LEVERAGING AFFINITY CYCLE CONSISTENCY TO ISOLATE FACTORS OF VARIATION IN LEARNED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying the dominant factors of variation across a dataset is a central goal of representation learning. Generative approaches lead to descriptions that are rich enough to recreate the data, but often only a partial description is needed to complete downstream tasks or to gain insights about the dataset. In this work, we operate in the setting where limited information is known about the data in the form of groupings, or set membership, and the task is to learn representations which isolate the factors of variation that are common across the groupings. Our key insight is the use of *affinity cycle consistency* (ACC) between the learned embeddings of images belonging to different sets. In contrast to prior work, we demonstrate that we can work with significantly fewer constraints on the factors of variation, and in some cases, with almost no constraints on the factors of variation for part of the data. By curating datasets from Shapes3D, we quantify the effectiveness of ACC through mutual information between the learned representations and the known image generative factors. In addition, we demonstrate the applicability of ACC to the tasks of digit style isolation and synthetic-to-real object pose transfer.

1 INTRODUCTION

Isolating desired factors of variation in a dataset requires learning representations that retain information only pertaining to those desired factors while suppressing or being invariant to remaining "nuisance" factors. This is a fundamental task in representation learning which is of great practical importance for numerous applications. For example, image retrieval based on certain specific attributes (e.g. object pose, shape, or color) requires representations that have effectively isolated those particular factors.

In designing approaches for such a task, the possibilities for the structure of the learned representation are inextricably linked to the types of supervision available. As an example, complete supervision of the desired factors of variation provides maximum flexibility in easily obtaining fully disentangled representation, where there is a simple and interpretable mapping between its elements and the factors of the variation (Bengio et al., 2013). However, such supervision is an unrealistic goal for most tasks since many common factors of variation in image data, such as 3D pose or lighting, are difficult to annotate at scale in real-world settings. At the other extreme, unsupervised representation learning makes the fewest limiting assumptions about the data but also does not allow any control over the discovered factors of variation.

The challenge is in designing a learning process that best utilizes the supervision that can be realistically obtained in different real-world scenarios. In this paper, we consider weak supervision in the form of set membership (Kulkarni et al., 2015; Denton & Birodkar, 2017). Specifically, this weak set supervision assumes only that we can curate subsets of training data where only the desired factors of variation to be isolated vary, and the remaining nuisance factors are fixed to same values. We will refer to the factors that vary within a set as the *active* factors, and those that have fixed and same values as *inactive*.

To illustrate this set supervision, consider the problem of isolating 3D object pose from images belonging to an object category (say, car images). The weak set supervision assumption can be satisfied by simply imaging each object from multiple viewpoints. Note, this would not require

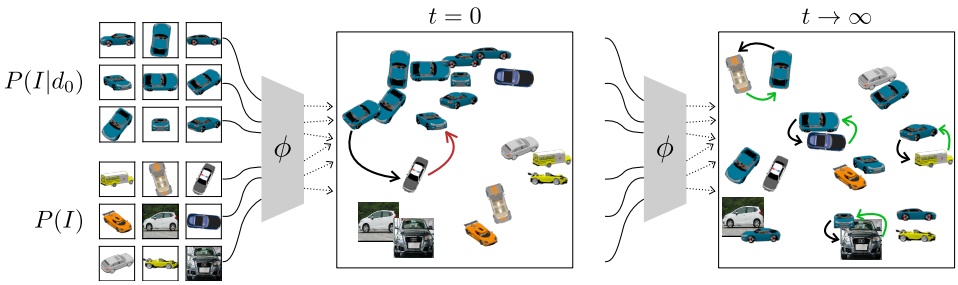

Figure 1: *Affinity cycle consistency (ACC) yields embeddings which isolate factors of variation in a dataset $P(I)$. It leverages weak supervision in the form of set membership, such as in the set of images $P(I|d_0)$ rendered around a given synthetic car (top, left). The method seeks correspondence between sets of inputs by a soft nearest neighbor cycle consistency loss, and does so by identifying factors that vary within each set and suppressing factors which do not. We show that ACC isolates nontrivial factors of variation, such as pose in the example above, even when only one of the sets has been grouped. Importantly, this allows us to factor in data for which there is no supervision at all, such as the real car images. The learned representations, like the ones on the right ($t \to \infty$), contain only the isolated factor of variation (contrast the alignment here with the untrained representations shown in the middle, $t = 0$).*

consistency or correspondence in viewpoints across object instances, nor any target pose values attached to the images. In practice, collecting multiple views of an object in a static environment is much more reasonable than collecting views of different objects with identical poses.

In this paper we propose a novel approach for isolating factors of variation by formulating the problem as one of finding alignment between two sets with some common active factors of variation. Considering the application of synthetic-to-real object pose transfer, Figure 1 illustrates two sample sets of car images where pose is the only active factor in the first set $P(I|d_0)$ of synthetic car images and the second set $P(I)$ is comprised of both real and synthetic car images. Given these sets, without any other supervision, the aim is to automatically learn the embeddings that can find meaningful correspondences between the points in the two sets. The key idea behind our approach is a novel utilization of cycle consistency. A cycle consistent mapping can be described broadly as some non-trivial mapping that brings an input back to itself, and in our case the mapping is between sets of points in embedding space. We denote our application of cycle consistency as *affinity cycle consistency* (ACC) as it uses a differentiable version of soft nearest neighbors since the correspondence forming the cycle or not known apriori[1]. Further, no explicit pairwise correspondence between the input sets is needed; it is found by the loss. We posit that this process of finding correspondences is crucial to isolating the desired factors of variation: to match across sets, the representations must ignore commonality within a set (the inactive factors) and focus on the active factors common to both the sets. For example, ACC learned embeddings from the two sets of car images in Figure 1 can isolate the object pose factor as that is the common active factor across both the sets.

We also show how our ACC model can be generalized to the partial set supervision setting: ACC can learn to isolate factors of variation even when set supervision is provided for only one set, while the second set is virtually unrestricted. This has practical importance as it allows us to integrate unsupervised data during training. In Section 4.3 we show how this process can be applied to isolate 3D pose in real images without ever seeing any supervised real images during training.

In the following two sections we cover the related works and formally introduce our ACC method. Given the novelty of our approach for isolating factors of variation, we present a progression of experiments to help develop an intuition behind the technique as it operates in different scenarios. In Section 4.1 we evaluate ACC in various settings using the synthetic Shapes3D dataset where the latent factor values are known, allowing for a quantitative analysis. Later, in Section 4.2 we demonstrate the use of ACC in isolating handwritten digit style from its content (class id). In Section 4.3,

[1]This specific loss has been used previously in Dwibedi et al. (2019) to align different videos of the same action, and here we show this loss is much more general. We term it *affinity cycle consistency* as opposed to the prior work's terminology, *temporal cycle consistency*, to indicate as such.

we show how ACC can be applied in its most general form to isolate 3D object pose in real images with a training process that combines a collection of set-supervised synthetic data with unsupervised real images. We conclude with a discussion and analysis.

2 RELATED WORK

Disentangled representations. Most approaches toward disentangled respresentations are unsupervised, and are generally based on generative modeling frameworks such as variational autoencoders (Kingma & Welling, 2014) or generative adversarial networks (Goodfellow et al., 2014). The VAE is a latent variable model that encourages disentanglement through its isotropic Gaussian prior, which is a factorized distribution. Numerous variations of the VAE have been proposed to further disentanglement, and these include β-VAE (Higgins et al., 2017), β-TCVAE (Chen et al., 2018), FactorVAE (Kim & Mnih, 2018), DIP-VAE (Kumar et al., 2018), JointVAE (Dupont, 2018), and ML-VAE (Bouchacourt et al., 2018). InfoGAN (Chen et al., 2016) encourages an interpretable latent representation by maximizing mutual information between the input and a small subset of latent variables. In Hu et al. (2018) adversarial training is combined with mixing autoencoders. In Locatello et al. (2019) it is shown that true unsupervised disentanglement is impossible in generative models, and inductive biases or implicit supervision must be exploited. Supervision has been incorporated in different ways. Graphical model structures are integrated into the encoder/decoder of a VAE model to allow for partial supervision (Siddharth et al., 2017). In Sanchez et al. (2020) disentanglement without generative modeling is proposed, but training requires pairs of images sharing the desired factor of variation. In contrast to these approaches, we learn an entangled representation where the isolated factors of variation are encoded in most or all its elements.

Cycle consistency. Often, cycle consistency has been used as a constraint for establishing point correspondences on images (Zhou et al., 2016; Oron et al., 2016) or 3D pointcloud (Yang et al., 2020; Navaneet et al., 2020). In a different setting, the time window between the image frames of Atari games can be learned using a discrete version of cycle consistency (Aytar et al., 2018). In contrast to Aytar et al. (2018), we use a discriminative approach and do not recover a disentangled representation. Cycle consistency has also been used in disentangling factors of variation with variational autoencoders using weak supervision in the form of set supervision (Jha et al., 2018). The most closely related work to ours is the work on temporal cycle consistency (TCC) (Dwibedi et al., 2019), where a differential soft nearest-neighbor loss is used to find correspondences across time in multiple videos of the same action. Our approach differs in two key ways. First, we generalize the approach to a broader class of problems where the data does not provide an ordering and is less likely to permit a 1-1 correspondence between sets. Second, and most significantly, we show the fundamental constraint of TCC, that both sets must share common active factors of variation, can be relaxed to allow one unconstrained set (no inactive factors), which allows for incorporating training data with no set supervision.

Weak supervision. For recovering semantics, there exists several self-supervision methods (Chen et al., 2020; Misra & van der Maaten, 2019) that rely on elaborate data augmentation and self-supervision tasks such as jigsaw and rotations. Augmentation can be effective if it is known how factors of variation act on the image space, but this is only true for some fully observable factors such as 2D position and orientation. This restriction similarly applies to models that bake transformations into the architecture, as with spatial transformers (Hinton et al., 2011) or capsules (Jaderberg et al., 2015). Often, we use fully self-supervised methods for recovering semantics necessary for downstream tasks. In this work, we are interested in isolating geometric factors of variation such as pose that are difficult to annotate. In order to do this, we rely on set supervision (Kulkarni et al., 2015; Mathieu et al., 2016; Cohen & Welling, 2015). In Cohen & Welling (2015) the latent representations of the training images are optimized limiting view synthesis to only objects seen at training time. In practice, getting full supervision with ground truth parameters for geometric entities such as ligthing and pose will be challenging. On the other hand, one can often capture videos where we fix one or more of these factors of variation, and allow the others to vary. This form of set supervision is a good tradeoff between labor-intensive manual annotation required for fully supervised methods, and fully self-supervised methods.

3D pose aware representations. An important factor of variation for many image tasks is 3D object pose, and not surprisingly there have been attempts to learn representations which encode this property. The $SO(3)$-VAE (Falorsi et al., 2018) places a uniform prior on the 3D rotation group

$SO(3)$, which allows learning manifold-valued latent variables. Latent representations that are pose-equivariant have been proposed in Worrall et al. (2017); Rhodin et al. (2018), and this allows for pose to be directly transformed in the latent space. Generative techniques for pose disentanglement include Kulkarni et al. (2015); Yang et al. (2015a). In Kulkarni et al. (2015) a simplistic experimental setting is considered (fewer factors of variation with synthetic or grayscale images, and disentangling only the 1D azimuth angle).

3 Affinity Cycle Consistency

Notation: We are given a finite set of n training samples $\mathcal{S} = \{x_1, \ldots, x_n\}$, where $x_i \in \mathcal{X}$, and $\mathcal{X}$ denote the instance space with dimensions D. We are also given weak supervision in the form of set membership where we have groups of images with the same inactive factors of variation. Let $\phi(x, w) : \mathcal{X} \to \mathbb{R}^E$ be the function that maps the input vector to an embedding in E-dimensional space. We use a affinity cycle consistency loss on the embeddings $\ell : \mathbb{R}^E \to \mathbb{R}_+$ to learn the parameters: $w^* = \arg\min_w \, {}_{x \sim \mathcal{X}} \ell(\phi(x, w_\phi))$.

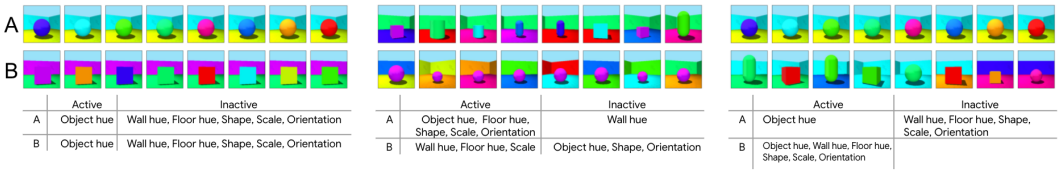

Figure 2: ***Sets with different active and inactive factors of variation from the Shapes3D dataset*** *(Burgess & Kim, 2018). On the left, we have the same active factor of variation for both sets. In the middle, we have some overlapping active factors of variation. On the right, we have no known inactive factor of variation for set B. We find* ACC *to be effective in all these three scenarios.*

Mini-batch construction using set-membership: Our goal is to learn $\phi : \mathcal{X} \to \mathbb{R}^E$ such that our learned representations isolate certain desired factors of variation and are invariant to certain other nuisance factors of variation. In order to learn such representations, we use cycle consistency loss, that allows us to learn the embedding function $\phi(x, w_\phi)$ by striving to obtain 1-1 mapping between two sets of points in the embedding space (Dwibedi et al., 2019). Our key idea is to learn embeddings with desired factors of variation by curating images in the two sets accordingly for mini-batch construction. Let us refer to these two sets of images by $A \subset \mathcal{X}$ and $B \subset \mathcal{X}$. In Figure 2, we show sets with several active and inactive factors of variation from the Shapes3D dataset (Burgess & Kim, 2018). In practice, we observe the following scenarios to be effective:

- The first set always uses at least one inactive factor of variation. Both the first and the second sets can have identical active and inactive factors of variation as shown in Figure 2(left), but this is not a necessary condition.
- We observe ACC to work under a wide variety of settings, and the only requirement we can emphasize is that it requires one or more common active factors of variation between the two sets as shown in Figure 2(middle). For example, if we need to learn pose-aware embeddings, then the sets A and B should both possess pose as one of the active factors of variation. If we need to exclusively isolate pose as the only factor of variation in the learned embeddings, then all other factors of variation should be inactive in both the sets.
- The second set can be completely unconstrained without any known inactive factors of variation, but this would require the first set to have fewer active factors of variation as shown in Figure 2(right).

Let the images in each of these sets be given by $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_m\}$, respectively. Let us denote the associated embeddings as $L = \{l_1, \ldots, l_n\}$ and $M = \{m_1, \ldots, m_n\}$, where $l_i = \phi(a_i, w)$ and $m_i = \phi(b_i, w)$. Let $d(x, y)$ denote cosine or squared L2 distance between the points in the embedding space. The notion of cycle consistency is used in many different contexts, and we use the following definition.

Definition 1 (***Cycle consistency***) *Given two sets of points $L = \{l_1, \ldots, l_n\}$ and $M = \{m_1, \ldots, m_n\}$, we say that $l_i \in L$ is cycle consistent if $l_i = \arg\min_{l \in L} d(m, l)$ where $m = \arg\min_{m \in M} d(l_i, m)$.*

In other words, we say that a point $l_i \in L$ is cycle consistent if its nearest neighbor $m \in M$ also has l_i as the nearest neighbor. In order to use cycle consistency as our loss function, we will use a differentiable formulation of the cycle consistency loss. To get a differentiable loss, we approximate the nearest neighbor with soft nearest neighbor (Goldberger et al., 2004), which has also been used in many applications (Movshovitz-Attias et al., 2017; Rocco et al., 2018; Snell et al., 2017).

Definition 2 (***Soft nearest neighbor***) *Given two sets of points $L = \{l_1, \ldots, l_n\}$ and $M = \{m_1, \ldots, m_n\}$, the soft nearest neighbor of $l_i \in L$ in the set M is given by $\tilde{m} = \sum_{j=1}^{n} \alpha_j m_j$, where $\alpha_j = \frac{e^{d(l_i, m_j)/\tau}}{\sum_{k=1}^{M} e^{d(l_i, m_k)/\tau}}$ and τ is the temperature parameter.*

We first compute the soft nearest neighbor for $l_i \in L$ as $\tilde{m} = \sum_{j=1}^{n} \alpha_j m_j$. To satisfy the cyclic consistency constraint, l_i should cycle back to itself. In order to enforce this, the nearest neighbor of $\tilde{m}$ should be l_i. In the differentiable formulation, we map the problem of cycling back to l_i as a classification problem (Dwibedi et al., 2019). Given that we have n points in L, we compute n logits given by $o_j = -d(\tilde{m}, l_j), \forall j \in \{1, \ldots, n\}$ and let $\hat{s} = \mathrm{softmax}(o)$. Let s denote the 1-hot encoding where $s_j = 1$ and zero otherwise. Now the affinity cyclic consistency (ACC) constraint is enforced as the cross-entropy loss function $\sum_{j=1}^{n} s_j log(\hat{s}_j)$.

Double augmentation: We have noticed that the two sets can have many active factors of variation. Ideally, we would like to have as few active factors of variation as possible to learn embeddings that isolate few selective factors of variation. The key idea behind our double augmentation scheme is simple. Instead of enforcing cycle consistency from an image $a_i \in A$ to its soft nearest neighbor and back to a_i, we allow the cycling back constraint to start from an augmented image a_i' and return back to another augmented image a_i''. By carefully selecting the augmentations a_i' and a_i'' along certain active factors of variation, we can learn embeddings that can be invariant to certain nuisance active factors of variation. We found the double augmentation to be critical in our pose transfer experiments involving real and synthetic cars in section 4.3.

4 Experiments

We show that ACC can isolate desired factors of variation in three settings: Shapes3D (Burgess & Kim, 2018), digit style isolation, and object 3D pose transfer from synthetic to real images.

4.1 Systematic Evaluations on Shapes3D

We quantitatively analyze ACC with the synthetic Shapes3D (Burgess & Kim, 2018) dataset, where we can freely control the active and inactive factors of variation in the two sets. Sample images from Shapes3D are shown in Figure 3 and consist of a geometric primitive with a floor and background wall. There are six factors of variation in the dataset: three color factors (wall hue, object hue and floor hue) and three geometric primitive factors (scale, shape and orientation). The full dataset consists of every possible combination of these six discrete generative factors.

In order to train ACC with a particular generative factor inactive, we randomly sample from among its possible values and hold it fixed across a set of inputs, while sampling uniformly across the remaining factors to generate A and B sets of size 32. For example, in Figure 3a, two sets of inputs are shown with wall hue as an inactive factor. We train convolutional networks which embed the 64x64 RGB images into two dimensions, and visualize one set of learned representations in Figure 3b. From the plots, the mutual information between object and floor hues and the learned embedding is qualitatively apparent. We measure the mutual information $I(U; G)$ (Saxe et al., 2019) between each of the generative factors and the embeddings, and repeat each experiment for 50 random seeds. In each subplot of Figure 3c, different factors are made active during training. We show the (nonzero) mutual information present even in the output from an untrained network, as well as the result of training without set supervision. In this baseline, all generative factors are active during the course of training and nothing is suppressed in the learned embeddings. Interestingly, absent any supervision at all, there arises a significant distinction between the hue factors and the others, presumably out of saliency with respect to the network's capabilities.

The subsequent subplots in Figure 3c show training with various generative factor(s) rendered inactive, marked by a shaded box. Clear trends arise which allow a more precise definition of active factor isolation: information with respect to the inactive factors is noticeably suppressed while information about the remaining factors is enhanced in the learned representation. In other words, given

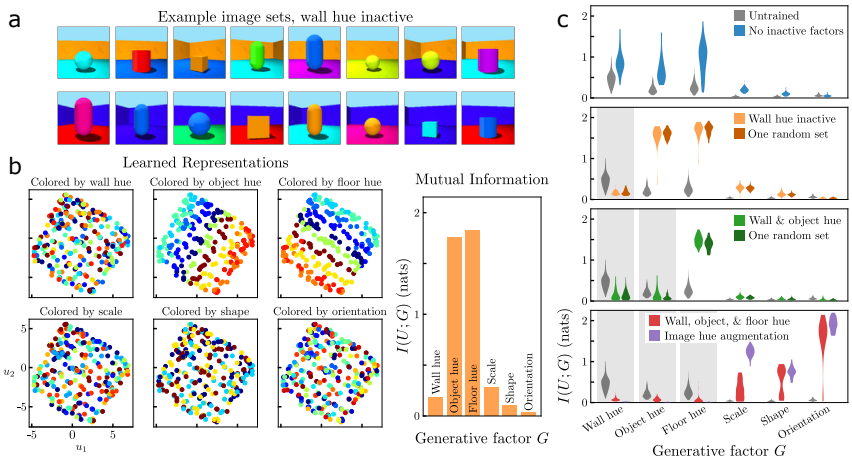

Figure 3: ***Analysis on Shapes3D. (a)*** *An example pair of sets for training where wall hue is the only inactive factor in each set.* ***(b)*** *We train ACC on pairs of image stacks as in* ***(a)****, and embed to 2D. In each of the six plots we display the same 256 embeddings, with each embedded image colored by its value for a each generative factor. In the bar chart we measure the mutual information between each of the factors and the learned representation.* ***(c)*** *We repeat the experiments 50 times each and display the spread in mutual information values as a violin plot. In each subplot a different subset of generative factors is used to group the data for training. We include in each subplot the results from an untrained network in black. In the bottom subplot we compare to a contrastive method which augments the overall hue of each input image twice and uses the two versions as a positive pair. In each setting, ACC successfully suppresses information about inactive factor(s) of variation and enhances information about active factors.*

weak set supervision over one factor of variation, ACC leads to learned representations which isolate factors of variation about which nothing was known beforehand. When all three hue factors are inactive we see scale, shape, and orientation feature most prominently in the learned representations, seemingly because the 'easy' hue factors have all been suppressed.

Additionally, we compare to setups where only one set A has supervision (active factors) and the other set B consists of random samples over the entire dataset. Strikingly, there is no significant change in the information content of the generative factors in the learned representations, meaning ACC performs just as well in a much looser setting. Finally, in the bottom subplot of Figure 3c, we compare ACC results where all hue factors were inactive, to the case of hue double augmentation removing all hue information from the learned embeddings. While the factor isolation effects are more pronounced, we emphasize that augmenting away nuisance factors is often not possible. This is the case in the middle two subplots, where augmenting hue would suppress all three hue factors indiscriminately and inhibit information with respect to the active variables.

4.2 Digit Style Isolation

Handwritten digits, such as those from MNIST (LeCun & Cortes, 1998), have two main factors of variation: content and style. Here, content refers to the class of the digit (say, 2 vs. 8) and the rest of the factors of variation can be referred to as style (such as stroke width, orientation, writing style etc.). In this experiment, our aim is to learn an embedding that can isolate digit style while being invariant to the digit class, with only set supervision on the digit class. For this, we group images by class into sets of size 64 and embed to 8D using a convolutional network. We held-out images of 9 from training sets in order to test the invariance of the learned embeddings to digit class.

Figure 4 (left) shows two-dimensional PCA plots of the learned embeddings next to those of untrained embeddings. The PCA plots clearly indicate the stronger correlation between digit style (width, orientation etc.) and the learned embeddings in comparison to those of untrained embeddings. As further analysis, we use test digits from each of the 10 classes to retrieve the nearest neighbor digits in other classes in Figure 4 (right). We compare to retrieval using embeddings learned with Jha et al. (2018), a VAE-based generative approach which utilizes set supervision as well. With-

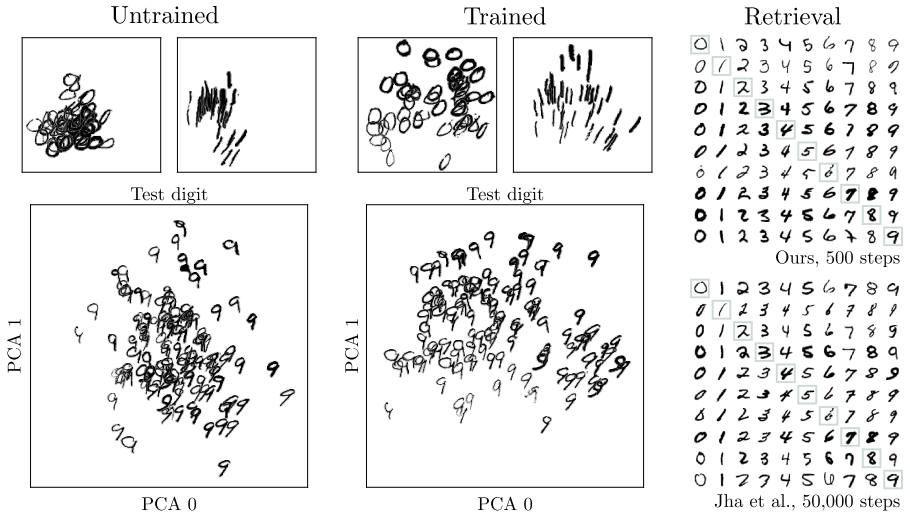

Figure 4: ***Style isolation in MNIST.*** *We define digit style as a combination of all factors of variation independent of the class. After training ACC with set supervision on the digit class (i.e. class label is the inactive factor within each set), we evaluate how well the style factors are isolated. We visualize embeddings from the MNIST test set using the top two PCA dimensions (accounting for more than 0.99 of the total variance). Before the network is trained, all the embeddings are clustered together (top left), while after training (top middle) the embeddings are fanned out by style factors, primarily thickness and slant in this case (this can be observed both within and across digits). All examples of the digit 9 are held out at training. The plots on the bottom show similar arrangement based on thickness and slant, confirming ACC generalizes to unseen invariant factor values. On the right we show retrieval results, where the boxed images along the diagonal are queries and the other images in each row are the nearest embeddings for each digit class (all query and retrieved images are from the test set). We see ACC retrieves images closer in style than CC-VAE (Jha et al., 2018), a generative approach which also utilizes set supervision. Notably, ACC also needed to train 100x fewer steps which highlights a benefit of our discriminative approach compared to generative models.*

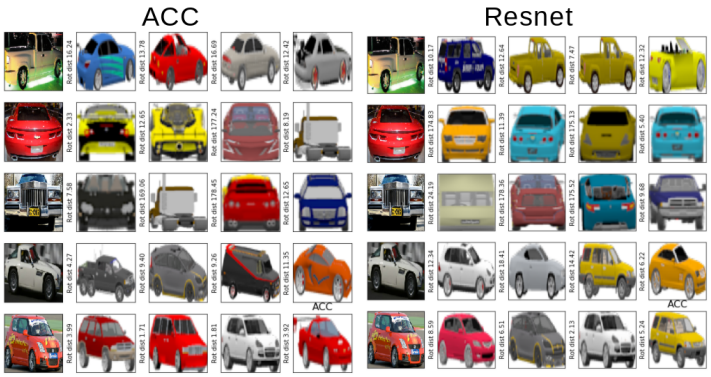

Figure 5: ***Retrieval results from ACC and ResNet embeddings.*** *For each real query image, we show alongside it the four nearest neighbors in embedding space, and the relative error from the ground truth pose. Note how ACC embeds images nearby which are often visually quite different, in contrast to the ResNet output. This serves as qualitative evidence that pose is being effectively isolated in the embeddings.*

out having to learn a full description of the data, ACC yields style-correlated embeddings 100 times faster than the related generative approach, with matching training protocol. This demonstrates the superior potential of ACC in isolating digit style without using any explicit supervision on styles.

4.3 Object Pose Transfer from Synthetic to Real Images

We showcase the unique capabilities of ACC on the task of 3D pose estimation in the setting with an abundance of real and synthetic images but without pose annotations. The goal is to make use of the unannotated real images by incorporating them into the training process and leveraging ACC's flexibility with respect to the inactive factors of variation in the second set. The baseline case for this scenario is to align synthetic images by pose, using set supervision over all other (inactive) factors of variation. We propose to incorporate real images under the assumption that ACC will be able to bridge the domain difference to isolate pose regardless of whether the image is real or synthetic.

We use the dataset included in KeypointNet (Suwajanakorn et al., 2018), which consists of renderings of ShapeNet (Chang et al., 2015) 3D models from viewpoints which are randomly distributed over the upper hemisphere. Set supervision is provided by grouping images according to their source 3D model. Other factors of variations such as object texture, lighting are also fixed (i.e. only viewpoint is an active factor within each set). We pair the synthetic images with real images from the CompCars (Yang et al., 2015b) and Cars196 (Krause et al., 2013) datasets for the car category, and 1000 images from the Pascal3d+ (Xiang et al., 2014) training split for chairs. All of the images are tight cropped and resized to 128x128.

We incorporate real images slowly into set B after training for 10,000 iterations with only synthetic images – grouped by ShapeNet model in set A and randomly sampled across all models and viewpoints in set B – before inserting 5.5% real images into each set B and training for another 10,000 iterations. Additionally, we found it beneficial to suppress nuisance active factors of variation in the images, such as the precise position of the object in the frame, by optimizing the double augmentation loss. For the double augmentation scheme explained in Section 3, we randomly augment each image twice with a variety of cropping, recoloring, and edge enhancement via Sobel filter. The network builds off the base of an ImageNet-pre-trained ResNet (He et al., 2015) with a few additional layers before embedding to 64 dimensions. As the similarity measure for ACC, we find cosine similarity with $\tau = 0.1$ outperforms L2 distance.

We evaluate on the images in Pascal3D+ by using nearest neighbor lookup between the embedded test image and a dictionary of 1800 rendered synthetic images with ground-truth pose, recovering pose without any annotations from real images. As baselines, we compare to the 16,384-dimensional flattened output from the ResNet base network, as well as to the embeddings learned with Jha et al. (2018). Quantitative results are shown in Table 1 and some retrieval results are shown in Figure 5. We additionally report a flip invariant metric to highlight how well ACC performs modulo flip symmetry, which can be particularly difficult for cars. We consistently outperform the baselines and the gap is significant in many cases and especially for chairs, showing the efficacy of ACC.

	Cars				Chairs			
			Flip invariant: $\min(\theta, 180^\circ - \theta)$				Flip invariant: $\min(\theta, 180^\circ - \theta)$	
	Med Err (°) ↓	Acc@30° ↑	Med Err (°) ↓	Acc@30° ↑	Med Err (°) ↓	Acc@30° ↑	Med Err (°) ↓	Acc@30° ↑
ResNet (pre-trained)	16.0	**0.65**	10.1	0.93	71.0	0.30	32.5	0.46
CCVAE (Jha 2018)	54.8	0.26	30.4	0.48	79.5	0.19	41.3	0.35
ACC (Ours)	**14.1**	**0.65**	**8.4**	**0.95**	**35.1**	**0.47**	**25.7**	**0.55**

Table 1: ***Pose estimation results on Pascal3D+.*** *Median error and accuracy metrics on Pascal3D+ car and chair test datasets. We obtain the pose with nearest neighbor lookup into 1800 synthetic images with GT pose, using different embeddings shown in different rows. ACC outperforms the VAE-based approach and the high dimensional ResNet embeddings.*

5 Discussion

In this work, we show how affinity cycle consistency can be utilized to isolate factors of variation provided only weak set supervision. Through extensive experiments on synthetic and real data, we show the technique can be applied in a number of different set supervision scenarios, including one where only a subset of the data requires any set supervision at all. This case is particularly important as it allows us to train with unsupervised real image data, and we validate this with promising experiments on the challenging problem of isolating 3D object pose in real images.

REFERENCES

Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI*, February 2018.

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

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2610–2620, 2018.

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

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2172–2180, 2016.

Taco S. Cohen and Max Welling. Transformation properties of learned visual representations. In *International Conference on Learning Representations (ICLR)*, 2015.

Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pp. 4414–4423, 2017.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 710–720. Curran Associates, Inc., 2018.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. *CoRR*, 2019.

Luca Falorsi, Pim de Haan, Tim R. Davidson, Nicola De Cao, Maurice Weiler, Patrick Forré, and Taco S. Cohen. Explorations in homeomorphic variational auto-encoding. *CoRR*, abs/1807.04689, 2018.

Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, Cambridge, MA, USA, 2004. MIT Press.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)*, pp. 44–51, 2011.

Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3399–3407, 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2017–2025, 2015.

Ananya Harsh Jha, Saket Anand, Maneesh Singh, and V. S. R. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

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.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pp. 2539–2547, 2015.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

Francesco Locatello, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

Michael Mathieu, Junbo Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. Disentangling factors of variation in deep representations using adversarial training. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 5047–5055, 2016.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations, 2019.

Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. *CoRR*, abs/1703.07464, 2017. URL http://arxiv.org/abs/1703.07464.

K L Navaneet, Ansu Mathew, Shashank Kashyap, Wei-Chih Hung, Varun Jampani, and R. Venkatesh Babu. From image collections to point clouds with self-supervised shape and pose networks, 2020.

Shaul Oron, Tali Dekel, Tianfan Xue, William T. Freeman, and Shai Avidan. Best-buddies similarity - robust template matching using mutual nearest neighbors, 2016.

Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Neighbourhood consensus networks. *CoRR*, abs/1810.10510, 2018. URL http://arxiv.org/abs/1810.10510.

Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *The European Conference on Computer Vision (ECCV)*, 2020.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, dec 2019. doi: 10.1088/1742-5468/ab3985. URL https://doi.org/10.1088%2F1742-5468%2Fab3985.

N. Siddharth, Paige Brooks, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5925–5935, 2017.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017. URL http://arxiv.org/abs/1703.05175.

Supasorn Suwajanakorn, Noah Snavely, Jonathan Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Interpretable transformations with encoder-decoder networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond PASCAL: A benchmark for 3d object detection in the wild. In *2014 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 75–82, March 2014.

Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NIPS)*, pp. 1099–1107. Curran Associates, Inc., 2015a.

Lei Yang, Wenxi Liu, Zhiming Cui, Nenglun Chen, and Wenping Wang. Mapping in a cycle: Sinkhorn regularized unsupervised learning for point cloud shapes, 2020.

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 (CVPR)*, June 2015b.

Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency, 2016.