranging from 50% on C-dSprites to 336% on dSprites. More detailed results can be found in Table 6, Fig. 12, and Fig. 13 in Appendix 7. Taking into account all datasets and metrics, the discrete VAE improves over its Gaussian counterpart in 31 out of 36 cases.

Comparison of the Total Correlation Regularizing Models. For each VAE, we choose the same 6 values of hyperparameter γ for the total correlation penalizing method and train 50 copies, respectively. The right side of Fig. 6 depicts the comparison of FactorVAE and FactorDVAE w.r.t. the MIG metric. The discrete model achieves a better score for all datasets but SmallNORB with median improvements ranging from 8% on C-dSprites to 175% on MPI3D.

³ The implementations and Appendix are at <https://github.com/david-friede/lddr>.

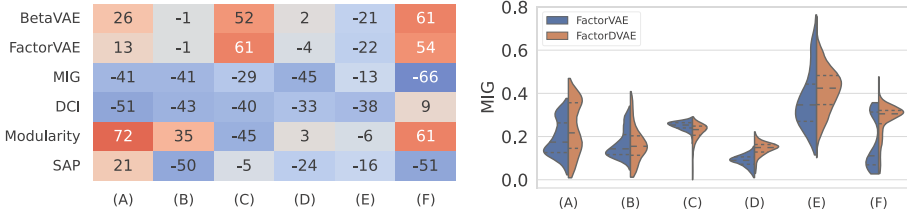


Fig. 6. Disentangling properties of FactorDVAE on different datasets: dSprites (A), C-dSprites (B), SmallNORB (C), Cars3D (D), Shapes3D (E), MPI3D (F). **Left:** The Spearman rank correlation between various disentangling metrics and Gap_{ST} of D-VAE and FactorDVAE combined. A small Gap_{ST} indicates high disentangling scores for most datasets regarding the MIG, DCI, and SAP metrics. **Right:** A comparison of the total correlation regularizing Gaussian and the discrete model w.r.t. the MIG metric. The discrete model yields a better score for all datasets but SmallNORB with median improvements ranging from 8% on C-dSprites to 175% on MPI3D.

6.2 Match State-of-the-Art Unsupervised Disentanglement Methods

Current state-of-the-art unsupervised disentanglement methods enrich Gaussian VAEs with various regularizers encouraging disentangling properties. Table 1 depicts the MIG scores of all methods as reported in [25] utilizing the same architecture as us. FactorDVAE achieves new state-of-the-art MIG scores on all datasets but SmallNORB, improving the previous best scores by over 17% on average. These findings suggest that incorporating results from the disentanglement literature might lead to even stronger models based on discrete representations.

6.3 Unsupervised Selection of Models with Strong Disentanglement

A remaining challenge in the disentanglement literature is selecting the hyperparameters and random seeds that lead to good disentanglement scores [27]. We propose a model selection based on an unsupervised score measuring the discreteness of the latent space utilizing Gap_{ST} from Sect. 3.4. The left side of Fig. 6 depicts the Spearman rank correlation between various disentangling metrics and Gap_{ST} of D-VAE and FactorDVAE combined. Note that the unregularized D-VAE model can be identified as a FactorDVAE model with $\gamma = 0$. A small Straight-Through Gap corresponds to high disentangling scores for most datasets regarding the MIG, DCI, and SAP metrics. This correlation is most vital for the MIG metric. We anticipate finding good hyperparameters by selecting those models yielding the smallest Gap_{ST} . The last row of Table 1 confirms this finding. This model selection yields MIG scores that are, on average, 22% better than the median score and not worse than 6%.

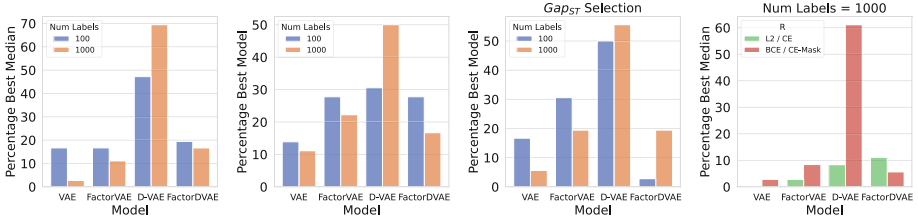


Fig. 7. The percentage of each semi-supervised method being the best over all datasets and disentanglement metrics for different selection methods: median, lowest R_s , lowest Gap_{ST} , median for 1000 labels. The unregularized discrete method outperforms the other methods in semi-supervised disentanglement task. Utilizing the masked regularizer improves over the unmasked one.

6.4 Utilize Label Information to Improve Discrete Representations

Locatello et al. [27] employ the semi-supervised regularizer R_s by including 90% of the label information during training and utilizing the remaining 10% for a model selection. We also experiment with a model selection based on the Gap_{ST} value. Figure 7 depicts the percentage of each semi-supervised method being the best over all datasets and disentanglement metrics. The unregularized discrete method surpasses the other methods on the semi-supervised disentanglement task. The advantage of the discrete models is more significant for the median values than for the model selection. Utilizing Gap_{ST} for selecting the discrete models only partially mitigates this problem. Incorporating the number of unique variations by utilizing the masked regularizer improves the disentangling properties significantly, showcasing another advantage of the discrete latent space. The quantiles of the discrete models can be found in Table 7 in Appendix 7.

6.5 Visualization of the Latent Categories

Prior work uses latent space traversals for qualitative analysis of representations [5, 16, 19, 50]. A latent vector $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ is sampled, and each dimension z_i is traversed while keeping the other dimensions constant. The traversals are then reconstructed and visualized. Unlike the Gaussian case, the D-VAE’s latent space is known beforehand, allowing straightforward traversal along the categories. Knowing the number of unique variations lets us use masked attention to determine the number of each factor’s categories, improving latent space interpretability. Figure 8 illustrates the reconstructions of four random inputs and latent space traversals of the semi-supervised D-VAE utilizing masked attentions. While the reconstructions are easily recognizable, their details can be partially blurry, particularly concerning the object shape. The object color, object size, camera angle, and background color are visually disentangled, and their categories can be selected straightforwardly to create targeted observations.

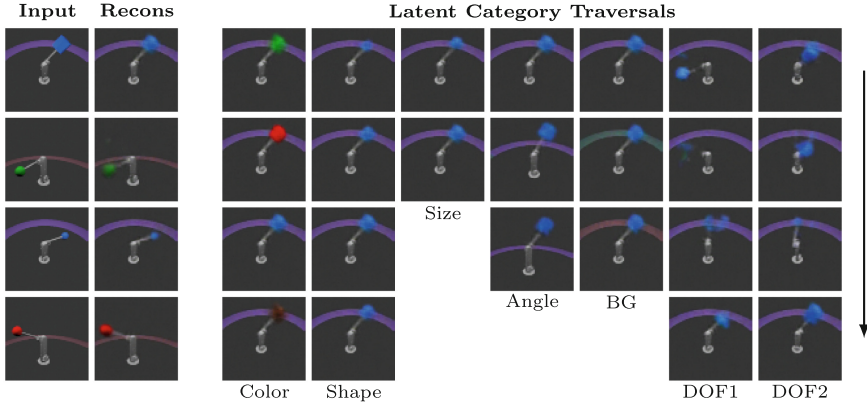


Fig. 8. Reconstructions and latent space traversals of the semi-supervised D-VAE, utilizing masked attentions with the lowest R_s value. The masked attention allows for the incorporation of the number of unique variations, such as two for the object size. We visualize four degrees of freedom (DOF), selected equidistantly from the total of 40. **Left:** The reconstructions are easily recognizable, albeit with blurry details. **Right:** The object color, size, camera angle, and background color (BG) are visually disentangled. The object shape and the DOF factors remain partially entangled.

7 Conclusion

In this study, we investigated the benefits of discrete latent spaces in the context of learning disentangled representations by examining the effects of substituting the standard Gaussian VAE with a categorical VAE. Our findings revealed that the underlying grid structure of categorical distributions mitigates the rotational invariance issue associated with multivariate Gaussian distributions, thus serving as an efficient inductive prior for disentangled representations.

In multiple experiments, we demonstrated that categorical VAEs outperform their Gaussian counterparts in disentanglement. We also determined that the categorical VAE provides an unsupervised score, the Straight-Through Gap, which correlates with some disentanglement metrics, providing, to the best of our knowledge, the first unsupervised model selection score for disentanglement.

However, our study has limitations. We focused on discrete latent spaces, without investigating the impact of vector quantization on disentanglement. Furthermore, the Straight-Through Gap does not show strong correlation with disentanglement scores, affecting model selection accuracy. Additionally, our reconstructions can be somewhat blurry and may lack quality.

Our results offer a promising direction for future research in developing more powerful models with discrete latent spaces. Such future research could incorporate findings from the disentanglement literature and potentially develop novel regularizations tailored to discrete latent spaces.

References

1. Adel, T., Ghahramani, Z., Weller, A.: Discovering interpretable representations for both deep generative and discriminative models. In: International Conference on Machine Learning, pp. 50–59. PMLR (2018)
2. Arcones, M.A., Gine, E.: On the bootstrap of U and V statistics. *Ann. Statist.* **20**, 655–674 (1992)
3. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1798–1828 (2013)
4. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. [arXiv:1308.3432](https://arxiv.org/abs/1308.3432) (2013)
5. Burgess, C.P., et al.: Understanding disentangling in β -vae. [arXiv:1804.03599](https://arxiv.org/abs/1804.03599) (2018)
6. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems, vol. 31, pp. 2615–2625 (2018)
7. Creager, E., et al.: Flexibly fair representation learning by disentanglement. In: International Conference on Machine Learning, pp. 1436–1445. PMLR (2019)
8. Dupont, E.: Learning disentangled joint continuous and discrete representations. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
9. Eastwood, C., Williams, C.K.: A framework for the quantitative evaluation of disentangled representations. In: International Conference on Learning Representations (2018)
10. Friede, D., Niepert, M.: Efficient learning of discrete-continuous computation graphs. In: Advances in Neural Information Processing Systems, vol. 34, pp. 6720–6732 (2021)
11. Gondal, M.W., et al.: On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
12. Grathwohl, W., Choi, D., Wu, Y., Roeder, G., Duvenaud, D.: Backpropagation through the void: optimizing control variates for black-box gradient estimation. In: International Conference on Learning Representations (2018)
13. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering Atari with discrete world models. [arXiv:2010.02193](https://arxiv.org/abs/2010.02193) (2020)
14. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. [arXiv:2301.04104](https://arxiv.org/abs/2301.04104) (2023)
15. Higgins, I., et al.: Towards a definition of disentangled representations. [arXiv:1812.02230](https://arxiv.org/abs/1812.02230) (2018)
16. Higgins, I., et al.: β -VAE: learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017)
17. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with Gumbel-softmax. In: International Conference on Learning Representations (2017)
18. Jeong, Y., Song, H.O.: Learning discrete and continuous factors of data via alternating disentanglement. In: International Conference on Machine Learning, pp. 3091–3099. PMLR (2019)
19. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning, pp. 2649–2658. PMLR (2018)
20. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: International Conference on Learning Representations (2013)

21. Klindt, D.A., et al.: Towards nonlinear disentanglement in natural data with temporal sparse coding. In: International Conference on Learning Representations (2021)
22. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. In: International Conference on Learning Representations (2017)
23. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 2, pp. II-104. IEEE (2004)
24. Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
25. Locatello, F., et al.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: International Conference on Machine Learning, pp. 4114–4124. PMLR (2019)
26. Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., Tschannen, M.: Weakly-supervised disentanglement without compromises. In: International Conference on Machine Learning, pp. 6348–6359. PMLR (2020)
27. Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., Bachem, O.: Disentangling factors of variations using few labels. In: International Conference on Learning Representations (2020)
28. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: a continuous relaxation of discrete random variables. In: International Conference on Learning Representations (2017)
29. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. [arXiv:1511.05644](https://arxiv.org/abs/1511.05644) (2015)
30. Nguyen, X., Wainwright, M.J., Jordan, M.I.: Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **56**(11), 5847–5861 (2010)
31. Ozair, S., Li, Y., Razavi, A., Antonoglou, I., Van Den Oord, A., Vinyals, O.: Vector quantized models for planning. In: International Conference on Machine Learning, pp. 8302–8313. PMLR (2021)
32. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press (2017)
33. Ramesh, A., et al.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. PMLR (2021)
34. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with VQ-VAE-2. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
35. Reed, S.E., Zhang, Y., Zhang, Y., Lee, H.: Deep visual analogy-making. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
36. Ridgeway, K., Mozer, M.C.: Learning deep disentangled embeddings with the F-statistic loss. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. *IEEE 2022*. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685 (2022)
38. Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., Mooij, J.M.: On causal and anticausal learning. In: International Conference on Machine Learning (2012)

39. Shayer, O., Levi, D., Fetaya, E.: Learning discrete weights using the local reparameterization trick. In: International Conference on Learning Representations (2018)
40. Shu, R., Chen, Y., Kumar, A., Ermon, S., Poole, B.: Weakly supervised disentanglement with guarantees. In: International Conference on Learning Representations (2020)
41. Sugiyama, M., Suzuki, T., Kanamori, T.: Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Ann. Inst. Stat. Math.* **64**(5), 1009–1044 (2012)
42. Träuble, F., et al.: On disentangled representations learned from correlated data. In: International Conference on Machine Learning, pp. 10401–10412. PMLR (2021)
43. Tschannen, M., Bachem, O., Lucic, M.: Recent advances in autoencoder-based representation learning. [arXiv:1812.05069](https://arxiv.org/abs/1812.05069) (2018)
44. Tucker, G., Mnih, A., Maddison, C.J., Lawson, J., Sohl-Dickstein, J.: REBAR: low-variance, unbiased gradient estimates for discrete latent variable models. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
45. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
46. Van Steenkiste, S., Locatello, F., Schmidhuber, J., Bachem, O.: Are disentangled representations helpful for abstract visual reasoning? In: Advances in Neural Information Processing Systems, vol. 32 (2019)
47. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
48. Watanabe, S.: Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **4**(1), 66–82 (1960)
49. Watters, N., Matthey, L., Borgeaud, S., Kabra, R., Lerchner, A.: Spriteworld: a flexible, configurable reinforcement learning environment (2019)
50. Watters, N., Matthey, L., Burgess, C.P., Lerchner, A.: Spatial broadcast decoder: a simple architecture for learning disentangled representations in VAEs. [arXiv:1901.07017](https://arxiv.org/abs/1901.07017) (2019)
51. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement Learning*, pp. 5–32 (1992)