DAVA: DISENTANGLING ADVERSARIAL VARIATIONAL AUTOENCODER

Anonymous authors
Paper under double-blind review

ABSTRACT

The use of well-disentangled representations poses many advantages for downstream tasks, e.g. increasing sample efficiency, or enabling interpretability. Their quality is, however, determined to a large extent by the choice of dataset-specific hyperparameters, most notably the regularization strength. To address the issue, we introduce DAVA, a novel training procedure for variational auto-encoders that alleviates the issue of hyperparameter selection at the cost of a comparatively small overhead. We compare DAVA against models with optimal choice of hyperparameters. Without any hyperparameter tuning, DAVA is competitive across a diverse range of commonly used datasets. Underlying DAVA, we discover a necessary condition for unsupervised disentanglement we call *PIPE*. We demonstrate the ability of *PIPE* to positively predict performance of downstream models in abstract reasoning. We also exhaustively examine correlations with existing supervised and unsupervised metrics.

1 INTRODUCTION

Real-world data is generally highly structured, and full with symmetry and patterns. This implies that there exists a lower-dimensional set of ground truth factors that is able to explain a significant part of variation present in real-world data. The goal of disentanglement learning is to recover these factors, so that changes in a single ground truth factor are only reflected in a single latent dimension of a model (see Figure 1 for an example). Such an abstraction allows for more efficient reasoning (Van Steenkiste et al., 2019) and improved interpretability (Higgins et al., 2017a). It further shows positive effects on zero-shot domain adaption (Higgins et al., 2017b) and data efficiency (Duan et al., 2020).

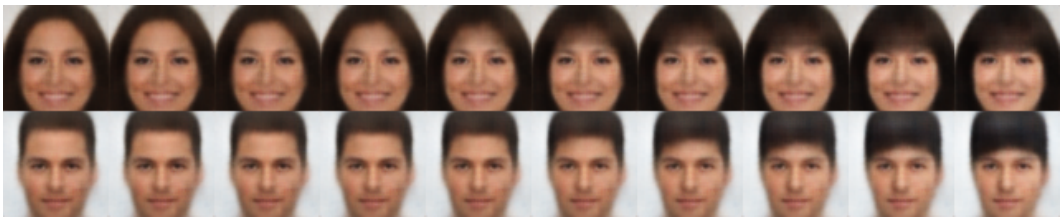


Figure 1: Latent traversals of a single latent dimension (hair fringes) of DAVA trained on *CelebA*. DAVA visibly disentangles the fringes from all other facial properties.

If the generative ground-truth factors are known and labeled data is available, one can train a model in a supervised fashion to extract the ground-truth factors. What if the generative factors are unknown, but one still wants to profit from the aforementioned benefits for a downstream task? This can be necessary if the amount of labeled data for the downstream task is limited or training is computationally very expensive. Learning disentangled representations in an unsupervised fashion is generally impossible without the use of some priors (Locatello et al., 2019b). These priors can be present both implicitly in the model architecture and explicitly in the loss function (Tschannen et al., 2018). An example of such a prior present in the loss function is low total correlation between latent variables of a model (Chen et al., 2018; Kim & Mnih, 2018). Reducing the total correlation has been shown to have a positive effect on disentanglement (Locatello et al., 2019b). Unfortunately, as we

show in more detail in this work, it is highly dataset-specific by how much total correlation should be reduced to achieve good disentanglement. The optimal hyperparameter setting for one dataset might yield bad results on another dataset. To optimize regularization strength, we need a way to evaluate disentanglement quality.

So how can we identify well disentangled representations? The evaluation of representation quality, even given labeled data, is no easy task. Perhaps as an instance of unfortunate nomenclature, the often-used term "ground-truth factor" implies the existence of a canonical set of orthogonal factors. However, often there exist multiple equally valid sets of ground truth factors, for example affine transformations of coordinate axes spanning a space, different color representations, or various levels of abstraction for group properties. This poses a problem for supervised disentanglement metrics as they fix the ground truth factors for evaluating a representation, judging the models too harshly if they learned another equally valid representation. Furthermore, acquiring labeled data in a practical setting is usually a costly endeavor. The previously stated reasons hinder the usability of supervised metrics for model selection.

In this work, we overcome the above mentioned limitations both for learning and evaluating disentangled representations. Our improvements are based on a rather lucid idea: If a VAE learned a truly disentangled representation, a discriminator network should not be able to distinguish reconstructed samples from samples generated by the decoder from a different latent distribution. Depending on the prior of low total correlation, we use a latent distribution where we sample for each dimension independently, calling it the factorial posterior latent distribution. Quantifying the ease of such separation yields a metric independent of specific choice of ground truth factors and reconstruction quality. We show that the similarity of the two distributions is a necessary condition for disentanglement. Further, we can use this property at training time by introducing an adversarial loss into classical training of VAEs. To achieve this, we introduce a discriminator network into training and use the decoder of the VAE as generator. During training we control the weight of the adversarial loss and correspondingly adjust the capacity of the latent space information bottleneck, inspired by (Burgess et al., 2017). By doing so, we allow the model to increase the complexity of its representation as long as it is able to disentangle.

This dynamic training procedure resolves the issue of dataset-specific hyperparameters and allows our approach to reach competitive disentanglement on a diverse set of commonly used datasets without hyperparameter tuning.

Our contributions are as follows:

- We identify a novel unsupervised aspect of disentanglement called *PIPE* and demonstrate its usefulness in a metric with correlation to supervised disentanglement metrics as well as a downstream task.
- We propose an adaptive adversarial training procedure (DAVA) for variational auto-encoders, resolving the common issue of disentanglement performance being largely dependent on dataset-specific regularization strength.
- We provide extensive evaluations on different commonly used disentanglement datasets to support our claims.

2 Related Work

2.1 Model Architectures

The β-VAE by Higgins et al. (2017a) is a cornerstone model architecture for disentanglement learning. The loss function of the β-VAE, the evidence lower bound (ELBO), consists of a reconstruction term and a KL-divergence term weighed by β, forcing the aggregated posterior latent distribution to closely match the prior distribution. The KL-divergence term seems to be encouraging disentanglement as shown in (Rolinek et al., 2019). The β-TCVAE architecture proposed by Chen et al. (2018) decomposes the KL divergence term of the ELBO further into an index-code mutual information, a total correlation and a dimension-wise KL term. They are able to show that it is indeed the total correlation that encourages disentanglement and propose a tractable but biased Monte Carlo estimate. Similar in spirit is the FactorVAE architecture (Kim & Mnih, 2018) using the density ratio trick with an adversarial network to estimate total correlation. The AnnealedVAE architecture (Burgess

et al., 2017) builds on a different premise, arguing that slowly increasing the information bottleneck capacity of the latent space leads to the model learning new latent dimensions step by step and therefore disentangling them. We will use a similar but optimized approach for DAVA.

2.2 Introduction of Adversarial Training to Autoencoders

When combining autoencoders (AEs) with an adversarial setting, one can connect the adversarial network either on the latent space or on the output of the decoder. Makhzani et al. (2016) proposed an adversarial AE (AAAE) that uses an adversarial discriminator network on the latent space of an AE to match its aggregated posterior to an arbitrary prior. There, the encoder of the AE acts as the generator of the GAN and is the only part of the AE that gets updated with respect to the discriminator's loss. This kind of training is strongly connected to VAE training, with the adversarial loss taking on the role of KL divergence in the classical VAE training objective but without the constraint of a gaussian prior. The previosuly mentioned FactorVAE (Kim & Mnih, 2018) also implements an adversarial loss on the latent space to reduce total correlation. Larsen et al. (2016) proposed using a discriminator network with the decoder of the VAE acting as generator to improve the visual fidelity of the generated images, but with no focus on disentanglement. The difference to DAVA is that they used the discriminator on the real observations, while we propose a discriminator that only sees observations generated by the decoder of the VAE. Zhu et al. (2020) introduce an recognition network based loss on the decoder that encourages predictable latent traversals. In other words, given a pair of images where all but one latent dimensions are kept constant, the recognition network should be able to predict in which latent dimension the change occurred. Applied on top of a baseline VAE model, this loss slightly improves the disentanglement performance of the baseline VAE. This is the only work that actively mention an adversarial loss in a VAE based setting to improve disentanglement, even though their work mostly focuses on the GAN setting.

2.3 Measuring Disentanglement

All supervised metrics have in common that they are dependent on a canonical factorization of ground truth factors. Given access to the generative process of a dataset, the FVAE metric (Kim & Mnih, 2018) can be used to evaluate a model. It first creates a batch of data generated by keeping aforementioned ground truth factor fixed and randomly sampling from the other ground truth factors. It then uses a majority vote classifier to predict the index of a ground truth factor given the variance of each latent dimension computed over said batch. Without access to the generative process, but given a number of fully labeled samples, one can use metrics like DCI Disentanglement by Eastwood & Williams (2018) and MIG by Chen et al. (2018). While DCI in essence assesses if a latent variable of a model captures only a single ground truth factor, MIG evaluates if a ground truth factor is captured only by a single variable. As a result, DCI compared to MIG does not punish multi-dimensional representations of a single ground truth factor, for example the RGB model of color or a sine/cosine representation of orientation.

There exists a small number of unsupervised disentanglement metrics. The unsupervised disentanglement ranking (UDR) by Duan et al. (Duan et al., 2020) evaluates disentanglement based on the assumption that a representation should be disentangled if many models, differently initialized but trained with the same hyperparameters, learn the same representation. To achieve this, they compute pairwise similarity (up to permutation and sign inverse) of the representations of a group of models. The score of a single model is the average of its similarity to all the other models in the group. This works relatively well as reflected by positive correlations with supervised metrics and downstream experiments. UDR does not assume any fixed set of ground truth factors. Nevertheless, a weakness of this approach is that it does not recognize similarity of a group of models that each learn a different bijective mapping of the ground truth factors. The latent variation predictability (VP) metric by Zhu et al. (2020) is based on the assumption that if a representation is well disentangled, it should be easy for a recognition network to predict which variable was changed given two input images with a change in only one dimension of the representation. Easiness is quantified by evaluating the accuracy of the recognition network in a few-shot learning problem, where higher accuracy indicates a better disentangled model. An advantage of the VP metric is its ability to evaluate GANs as it is not dependent on the model containing an encoder. In comparison to our proposed metric, both the VP metric and the UDR metric are dependent on the size of the latent space or need to define additional hyperparameters to recognize inactive dimensions. Estermann et al. (2020) showed that the UDR

has a strong positive bias for low-dimensional representations, as low-dimensional representations are more likely to be similar to each other. One could see the same issue arise for the VP metric, as accuracy for the recognition network will likely be higher when there is only a low number of change-inducing latent dimensions available to choose from.

2.4 Applications of Disentangled Representations

The effects of disentangled representations on down-stream task performance are manifold. Van Steenkiste et al. (2019) showed that for abstract reasoning tasks, disentangled representations enabled quicker learning. We show in Section 4.2 that the same holds for our proposed metric. In work by Higgins et al. (2017b), disentangled representations provided improved zero-shot domain adaption for a multi-stage reinforcement-learning (RL) agent. The UDR metric (Duan et al., 2020) correlated well with data efficiency of a model based RL agent introduced by Watters et al. (Watters et al., 2019). Disentangled representations further seem to be helpful in increasing the fairness of down-stream prediction tasks (Locatello et al., 2019a). Contrary to previous qualitative evidence (Eslami et al., 2018; Higgins et al., 2018), Montero et al. (Montero et al., 2020) present quantitative findings indicating that disentanglement does not have an effect on out-of-distribution (OOD) generalization. Recent work by Schott et al. (Schott et al., 2022) supports the claims of Montero et al., concluding that disentanglement shows improvement in downstream task performance but not so in OOD generalization.

3 Posterior Indifference Projection Equivalence

We first introduce the notation of the VAE framework, closely following the notation used by Kim & Mnih (2018). We assume that observations $x \sim \tilde{D}$ are generated by an unknown process based on some independent ground-truth factors. The goal of the encoder is to represent x in a latent vector $z \in \mathbb{R}^d$. We introduce a Gaussian prior $p(z) = \mathcal{N}(0, I)$ with identity covariance matrix on the latent space. The variational posterior for a given observation x is then $q_\theta(z|x) = \prod_{j=1}^{d} \mathcal{N}(z_j|\mu_j(x), \sigma_j^2(x))$, where the encoder with weights θ outputs mean and variance. The decoder with weights ϕ projects from the latent space z back to observation space $p_\phi(x|z)$. We can now define the distribution $q(z)$.

Definition 3.1 (EP)**.** The empirical posterior (EP) distribution $q(z)$ of a VAE is the multivariate distribution of the latent vectors z over the data distribution $\tilde{D}$. More formally, $q(z) = \mathbb{E}_{x\sim\tilde{D}}[q_\theta(z|x)]$. We can reconstruct an observation $x \sim \tilde{D}$ the following way: We sample $z \sim q_\theta(z|x)$ and then get the reconstruction $\hat{x} \sim p_\phi(x|z)$. We informally call observations generated by this process reconstructed samples and denote them as $\hat{x}$.

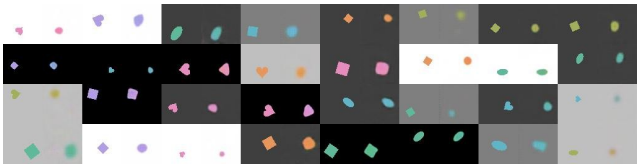
The decoder is not constrained to only project from $q(z)$ to observation space. We can sample observations from the decoder by using different distributions on z. We define a particularly useful distribution.

Definition 3.2 (FP)**.** The factorial posterior (FP) distribution $\bar{q}(z)$ of a VAE is a multivariate distribution with diagonal covariance matrix. We define it as the product of the marginal EP distrbution: $\bar{q}(z) = \prod_{j=1}^{d} q(z_j)$. We can use the decoder to project $z \sim \bar{q}(z)$ to observation space $\tilde{x} \sim p_\phi(x|z)$. We informally call observations created by this process generated samples and denote them as $\tilde{x}$.

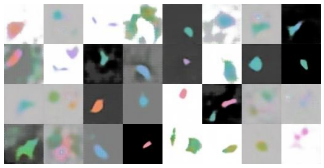
We can now define the data distributions that arise when using the decoder to project images from either the EP or the FP.

Definition 3.3 ($\tilde{D}_{EP}, \tilde{D}_{FP}$)**.** $\tilde{D}_{EP}$ is generated by the decoder projecting observations from the EP latent distribution $q(z)$, i.e. reconstructed samples. $\tilde{D}_{EP} = \mathbb{E}_{z\sim q(z)}[p_\phi(x|z)]$. $\tilde{D}_{FP}$ is generated by the decoder projecting observations from the FP distribution $\bar{q}(z)$, i.e. generated samples. $\tilde{D}_{FP} = \mathbb{E}_{z\sim\bar{q}(z)}[p_\phi(x|z)]$.

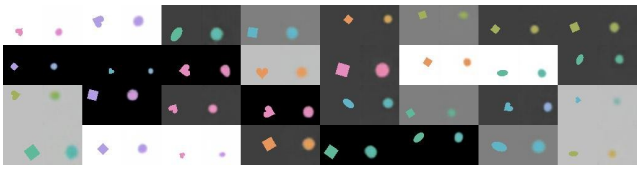
We can now define the core concept of this paper. We look at the similarity of two data distributions generated by the decoder, $\tilde{D}_{EP}$ and $\tilde{D}_{FP}$.



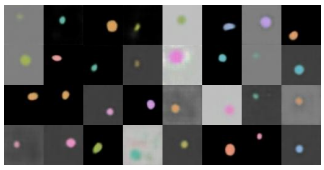
(a) Reconstructed samples.



(b) Generated samples.



(c) Reconstructed samples.



(d) Generated samples.

Figure 2: Displayed are two models from `disentanglement-lib` (Locatello et al., 2019b) with different disentanglement performance. The top part shows a badly disentangled model whereas the model on the bottom is disentangled comparably well. In (a) and (c), odd columns show images sampled from the *AbstractDSprites* dataset, even columns show the respective models reconstruction. (b) and (d) show images generated by the decoder of the respective model given the factorial posterior distribution. While the model on top achieves better reconstructions (a), its generated samples (b) are visibly out of distribution. The bottom model ignores shape in its reconstructions (c), but its generated samples (d) capture the true data distribution more accurately.

Definition 3.4 (PIPE)**.** The posterior indifference projection equivalence (*PIPE*) of a VAE represents the similarity of the data distributions $\tilde{D}_{EP}$ and $\tilde{D}_{FP}$. In other words, *PIPE* is a measure of the decoder's indifference to the latent distribution it projects from.
$PIPE(\theta, \phi) = \omega(\mathbb{E}_{z\sim q(z)}[p_\phi(x|z)], \mathbb{E}_{z\sim \bar{q}(z)}[p_\phi(x|z)]$, where ω is a general similarity measure.

Unsupervised disentanglement learning without any inductive biases is impossible as has been proven by Locatello et al. (2019b). Given a disentangled representation, it is always possible to find a bijective mapping leading to a representation which is entangled. Without knowledge of the ground truth factors, it is impossible to differentiate the two representations. We argue that *PIPE* is necessary for a disentangled model, even when it is not sufficient. Assume that a model has learned a disentangled representation, but $\tilde{D}_{EP}$ is not equivalent to $\tilde{D}_{FP}$. There are two cases in which this could occur. The first option is that the latent dimensions of the model are not independent. This violates the independence assumption of the ground truth factors, consequently the model cannot be disentangled. The second option is that the model has not learned a representation of the ground truth factors and generates samples which are not represented by the ground truth factors. This model would also not be disentangled. Therefore we conclude that *PIPE* is a necessary condition for disentanglement.

In Figure 2 we present reconstructed and generated samples of two different models to support our chain of reasoning. Generated samples of the entangled model (Figure 2 (b)) look visibly out of distribution compared to the reconstructed samples. Generated samples of the disentangled model (Figure 2 (d)) seem to be equivalent to the distribution of reconstructed samples.

4 PIPE Metric

We propose a way to quantify the similarity metric ω of *PIPE* with a neural network and call this the *PIPE* metric. Given a model $\mathcal{M}$ with encoder weights θ, decoder weights ϕ, and corresponding $\tilde{D}_{EP}, \tilde{D}_{FP}$.

1. Create a set of observations S_{EP} by sampling from $\tilde{D}_{EP}$. Informally, these are the reconstructed samples.
2. Create a set of observations S_{FP} by sampling from $\tilde{D}_{FP}$. Informally, these are the generated samples.

3. Randomly divide $S_{EP} \cup S_{FP}$ into a train and a test set.
4. Train a discriminator network on the train set.
5. Evaluate accuracy acc of the discriminator on the test set.
6. Since a random discriminator will guess half of the samples accurately, we report a score of $2 \cdot (1 - acc)$, such that 1 is the best and 0 is the worst score.

We train the discriminator network for $10,000$ steps to keep the distinction of S_{EP} and S_{FP} sufficiently difficult. We further use a uniform factorial distribution instead of FP for a slight improvement in performance. Details on the implementation can be found in Appendix B.

4.1 Relation to Existing Metrics

To classify the performance of the *PIPE* metric, we evaluate correlations with supervised metrics on a diverse set of commonly used datasets. We namely consider *Shapes3D* (Burgess & Kim, 2018), *AbstractDSprites* (Van Steenkiste et al., 2019) and *Mpi3d Toy* (Gondal et al., 2019). All datasets are part of `disentanglement-lib` (Locatello et al., 2019b), which we used to train the models we evaluated our metric on. More details on the implementation of the *PIPE* metric can be found in Appendix B. Results are displayed in Figure 3. They show that the *PIPE* metric correlates positively with existing supervised metrics DCI, MIG and FVAE, surpassing the correlations of the unsupervised baselines. In particular on the challenging dataset *Mpi3d Toy*, it performs much better, keeping a positive correlation. This lets us conclude that *PIPE* indeed captures a new, unsupervised aspect of disentanglement.

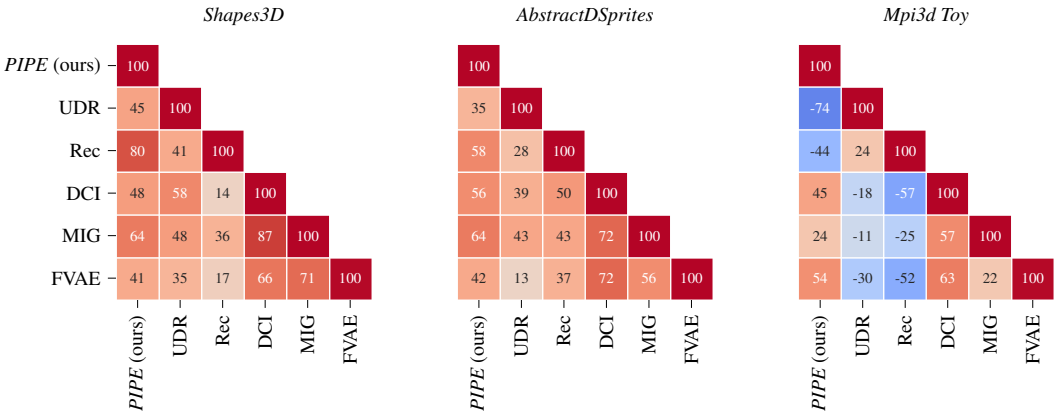


Figure 3: Spearman rank correlation between various metrics on three different datasets. Correlations take values in the range $[-1, 1]$, where 1 means perfect correlation, 0 means no correlation, and negative values mean anti-correlation. For better readability, we multiplied all values by 100. *PIPE* and UDR are the corresponding unsupervised metrics. We include reconstruction loss Rec as a trivial unsupervised baseline. DCI, MIG and FVAE are the corresponding supervised metrics. Low or even negative correlations of our metric with UDR show that our metric captures a different aspect of disentanglement. Correlation of our metric with supervised disentanglement metrics is mostly consistent across datasets. We note that the correlation with supervised metrics on the *Mpi3d Toy* dataset changes direction for both UDR and Rec. The *PIPE* metric seems to capture a more universal aspect of disentanglement and does not suffer from this issue.

4.2 Correlation with Downstream Task Performance

To further quantify the usefulness of the *PIPE* metric, we analyze the predictive performance of the metric for an abstract reasoning downstream task. For the downstream task, a VAE is trained to learn a disentangled representation of a dataset. The downstream model then only gets access to this representation then trying to solve the downstream task. Van Steenkiste et al. (2019) evaluated different supervised disentanglement metrics in terms of predictive performance of accuracy in an

abstract reasoning task on the datasets *Shapes3D* and *AbstractDSprites*. They showed that good disentanglement is an indicator of better accuracy during the early stages on training. We reproduce their experiment on a reduced scale by only considering up to 6,000 training steps. As can be seen in Figure 4, our metric is on par with supervised metrics and clearly outperforms the unsupervised baselines. More importantly this means that *PIPE* is a positive predictor of accuracy of downstream models in a few-sample setting and is therefore a desirable property for unsupervised model selection.

	Shapes3D			*AbstractDSprites*	
	FactorVAE	β-TCVAE		FactorVAE	β-TCVAE
PIPE (ours)	42	83	Unsupervised (For comparison)	63	78
UDR	39	58		25	50
Rec	0	46		38	71
DCI	76	78	Supervised (For context)	81	72
MIG	79	85		77	75
FVAE	69	85		79	50

Figure 4: Spearman rank correlation between various disentanglement metrics and down-stream accuracy of the abstract visual reasoning task (Van Steenkiste et al., 2019) after 6,000 training steps for FactorVAE and β-TCVAE. The same metrics as in Figure 3 are evaluated. Correlation for the unsupervised metrics are more sensitive to model architecture than the supervised ones. *PIPE* shows higher correlation with downstream performance than UDR and clearly outperforms the Rec baseline.

5 Training Procedure for Variational Auto-encoders (DAVA)

In the previous chapter we have established that PIPE is desirable for a model. We now present how to encourage this property at training time by designing an adversarial training procedure. We train a discriminator and a VAE at the same time. The discriminator needs to differentiate $\tilde{D}_{EP}$ from $\tilde{D}_{FP}$, which is achieved with the following loss:

$$\mathcal{L}_{dis} = \mathbb{E}_{\hat{x}\sim\tilde{D}_{EP}}[\log(Dis(\hat{x})] + \mathbb{E}_{\tilde{x}\sim\tilde{D}_{FP}}[\log(1 - Dis(\tilde{x}))].$$

The loss function for the VAE is compromised of a reconstruction term with the weighted negative objective of the discriminator:

$$\mathcal{L}_{adv} = \mathcal{L}_{rec} - \mu\mathcal{L}_{dis}.$$

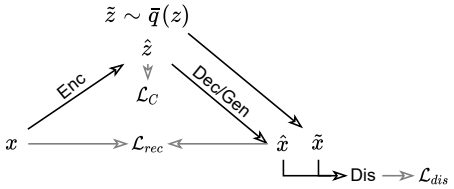


Figure 5: Overview of the composition of DAVA. A discriminator attached to the decoder of a VAE. The goal for the discriminator is to differentiate samples coming from the EP from those coming from the FP.

with reconstruction loss

$$\mathcal{L}_{rec} = \mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)]$$

While this works in practice, it faces the issue of the weight μ being very dataset-specific. It is also closely related to FactorVAE, with the difference that FactorVAE applies the adversarial loss to the latent space and encoder only. Let us consider an alternative approach. Work by Burgess et al. (2017) looked at the KL-Divergence from an information bottleneck perspective. They proposed using a controlled latent capacity increase by introducing a loss term on the deviation of some goal capacity C that increases during training:

$$\mathcal{L}_C = |\mathrm{KL}(q(z|x)||p(z)) - C|$$

Such an loss provides a disentanglement prior complementary to the minimizing total correlation prior used β-TCVAE and FactorVAE. Unfortunately, it too depends on specific choice of hyperparameters, mainly the max capacity C and the speed of capacity increase. We now demonstrate how to

incorporate both the adversarial loss and the controlled capacity increase into a single loss function. We call this approach disentangling adversarial variational auto-encoder DAVA (also see Figure 5):

$$\mathcal{L}_{DAVA} = \mathcal{L}_{rec} - \gamma\mathcal{L}_C - \mu\mathcal{L}_{dis}$$

The main strength of DAVA is its ability to dynamically tune C and μ by using the accuracy of its discriminator. This yields a model that performs well on a diverse range of datasets. We now outline the motivation and the method for tuning C and μ during training.

The goal of the controlled capacity increase in the AnnealedVAE architecture is to allow the model to learn individual factors at a time into individual latent dimensions. The order in which they are learned corresponds to their respective contribution to the reconstruction loss. In DAVA, we want to encourage the model to *learn new factors by increasing C as long as it has learned a disentangled representation of the currently learned factors*. This is the case when the discriminator cannot distinguish $\tilde{D}_{EP}$ from $\tilde{D}_{FP}$. As soon as the accuracy of the discriminator increases, we want to stop increasing C and increase the weight of the adversarial loss to ensure that the model does not learn any new factors of variation while helping it to disentangle the ones it has picked up into individual dimensions. As can be seen in Figure 6, this results in dataset-specific schedules of C over the course of the training. Algorithm 1 in Appendix C.1 describes the training of DAVA in more detail.

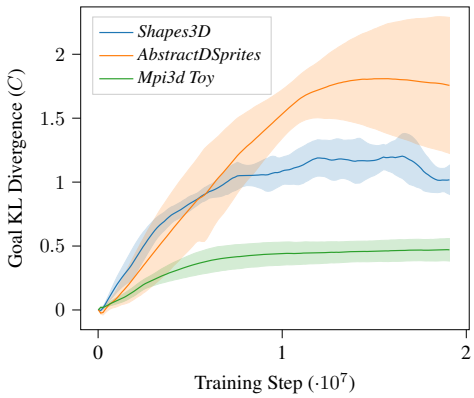


Figure 6: Information bottleneck capacity C during training of DAVA for different datasets. Depending on the complexity and structure of the dataset as well as random initialization of the model, a different schedule of C is necessary to achieve best performance. The shaded area denotes one standard deviation across 5 random seeds.

5.1 Experiments

To quantify the performance of DAVA, we perform an extensive study including multiple model architectures and datasets. We include the best disentangling models according to (Locatello et al., 2019b) as baseline, namely the β-TCVAE (Chen et al., 2018) and FactorVAE (Kim & Mnih, 2018). We further include AnnealedVAE (Burgess et al., 2017), as it relates to DAVA too. As discussed earlier, the performance of each of these approaches is highly dependent on the regularization strength. For this reason, we perform a grid-search over a range of values curated by roughly following the work of Locatello et al. (2019b) and adapted where necessary to ensure best possible performance. The exact hyperparameters considered can be found in appendix C.1. The training of DAVA does not rely on such dataset-specific regularization, therefore **all hyperparameters** for the training of DAVA were kept **constant** across **all** considered datasets. All other hyperparameters closely follow (Locatello et al., 2019b) and were kept consistent across all architectures (baseline and DAVA) and are also reported in appendix C.1. We run 5 different random seeds for each configuration to establish statistical significance. We provide summarized results for the most insightful datasets in Tables 1, 2 and 3.

Table 1: Results *Shapes3D*

Architecture	MIG	DCI	FVAE	PIPE
β-TCVAE	0.26±0.06	0.50±0.03	0.76±0.03	0.11±0.03
FactorVAE	0.20±0.10	0.41±0.07	0.78±0.04	0.22±0.04
AnnealedVAE	0.52±0.02	0.70±0.02	**0.92±0.01**	0.26±0.03
Ours	**0.62±0.05**	**0.78±0.03**	0.82±0.03	**0.61±0.04**

For the baselines we report the performance of the mean hyperparameter setting according to the average score of all considered metrics. The reader is advised to consult Appendix D for complete

Table 2: Results *AbstractDSprites*

Architecture	MIG	DCI	FVAE	PIPE
β-TCVAE	0.12±0.01	0.18±0.01	0.45±0.02	0.19±0.01
FactorVAE	0.12±0.02	0.19±0.02	0.55±0.03	0.20±0.03
AnnealedVAE	0.15±0.01	0.21±0.01	0.54±0.02	0.18±0.01
Ours	**0.23±0.04**	**0.27±0.05**	**0.67±0.05**	**0.35±0.03**

Table 3: Results *Mpi3d Toy*

Architecture	MIG	DCI	FVAE	PIPE
β-TCVAE	0.11±0.02	0.23±0.01	0.39±0.02	0.09±0.01
FactorVAE	0.02±0.01	0.13±0.00	0.38±0.02	(0.99±0.10)
AnnealedVAE	0.07±0.03	0.23±0.02	**0.50±0.02**	0.02±0.01
Ours	**0.12±0.09**	**0.30±0.03**	0.41±0.04	**0.21±0.03**

results of all evaluated datasets in order to get a broader picture of the performance of each approach. Appendix D also contains the results of the best performing hyperparameter choice for each baseline model, demonstrating their variance across datasets. Note that the PIPE scores of FactorVAE in Table 3 are exploiting a design decision discussed in depth in Appendix D. The decoders of these models mapped all latent values to a single image, therefore we exclude their scores. The results showcase the strong consistent performance of DAVA across different datasets. Compared to the baselines, DAVA is able to deliver improvements of up to 50% for the supervised metrics, and over 100% for *PIPE*. This allows DAVA to be used in practice as consistent performance can be expected.

We further validate DAVA on three real-world dataset. CelebA (Liu et al., 2015) is dataset commonly used in the disentanglement literature. It compromises of over 200,000 images of faces of celebrities. As shown in Figure 7, DAVA achieves a competitive FID while still having a comparatively high *PIPE* metric score. This means that DAVA accomplishes a similar visual fidelity of its generated samples as the best baseline methods while better disentangling its latent space. Results for datasets *Mpi3d Real* (Gondal et al., 2019) and *CompCars* (Yang et al., 2015) can be found in Appendix D and E, respectively.

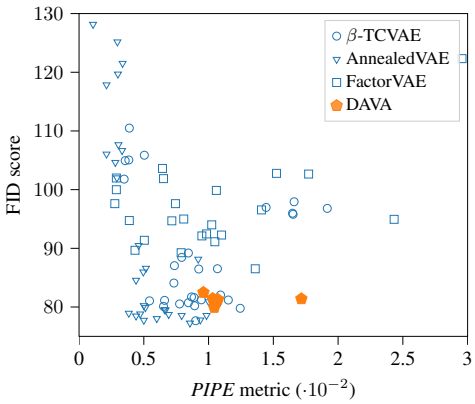


Figure 7: The FID score (Heusel et al., 2017) measures the distance of two different distributions of images and is commonly used in GAN literature to evaluate the quality of the generator. A low FID indicates that a model generates visually compelling images. We evaluate the FID of image samples from $\tilde{D}_{FP}$ of the respective models versus the original image. We note that DAVA achieves a competitive FID with comparatively good *PIPE*. We removed a small number of bad performing outliers of the baseline models to improve readability.

6 Conclusions

In this work, we presented a new way to understand and formalize disentangled representations from an unsupervised point of view. The main idea is that the distribution of reconstructed samples of a VAE should be similar to the distribution of generated samples. We quantify this notion in a new unsupervised disentanglement metric called *PIPE* metric to evaluate disentangled representations. Without access to ground truth, *PIPE* provides results in line with supervised metrics. Furthermore, we apply the same idea to DAVA, a novel VAE training procedure. DAVA can self-tune its regularization strength and produces results competitive with existing models that require more dataset-specific hyperparameter tuning. Robust unsupervised disentanglement performance is an important requirement for applicability in practice. With DAVA, we provide a method that can be used by practitioners with only a relatively small amount of compute needed.

REFERENCES

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

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *Workshop on Learning Disentangled Representations at the 31st Conference on Neural Information Processing Systems*, 2017.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. 2020.

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

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

Benjamin Estermann, Markus Marks, and Mehmet Fatih Yanik. Robust disentanglement of a few factors at a time. *Advances in Neural Information Processing Systems*, 33, 2020.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Irina Higgins, Loic 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. 2017a. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pp. 1480–1490. PMLR, 2017b.

Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bošnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. In *International Conference on Learning Representations*, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pp. 1558–1566. PMLR, 2016.

Yann LeCun, Fu Jie Huang, and Leon Bottou. 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.*, volume 2, pp. II–104. IEEE, 2004.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, 32, 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.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *Workshop track - ICLR*, 2016.

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

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2020.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in neural information processing systems*, 28, 2015.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2019.

L. Schott, J. von Kügelgen, F. Träuble, P. Gehler, C. Russell, M. Bethge, B. Schölkopf, F. Locatello, and W. Brendel. Visual representation learning does not generalize strongly within the same domain. In *10th International Conference on Learning Representations (ICLR)*, April 2022. URL `https://openreview.net/forum?id=9RUHPlladgh`.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. In *Third workshop on Bayesian Deep Learning (NeurIPS 2018)*, 2018. URL `http://www.nari.ee.ethz.ch/pubs/p/autoenc2018`.

Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *Advances in Neural Information Processing Systems*, 32, 2019.

Nicholas Watters, Loic Matthey, Matko Bosnjak, Christopher P Burgess, and Alexander Lerchner. Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration. *arXiv preprint arXiv:1905.09275*, 2019.

Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3973–3981, 2015.

Xinqi Zhu, Chang Xu, and Dacheng Tao. Learning disentangled representations with latent variation predictability. In *European Conference on Computer Vision*, pp. 684–700. Springer, 2020.