INDUCTIVE-BIASES FOR CONTRASTIVE LEARNING OF DISENTANGLED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning disentangled representations is a core machine learning task. It has been shown that this task requires inductive biases. Recent work on class-content disentanglement has shown excellent performance, but required generative modeling of the entire dataset, which can be very demanding. Current discriminative approaches are typically based on adversarial-training and do not reach comparable accuracy. In this paper, we investigate how to transfer the inductive-biases implicit in generative-approaches to contrastive methods. Based on our findings we proposed a new, *non-adversarial* and *non-generative* method named ABCD: Augmentation Based Contrastive Disentanglement. ABCD uses contrastive representation learning relying only on content-invariant augmentations to achieve domain-disentangled representations. The discriminative approach, makes ABCD much faster to train relative to other generative approaches. We evaluate ABCD on image translation and retrieval tasks, and obtain state-of-the-art results.

1 INTRODUCTION

The task of learning domain-invariant representations of data is key in machine learning as it has many important downstream applications. Some of those include: cross-domain matching, synthesizing analogies across domains (image translation), domain adaptation and generalization, learning to make fair decisions etc. The required representations must satisfy two goals: i) invariance: the representation of a sample must not reveal the domain from which it was collected ii) alignment: similar representations should correspond to similar groundtruth hidden attributes. It is intuitive that learning such representations is a discriminative task which does not require generative modeling of the data. For example, while looking at an image of a car, a human can immediately infer its type, camera pose and color, without reconstructing every image pixel. Counter-intuitively, the best performing approaches for learning domain-invariant representations are generative and typically use some form of variational-autoencoder (VAE). These methods reported strong results on multiple benchmark datasets (Gabbay & Hoshen, 2020b; Bouchacourt et al., 2018; Denton et al., 2017). Although discriminative approaches based on adversarial-training have been proposed, the representations that they learn have typically not equalled those of generative approaches. The parameter-sensitivity of adversarial training make such approaches tricky to train, which may explain the performance gap.

We begin with the observation that although generative, VAE-based approaches are guaranteed to learn disentangled representations (under some restrictive conditions), they are not guaranteed to learn aligned representations across different domains. Remarkably, in practice, generative models often learn aligned representations. As this is not enforced by the objective, we hypothesize this is due to inductive biases implicit in generative models. To test our hypothesis, we first determine the inductive biases of generative models. We perform experiments that evaluate the invariance of autoencoder models to different image transformations. Our findings reveal that the list of invariances differs between autoencoders and datasets, some invariances are shared by all models trained on multiple datasets. Particularly, the most preserved invariances are: blur and high contrast and saturation transformations. We therefore hypothesize that these transformations play a key role in the ability of generative models to learn aligned invariant representations.

To test our hypotheses, we adapt contrastive learning for domain-invariant representation learning using the above inductive biases. First, we show that just the denominator of the contrastive ob-

jective is sufficient for learning domain invariant representation by enforcing the representation of every image is far from all other images in the same domain. Unfortunately, it is insufficient for learning aligned representations. To allow domain alignment, we use the inductive biases of the generative models. Specifically, for every image, we learn a representation that is similar to that of its augmented version using the transformations to which autoencoders were invariant. We show the choice of invariance is critical and that using the standard transformations (e.g. SimSiam augmentations) results in poor alignment or poor disentanglement.

We therefore introduce a new approach named ABCD: **A**ugmentation **B**ased **C**ontrastive **D**isentanglement. We find that beyond the modifications to the standard contrastive objective mentioned above, ABCD enjoys the best of both worlds as it is i) non-generative: it does not require reconstruction of every pixel in the training set - which complicates and slows-down the training process. ii) discriminative but non-adversarial: the optimization is simple and does not suffer from the sensitive parameter tuning that plague discriminative, adversarial approaches.

We evaluate our method at various levels: i) direct measurement of the disentanglement and alignment of the learned representation ii) downstream tasks - cross-domain image translation and retrieval. We show that our method learns domain invariant representations that are aligned across domains. When compared to generative approaches, ABCD is faster to train as it does not require training a generator. ABCD is shown to achieve state-of-the-art performance on unsupervised image translation and retrieval tasks.

Our contribution include:

1. Developing an understanding of the inductive biases of generative models responsible for their strong domain alignment performance.
2. A new contrastive method that enjoys the inductive biases of generative models while being non-generative and non-adversarial.
3. An evaluation of our approach both at the representation level and also on downstream tasks.

2 Related Work

Learning class-content disentangled representations. The task of separating between labeled and unlabelled attributes as been extensively researched. The objective is to learn a representation for the unlabelled attributes which is: i) independent of the labeled attributes. ii) informative on the unlabelled attributes. Several methods use adversarial training (Denton et al., 2017; Szabó et al., 2018; Mathieu et al., 2016). Other methods use other non-adversarial approaches, e.g. cycle consistency (Harsh Jha et al., 2018), group accumulation (Bouchacourt et al., 2018) or latent optimization (Gabbay & Hoshen, 2020a; 2021b). All the above methods are generative and require reconstruction of the entire training datasets. Here, we propose a discrminative approach that does not require learning to reconstruct the dataset - which is much faster and less computationally demanding.

Contrastive representation learning. Over the last several years, significant progress in self-supervised representation learning was achieved by methods relying on pairs of augmented samples. Most recent methods use the constraint that the neural representations of different augmentations of the same image should be equal. Non-contrastive methods Chen & He (2020); Grill et al. (2020); Richemond et al. (2020) use the above constraint with various other tricks for learning representations. As the above formulation is prone to collapse, contrastive methods Ye et al. (2019); Hjelm et al. (2019); Wu et al. (2018); van den Oord et al. (2018); Hjelm et al. (2019); He et al. (2020); Chen et al. (2020c); Misra & Maaten (2020); Chen et al. (2020a;b) add an additional uniformity constraint that prohibits collapse of the representation to a single point. Our method adapts the contrastive objective for the task of class-content disentanglement.

Contrastive approaches for disentanglement. Recently, Zimmermann et al. (2021) proposed a seminal approach for contrastive learning of disentangled representations. They tackle the ambitious setting of unsupervised disentanglement, and therefore make strong assumptions on the distribution of the true factors of variation as well as requiring temporal sequences of images at training time. Our method applies to the different (and less ambitious) setting of class-content disentanglement - where we assume class supervision on the training data but do not require image sequences or

making particular assumptions on the evolution of unlabeled true factors. Our technical approaches are consequently very different.

Applications of disentangled representations. Learning disentangled representations has many applications including: controllable image generation (Zhu et al., 2018), image manipulation (Gabbay & Hoshen, 2020b; 2021a; Wu et al., 2021) and domain adaptation (Peng et al., 2019). Furthermore, it is believed that better disentangled representations will have future impact on model interpretability (Hsu et al., 2017), abstract reasoning (van Steenkiste et al., 2019) and fairness (Creager et al., 2019). . In this work, we concentrate on application to cross-domain translation and retrieval.

3 Unraveling the Inductive Biases of Generative Disentanglement Models

We receive as input a set of training samples $x_1, x_2..x_N$. Each training sample x has labelled attributes y and also has unlabelled attributes u which are not correlated to y. In this paper, we assume that the labeled attribute y is a single, categorical variable. The objective is to learn an encoder E, which encodes each image x as code $c = E(x)$. We require the code c to satisfy two requirements: i) *Disentanglement*: there should not exist a function that can predict the labelled attribute y given the representation c, in other words, the representation should not be informative of the labelled attribute. ii) *Alignment*: there should exist a function that can predict u given code c - in other words the representation c should be informative of the unlabelled attributes.

3.1 Disentanglement objectives do not ensure unknown attribute is identifiable

It has been established by Locatello et al. (2019) that any disentanglement method must have some source for inductive bias for the disentanglement to be possible. As the class-content disentanglement setting has labeled examples, it may be hoped this should enable recovery of the unlabeled attributes. Indeed, previous research confirmed that generative models have been empirically successful at learning disentangled representations. In this section, we will argue that standard class-content disentanglement objectives do not provide enough guidance for learning aligned-disentangled representations and therefore that inductive bias is necessary.

Both VAE and GAN-based disentanglement methods, learn a representation c that satisfies two properties: i) the representation is independent of the class, $p(c|y) = p(c)$ ii) there exists a function G, such that $x = G(c, y)$ for every image x. Most methods also force $p(c|y) = N(0, I)$. Although this ensures independence from y, we explain this does not force identifiability of u given c. As a simple demonstration, we will show an unidentifiable case that satisfies the two requirements above. Assume that $p(u) = N(0, I)$ ($u \in \mathbb{R}^d$) and that we learned representations c s.t. $c = u$ for images with $y = 0$ and $c = Pu$ for images with $y = 1$ (where P is a permutation matrix). It is clear that $p(c|y) = p(c)$. Also, as we assume there exists a function G^* s.t. $x = G^*(y, u)$, it is easy to construct a function $x = G(y, c) = G^*(y, (P^y)^T c)$. However, given c and without knowledge of y, it is not possible to recover u (as it may be either c or $P^T c$ depending on the sign of y). This shows that the objective by itself, is insufficient for learning a representation c that has an injective mapping to the unknown attribute u.

3.2 Investigating the Inductive Biases of Generative Models

In this section, we investigate the inductive biases of generators. We only investigate one class of possible inductive biases, invariance of generator to particular image transformations. We propose the following experiment: i) train an autoencoder AE on an image dataset without any augmentations s.t. $\min_{AE} \sum_{x \in \mathcal{X}} \|x - AE(x)\|^2$, where $\mathcal{X}$ is the training set. ii) transform the original images from the test set of the dataset with a range of image augmentations T iii) evaluate the invariance of the outputs of the autoencoder. We use the following two invariance metrics f_{unnorm}, f_{norm} for evaluating how much the distance between the original and transformed images change when evaluated on autoencoder outputs.

$$f_{unnorm} = dist(AE(x), AE(f(x))) \tag{1}$$

Table 1: An evaluation of the invariance of autoencoders to different image transformations

	Average	
	f_{norm}	f_{unnorm}
Horizontal Flip	0.868	0.2489
Vertical Flip	0.791	0.3125
Low Contrast	1.197	0.1503
Low Brightness	0.759	0.1666
Color Rotation	0.876	0.1020
Random Erase []	0.796	0.1967
Affine Transformation	**0.554**	0.2838
GrayScale	0.787	**0.0570**
Crop	0.778	0.2769
High Brightness	0.806	**0.0759**
High Contrast	**0.433**	**0.0572**
High Saturation	**0.579**	**0.0261**
Gaussian Blurring	**0.285**	**0.0816**

$$f_{norm} = \frac{dist(AE(x), AE(f(x)))}{dist(x, f(x))} \tag{2}$$

We use the perceptual loss as the distance function. If an autoencoder is invariant to a particular transformation, both metrics should be be small. The normalized metric is sensitive to smaller transformation, and the unnormalized metric is sensitive to larger transformation.

We conducted the experiment on three datasets: Cars3D Krause et al. (2013), CelebA Liu et al. (2015) and Edges2Shoes (shoes only) Yu & Grauman (2014). The 14 augmentations from the TorchVision library were evaluated. The full results are presented in the appendix. Here, we present the metrics averaged over the three datasets. We observe that autoencoders are highly invariant to blur, high-saturation and high-contrast. They are mostly equivariant to horizontal flipping, and color changes. As these are the inductive biases of generative methods, it suggests that providing these biases to discriminative methods can potentially transfer some of the attractive qualities of generative methods.

4 ABCD: A Contrastive Method for Representation Disentanglement

In this section we introduce ABCD, a new, discriminative approach for class-content disentanglement.

As explained in Sec. 3, disentanglement methods learn representations c that are disentangled from the labeled attribute y s.t. $p(c|y) = p(c)$. Although typically adversarial or VAE objectives are used, here we propose to use a contrastive objective. It was shown by Wang & Isola (2020) that the denominator of the contrastive objective $\sum_j -\log(\sum_i \mathbf{1}_{i \neq j} e^{sim(E(x_i), E(x_j))})$ encourages the learned feature space of the encoder E to be uniformly distributed on the unit sphere. We propose to use this objective to learn an encoder E that learns a disentangled representation c for an image x. The key is to apply the contrastive objective for the images of each class y separately (but share the same encoder for all classes) - this ensures representations c of each class y are distributed uniformly on the unit sphere. As $p(c|y)$ are equal for all values of y, we have $p(c|y) = p(c)$ and c is independent of the class. Additionally, as for each image in the training set there exists a unique combination of c and class y, it is possible in-principle to construct a function such that $x = G(y, c)$. The representations learned in this fashion therefore satisfy standard disentanglement objectives.

We conduct an experiment to evaluate the learned representations of the SmallNORB datasets, where the labeled attribute y is the object type while the unlabeled attribute u is the object pose. After learning the encoder E, we compute the representation $c = E(x)$ for every image x. We compute a deep classifier that attempts to predict u from c and another that attempts to predict y from c. The results are presented in Tab. 4. This shows that although the learned representations are disentangled,

	Domain Accuracy (↓)	Content Mean Accuracy (↑)
Majority	0.020	0.110
SimSiam Transformations	0.251	0.637
Negatives From Random Classes	0.790	0.697
No Positive Pairs	0.034	0.230
C-Invariant Only	0.045	0.597

Table 2: Disentanglement and alignment metrics on SmallNORM for different versions of the objective. We see that using many transformations or using negative pairs from multiple classes severely hurts disentanglement. On the other hand, not using positive pairs hurt representation alignment. Our full objective enjoys the best of both worlds, with excellent disentanglement and alignment.

they do not uniquely identify u. It is apparent that the trivial contrastive formulation above does not provide the inductive biases required for learning identifiable representations.

To transfer the inductive biases from generative models to our contrastive formulation, we enforce the invariance of the learned representations to the transformations that generative models were found to be invariant to. Specifically, we add images augmented by blur, high contrast and color saturation as positive examples. The objective becomes:

$$\mathcal{L}_{contrastive}(x_i) = \frac{e^{sim(E(x_i),E(f(x_i)))}}{\sum_j -\log(\sum_i \mathbf{1}_{y_i=y_j} e^{sim(E(x_i),E(x_j))})} \tag{3}$$

Where d_i is the domain from which x_i is drawn from and f is randomly selected of the four augmentations listed above. We rerun the experiment above, now using the transferred inductive biases. The results are presented in Tab. 4. We now see that the representations remain disentangled, but they are now also informative of the unknown attribute u i.e. the pose can now be predicted from the learned representation. We conduct a further experiment, where instead of using negative examples with the same class, we use negative examples from the entirer mini-batch (across all classes). Results on SmallNORB when negative examples from all classes are used, are shown in Tab. 4. This illustrates that our modification to the denominator is key for making our approach work.

A key aspect of our approach is using transformations to which generative models were found to be invariant. It is imperative to investigate whether standard augmentations e.g. those used in SimSiam (or other augmentation-based representation leanring methods would suffice). To test this hypothesis, we repeated the same experiment as above but with all the augmentations used by SimSiam rather than the three invariant transformations from Sec. 3.2. We report the results in Tab. 4. We can see that using transformations to which generators are not invariant hurts disentanglement. To understand how including bad transformations can hurt performance, let us assume that a transformation changes the content, it would exclude content information from being included in the code. However, as the class is also excluded from the code by the uniformity constraint, it will not be possible to satisfy the contrastive objective causing reduced performance. This will be expressed either in reduced disentanglement or in reduced alignment.

To summarize, we train an encoder that takes in an image x and returns code c. The encoder is trained using the contrastive objective in Eq. 3. Although, at first sight, our objective might appear very similar to the standard contrastive objective, there are two key differences: i) the negative examples in the denominator are only taken from the same class as the target image, rather than all images. We showed theoretically and empirically that this simple modification is critical. ii) the augmentations used correspond to the three transformations that generators are invariant to. This was also shown to be critical for the performance of the method.

5 Experiments

In this section, we evaluate our method against generative and adversarial approaches. In Sec. 5.2, we evaluate the disentanglement and alignment of the learned representations. In Sec. 5.3, we evaluate performance on downstream tasks, specifically, cross-domain translation and retrieval.

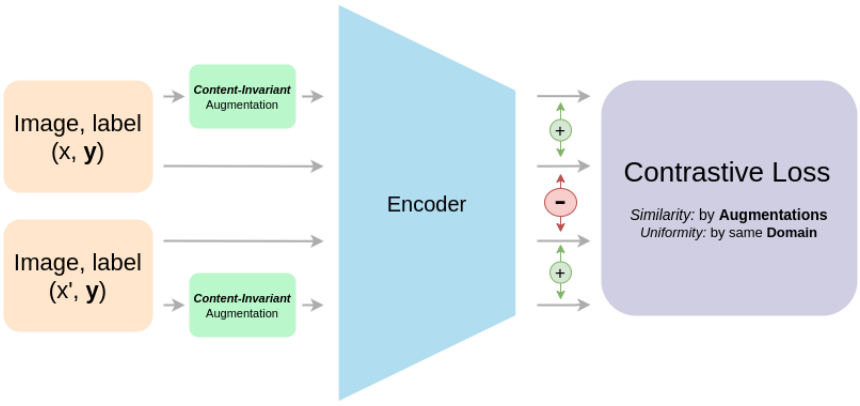

Figure 1: ABCD's Architecture. We train a single encoder for all domains. The objective function encourages the representation of every image to be as far as possible from those of other images from the same domain - this encourages domain disentanglement. Additionally, the objective encourages the representation of different augmentations of the same image to be the same. This encourages representation alignment. The choice of augmentations determines the inductive bias of the method.

5.1 Implementation Details

Architecture. We use a ResNet18 encoder. In line with other methods such as LORD (that uses a perceptual loss), we use ImageNet pretrained weights.

Optimization hyperparameters. We use a learning rate of 0.001. For SmallNORB and Cars3D we train our method for 200 epochs, using a batch size of 512, composed from 32 images drawn from 16 different classes. Since the classes in the CelebA dataset are smaller we use 16 classes and 8 samples from each one in each batch.

Temperature. We tune our temperature constant for the contrastive learning between the values of 0.1 to 0.3. We use $0.1, 0.2, 0.3$ for CelebA, SmallNORB and Cars3D accordingly.

Baselines. We implement ML-VAE, DrNET using their default parameters. We tried to replace their encoders by Resnet18 but this resulted in degraded performance. We therefore report their best results. We train LORD's second stage encoder using a Resnet18 as well. We trained it for 200 epochs for CelebA and Cars3D, and for 300 epochs on SmallNORB (as 200 were not sufficient for convergence).

Augmentations. As mentioned in Sec. 4, we used Gaussian blurring, high contrast and high saturation transformations as our positive augmentation.

5.2 Direct Representation Evaluations

In this section, we conduct direct evaluations of the learned representations.

Experimental setup. We evaluate the two key aspects of the representation, namely: i) disentanglement - prediction accuracy of the domain y, given the code c. Low accuracy would reflect a high degree of disentanglement. ii) alignment - prediction accuracy of the hidden attribute u given the code c. Note that this metric requires having groundtruth labels for the hidden attributes, which is typically available for synthetic datasets such as Cars3D and SmallNorb, but not for real datasets like CelebA. We therefore provide this metric for the synthetic datasets only. We conducted the experiment for our method and LORD. As well as DrNet and ML-VAE that represent adversarial and non-adversarial baselines.

Results. We report results on Cars3D, SmallNorb and CelebA. We observe that on Cars3D, both our method and LORD achieved excellent (nearly perfect performance). This is expected, as this dataset is relatively simple. We can see however that ML-VAE and DrNet did not perform as well on this dataset. This is inline with the results reported in LORD. On SmallNorb, our method was able to achieve disentangled representations whereas none of the other methods could. Note this

SmallNorb benchmark is the original version and not the simplified version developed in the LORD paper. In this setting, the object category only is known whereas both pose and lighting are unknown. The poor disentanglement of other methods allows them to include more information on the unknown attributes in the code. However, it is clear that our method provides a better tradeoff between disentanglement and content alignment than the alternative methods (as it is the only one that allows good disentanglement). Finally, on celebA we provide better disentanglement than the competing methods. As there are no groundtruth labels on the unknown attributes in celebA, we did not provide this analysis.

Table 3: Content Disentanglement ($\downarrow$) (Content to Domain) and Representation Quality ($\uparrow$) (Average Prediction Accuracy). For CelebA there is no groundtruth for the unknown attributes.

	Cars3D		**SmallNorb**		**CelebA**
	Domain($\downarrow$)	Factors($\uparrow$)	Domain($\downarrow$)	Factors($\uparrow$)	Domain($\downarrow$)
Majority	0.005	-	0.021	-	$1.5 \cdot 10^{-3}$
LORD	0.009	0.941	0.46	0.707	0.019
DrNet	0.505	0.912	0.953	0.914	0.092
ML-VAE	0.709	0.931	0.982	0.946	0.139
Ours	0.008	0.955	0.045	0.597	0.007

5.2.1 Training time analysis

We provide a training time comparison of the training between our method and the current SOTA, LORD (Gabbay & Hoshen, 2020b). Both algorithms were run on a single NVIDIA-QuadroRTX-6000 for 200 epochs for all datasets. For LORD, we present 2 different timings, the end of the latent optimization stage, and the end of the amortized stage. Results are presented in Tab. 4. We can observe that our method is **an order of magnitude faster** than LORD.

Table 4: Training Times ($\downarrow$) In Hours

	Cars3D	**SmallNorb**	**CelebA**
LORD	4.7/7.7	6.5/15.5	80/160
Ours	**0.7**	**1.7**	**12**

5.3 Downstream Application

5.3.1 Image Translation

Experimental setup. Although the objective of our method is to learn strong representations rather than image generation, we provide some qualitative image translation results. For each image set, we extract the domain y (object category) from the left, while the unlabeled attributes (typically pose or lighting) are taken from the top row. We presented results for our method and LORD.

Results. We observe that LORD and our method achieve excellent results on Cars3D. We see however that LORD fails on SmallNorb. Although it is able to transfer the pose, it fails to transfer the lighting. On the other hand, our method is able to extract the correct representations from the rerlevant images.

5.3.2 Cross Domain Retrieval

In this section we demonstrate the performance of our method on a discriminative downstream task.

Experimental setup. We evaluate the cross-domain retrieval task. Given an image from one domain, and a set of images from another domain, our objective is to recover the image whose unlabelled attributes are most similar to those of the target image. We evaluate the performance of our learned encoder against those of the competing disentanglement methods: LORD, DrNet and ML-VAE. We compute results on Cars3D and on SmallNorm. We did not provide quantitative results on

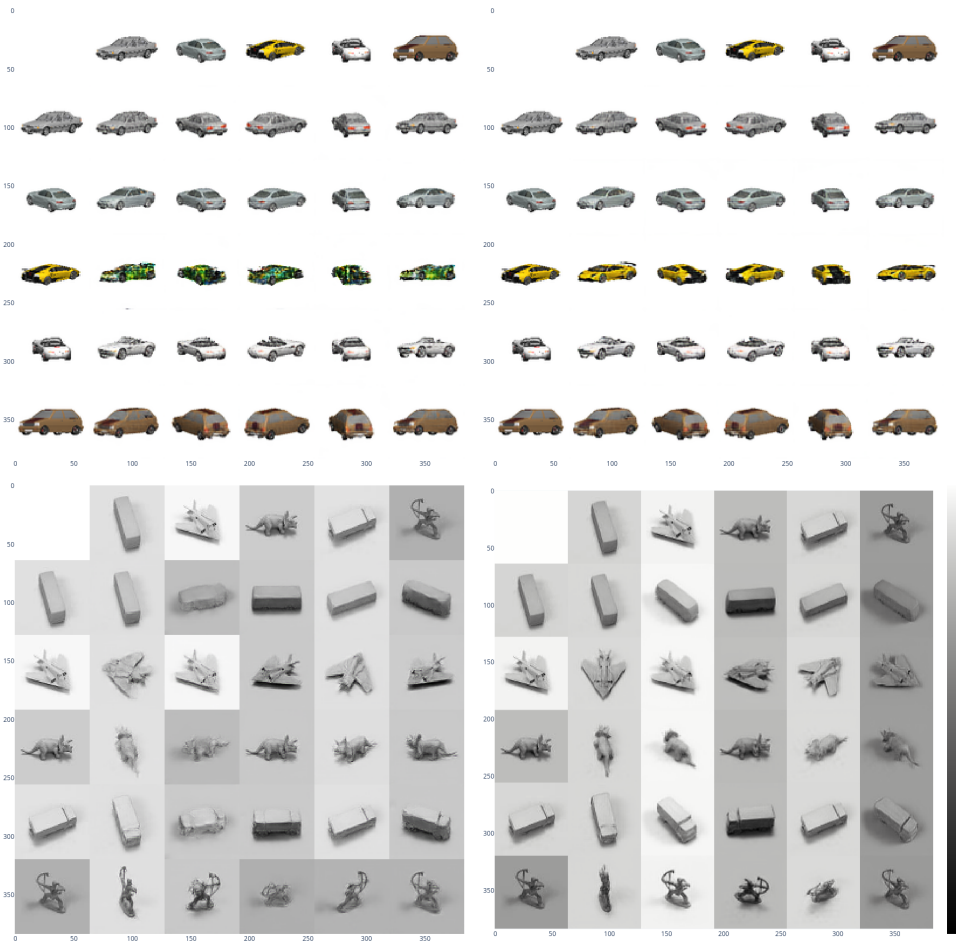

Figure 2: Image Translation results of LORD & ABCD. Both methods perform very well on Cars3D. LORD fails to transfer the lighting in SmallNorb while our method succeeds.

Table 5: Cross-domain retrieval accuracy (%) (↑)

	Cars3D		**SmallNORB**	
	Top-1	Top-5	Top-1	Top-5
Majority	0.010	0.052	0.001	0.005
ML-VAE (Bouchacourt et al., 2018)	0.529	0.851	0.061	0.192
DrNet (Denton et al., 2017)	0.642	0.940	0.053	0.164
LORD (Gabbay & Hoshen, 2020b)	0.823	0.956	0.047	0.136
Ours	**0.920**	**0.967**	**0.083**	**0.224**

celebA as its unknown attributes are not labeled. We evaluate the methods by their top-1 and top-5 retrieval performance.

Results. Our quantitative evaluation is presented in Tab. 7. We can see that our method dominated all other methods on all metrics.

Figure 3: Image Translation results of LORD & ABCD. The leftmost images are the targets, the rightmost three and the top-3 retrievals. We can see that results on Cars3d are good for both methods. On the other hand, our method dominates LORD on SmallNorb retrieval.

6 Discussion and Conclusion

We presented a discriminative, non-adversarial method for learning disentangled and aligned representations. This was achieved by transferring the inductive biases of generative models, to a contrastive learning approach. We made several important modifications to the contrastive loss, and found they are critical for our methods to work. We evaluated our method and found that it indeed learns disentangled and aligned representations. Our method is about an order of magnitude faster than competing approaches. It was also found to achieve better results than strong baselines on several tasks and datasets.

Naturally, our method has several limitations that can be addressed in future work. We discuss some of those below:

Non-generative inductive biases. Our method currently replicates the inductive biases of generative models. It therefore does not have other useful inductive biases which generative models do not have. By designing new augmentation, future work may be able to extend the range of inductive biases.

Batch-size sensitivity. Our method is based on SimCLR, whose performance is positively correlated with the batch-size. Future work may investigate using other framework e.g. MoCo-v2 that have a reduced dependence on the batch size.

REFERENCES

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2020.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, 2019.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pp. 4414–4423, 2017.

Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *ICLR*, 2020a.

Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *International Conference on Learning Representations (ICLR)*, 2020b.

Aviv Gabbay and Yedid Hoshen. Scaling-up disentanglement for image translation. *arXiv preprint arXiv:2103.14017*, 2021a.

Aviv Gabbay and Yedid Hoshen. Scaling-up disentanglement for image translation. In *ICCV*, 2021b.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *ECCV*, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pp. 1878–1889, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. http://mmlab.ie.cuhk.edu.hk/projects/celeba.html. In *ICCV*, 2015.

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, 2019.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, 2016.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pp. 5102–5112. PMLR, 2019.

Pierre H. Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, and Michal Valko. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. *ICLRW*, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pp. 14245–14258, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Zhirong Wu, Yuanjun Xiong, Stella Yu, , and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, 2019.

Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 192–199, 2014.

Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *Advances in neural information processing systems*, pp. 118–129, 2018.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *ICML*, 2021.

A Appendix

A.1 Inductive Bias Analysis

Table 6: An evaluation of the invariance of autoencoders to different image transformations

	Cars3D	
	Norm. Invariance (2)	**Invariance (1)**
Horizontal Flip	0.861	0.1212
Vertical Flip	0.853	0.1991
Low Contrast	1.031	0.1199
Low Brightness	0.749	0.1978
Color Rotation	1.118	0.0725
Random Erase []	0.728	0.1994
Affine Transformation	0.440	0.3143
GrayScale	0.957	0.0479
Crop	0.671	0.2107
High Brightness	0.355	0.0372
High Contrast	0.312	0.0356
High Saturation	0.337	0.0095
Gaussian Blurring	0.260	0.0724

Table 7: An evaluation of the invariance of autoencoders to different image transformations

	Edges2Shoes	
	Norm. Invariance (2)	**Invariance (1)**
Horizontal Flip	0.868	0.3387
Vertical Flip	0.770	0.3458
Low Contrast	1.901	0.2022
Low Brightness	1.098	0.2265
Color Rotation	1.015	0.0829
Random Erase []	0.884	0.1994
Affine Transformation	0.602	0.3143
GrayScale	0.939	0.0502
Crop	0.808	0.2879
High Brightness	1.188	0.0956
High Contrast	0.454	0.0499
High Saturation	0.768	0.0243
Gaussian Blurring	0.417	0.1079

Table 8: An evaluation of the invariance of autoencoders to different image transformations

	CelebA	
	Norm. Invariance (2)	**Invariance (1)**
Horizontal Flip	0.876	0.2869
Vertical Flip	0.751	0.3925
Low Contrast	0.659	0.1289
Low Brightness	0.429	0.0755
Color Rotation	0.496	0.1507
Random Erase []	0.776	0.1968
Affine Transformation	0.619	0.3121
GrayScale	0.464	0.0729
Crop	0.854	0.3322
High Brightness	0.875	0.0950
High Contrast	0.534	0.0862
High Saturation	0.632	0.0445
Gaussian Blurring	0.177	0.0644