A Contrastive Objective for Learning Disentangled Representations

Jonathan Kahana and Yedid Hoshen

Hebrew University Of Jerusalem, Jerusalem, Israel

Abstract. Learning representations of images that are invariant to sensitive or unwanted attributes is important for many tasks including bias removal and cross domain retrieval. Here, our objective is to learn representations that are invariant to the domain (sensitive attribute) for which labels are provided, while being informative over all other image attributes, which are unlabeled. We present a new approach, proposing a new domain-wise contrastive objective for ensuring invariant representations. This objective crucially restricts negative image pairs to be drawn from the same domain, which enforces domain invariance whereas the standard contrastive objective does not. This domain-wise objective is insufficient on its own as it suffers from shortcut solutions resulting in feature suppression. We overcome this issue by a combination of a reconstruction constraint, image augmentations and initialization with pre-trained weights. Our analysis shows that the choice of augmentations is important, and that a misguided choice of augmentations can harm the invariance and informativeness objectives. In an extensive evaluation, our method convincingly outperforms the state-of-the-art in terms of representation invariance, representation informativeness, and training speed [1]. Furthermore, we find that in some cases our method can achieve excellent results even without the reconstruction constraint, leading to a much faster and resource efficient training.

1 Introduction

Representing the attributes of an image that are independent of its domain (e.g. imaging modality, geographic location, sensitive attribute or object identity) is key for many computer vision tasks. For instance, consider the following toy example: assume that we observe images of faces, each image is specified by the identity and pose but only labels of the identity are provided. The goal is to learn a representation that captures the unlabeled pose attribute, and carry no information about the identity attribute. This task has many other applications, including: learning to make fair decisions, cross domain matching, model anonymization, image translation etc. It is a part of the fundamental machine learning problem of representation disentanglement. We note that the most ambitious disentanglement setting, i.e. unsupervised disentanglement where no labels are provided, was proven by Locatello et al. [25] to be impossible without

[1] Our Code is available at https://github.com/jonkahana/DCoDR.

inductive biases. Luckily, our setting is easier than unsupervised disentanglement as the domain label is provided for all training images. This setting has attracted much research e.g. DRNET [10], ML-VAE [1] and LORD [13].

We begin by defining the desired properties for domain disentanglement. This task has two objectives: i) *Invariance*: the learned representation should be invariant to the domain ii) *Informativeness*: the learnt representation should include the information about all of the attributes which are independent of the domain. The invariance requirement is challenging, but it can *in-principle* be directly optimized as the domain label is provided, e.g. using an adversarial discriminator. The informativeness requirement, however, is not generally possible to directly optimize without additional inductive biases as the attributes are unlabeled. This was theoretically demonstrated by [20,40]. Nonetheless, recent methods have been able to achieve meaningful representations in many cases, by enforcing a reconstruction term, which optimizes a related objective.

We present a new method, **DCoDR**: **D**omain-wise **Co**ntrastive **D**isentangled **R**epresentations, that significantly improves both representation domain invariance and informativeness. To enforce the domain invariance, we propose a per-domain contrastive loss, that requires the representations of each domain to be uniformly distributed across the unit sphere. Differently from standard contrastive losses [4], our objective only considers negative examples from the *same* domain. As shown in Sec. 5.2, this seemingly simple change is crucial for learning domain invariant representations. Unfortunately, we find that encoders which satisfy this invariance constraint alone, are often uninformative over the desired attributes. This is a case of the documented phenomenon of *feature suppression* [24,6,32]. In line with previous methods [13,10,1], we optimize the informative-

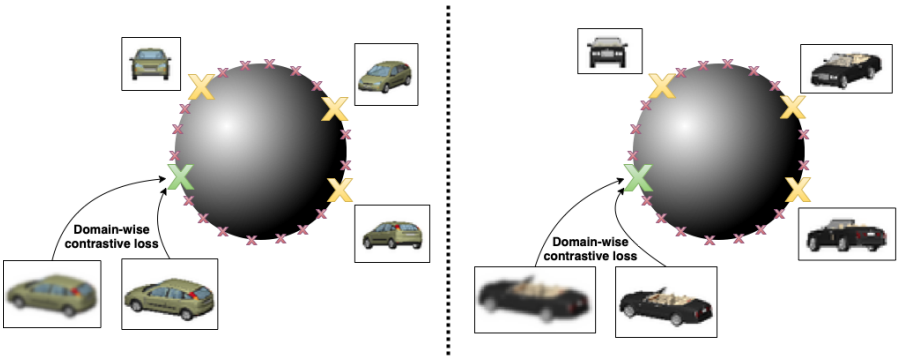

Fig. 1: An illustration of our method. The representations are domain invariant as the representations of each domain follow a spherically uniform distribution (encouraged by our domain-wise contrastive objective). Image augmentations (here Gaussian blurring) are used to assign similar images to nearby representations which indirectly improves informativeness. The reconstruction objective and encoder pre-trained weights initialization are not shown in this diagram.

ness of the representations indirectly by a reconstruction constraint. As we find this may be insufficient for learning informative representations in some cases, we propose two other techniques: i) Similarly to several self-supervised objectives (e.g. the one in SimCLR [4]), we enforce representations of images to be similar to those of their augmentations. Despite being common among self-supervised methods, we show that standard choices of augmentations (specifically, those used by SimSiam [8]) can harm the domain invariance of the representation. We analyse the effectiveness of different augmentations for domain invariant representation learning. ii) Initializing the image encoder using weights pre-trained with self-supervision on an external dataset, which we empirically find to learn both more informative and invariant representations.

We evaluate our method on five popular benchmarks. Our method significantly exceeds the state-of-the-art in terms of invariance and informativeness. We investigate a fully discriminative version and find that in many cases it is competitive with the previous state-of-the-art while being much faster.
A summary of our contributions:

1. A non-adversarial and non-generative, domain invariance objective.
2. Analysing the benefits and pitfalls of image augmentations for informativeness and domain invariance of the learned representations.
3. A new approach, DCoDR, which significantly outperforms the state-of-the-art in domain invariant representation learning.
4. A discriminative only variant, which is 5X faster than existing approaches.
5. An extensive evaluation on five datasets.

2 Related Work

Learning domain disentangled representations. Much research was done on separating between labeled and unlabelled attributes. Several methods use adversarial training [11,34,26]. Other methods use non-adversarial approaches, e.g. cycle consistency [16], group accumulation [1] or latent optimization [13,14]. Our method improves upon this body of work.

Contrastive representation learning. 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 [8,15,31] use the above constraint with various other tricks for learning representations. As the above formulation is prone to collapse, contrastive methods [38,18,36,29,18,17,7,27,4,5] add an additional uniformity constraint that prohibits collapse of the representation to a single point. We propose a per-domain contrastive objective, tailored for domain disentanglement.

Contrastive approaches for disentanglement. Recently, Zimmerman et al. [42] 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 domain disentanglement - assuming domain labels for training data, but not having image sequences or strong 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 [41], image manipulation [13,14,37] and domain adaptation [30]. Furthermore, it is believed that better disentangled representations will have future impact on model interpretability [19], abstract reasoning [33] and fairness [9].

3 Domain Invariant Representation Learning

3.1 Preliminaries

We receive as input a set of training samples $\mathcal{X}_t = \{x_1, x_2, .., x_N\}$. Each training sample $x \in \mathcal{X}_t$ has a labeled domain d and unlabelled attributes y which are uncorrelated to d. We assume that the labeled domain d is a single categorical variable. The objective is to learn an encoder E, which encodes each image x as code $z = E(x)$ satisfying the criteria in Sec. 3.2.

3.2 Criteria

The domain disentanglement task requires satisfying the following two criteria:

Invariance: We require that the representation z should not be predictive of the domain d. This can be written as:

$$P(d|z) = P(d) \tag{1}$$

Informativeness: We require that the representation z should encapsulate as much information on attributes y as possible. Note that z cannot hold more information about y than the original image x, as there exists a deterministic encoder E which maps x to z. It therefore follows by the data processing inequality, that the maximally informative representation z should be as informative as the original image about the attributes y:

$$I(y, z) = I(y, x) \tag{2}$$

In our setting, only the domain labels d are provided but not the attribute labels of y. The objective in Eq. 2 cannot therefore be optimized directly. Saying that, in line with previous methods, we optimize informativeness by training a conditional generator through a reconstruction objective. Unlike previous methods, we use additional techniques which increase informativeness significantly. Our proposed approach will be detailed in Sec. 4.3.

3.3 Existing Approaches for Invariance Optimization

Current methods optimize the invariance criterion using two main approaches:

Adversarial methods [10]. Many disentanglement methods rely on adversarial domain confusion constraints to ensure representation invariance. They are often written in the following form:

$$L_{adv} = \max_{D} \ell_{CE}(D(E(x)), d) \tag{3}$$

Where ℓ_{CE} is the cross-entropy loss. The discriminator D measures how informative the representation $z = E(x)$ is over the original domain d. An encoder that satisfies this constraint will indeed be domain invariant $P(d|z) = P(d)$. Unfortunately, adversarial training is challenging and the optimization often fails to minimize this loss perfectly.

Variational-autoencoders (VAE) [1,13]. Given the weaknesses of adversarial methods, variational methods were proposed that ensure the representations are normally distributed $P(z|d) = N(0; I)$. The encoder in this case outputs the parameters of a Gaussian distribution of the posterior $p(z|x)$. Using the ELBO criterion, the objective becomes:

$$L_{vae} = \ell_{KL}(E(x), N(0, I)) \tag{4}$$

However, LORD [13] found that simply optimizing this criterion does not converge to disentangled representations. Furthermore, they showed that randomly initialized encoders are highly entangled and variational losses were insufficient for removing this entanglement. Instead, they suggested using latent optimization rather than deep encoders at first, for directly learning the representation z of each training image x. This indeed improves the domain invariance of the representations, but is more sensitive to hyper-parameter choices. It also requires an inconvenient second stage for learning an image to representation encoder.

4 DCoDR: Learning Domain-wise Contrastive Disentangled Representation

4.1 Overview

We introduce a new approach, DCoDR, for learning informative, domain invariant representations. In Sec. 4.2, a new per-domain contrastive loss is proposed to enforce invariance directly. It does not, by itself, require the representation to be maximally informative. To overcome this issue, we optimize informativeness indirectly by *reconstruction* and *image augmentation* objectives as well as *encoder pre-trained weights initialization*. We investigate an additional, fully discriminative variant of our method, which is much faster than existing methods at the price of lower informativeness.

4.2 Representation Invariance with Domain-wise Contrastive Losses

Learning an invariant representation requires the domain d to be unpredictable from the learned representation z. We present a non-adversarial method for encouraging domain invariance. Our approach enforces the probability distribution of representations z to follow a uniform spherical distribution (denoted U_S) regardless of the domain d: $P(z|d) = U_S$. It follows from Bayes' law that the representation z does not provide any information about the domain, $\forall z : P(d|z) = P(d)$. This also yields that mutual information between the domain and representation is zero $I(d, z) = 0$.

The above analysis requires that $P(z|d) = U_S$ for every domain d. We do so by training a separate contrastive loss for every domain d. It was highlighted by Wang and Isola [35] that the denominator of the contrastive objective encourages the representations follow a uniform spherical distribution. Learning a contrastive loss *separately* over image representations from different domains, ensures that the representations z are distributed as U_S regardless of the domain d. For an image x from domain d, this can be written as follows:

$$\mathcal{L}_{inv}(x, d) = \log \sum_{(x',d')\in\mathbb{X}} \mathbf{1}_{d'=d} e^{sim(E(x'),E(x))} \tag{5}$$

sim is a similarity function, cosine similarity in our case. The objective only considers image pairs drawn from the *same* domain. Unlike previous methods in Sec. 3.3 (e.g. [10,28,13]), it does not rely on adversarial or variational approximations.

4.3 Improving Representation Informativeness

Beyond invariance, the representations z should encapsulate the information about all of the image attributes y except the domain label d. In Eq. 2 this was shown to imply $I(y, x) = I(y, z)$. We cannot directly optimize this constraint, as the attributes y for image x are not provided in our setting. In line with previous methods presented in Sec. 3.3, we optimize the informativeness indirectly by a reconstruction constraint. Furthermore, we present two algorithmic choices that empirically further increase informativeness significantly.

Reconstruction: Reconstruction constraints are an established way to improve the informativeness of the representation d. They have been used in many previous methods [10,1,13]. In line with previous methods, we include a reconstruction constraint in our method. Specifically, we learn a conditional generator G that takes as input the domain d and representation z and outputs an image $G_d(z)$. The reconstruction objective requires that the output image is as close as possible to the input image x. The difference between the reconstruction and original images is measured using the function ℓ. In practice, we use the same perceptual loss as in LORD [13] in Eq. 6

$$L_{rec} = \sum_{d\in\mathcal{D}} \sum_{x\in\mathcal{X}_d} \ell_{perc}(G_d(E(x)), x) \tag{6}$$

Augmentations: Contrastive objectives are susceptible to shortcut solutions that lower informativeness, also known as *feature suppression* [32]. This occurs by (inadvertently) learning an encoder that maps nuisance image attributes (or noise) to the spherical uniform distribution. This representation ignores the other image attributes, therefore being insufficiently informative. Ensuring that image augmentations have similar representations to the original image can help reduce this collapse, for suitably well selected augmentations:

$$\mathcal{L}_{aug} = \sum_{x \in \mathcal{X}_t} -sim(E(A_1(x)), E(A_2(x)) \tag{7}$$

Where $A_1(x)$ and $A_2(x)$ are two random augmentations of image x. Unfortunately, poorly selected augmentations can make the representation z invariant to the desired attributes y, which is harmful. E.g. when y is pose, and the augmentation is horizontal flip, the representation z will be invariant to flip direction, therefore less informative over the pose. To test this hypothesis, we trained our method's discriminative variant, DCoDR-norec, using blur or flip augmentations on Cars3D. We measure each metric as explained in Sec. 5.2. Tab. 1 shows flipping significantly reduced the informativeness.

Table 1: Evaluation of our method's discriminative variant, DCoDR-norec, with 2 different augmentations on Cars3D.

	Inv.	Inform.
Blur	0.002	0.960
H. Flip	0.003	0.725

It is clear from the discussion above that augmentations can be highly desirable for improving informativeness, while their choice is important. We discovered that the standard augmentations used by state-of-the-art contrastive methods e.g. [4,15,8] are not optimal for our task. The reason is that they are designed to keep information only about the object's 'class' while being invariant to all other attributes. This may, in some cases, also require invariance on the attributes of interest y. Instead, we selected a much smaller set of augmentations which we empirically show to be effective on a set of datasets that we considered. There selected augmentations are: i) Cropping, ii) Gaussian Blurring, iii) Increase of contrast iv) Increase in saturation. For Edges2Shoes [39] dataset, we find it more effective to include gaussian blurring alone. In Sec. 5.2 we show the selected augmentations significantly outperform the standard set of SimSiam [8].

Encoder Initialization with Unsupervised Pre-Trained Weights: Although the constraints proposed in this section are effective for learning domain disentangled representations, we empirically find they are not always sufficient. In order to improve generalization [12], we propose to initialize the encoder with the weights of a network pre-trained in an **unsupervised** manner (MoCo-V2 [7]) on the ImageNet dataset. Using the inductive bias from pre-trained weights in this setting is common, e.g. LORD [13] uses an ImageNet pre-trained perceptual loss. Note, this initialization is not beneficial for LORD as it does not use an encoder in the first stage.

4.4 Our Complete Method: DCoDR

DCoDR optimizes the combination of the 3 objectives presented in this section:

$$\min_{E,G} L_{DCoDR} = L_{inv} + L_{aug} + L_{rec} \tag{8}$$

We use the augmentations from Sec. 4.3. We initialize the encoder E with the weights of an MoCo-V2 encoder pre-trained on ImageNet (without labels).

Discriminative DCoDR (DCoDR-norec) We present a discriminative variant of our method, by simply dropping the reconstruction constraint:

$$\min_{E} L_{DCoDR-norec} = L_{inv} + L_{aug} \tag{9}$$

The lack of a reconstruction constraint, makes this variant typically learn less informative representations than DCoDR. However, as this variant does not train a generator, it is several times faster than DCoDR which by itself is considerably faster than previous state-of-the-art LORD.

4.5 Differences From SimCLR

Although a part of our method is motivated by the SimCLR [4] objective, it is significantly different and attempts to obtain satisfy different criteria compared to SimCLR (the first 3 apply for DCoDR-norec as well):

- **Domain-wise Loss.** DCoDR learns a contrastive loss over each domain separately whereas SimCLR learns a single loss over all the data.
- **Choice of Augmentations.** DCoDR learns a reduced set of augmentations rather than the standard set used in SimCLR.
- **Pre-Training.** DCoDR initializes the encoders weights by *unsupervised* pre-training on ImageNet using of MoCo-V2 [7], which does *not* use any labels.
- **Reconstruction** DCoDR uses a reconstruction term for increasing the informativeness of its representations, which does not exist in SimCLR.

Tab. 2 and 3 show that although the differences from SimCLR might look superficially simple, each of them is essential for the success of our method, on the described domain disentanglement setup.

5 Experiments

In this section, we evaluate our method against (variational and adversarial) state-of-the-art domain disentanglement approaches. We evaluate the invariance and informativeness of the learned representations. We then demonstrate cross domain retrieval of our method compared to the other baselines in Sec. 5.3.

Benchmark Datasets. We report results on Cars3D [21], SmallNorb [22], Shapes3D [3], CelebA [23] and Edges2Shoes [39]. All datasets are used in 64x64 resolution. Due to the large number of samples in the full Shapes3D and the limited variation between them, we randomly sample $50,000$ images for training, while keeping the test set size at 10% of the original size.

5.1 Implementation Details

Architecture and optimization. We used a ResNet50 encoder, trained for 200 epochs using a batch size of 128. Each batch was composed from 32 images drawn from 4 different classes. In line with other methods e.g. LORD, the reconstruction loss is computed using a VGG based perceptual loss pre-trained on ImageNet.

Baselines. We use the default parameters of ML-VAE [1] and DRNET [10]. We tried to replace their encoders by larger ResNet architectures but it resulted in degraded performance. We therefore kept the original architectures and hyperparameters for all runs. We use a ResNet50 architecture for LORD's second stage and SimCLR's encoders, training each for 200 epochs. We do not compare to OverLORD [14] as in our evaluated datasets it is exactly the same as LORD.

Augmentations. As mentioned in Sec. 4.3, we used cropping, Gaussian blurring, high contrast and high saturation transformations as our positive augmentations, except for Edges2Shoes where we use only Gaussian blurring.

5.2 Representation Evaluation

Experimental Setup. For each dataset, we evaluate both *Invariance* and *Informativeness* of the representations. To do so, we train a deep classifier to predict all image attributes from the learned representations, including the domain d and the other factors y. For the synthetic datasets, we compute each of the two objectives over each factor separately. Since some of the datasets have multiple factors, we present the average of the informativeness over all factors, while the full results are presented in the SM. For CelebA we use the location of the 68 landmarks [2] as the uncorrelated attribute. As the landmarks are numeric rather than categorical, we train an MLP regression model to predict the landmark locations. We measure the L_1 error of the MLP regressor where lower errors are better. To understand how far the results are from the theoretical limit, we present the frequency of the most common domain value as a lower bound on the invariance. Note that since we use a probabilistic estimator to evaluate our metrics, in some cases (especially when performance is close to optimal limit) the invariance may be slightly lower than the theoretical limit. This can happen when the classifier slightly overfits its training data, hence the small gap. To ensure a fair comparison is made, we train each classifier with several regularization strengths and present the one that is able to generalize best.

Results. The results on all datasets are presented in Tab. 2. We observe that on Cars3D, even though LORD is a strong baseline, both discriminative and complete variants of DCoDR are able to surpass it, and achieve nearly perfect results. ML-VAE, DRNET and SimCLR do not perform as well on this dataset, inline with the observations in [13]. On SmallNorb, it is clear that LORD fails to disentangle the domain. Both our methods outperforms it on both metrics, achieving much more disentangled representations than any other method. As the representations learned by ML-VAE and DRNET are not domain invariant, they have higher informativeness but do not satisfy the main requirement of disentanglement. Note that we used the original version of the SmallNorb

Table 2: Content Invariance (↓) (Content to Domain) and Representation Quality (↑) (Average Prediction Accuracy). For CelebA we use extracted landmarks as attributes, and compute the regression L1 (↓) error.

	Cars3D		**SmallNorb**		**Shapes3D**		**CelebA**	
	Inv. ↓	Inf. ↑	Inv. ↓	Inf. ↑	Inv. ↓	Inf. ↑	Inv. ↓	L1 ↓
SimCLR	0.885	0.443	0.956	0.758	1	0.99	0.116	1.286
LORD	0.009	0.940	0.393	0.670	0.703	0.995	0.019	0.862
DRNET	0.504	0.909	0.953	0.899	0.892	1	0.084	0.795
ML-VAE	0.697	0.930	0.968	0.944	0.999	1	0.136	0.723
DCoDR-norec	0.005	0.970	0.071	0.730	0.246	0.997	0.015	1.127
DCoDR	0.005	0.980	0.143	0.785	0.245	0.999	0.017	0.858
Optimal	0.005	1	0.021	1	0.251	1	0.002	0

benchmark rather than the simplified version presented in the LORD paper. In this setting, the domain is defined as the object category alone whereas *both* pose and lighting are unknown. On Shapes3D, again both variants of DCoDR achieve almost perfect results while all other methods suffer from lack of domain invariance. LORD achieves very limited invariance while ML-VAE, DRNET and SimCLR learn representations that are not invariant at all. CelebA is challenging for our per-domain contrastive loss, as it contains very few images per each domain, meaning the estimation of a uniform distribution for each domain is limited. That being said, we observe DCoDR performs better than LORD. It has an additional advantage over LORD of not requiring 2-stage optimization. CelebA is a failure case for our discriminative variant. Although presenting stronger invariance than the other methods, it is not sufficiently informative.

Generally, DCoDR demonstrated state-of-the-art results in invariance and informativeness. In some cases (e.g. CelebA), DCoDR-norec fails to learn sufficiently informative representations, while being more invariant than previous methods as well as DCoDR itself. We emphasize that a key advantage of DCoDR-norec is its training time, as shown in Sec. 5.4.

Ablation Study We ablate our method on the SmallNorb dataset (Tab. 3). First, we observe that removal of the unsupervised MoCo-V2 pre-trained weight initialization significantly hurts all metrics. Removal of per-domain negative pairs i.e. using a single contrastive loss for all domains (the loss used in SimCLR), makes the representations entangled. We also tested removing the positive augmentations, using the objective in Eq. 5. Removing the positive augmentations has different effects in to DCoDR and DCoDR-norec. DCoDR's informativeness was reduced while invariance improved. DCoDR-norec fails without the positive augmentations as they are its only objective that enforces informativeness. Lastly, we consider the standard set of augmentations used in SimSiam [8]. This choice significantly harms both invariance and informativeness in both variants.

Table 3: Ablation analysis on SmallNorb.

	DCoDR		**DCoDR-norec**	
	Inv. (↓)	Inform. (↑)	Inv. (↓)	Inform. (↑)
No Domain Negatives	0.863	0.829	0.879	0.754
No Positive Augmentations	0.057	0.555	0.021	0.166
No Pre-Training	0.253	0.701	0.298	0.716
SimSiam [8] Augmentations	0.244	0.643	0.246	0.658
Complete Method	0.143	0.785	0.071	0.730
Optimal	0.020	1	0.020	1

5.3 Cross Domain Retrieval Evaluation

Experimental Setup. To evaluate cross-domain retrieval, we first extract representation $z = E(x)$ for each image x from domain d in the test set. Than, we retrieve its nearest neighbors (using L_2 distance) from each domain d' so that $d' \neq d$ and average the results over all domains. Finally, results are averaged over all test images. We present both quantitative and qualitative analyses. For our quantitative analysis, we use the labels of the attributes y for deciding weather a match was found or not. Since many attributes are naturally ordered we would like to consider more than just perfect matches in all attributes. To do so, we allow a match for small changes in some numeric attributes, as detailed in the SM. Here we present the accuracy of matching over all attributes. The accuracy of matching individual attributes is presented in the SM. We also visually present the 5 nearest-neighbor images for several test set images - using the representations learned by our and baseline methods. Here, we search for neighbors *in all domains at once*, contrary to the previous quantitative retrieval evaluation, which was performed for each domain separately, then averaged. The analysis highlights leakage of domain information in the representations.

Quantitative Analysis. Our numerical retrieval results are presented in Fig. 4. Similarly to the earlier probing experiments on Cars3D, LORD achieves the highest retrieval scores among all the baseline methods on this dataset. Our

Table 4: Retrieval Accuracies Comparison.

	Cars3D	SmallNorb	Shapes3D	Edges2Shoes
SimCLR	0.07	0.02	<0.01	0.40
LORD	0.88	0.06	0.76	0.66
DRNET	0.64	0.09	0.86	0.66
ML-VAE	0.50	0.06	0.63	0.65
DCoDR-norec	0.96	0.22	0.99	0.41
DCODR	**0.97**	**0.26**	**1**	**0.90**

method, convincingly outperforms it, both with and without reconstruction. DRNET and ML-VAE achieve acceptable results, but underperform DCoDR and LORD due to their lack of invariance. SimCLR fails to retrieve accurately since, as Tab. 2 suggests, it prefers representing the domain over the pose. SmallNORB is a much harder task, all baseline methods struggle on this dataset achieving poor retrieval accuracy. We showed in Tab. 2 that these methods have high informativeness and poor invariance on this dataset. This shows that invariance is important for succeeding in cross domain retrieval. DCoDR (with and without reconstruction) is able to retrieve much better matches as it is considerably less biased by the domain. This is backed up by the qualitative analysis of SmallNorb in Fig. 2. Results on Shapes3D describe a similar case. Although all methods achieve strong informativeness, DCoDR and DCoDR-norec only are able to retrieve nearly perfect matches due to domain invariance. Surprisingly, on this dataset DRNET was able to retrieve strong matches from different domains, despite not being domain invariant at all. Finally, Edges2Shoes showcases a failure of DCoDR-norec, as the augmentations do not provide a strong enough inductive bias for learning informative representations. Saying that, when given the inductive bias of a generator, DCoDR exceeds previous methods significantly.

Qualitative Analysis. We present retrieval results on SmallNorb [22] and Edges2Shoes [39] datasets in Fig. 2 and 3 respectively. We present DCoDR-norec on SmallNorb, and DCoDR on Edges2Shoes (as the reconstruction loss is needed there). On SmallNorb, DRNET and ML-VAE retrieve images from the same domain at the expense of changing the pose, achieving poor retrieval results. While LORD does select images from other domains, the domains are typically similar to the source. DCoDR-norec retrieves images from a variety of domains while preserving the pose. Both LORD and DCoDR-norec struggle with

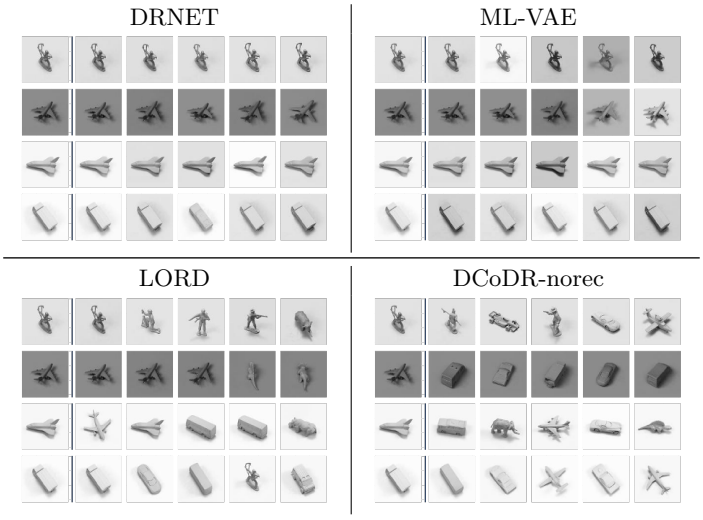

Fig. 2: Retrieval Examples From SmallNorb.

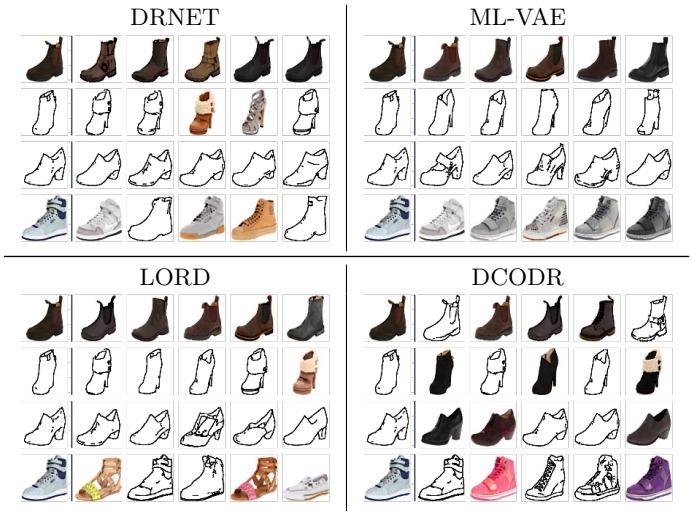

Fig. 3: Retrieval Examples From Edges2Shoes.

Table 5: Training Times (↓) In Hours.

	Cars3D	SmallNorb	Shapes3D (50K)	CelebA
LORD	7.5	15.5	18	160
DCODR	5.5	9.5	11.5	30
DCoDR-norec	**1.5**	**3.5**	**3.5**	**9**

180° flips. For Edges2Shoes, ML-VAE clearly shows lack of domain invariance. DCoDR retrieves more accurate images than DRNET and LORD.

5.4 Runtime Comparison

We compared our method's runtime with LORD [13], the top baseline. All methods were run on a single NVIDIA-RTX6000 for 200 epochs for all datasets (note that LORD has two stages). Results are presented in Tab. 5. Both DCoDR and DCoDR-norec are faster than LORD. DCoDR-norec is **5-10 times faster** than LORD as it does not train a generator nor require perceptual loss computation.

6 Discussion

The mismatch between conditional reconstruction constraints and informativeness. By the data processing inequality, the existence of a deterministic mappings $x = G_{d_{true}}(z)$ (and accordingly $x = E(z)$) implies that $I(y; x|z, d_{true}) = 0$. In other words, all the information about y which exists in x, exists in the combination of z and the domain label d_{true} as well. Note this

does not imply $I(y; x|z, d') = 0$ for any domain d' but only for the true domain of x, $d' = d_{true}$. To be equivalent to Eq. 2, it was shown by [20] and [40] that this requires another property from the representations which is *alignment*. Meaning, $p(y|z, d) = p(y|z, d') = p(y|z)$, where d and d' are two different domains and $p(y|z, d)$ is the PDF of y's values given the representation z under domain d. Alignment is not guaranteed without additional inductive biases but in practice learned representations are often well aligned.

Inductive bias of generators. We presented a discriminative variant that, in some cases, competes with the top domain disentanglement methods, which are generative. We believe the reason for the success of conditional generator based methods is two-fold: i) a regularization effect caused by the difficulty of conditional generator training, pushing the representations of different domains to be more aligned. ii) invariance of generators to various image transformations. DCoDR-norec presented *partial* improvements in these two aspects. Pre-trained weights are used for initialization, we hypothesize this acts as a regularizer although not as strong as a conditional generator. Image augmentations are used, *most of which* are encapsulated in the invariance of generators. To test the invariance of generators to different augmentations, we performed an experiment where we trained autoencoders on several datasets and compared their reconstruction for images with and without augmentations. This motivated our choice of augmentations. For more details, see the SM. Despite DCoDR-norec showing promising results of in some cases, we find that all components of our method are needed for sufficient informativeness. We expect that future research will find other augmentations which will result in further improvements.

Limitations. Our method has a few limitations which we leave for future work:

(i) *Discrete domains.* As required by our per-domain invariance objective.

(ii) *Pre-training.* We showed in Sec. 5.2 that using unsupervised pre-training (MoCo-V2 trained on ImageNet) significantly improves both invariance and informativeness. Although requiring an external dataset is a limitation, we do not believe it is a very serious one for two reasons. Firstly, previous methods, e.g. LORD, often use supervised pre-trained features in their perceptual loss as well. Secondly, these weights are available to all and identical for all new datasets.

(iii) *Image-specific augmentations.* Our method rely on image augmentations, which are not always transferable to other modalities e.g. audio or text. Nevertheless, we believe other helpful augmentations can be found in each modality.

7 Conclusion

We presented a new approach for learning domain disentangled representations from images. It uses a per-domain contrastive loss, a reconstruction objective, image augmentations and self-supervised pre-trained encoder initialization. Our method demonstrated results that are better in both invariance and informativeness metrics over the state-of-the-art.

Acknowledgement Jonathan Kahana was partially supported by grants from the Israeli Prime Minister Office and the Council for Higher Learning.

References

1. Bouchacourt, D., Tomioka, R., Nowozin, S.: Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: AAAI (2018) 2, 3, 5, 6, 9
2. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017) 9
3. Burgess, C., Kim, H.: 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/ (2018) 8
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: ICML (2020) 2, 3, 7, 8
5. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS (2020) 3
6. Chen, T., Luo, C., Li, L.: Intriguing properties of contrastive losses. Advances in Neural Information Processing Systems **34** (2021) 2
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 3, 7, 8
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2020) 3, 7, 10, 11
9. Creager, E., Madras, D., Jacobsen, J.H., Weis, M.A., Swersky, K., Pitassi, T., Zemel, R.: Flexibly fair representation learning by disentanglement. In: International Conference on Machine Learning (2019) 4
10. Denton, E., Birodkar, V.: Unsupervised learning of disentangled representations from video. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 4417–4426 (2017) 2, 5, 6, 9
11. Denton, E.L., et al.: Unsupervised learning of disentangled representations from video. In: Advances in neural information processing systems. pp. 4414–4423 (2017) 3
12. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research **11**(19), 625–660 (2010), http://jmlr.org/papers/v11/erhan10a.html 7
13. Gabbay, A., Hoshen, Y.: Demystifying inter-class disentanglement. In: International Conference on Learning Representations (ICLR) (2020) 2, 3, 4, 5, 6, 7, 9, 13
14. Gabbay, A., Hoshen, Y.: Scaling-up disentanglement for image translation. In: ICCV (2021) 3, 4, 9
15. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020) 3, 7
16. Harsh Jha, A., Anand, S., Singh, M., Veeravasarapu, V.: Disentangling factors of variation with cycle-consistent variational auto-encoders. In: ECCV (2018) 3
17. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 3
18. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019) 3

19. Hsu, W.N., Zhang, Y., Glass, J.: Unsupervised learning of disentangled and interpretable representations from sequential data. In: Advances in neural information processing systems. pp. 1878–1889 (2017) 4
20. Johansson, F.D., Sontag, D., Ranganath, R.: Support and invertibility in domain-invariant representations. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 527–536. PMLR (2019) 2, 14
21. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013) 8
22. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2**, II–104 Vol.2 (2004) 8, 12
23. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 8
24. Li, T., Fan, L., Yuan, Y., He, H., Tian, Y., Katabi, D.: Information-preserving contrastive learning for self-supervised representations. CoRR **abs/2012.09962** (2020), https://arxiv.org/abs/2012.09962 2
25. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019) 1
26. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: NIPS (2016) 3
27. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: CVPR (2020) 3
28. Moyer, D., Gao, S., Brekelmans, R., Steeg, G.V., Galstyan, A.: Invariant representations without adversarial training. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 9102–9111 (2018) 6
29. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 3
30. Peng, X., Huang, Z., Sun, X., Saenko, K.: Domain agnostic learning with disentangled representations. In: International Conference on Machine Learning. pp. 5102–5112. PMLR (2019) 4
31. Richemond, P.H., Grill, J.B., Altché, F., Tallec, C., Strub, F., Brock, A., Smith, S., De, S., Pascanu, R., Piot, B., Valko, M.: Byol works even without batch statistics. arXiv preprint arXiv:2010.10241 (2020) 3
32. Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., Sra, S.: Can contrastive learning avoid shortcut solutions? Advances in Neural Information Processing Systems **34** (2021) 2, 7
33. van Steenkiste, S., Locatello, F., Schmidhuber, J., Bachem, O.: Are disentangled representations helpful for abstract visual reasoning? In: Advances in Neural Information Processing Systems. pp. 14245–14258 (2019) 4
34. Szabó, A., Hu, Q., Portenier, T., Zwicker, M., Favaro, P.: Challenges in disentangling independent factors of variation. ICLRW (2018) 3
35. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020) 6

36. Wu, Z., Xiong, Y., Yu, S., , Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018) 3
37. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
38. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: CVPR (2019) 3
39. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 192–199 (2014) 7, 8, 12
40. Zhao, H., Des Combes, R.T., Zhang, K., Gordon, G.: On learning invariant representations for domain adaptation. In: International Conference on Machine Learning. pp. 7523–7532. PMLR (2019) 2, 14
41. Zhu, J.Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., Freeman, B.: Visual object networks: Image generation with disentangled 3d representations. In: Advances in neural information processing systems. pp. 118–129 (2018) 4
42. Zimmermann, R.S., Sharma, Y., Schneider, S., Bethge, M., Brendel, W.: Contrastive learning inverts the data generating process. In: ICML (2021) 3