Learning ABCs: Approximate Bijective Correspondence for isolating factors of variation with weak supervision

Kieran A. Murphy*
Univ. of Pennsylvania
kieranm@†

Varun Jampani
Google Research
varunjampani@‡

Srikumar Ramalingam
Google Research
rsrikumar@‡

Ameesh Makadia
Google Research
makadia@‡

Abstract

Representational learning forms the backbone of most deep learning applications, and the value of a learned representation is intimately tied to its information content regarding different factors of variation. Finding good representations depends on the nature of supervision and the learning algorithm. We propose a novel algorithm that utilizes a weak form of supervision where the data is partitioned into sets according to certain inactive *(common) factors of variation which are invariant across elements of each set. Our key insight is that by seeking correspondence between elements of different sets, we learn strong representations that exclude the inactive factors of variation and isolate the* active *factors that vary within all sets. As a consequence of focusing on the active factors, our method can leverage a mix of set-supervised and wholly unsupervised data, which can even belong to a different domain. We tackle the challenging problem of synthetic-to-real object pose transfer, without pose annotations on anything, by isolating pose information which generalizes to the category level and across the synthetic/real domain gap. The method can also boost performance in supervised settings, by strengthening intermediate representations, as well as operate in practically attainable scenarios with set-supervised natural images, where quantity is limited and nuisance factors of variation are more plentiful. Accompanying code may be found on github.*

1. Introduction

A good representation is just as much about what it excludes as what it includes, in terms of factors of variation across a dataset [52]. Control over the information content of learned representations depends on the nature of available supervision and the algorithm used to leverage it. For example, full supervision of desired factors of variation provides maximum flexibility for fully disentangled representations, as an interpretable mapping is straightforward to obtain between elements and the factors [3, 19]. However, such supervision is often unrealistic since many common factors of variation, such as 3D pose or lighting in image data, are difficult to annotate at scale in real-world settings. On the other hand, unsupervised learning makes the fewest limiting assumptions about the data but does not allow control over the discovered factors [30]. Neither extreme, fully supervised or unsupervised, is practical for many real-world tasks.

As an alternative, we consider weak supervision in the form of set membership [9, 24], used in prior works though often only informally defined. To be specific, we assume access to subsets of training data within which some *inactive* factors of variation have fixed values and the remaining *active* factors freely vary for different elements of the subset. For example, consider the images of a synthetic car in set $\mathcal{A}$ of Fig. 1. All images in this set share common values for factors of variation relating to the specific car instance, and the only actively varying factor is the car's orientation in the image. Set membership is the only information; there are no annotations on any factors of variation. In many complex tasks that are beyond the scope of categorical classification, set supervision serves as a more flexible source of information for operating on factors of variation across a dataset.

Many techniques designed to utilize set supervision exploit correspondence across data that match in desired factors of variation [7, 54]. For instance, if images of cars with the same 3D pose are grouped together (i.e. the inactive factor in each set is pose), then a straightforward training objective that maps images within groups to similar embeddings and images from different groups to dissimilar embeddings will have successfully isolated pose. However, in this scenario and more generally, this variant of set supervision is often prohibitive to obtain: in our example it requires identifying images of different cars from exactly the same viewpoint.

A more readily available form of set supervision is where the desired factors are active in each set. Continuing the example, such supervision can be obtained by simply imaging each car from multiple viewpoints (as in set $\mathcal{A}$ in Fig. 1).

*Work done as part of the Google AI Residency program.
†seas.upenn.edu
‡google.com

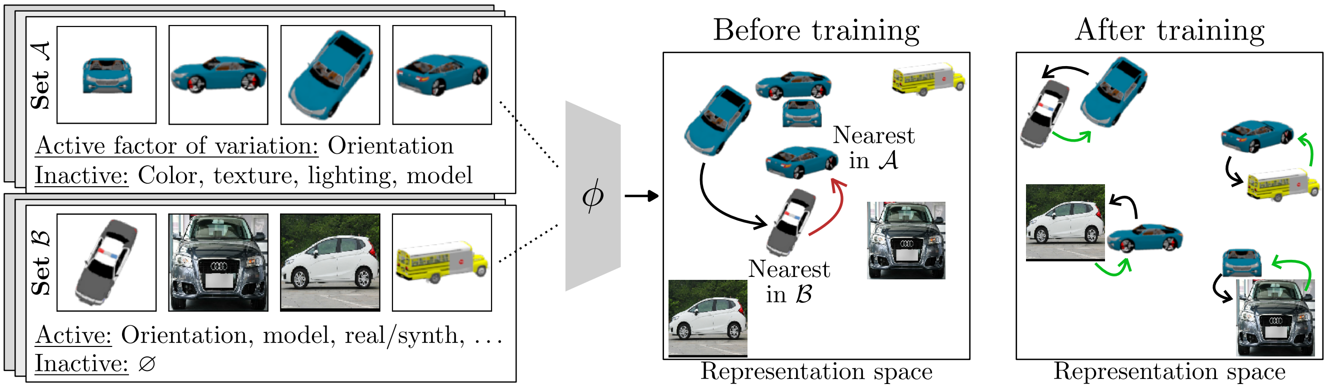


Figure 1. ***Approximate bijective correspondence (ABC).*** *Leveraging weak set supervision—merely groupings of data within which certain factors of variation are invariant—ABC isolates factors of variation which actively vary across sets. The images in set $\mathcal{A}$ (left) actively vary by only the orientation of the rendered car. We claim that if one-to-one correspondence can be found between $\mathcal{A}$ and $\mathcal{B}$, for all possible pairs $\mathcal{A}$ and $\mathcal{B}$, it must leverage orientation. We find this to be true even when only one of the sets in each pair is set-supervised. Importantly, this allows the incorporation of out-of-domain data with no supervision of any sort, such as the images of real cars in $\mathcal{B}$. By training a neural network ϕ with a loss that measures correspondence in representation space by the degree to which the nearest neighbor in $\mathcal{B}$ of a point in $\mathcal{A}$ (black arrow) is paired up with the same point in $\mathcal{A}$ (green arrow) or a different point in $\mathcal{A}$ (red arrow, middle), the learned representations (right) isolate the active factor of variation, orientation.*

This does not require correspondence in viewpoints across object instances, nor any pose values attached to the images. However, isolating the active factors (pose in this example) from set supervision is much harder, as there is no explicit correspondence in the desired factor (i.e., no matching images with identical pose information).

In this work, our goal is to operate in this more practical set-supervised setting, but the lack of correspondence in the desired active factors makes a solution difficult. To this end, we propose a novel approach, *approximate bijective correspondence* (ABC), which isolates the active factors through the process of finding correspondence between elements of different sets. To consistently yield correspondence across sets, learned representations must ignore invariant information within a set (inactive factors) and focus on active factors common to all sets. A powerful consequence is the capability to incorporate sets with extraneous active factors, including wholly unsupervised and even out-of-domain data (e.g., set $\mathcal{B}$ in Fig. 1), as long as one of the sets is more constrained (set $\mathcal{A}$ in Fig. 1). In the example of Fig. 1, ABC-learned embeddings isolate orientation, the common active factor across every pair of sets during training.

In our approach, each element of a set is paired with a cooresponding proxy element of another set constructed with a differentiable form of nearest neighbors [10, 14, 34, 40, 46]. The two serve as a positive pair for use in a standard contrastive (InfoNCE) loss [53]. We find that the same desirable properties of learned representations that optimize InfoNCE on explicitly provided positive pairs—namely, *alignment*, where differences within positive pairs are ignored, and *uniformity*, where maximal remaining information is retained [54, 57]—can be utilized to guide a network to find useful correspondences on its own. The key strengths of ABC are the following:

- **Isolates factors inaccessible to related methods.** ABC isolates the *active* factors of variation in set-supervised data, and suppresses the inactive factors.
- **Mixed-domain learning.** The ability to incorporate unsupervised data with extraneous factors of variation allows ABC to learn representations which bridge domain gaps with entirely unsupervised data from one domain.
- **Faster training.** ABC is much faster than alternative routes to isolating active factors from set-supervised data, all of which require learning the inactive factors as well.

We analyze the method and its strengths through experiments on a series of image datasets including Shapes3D [4] and MNIST [25]. In its fullest form, ABC addresses the challenging task of pose estimation in real images by meaningfully utilizing entirely unsupervised real images with set-supervised synthetic images, bridging the domain gap from synthetic to real. Our experiments show that ABC presents a viable path to learning 3D pose embeddings of real images of unseen objects without having access to any pose annotations during training. We conclude by training ABC with set-supervised real images, including one scenario matching the hypothetical example of images of cars taken from multiple viewpoints. ABC successfully isolates active factors of variation out of the many nuisance factors of variation common to natural images, all with access to only a limited quantity of training examples.

2. Related work

Isolating factors of variation. Recent work [30] has shown unsupervised disentanglement of latent factors to be impossible without incorporating some sort of supervision or inductive bias, spurring research into the best that can be achieved with different forms of supervision [29, 44, 45, 54]. A more realistic goal is the isolation of a subset of factors of variation, where learned representations are informative with respect to those factors and not others, with no guarantees about the structure of these factors in latent space.

Set supervision. Often, data is readily grouped into sets according to certain factors of variation without requiring explicit annotation on the factors. Generally, the methods harnessing information present in such groupings either **(i)** learn all factors and partition the representation such that one part is invariant across sets and the remaining part captures the intra-set (*active*) variation [8, 12, 21, 24, 32, 41], or **(ii)** learn the factors which are invariant (*inactive*) across sets [7, 51, 52, 55]. The methods of **(i)** almost always employ generative models, with the exception of [41], which grants it $6\times$ faster training over the VAE-based approach of [21]; the downside is the method of [41] requires seven networks and a two-stage, adversarial training process to learn first the inactive and then the active partitions of the representation. The methods of **(ii)** generally create subsets of data via augmentation [7, 16, 59] or pretraining tasks [33], or leverage multiple views of the same scene [43, 51], where semantic information is the target of training and is taken to be invariant across sets. By contrast, ABC directly learns *active* factors of variation across sets, offering a faster and simpler alternative to methods in **(i)** and tackling problems which are currently unassailable by methods in **(ii)**.

Videos, images, and point clouds are common forms of data which naturally offer set supervision. Approaches to find correspondence between frames of related videos, first using a discrete form of cycle consistency [1] and later a differentiable form [10], helped inspire this work. Cycle consistency has also been used to establish point correspondences in images [37, 63] and 3D point clouds [35, 36, 60]. In contrast to methods focusing on specific applications such as action progression in videos [10, 15] or robotics simulations [62], we present a general approach applicable to a broad class of problems.

Pose estimation and domain transfer. Although 3D pose estimation of objects in real images is an actively researched topic [27, 28, 31, 64], supervised pose estimation is difficult to deploy in practical scenarios due to the difficulty in obtaining accurate 3D pose labels at scale, and to annotation ambiguities caused by object symmetries. In light of the challenges posed by object symmetries, several methods attempt unsupervised learning of pose-aware embeddings rather than directly regressing absolute pose [48, 49]. In order to evaluate the learned representations, lookup into a codebook of images with known pose grants an estimate for each test image. Others have proposed to address domain transfer where models trained on synthetic but applied on real data [22, 39, 56]; however these methods operate in constrained settings such as where the same object instance is available at both test and train time (instance-based), or exploiting depth images or 3D models for inference. In contrast, our set-supervised method recovers pose embeddings on real images without using any pose annotations or seeing the same object instance at training time.

3. Methods

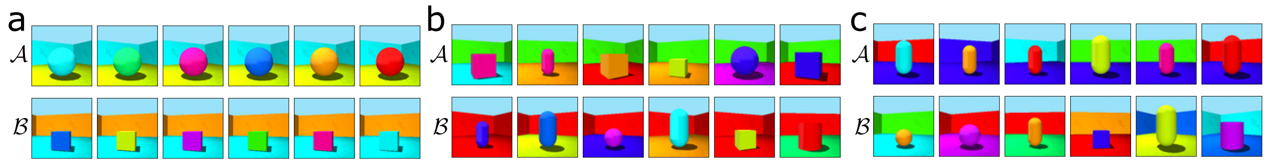
ABC uses set-supervised data, such that set membership is defined based on certain inactive factors; e.g., the data is grouped into sets such that all images in any given set have the same object class, making the object class an inactive factor. The basic idea of ABC is to consider all pairs of such sets (which have different values for the inactive factors of variation), and seek approximate correspondences among their elements through the learned representations. The guiding intuition is that this can only be achieved if representations use information about the active factors of variation present in every set and exclude all other information.

To be more concrete, let us consider the pose isolation task introduced earlier. Assume that a latent description of each image in Fig. 1 consists of the make and model of the car, all specifics relating to appearance, and the pose of the car in the image. With set-supervised data where the car instance specifics are the inactive factors within each set and the only active factor is pose (e.g., Set $\mathcal{A}$ in Fig. 1), ABC will pair elements across two sets which have similar pose.

3.1. The ABC Algorithm

Setup and notation: We follow the setup and notation from [54], which uses a latent variable model for the theoretical modeling of self-supervised learning methods. Let us denote an input image as x from the observation space $\mathcal{X}$ and an associated latent code as z from the representational space $\mathcal{Z}$. As per the latent variable model, the observations can be generated from the latent code using an invertible function $x = f(z)$, with $z \sim p_z$. Without loss of generality, we assume that the latent vector z can be partitioned into inactive z_i and active z_a components such that all elements within each set share identical z_i. Let $\phi(x) : \mathcal{X} \to \mathbb{R}^E$ be the function that maps the input vector to an embedding u in E-dimensional space. Our goal is to learn this function so that u may be informative with respect to the active partition z_a of the true underlying latent code z.

Formation of pairs of sets for training: We either leverage natural groupings of images or curate images into sets by controlling for certain factors of variation during mini-



	a Inactive	a Active	b Inactive	b Active	c Inactive	c Active
$\mathcal{A}$	WH, FH, Sc, Sh, Or	OH	WH	OH, FH, Sc, Sh, Or	FH, Sh	WH, OH, Sc, Or
$\mathcal{B}$	WH, FH, Sc, Sh, Or	OH	WH	OH, FH, Sc, Sh, Or	$\varnothing$	WH, FH, OH, Sc, Sh, Or

WH = Wall hue   OH = Object hue   FH = Floor hue   Sc = Scale   Sh = Shape   Or = Orientation

Figure 2. ***ABC isolates active factors in a broad range of set supervision scenarios****. We show an example pair of sets $\mathcal{A}$ and $\mathcal{B}$ which could arise in each of three set supervision scenarios on the Shapes3D dataset [4]. **(a)** In the case with five inactive factors for each set, there is only one factor to isolate and use to find correspondence: object hue. **(b)** The sets can be much less constrained, here defined by only a single inactive factor. In contrast to **(a)**, all active factors may not be needed to find correspondence between every pair of sets $\mathcal{A}$ and $\mathcal{B}$. **(c)** One set can have extraneous active factors, or even be completely unconstrained. In this case, correspondence is only found through active factors common to both sets, meaning floor hue and shape would not be isolated. In all three scenarios, ABC isolates factors which actively vary in both sets even though no correspondence is known* a priori *between images with matching active factors.*

batch construction, where each mini-batch consists of two such sets. For example, we show example sets with different active and inactive factors of variation curated from the Shapes3D dataset [4] in Fig. 2. Values for the inactive factors are randomly sampled and held fixed for each set, with the active factors free to vary (Fig. 2a,b).

Approach: Let the pair of sets for a particular mini-batch be given by $\mathcal{A} = \{a_1, \ldots, a_n\}$ and $\mathcal{B} = \{b_1, \ldots, b_m\}$, respectively. Let us denote the associated embeddings as $\mathcal{U} = \{u_1, \ldots, u_n\}$ and $\mathcal{V} = \{v_1, \ldots, v_m\}$, where $u_i = \phi(a_i, w)$ and $v_i = \phi(b_i, w)$. Functionally, we parameterize ϕ with the same neural network (with weights w) for both $\mathcal{A}$ and $\mathcal{B}$. Let $s(u, v)$ denote a similarity metric between points in embedding space, with $s(u, v) = s(v, u)$. To create an end-to-end differentiable loss, we use the soft nearest neighbor [10, 14, 34, 40, 46] to establish correspondence.

Definition 1 (*Soft nearest neighbor*) *Given a point u and a set of points $\mathcal{V} = \{v_1, \ldots, v_m\}$, the soft nearest neighbor of u in the set V is given by $\tilde{u} = \sum_{j=1}^m \alpha_j v_j$, where $\alpha_j = \frac{\exp(s(u_i,v_j)/\tau)}{\sum_{k=1}^m \exp(s(u_i,v_k)/\tau)}$ and τ is a temperature parameter.*

We first compute the soft nearest neighbor for each $u_i \in \mathcal{U}$ as $\tilde{u}_i = \sum_{j=1}^m \alpha_j v_j$. A soft bijective correspondence between the two sets is quantified through an InfoNCE loss [53], averaged over every element in each of the sets.

Definition 2 (*Approximate Bijective Correspondence loss*) *The correspondence loss from $\mathcal{U}$ to $\mathcal{V}$ is given by $\mathcal{L}(\mathcal{U}, \mathcal{V}) = -\frac{1}{n}\sum_i^n \log \frac{\exp(s(u_i,\tilde{u}_i)/\tau)}{\sum_j^n \exp(s(u_j,\tilde{u}_i)/\tau)}$. The full loss is the sum, $\mathcal{L} = \mathcal{L}(\mathcal{U}, \mathcal{V}) + \mathcal{L}(\mathcal{V}, \mathcal{U})$.*

The temperature parameter τ sets a length scale in embedding space as the natural units for the loss. It is generally unimportant when using an unbounded similarity metric such as negative Euclidean distance (Supp.). By contrast, a metric like cosine similarity benefits from tuning τ.

Double augmentation: We introduce a modification to the correspondence loss which allows suppression of factors of variation which can be augmented. We assume a group of transforms H is known to leave desired factors of variation unchanged [6, 7, 19]. We randomly sample two transforms $h^{(1)}, h^{(2)} \in H$ per image per training step. Let $u_i^{(1)} = \phi(h^{(1)} a_i, w)$ and similarly for $u_i^{(2)}$. The soft nearest neighbor is found using $u_i^{(1)}$, and then the correspondence is evaluated using $u_i^{(2)}$. The correspondence loss becomes $\mathcal{L}(\mathcal{U}, \mathcal{V}) = -\frac{1}{n}\sum_i^n \log \frac{\exp(s(u_i^{(2)},\tilde{u}_i^{(1)})/\tau)}{\sum_j^n \exp(s(u_j^{(2)},\tilde{u}_i^{(1)})/\tau)}$. The effect is to make the representations $u_i^{(1)}$ and $u_i^{(2)}$ invariant to the augmented factors of variation.

In summary, we sample pairs of sets for every mini-batch and learn an embedding network ϕ that produces embeddings which minimize the ABC loss through correspondence between elements in the sets. For every element in a set, the soft nearest neighbor serves as the correspondent point in the opposite set. The correspondence loss taken over both sets measures how close the correspondence is to being bijective.

3.2. Extensions

The ABC method can be extended to incorporate both fully unsupervised and supervised data.

***ABC-X* for incorporating unsupervised data:** Only the active factors of variation common to *both* sets are useful for establishing correspondence. Information about one set's inactive factor of variation cannot help distinguish between elements of that set and therefore cannot help form correspondence with elements of another, even if the factor actively varies in the second set. This has the powerful consequence that ABC can work just as well when one of the sets in each pair is completely unconstrained, as in Figs. 1 and 2c. *Wholly unsupervised, and even out-of-domain data with additional active factors, can be utilized.* We denote

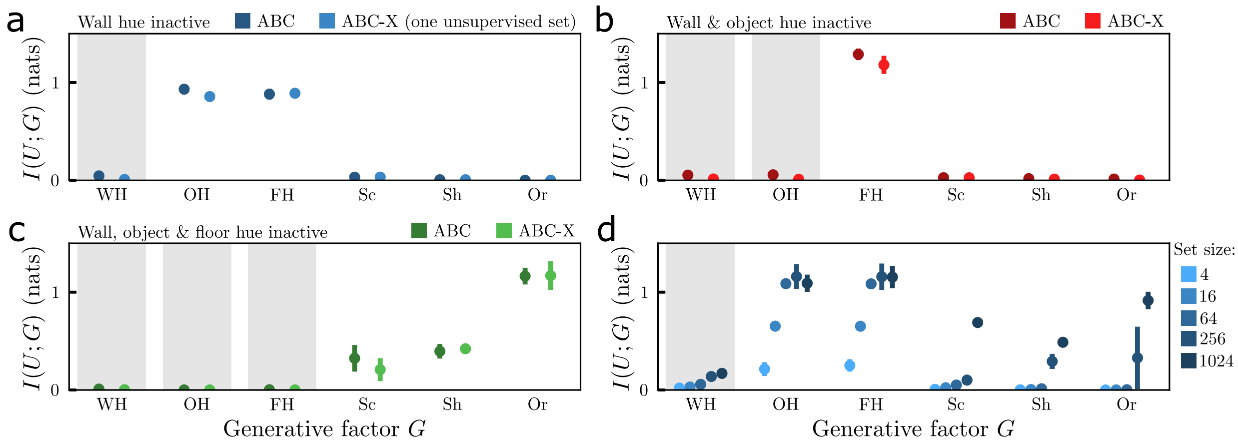


Figure 3. ***Factor isolation even with one unsupervised set; more factors isolated with larger set sizes during training.*** *We estimate the mutual information $I(U;G)$ between the learned representations and each of the generative factors using MINE [2]. Error bars display the standard deviation across ten random seeds. The inactive factors during training are indicated by shading.* ***(a-c)*** *We find the isolation of active factors to be unchanged when training with one of the two sets unsupervised (ABC-X).* ***(d)*** *Increasing the set size isolates more of the active factors of variation because finding correspondence requires more discerning power.*

this version of the method ABC-Extraneous, or *ABC-X*.

***ABC-S* for incorporating annotated data:** ABC can be organically applied to an intermediate representation space in a network trained with full supervision on a particular factor of variation, by training on a weighted sum of ABC and other losses. If set supervision is available with the supervised factor active, ABC can condition the intermediate representation space by isolating certain factors and suppressing others, and to incorporate unsupervised data. We denote this version of the method ABC-Supervised, or *ABC-S*.

3.3. ABC in the context of contrastive learning

While both ABC and self-supervised learning (SSL) methods such as SimCLR [7] use the InfoNCE loss on positive and negative pairs, a fundamental difference arises from how one acquires the pairs. In SSL a representation space is learned around explicitly provided positive pairs, obtained through augmentations known to affect certain factors while leaving others invariant. In ABC, a representation space is learned which also yields the positive pairs, as they are unknown *a priori* and must be formed by matching nearby embeddings across sets for every evaluation of the loss. ABC finds representations that produce good positive pairs, and does so by isolating the active factors, i.e., style, which would be inaccessible to general SSL methods. ABC can thus be seen as complementary to common SSL methods.

4. Experiments

We probe the method in four arenas. First, we leverage complete knowledge of generative factors in the artificial Shapes3D dataset [4] in order to vary the specifics of set supervision, and precisely illustrate ABC factor isolation by measuring the information content of learned representations. Second, we demonstrate a significant practical advantage of ABC—speed—by isolating style from class of MNIST digits [25]. Third, we tackle the challenge of pose estimation on real images with no pose annotations with ABC-X, utilizing only set supervision on synthetic images. Finally, with a limited quantity of set-supervised real images, ABC is shown to successfully isolate active factors of variation in the midst of many challenging nuisance factors. Implementation details and extended experiments may be found in the Supp.

4.1. Systematic evaluations on Shapes3D

Images from the Shapes3D dataset consist of a geometric primitive with a floor and background wall (Fig. 2). There are six factors of variation in the dataset: three color factors (wall, object and floor hue) and three geometric factors (scale, shape and orientation). Images were grouped with certain generative factors held inactive for each of many different training scenarios in Fig. 3; no augmentations were used.

We probed ABC-learned representations through the mutual information $I(U;G)$ between representations U and known latent factors G, estimated using MINE [2] and averaged over ten runs each. Deterministic networks generally preserve all information between input and output, so noise was added for a meaningful quantity $I(U+\eta;G)$, with $\eta \sim \mathcal{N}(0,\, \sigma^2)$ [11,42]. In the case where $s(u,v)$ is negative Euclidean distance, τ serves as a natural length scale of the correspondence loss so we used $\sigma = \tau$ (Supp.). We discuss noteworthy aspects of learned representations below.

All inactive factors are suppressed; a subset of active factors are isolated: In Fig. 3 information with respect to all inactive factors was suppressed, and a subset of active factors—not necessarily all—were isolated. Only when all three hue factors were inactive (Fig. 3c) were the geometric factors present in the learned representations. Presumably the hue factors are easier to learn and serve as shortcuts [52], allowing the representations to ignore other factors.

Semi-supervised ABC-X is equally effective: Correspondence is found through active factors common to both sets, which means if one set consistently has additional active factors, they will not be useful for optimizing the ABC loss. In semi-supervised scenarios with one set-supervised set per mini-batch and the other consisting of random samples over the entire dataset (e.g., Fig. 2c), ABC-X performed as well as ABC with full set supervision (Fig. 3a-c).

Increasing set size isolates more active factors: Intuitively, finding a one-to-one correspondence between sets with more elements requires more discerning power. In Fig. 3d, information in the learned representations about all active factors increased with the set size used during training. The set size effectively serves as the number of negative samples in the InfoNCE loss, and it has been found that more negative samples benefits contrastive learning [20].

4.2. Fast digit style isolation

Handwritten digits, such as from MNIST [25], have a natural partitioning of factors of variation into digit class (e.g., 2 or 8) and style (stroke width, slant, shape, etc.). Our goal is to learn style information generalized across digit class, without access to style annotations or images grouped with matching style. Images were grouped by class into sets of size 64 and embedded to $\mathbb{R}^8$; no augmentations were used.

ABC-learned embeddings of the digit 9—withheld during training—organized according to stroke thickness and slant (Fig. 4a), demonstrating generalization of isolated style information across digit classes. In Fig. 4b-d we retrieved the most similar digits of each class to a set of test digits. Without having to learn a full description of the data, ABC yielded style-informative embeddings orders of magnitude faster than related approaches.

4.3. Pose transfer from synthetic to real images

We next utilized ABC-X for object pose estimation in real images without pose annotations at training time. The goal was the effective isolation of pose information from set-supervised synthetic images, which generalizes to the category level and bridges the synthetic/real domain gap. The ability of ABC-X to handle extraneous active factors of variation in one set allowed the incorporation of unsupervised real images. This significantly extends ABC-X in Sec. 4.1 by introducing active factors of variation which do not exist in the synthetic domain (e.g. lighting effects, occlusions). The learned representations isolated pose, as the only factor actively varying across both sets in each training pair, while suppressing the additional domain-specific factors.

We used images of ShapeNet models [5] from viewpoints

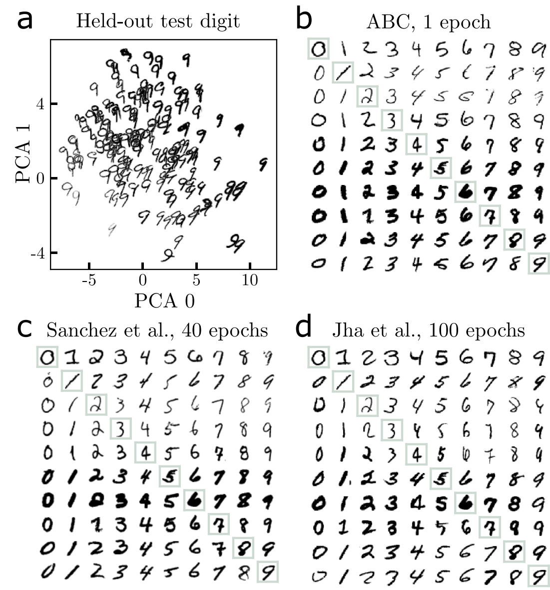


Figure 4. ***Fast style isolation from MNIST digits.*** *With digit class as the inactive factor during training, ABC isolated style.* ***(a)*** *Embeddings of the digit 9—withheld during training— fan out by thickness and slant, active factors common to all digit classes.* ***(b)*** *The boxed images along the diagonal were queries for retrieval from the test set; the other images in each row were the nearest embeddings per class. ABC isolated style information more than an order of magnitude faster than* ***(c)*** *the discriminative approach of [41] and* ***(d)*** *the VAE approach of [21].*

randomly distributed over the upper hemisphere [50]. Images were grouped in sets with their source 3D model inactive (as in set $\mathcal{A}$, Fig. 1). We gradually incorporated unsupervised real images from the CompCars [61] and Cars196 [23] datasets for the car category, and 1000 images from the Pascal3D+ [58] training split for chairs. We evaluated on the test split of Pascal3D+. All images were tight-cropped.

The augmentation loss (Sec. 3.2) helps bridge the domain gap by removing nuisance factors of variation which could shortcut the task of finding correspondence through pose [52]. Images were randomly augmented with cropping, recoloring, and painting the background with random crops from images of ImageNet-A [18], following augmentations used to bridge the synthetic/real domain gap in [48, 49]. Images were embedded to $\mathbb{R}^{64}$ using a few layers on top of an ImageNet-pre-trained ResNet50 [17]. We used cosine similarity with temperature $\tau = 0.1$ (ablations in Supp.).

4.3.1 Mixed-domain pose isolation

In the first experiment there were no pose annotations, for real nor synthetic images. The learned representations had no sense of absolute pose, but if pose information was successfully isolated then similar representations would have

	Dim ($\mathbb{R}^N$)	Cars				Chairs			
		Med (°) ↓	Acc. @10° ↑	Acc. @15° ↑	Acc. @30° ↑	Med (°) ↓	Acc. @10° ↑	Acc. @15° ↑	Acc. @30° ↑
CCVAE [21]	256	54.9	0.03	0.07	0.27	81.5	0.04	0.07	0.18
ML-VAE [12]	32	75.6	0.05	0.10	0.27	80.6	0.03	0.07	0.19
LORD [13]	128	71.3	0.09	0.15	0.32	89.8	0.03	0.05	0.15
ResNet	2048	85.3	0.07	0.14	0.28	80.7	0.04	0.07	0.19
ResNet-Intermediate	16,384	15.8	0.30	0.49	0.64	47.7	0.08	0.15	0.37
Set supervision w/ TCC loss [10]	64	23.1	0.14	0.29	0.59	58.3	0.09	0.16	0.40
Augmentation alone (with [7])	64	80.2	0.16	0.24	0.33	84.4	0.04	0.09	0.21
ABC	64	15.1	0.34	0.50	0.65	22.1	0.17	0.33	0.60
ABC-X	64	13.0	0.37	0.56	0.73	16.8	0.27	0.45	0.74

Table 1. ***Pose estimation with no pose annotations at training, set supervision on synthetic images.*** *Median error and accuracies (the fraction of errors better than the threshold value) on the Pascal3D+ car and chair test sets. Pose estimates were obtained through nearest neighbor lookup into a 'codebook' of 1800 synthetic images with associated GT pose; reported values are the average over ten random codebooks. The full ABC-X method—able to suppress augmentable nuisance factors of variation and to utilize unannotated real images during training—outperformed everything else, particularly in the difficult chair category.*

	Cars		Chairs	
	Med (°) ↓	Acc. @30° ↑	Med (°) ↓	Acc. @30° ↑
Liao et al. [28]	12.3	0.85	30.8	0.49
+ ABC-S	11.0	0.79	28.1	0.52
+ ABC-SX	9.3	0.87	26.0	0.55

Table 2. ***Leveraging pose annotations on synthetic images, wholly unsupervised real images.*** *ABC-X is effective as an additional loss term when the data consists of annotated synthetic images and unannotated real images. It provided a means to incorporate the latter which helped bridge the domain gap.*

Figure 5. ***Retrieval from ABC-X and ResNet-Intermediate.*** *Given a query image from the Pascal3D+ test set, we display the nearest neighbors in embedding space, from 1800 ShapeNet images and the Pascal3D+ train split. The accuracy and visual diversity of the ABC-X retrievals illustrate effective isolation of pose information generalized across the category and the synthetic/real domain gap.*

similar pose, regardless of the instance-specific details or domain of the image. To assign a pose estimate to each image of the test set, we found the most similar synthetic image (in representation space) out of a pool of 1800, unseen at training, each with associated ground-truth pose. We compare ABC with the VAE-based approaches of [21] and [12], the latent optimization method of [13], and to feature vectors of a pre-trained ResNet (Table 1). We found that an intermediate output (ResNet-Intermediate), though impractical due to its high dimensionality, is a surprisingly effective baseline.

While all methods were performant when tested in the synthetic domain (Supp.), most have no means of utilizing the unsupervised real images to bridge the domain gap and consequently performed poorly when tested on real images. Ablative comparisons illustrate the synergy of the components of ABC-X. Applying only the correspondence loss used in a limited setting of video alignment by [10] (TCC), we found reasonable performance on the car category but a failure to isolate pose in chairs. Suppressing irrelevant factors from the representations via augmentation without seeking correspondence did not isolate pose for either category. The incorporation of real images in ABC-X, ramped linearly to an average of 10% per set $\mathcal{B}$ by the end of training, boosted performance over ABC. Retrieval examples (Fig. 5) qualitatively illustrate the generalization across instance and domain-specific factors of variation.

4.3.2 Boosting cross-domain pose regression

We next seek to regress the pose of objects in the real domain given pose annotations in the synthetic domain only,

Figure 6. ***Active factor isolation with real images only.*** *We trained ABC on the Extended YaleB face dataset with only 28 different identities (**left**), and the EPFL Multi-View Car dataset with only 20 different car instances (**right**). We compare to ResNet feature vectors, and display the nearest three neighbors (per method) to the query image out of the remainder of the dataset. Orange outlines for YaleB indicate all active factors exactly match the query image.*

assuming synthetic images can be grouped by instance as in Sec. 4.3.1. Starting with the spherical regression framework of [28] we incorporated ABC-SX to condition an intermediate representation space, as described in Sec. 3.2. We trained on a weighted sum of a regression loss on the pose annotations for the synthetic images, and the ABC-SX loss used for Sec. 4.3.1 (though with a subset of the augmentations). In principle, any typical supervised pose regression network can be integrated with ABC-S. We specifically used [28] as it has shown superior performance on supervised pose benchmarks, and in particular training with synthetic data (created by RenderForCNN [47]) mixed with real images.

Even without real images during training, ABC-S improved performance by better conditioning the intermediate latent space (Table 2). A further boost for both categories resulted from a small amount of real images (2%) folded in to ABC-SX gradually over training. Thus ABC-SX can be advantageous in scenarios where there is more supervision available than set supervision, here serving to help bridge the real/synthetic domain gap by encouraging the suppression of factors of variation irrelevant to pose estimation.

4.4. Set supervision in the wild

We conclude with experiments demonstrating active factor isolation with ABC from real images only. The data is more limited in quantity and plagued by nuisance factors (such as complex backgrounds) than when the training data can incorporate a wealth of synthetic examples. The Extended YaleB Face dataset [26] has three dominant factors of variation: face identity (of which there are only 28), face pose, and illumination orientation. We trained the augmentation variant of ABC with illumination as the only active factor, and with both illumination and face pose as the active factors, and compare retrieval results with those from a ResNet feature vector in Fig. 6. Because both factors have a discrete set of possible values, the retrieval results can match the query's active factors perfectly. We highlight perfect retrieval results in orange; ABC was remarkably successful because it can overlook person identity to find images which match in the active factor(s), something the ResNet representations failed to do.

We also trained ABC on the EPFL Multi-View Car dataset [38], consisting of images of 20 cars on turntables with different rotation speeds, backgrounds, camera focal lengths, and rotation ranges: the hypothetical example from the Introduction. The visual disparity of ABC-retrieval images in Fig. 6 compared to those of ResNet embeddings demonstrates the success of ABC at suppressing the many inactive factors of variation in this challenging dataset.

5. Discussion

The pursuit of bijective correspondence offers a powerful new foothold into factors of variation in learned representations. ABC is significantly faster than related approaches because a full description of the data is not needed; indeed, not even all active factors of variation need be isolated. The size of sets during training and augmentation serve as additional control over which factors of variation get isolated. ABC is well-suited for domain transfer scenarios where an abundance of unannotated real data is accompanied by related synthetic data. By finding its own positive pairs for use in a contrastive learning loss, ABC complements existing approaches by isolating active factors in set-supervised data.
Limitations: The task of finding correspondence does not require isolating all active factors of variation with limited set sizes, making it vulnerable to undesired 'easy' factors. One should incorporate augmentations on nuisance factors if possible, and carefully analyze learned representations.
Societal impact: This work is intentionally broad in its scope, and we have emphasized intuition and insight wherever possible to improve accessibility of this research.

References

[1] Yusuf Aytar, Tobias Pfaff, David Budden, Tom Le Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube, 2018. 3

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 10–15 Jul 2018. 5

[3] 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. 1

[4] Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018. 2, 4, 5

[5] 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. 6

[6] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. 4

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1, 3, 4, 5, 7

[8] Taco S. Cohen and Max Welling. Transformation properties of learned visual representations. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[9] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423, 2017. 1

[10] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. *CoRR*, 2019. 2, 3, 4, 7

[11] Adar Elad, Doron Haviv, Yochai Blau, and Tomer Michaeli. Direct validation of the information bottleneck principle for deep nets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 5

[12] Bouchacourt et al. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI*, 2018. 3, 7

[13] Aviv Gabbay and Yedid Hoshen. Demystifying inter-class disentanglement. In *International Conference on Learning Representations (ICLR)*, 2020. 7

[14] 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. 2, 4

[15] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N. Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5548–5558, June 2021. 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 6

[18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 6

[19] Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230, 2018. 1, 4

[20] Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019*. ICLR, April 2019. 6

[21] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and V. S. R. Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders, 2018. 3, 6, 7

[22] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. SSD-6D: making rgb-based 3d detection and 6d pose estimation great again. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 1530–1538, 2017. 3

[23] 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. 6

[24] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015. 1, 3

[25] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998. 2, 5, 6

[26] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005. 8

[27] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An Analysis of SVD for Deep Rotation Estimation. In *Advances in Neural Information Processing Systems 34*, 2020. 3

[28] Shuai Liao, Efstratios Gavves, and Cees G. M. Snoek. Spherical regression: Learning viewpoints, surface normals and 3d rotations on n-spheres. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 7, 8

[29] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 2018. 3

[30] 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. 1, 3

[31] S. Mahendran, H. Ali, and R. Vidal. 3d pose regression using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 3

[32] 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*, page 5047–5055, 2016. 3

[33] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations, 2019. 3

[34] 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. 2, 4

[35] 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. 3

[36] Natalia Neverova, Artsiom Sanakoyeu, Patrick Labatut, David Novotny, and Andrea Vedaldi. Discovering relationships between object categories via universal canonical maps. In *CVPR*, 2021. 3

[37] Shaul Oron, Tali Dekel, Tianfan Xue, William T. Freeman, and Shai Avidan. Best-buddies similarity - robust template matching using mutual nearest neighbors, 2016. 3

[38] Mustafa Ozuysal, Vincent Lepetit, and Pascal Fua. Pose estimation for category specific multiview object localization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 778–785, 2009. 8

[39] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Domain Transfer for 3D Pose Estimation from Color Images without Manual Annotations. In *ACCV*, 2018. Oral presentation. 3

[40] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Neighbourhood consensus networks. *CoRR*, abs/1810.10510, 2018. 2, 4

[41] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *The European Conference on Computer Vision (ECCV)*, 2020. 3, 6

[42] 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. 5

[43] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, 2018. 3

[44] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020. 3

[45] N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip H.S. Torr. Learning disentangled representations with semi-supervised deep generative models. In *NeurIPS*, 2017. 3

[46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 4

[47] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 8

[48] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 6

[49] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3, 6

[50] 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. 6

[51] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3

[52] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020. 1, 3, 6

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 2, 4

[54] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 3

[55] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. Nestedvae: Isolating common factors via weak supervi-

sion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2020. 3

[56] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *The European Conference on Computer Vision (ECCV)*, August 2020. 3

[57] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 13–18 Jul 2020. 2

[58] 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)*, pages 75–82, March 2014. 6

[59] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021. 3

[60] Lei Yang, Wenxi Liu, Zhiming Cui, Nenglun Chen, and Wenping Wang. Mapping in a cycle: Sinkhorn regularized unsupervised learning for point cloud shapes, 2020. 3

[61] 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 2015. 6

[62] Qiang Zhang, Tete Xiao, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Learning cross-domain correspondence for control with dynamics cycle-consistency. In *International Conference on Learning Representations*, 2021. 3

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

[64] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3