

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

A Computational Account Of Self-Supervised Visual Learning From Egocentric Object Play

Permalink

<https://escholarship.org/uc/item/7rq102zj>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors

Sanyal, Deepayan

Michelson, Joel

Yang, Yuan

et al.

Publication Date

2023

Peer reviewed

A Computational Account Of Self-Supervised Visual Learning From Egocentric Object Play

Deepayan Sanyal, Joel Michelson, Yuan Yang, James Ainooson & Maithilee Kunda

Department of Computer Science

Vanderbilt University

Nashville, TN 37212, USA

{deepayan.sanyal, joel.p.michelson, yuan.yang, james.ainooson, mkunda}@vanderbilt.edu

Abstract

Research in child development has shown that embodied experience handling physical objects contributes to many cognitive abilities, including visual learning. One characteristic of such experience is that the learner sees the same object from several different viewpoints. In this paper, we study how learning signals that equate different viewpoints—e.g., assigning similar representations to different views of a single object—can support robust visual learning. We use the Toybox dataset, which contains egocentric videos of humans manipulating different objects, and conduct experiments using a computer vision framework for self-supervised contrastive learning. We find that representations learned by equating different physical viewpoints of an object benefit downstream image classification accuracy. Further experiments show that this performance improvement is robust to variations in the gaps between viewpoints, and that the benefits transfer to several different image classification tasks.

Keywords: infant learning; embodied vision; machine learning.

Introduction

In interacting with the real-world, an individual’s experience is highly connected from one instant to the next. If someone is holding a spoon at one moment, it is likely that they will still be holding the same spoon in the next, possibly at a slightly different distance and hand/head/spoon pose. This physical continuity serves to generate a multitude of different views of the held object. Furthermore, the physical act of holding the object informs the learner that the sequence of differing views is tied to the same object, i.e. a form of object permanence. Even if the observer does not know that an object is a spoon, they understand that the object is the same across multiple moments in time. In this paper, we study whether this embodied experience of seeing different views of an object, and knowing that the views correspond to the same object, can provide a useful form of self-supervisory signal to enable visual learning in computational models.

There is a rich body of research studying the links between motor development in infants and their perceptual and cognitive abilities. Bushnell and Boudreau (1993) proposed that the progressive development of different motor abilities in infants leads to different schedules for various kinds of perceptual inputs; these, in turn, cause a temporal difference in the appearance of various cognitive abilities. Further studies have elaborated on the links between different kinds of perceptual inputs in development and the appearance of various cognitive skills (Needham, 2000; Libertus & Needham, 2010; Schwarzer et al., 2013; Baumgartner & Oakes, 2011).

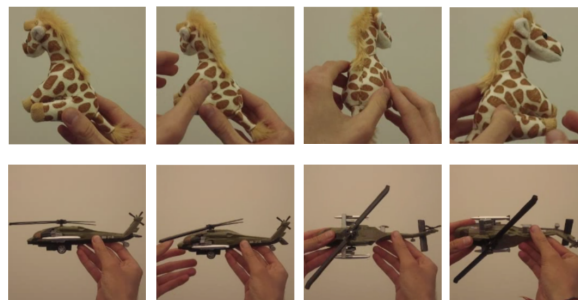


Figure 1: Visual experience during embodied object manipulation. Each row shows frames from an egocentric video of one object being manually rotated. Equating different physical views provides a strong learning signal.

Looking specifically at the ability to hold and manipulate objects, there is evidence that being able to perform hand-held object manipulations benefits several different cognitive abilities, such as learning nouns (Slone et al., 2019), visual understanding (Ruff, 1982; Soska et al., 2010) and understanding causality of actions (Rakison & Krogh, 2012). Recent research aiming to characterize infant visual experience using head-mounted cameras has found that first-person visual experience of manipulating objects during self-play constitutes a significant portion of the infants’ visual diets (Herzberg et al., 2022). In addition, there is considerable consistency in the distributions of object viewing experience across different cultures (Casey et al., 2022).

While previous studies have observed the importance of the embodied experiences generated by infants, the specific learning mechanisms that link these inputs (and their characteristics and distributions) to learning outcomes are not well understood. In this paper, we consider the visual experience that is generated during embodied manipulation of objects and propose a possible mechanism by which this experience helps develop good visual representations which support category learning. To do this, we use the SimCLR framework (Chen et al., 2020) which learns effective representations by maximizing the representational similarity between two differently-augmented versions of one image. This framework relies on instance-level similarity to learn representations, and a similar framework has recently been proposed to explain the representational goal of the visual system (Konkle & Alvarez, 2022). We hypothesize that natural visual expe-

rience provides stronger signals for this kind of learning. In this paper, we focus on the multi-view aspect of natural visual experience and show that access to different physical views of the same object leads to emergence of strong category structure.

Our work is also linked to research showing that temporal contiguity of visual experience can play a crucial role in learning invariant representations (Sprekeler, Michaelis, & Wiskott, 2007; Li & DiCarlo, 2010; Wood & Wood, 2018). Further, the development of such invariant object representations is not affected by reward (Li & DiCarlo, 2012), suggesting an unsupervised mechanism which regulates this kind of learning. For our part, we only consider the different views of an object that are generated during embodied manipulation of the object and show that equating these views presents a strong signal for category learning. Our contributions in this paper are:

- We demonstrate that representations learned by maximizing similarity between different physical views of the same object support strong performance on a subsequent classification task.
- We show that the representations are fairly robust to variations in the magnitude of difference between the paired object views utilized for learning.
- We demonstrate that these learned representations also successfully transfer to a diverse set of downstream classification tasks.

Related Work

There has been recent interest in using machine learning (ML) models to explain and understand different facets of human visual abilities as they relate to human visual experience. Bambach et al. (2018) used convolutional neural networks (CNN) to investigate the differences in the visual experiences of infants and adults and showed that an infants’ visual experience contains a more diverse range of views of objects, which lends itself to better object recognition performance. Stojanov et al. (2019) addressed the problem of learning object representations from incremental experience with individual objects and showed that repeated experiences with objects help ML models avoid problems related to catastrophic forgetting.

A recent work (Orhan et al., 2020) considered the problem of learning representations from infant headcamera recordings without explicit image labels. They used data from the SAYCam dataset (Sullivan et al., 2021), and showed that a learning signal based on temporal continuity enables learning representations that support image classification on the SAYCam and the Toybox datasets. While this work has similarities to our work, we focus on the visual experience that is generated during embodied object manipulation.

Other works have used CNNs to reason about the relationship between visual abilities in humans and limitations in visual experience; (Vogelsang et al., 2018) showed that CNNs can help explain deficits in configural face processing in chil-

dren born with congenital cataracts. Jang and Tong (2021) showed that while CNNs can be used to recreate differences between object and face processing, they do not yet account for robustness of adult vision to image blur.

Another relevant body of research is that of learning representations from visual data without explicit labels in the field of computer vision. Initial approaches for these methods used various pretext tasks such as image colorization (Zhang et al., 2016), predicting relative patches in images (Doersch et al., 2015), solving jigsaw puzzles (Noroozi & Favaro, 2016) and predicting rotations (Gidaris et al., 2018) to generate self-supervision. However, a recent body of work (Grill et al., 2020; Misra & Maaten, 2020; Chen et al., 2020) based on contrastive learning (Hadsell et al., 2006) has significantly outperformed those earlier approaches. Self-supervised approaches have also been applied to the problem of learning visual representations from videos (X. Wang & Gupta, 2015; J. Wang et al., 2020; Qian et al., 2021; Tschannen et al., 2020).

Our Approach

Dataset

Previous research has established differences between the distributional properties of infant visual experience and traditionally popular datasets used in the computer vision literature (Smith & Slone, 2017). Therefore, we used the Toybox (X. Wang et al., 2018) dataset, which was designed to contain more human-like continuous videos of egocentric handheld object manipulations. The dataset consists of 12 categories from 3 super-categories: household items (ball, cup, mug, spoon), animals (cat, duck, giraffe, horse) and vehicles (airplane, car, helicopter, truck). These 12 categories are among the most common early-learned nouns for children in the US (Fenson et al., 2007). For vehicle and animal categories, the objects in the dataset are either realistic, scaled-down models or toy objects. Fig 2 shows one object per category from the Toybox dataset.

The dataset consists of short videos, each of which shows



Figure 2: Examples of all 12 classes in the Toybox dataset: car, truck, helicopter, plane, ball, spoon, cup, mug, giraffe, horse, duck, cat. This figure shows full images; in our experiments, we used images cropped to their bounding boxes.

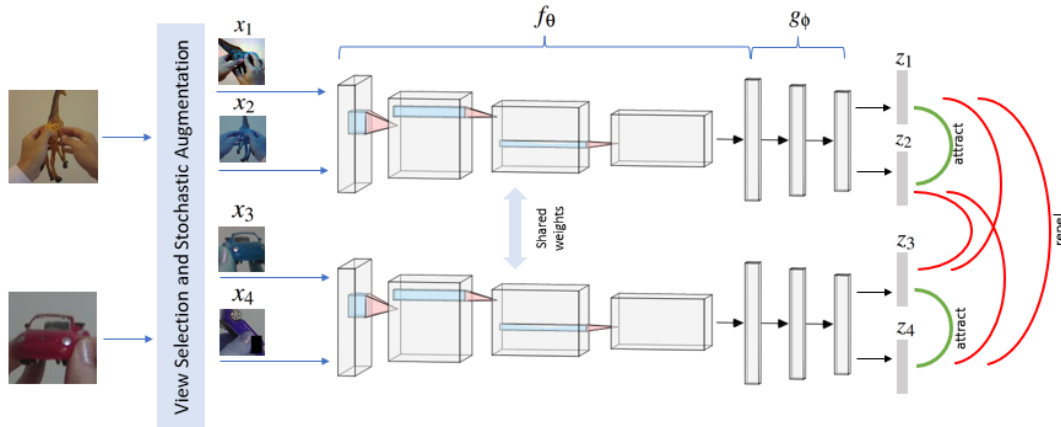


Figure 3: An overview of the learning framework with four images from a batch. Each of the four augmented images are run through the network and we obtain the feature vector z_i associated with each image. The contrastive learning signal then works by moving the positive image pairs closer together while pushing the negative image pairs further apart. The pairs of images linked by the green arcs represent the positive pairs, while the image-pairs linked by the red arcs represent the negative pairs.

one object being manipulated in one of several ways using an egocentric head-mounted wearable camera. The manipulations present in the dataset include systematic transformations, such as rotation and translation as well as random manipulations labeled as “hodgepodge” videos. Since our learning signal uses different viewpoints for each object, we use the 6 rotation videos (one around each axis in one direction) and the hodgepodge video. This gives us a total of 2520 videos for the 360 objects. Each video is about 20 seconds in length, and rotation videos contain two full revolutions around the specified axis and direction.

There are several interesting aspects of the Toybox dataset. First, since the objects are being manipulated by hand, the objects are often partially occluded by the subjects’ hands. Second, there are several views for each object, including a lot of non-canonical views. Third, unlike traditional ImageNet-style datasets which contain many thousands or millions of objects (with one image each), Toybox has images from a relatively small set of physical instances (30 objects per category) with a large number of images from each. Thus, it can be challenging for a learner to acquire category-general representations that are less sensitive to the idiosyncrasies of individual objects in the training data. However, these specific aspects of the dataset enable our experiments, since these properties also characterize the visual experience of infants.

Bounding box annotations at 1 fps are available for the rotation and the hodgepodge videos in the Toybox dataset. In order to maintain the original aspect ratios of the objects in the images, we extended the bounding boxes along their shorter dimension to match the size of the larger dimension. Cropping each image to this extended bounding box helps maximize the information content in the images while also preventing distortion of the images.

Method

SimCLR framework We use the paradigm of contrastive learning in our experiments, and particularly the SimCLR approach (Chen et al., 2020). The experiments progress in two steps:

1. *Self-supervised representation learning.* First, a CNN backbone (Lecun et al., 1998) is trained from scratch to learn image representations. During this phase of training, a base network f_θ is attached to a smaller projection network g_ϕ , and this combined network is trained using a self-supervised objective function.
2. *Representation evaluation using supervised learning.* In the second phase of training, called the linear evaluation phase, we throw away the projection network g_ϕ , the backbone f_θ is frozen and we attach a linear classifier fc on top of the backbone network. This linear classifier is then trained to perform image classification.

We now describe the learning signal used for training the network. During training, each minibatch M contains N pairs of images $\{x_{2i}, x_{2i+1}\}_{i=1}^N$. Each pair (x_{2i}, x_{2i+1}) forms a positive pair and all other image pairs (x_i, x_k) within M constitute the negative pairs. Each image is passed through the backbone and the projection network to obtain $z_i = g \circ f(x_i)$. The loss for one pair of positive images (x_i, x_j) is given by

$$l(i, j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

where $\text{sim}(u, v)$ represents the dot product $u \cdot v$, τ is the temperature parameter which modulates how sharp the similarity function is and $\mathbb{1}$ represents the indicator variable, which evaluates to 1 when $k \neq i$ and to 0 otherwise. The above loss function is called the NT-XEnt loss. For the entire minibatch, the loss function for all positive pairs is aggregated as:

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^N [l(2k, 2k+1) + l(2k+1, 2k)]$$

By minimizing the above loss function, the learning signal encourages the network to learn representations so that the positive image pairs are closer in the representation space, while the negative image pairs are further away. The effectiveness of the learning signal depends on the positive image pairs that are used. In the original paper (Chen et al., 2020), x_i and x_j are sourced from the same image with different amounts of stochastic image augmentation applied on them, thus telling the network to put differently augmented versions of the same image closer in the feature space compared to different images.

Modifications and Details¹ In our experiments, we investigate the extent to which having access to different physical views of the same object contributes to good representations through self-supervision. Thus, in addition to applying stochastic augmentations on the images, we vary the viewpoints from which the positive image pair are chosen. Thus, by equating these two different views, the underlying network learns to bring the representations of these views closer. Fig 3 provides an overview of our learning framework.

We use 27 objects from each Toybox class as the training set. During both phases of training, images from these 324 objects are used to train the network. Classification accuracies are reported on images from the remaining 3 objects. During the linear evaluation phase, in keeping with prior work, we use a randomly sampled 10% of the images to train the network. During both phases of training, we apply the following set of augmentations to all training images: color jitter, random grayscale, random crop, and random horizontal flip. No augmentations are applied to the images while calculating the accuracies. For our backbone f_θ , we use a ResNet-18 (He et al., 2016) and the projection head g_ϕ is a 2-layer neural network.

Experiment 1

As stated above, we vary the viewpoints that comprise each positive pair during training. In doing so, we are signalling that the different views are from the same object. To systematically study how this signal contributes to the learned representations, we use 5 different settings in our experiments:

1. **SimCLR + Self:** The positive pair is sourced from the same image frame with different image augmentations applied. This is the default setting for the SimCLR framework.
2. **SimCLR + Transform:** The Toybox dataset consists of 7 videos for each object. In this setting, the positive image pair are sourced from any one of those videos. Specifically, for every image in the dataset, we randomly sample another image from the same video to form the positive pair.

¹The code for these experiments can be found at: https://github.com/aivaslab/toybox_simclr

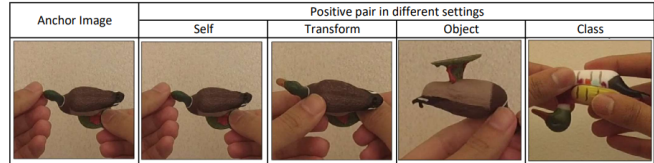


Figure 4: Image pairings used in different experiment settings. In all cases, the anchor image is paired with one other image. In the Self setting, the same image is reused. In the Transform setting, another image of the same object from the same video is selected as the pair. In the Object setting, the image pair can be any image from any of the videos of the object. In the Class setting, the only restriction is that the image pair need to belong to the same class. After the image pair is selected, stochastic image augmentation is applied to both to generate augmented images for learning.

3. **SimCLR + Object:** The positive pairs, in this setting, come from any videos of the same object.
4. **Supervised:** For baseline comparison, we train a network in a supervised setting on the training images from Toybox.
5. **SimCLR + Class:** As a second baseline, we use SimCLR with positive pairs formed by two images from any two objects of the same class. This setting uses the same information about category membership as the Supervised setting but modified to the SimCLR framework.

Fig 4 shows example image pairings used as positive pairs in these different settings. We observe that the difficulty of the self-supervised task increases from the Self setting to the Class setting as the visual dissimilarity between the positive image pair comes from a larger range.

Experimental Setting	Top-1 Accuracy
SimCLR + Self	46.54 (0.84)
SimCLR + Transform	73.83 (0.47)
SimCLR + Object	71.92 (1.13)
SimCLR + Class	69.13 (0.46)
Supervised	73.57 (1.42)

Table 1: Performance under different training settings. (Random guessing would yield 1/12, i.e., roughly 8.3% accuracy.) The best performance is shown by the learner in Transform setting and is comparable to the supervised learner. Accuracy drops off in the Object and Class settings. It is notable that the Transform setting exceeds the performance of the Self setting, which is the default for how SimCLR works. We report the mean and std over two runs with different random seeds.

Results Table 1 shows the results of our experiments in the different settings. We find that the default SimCLR setting achieves modest performance on the Toybox dataset. However, in both the Transform and the Object settings, the final accuracy approaches that of the supervised model. These accuracies show that the representations learned by equating

different views of the same object support good classification performance. What we find exciting in the results is that the Transform setting performs so well, despite learning from a weaker supervisory signal compared to the supervised model and the SimCLR models in the Object and Class settings. This seems to suggest that access to some form of viewpoint variation during training is extremely beneficial for the learned representations. We explore this more in Experiment 2.

The model trained in the Class setting did not perform as well as the Transform or the Object settings. This is likely because of the negative pairs: while we control which images form the positive pairs, the negative pairs are automatically decided during training. Because of this, several of the negative pairs are images from the same category. While this drawback is present in the other settings as well, the network seems to be able to handle them better in those settings. This robustness of the learning signal in the Transform and Object settings derives from the fact that in these cases, the chances of getting a negative pair which is more closely related than a positive pair are lower. Hence, the *false negative* pairs do not affect performance in these cases as much.

Experiment 2

In the previous experiment, we saw that the Transform model performs better than the Object model despite weaker learning signal from the positive pairs. In the current experiment, we wish to study how the visual dissimilarity between the images forming the positive pair affect the learned representations. We do this by carefully controlling the gap between the video frames which form the positive pairs. We focus on the *SimCLR+Transform* configuration in these experiments. We vary the gap between the frames in two settings: 1) Fixed: We fix the gap between the frames, i.e. we say that the two frames forming the positive pair have to be 2 or 4 seconds apart in the same Toybox video. 2) Range: We fix the maximum gap between the two frames, i.e. if we fix the gap to be 2s, the two frames can be 1s or 2s apart. In both settings, we increase the gap in steps of 2s from 0s to 10s and train the networks as described in the previous section. By varying the gap between frames, we can see how the distance in viewpoints for the positive pairs affects the learning performance. It should be noted that a gap of 0 in both settings corresponds to the *SimCLR+Self* model.

Results Table 2 shows our results for these experiments. We see that the *Range* setting seems to perform comparably with the *Fixed*, though it has more variation in the learning signal. This seems to indicate that there is enough variability that arises from the visual data itself which can lead to stronger learning. Further, we see that the performance in both settings remains in the same range even with decreasing gap between the positive pair.

To reduce the gap further, we used a version of the Toybox dataset sampled at 3fps. Since the bounding box annotations are done at 1fps, we use linear interpolation to obtain the an-

Gap b/w frames (seconds)	Setting	
	Fixed	Range
0	46.54 (0.84)	46.54 (0.84)
2	71.90 (0.84)	72.70 (0.89)
4	71.64 (0.43)	72.72 (0.51)
6	70.18 (2.09)	74.63 (1.04)
8	72.02 (0.52)	71.77 (0.62)
10	73.00 (0.42)	71.61 (0.32)

Table 2: Comparison of model performance using the *SimCLR + Transform* model in the Fixed and Range settings as the gap between frames is varied from 0 to 10 seconds. We report the mean and std over 2 runs.

Gap b/w frames (seconds)	Setting	
	Fixed	Range
0	48.94 (0.23)	48.94 (0.23)
0.67	74.05 (0.49)	72.04 (0.63)
1.33	71.27 (0.44)	72.16 (0.36)
2.00	70.73 (0.39)	70.64 (0.47)
2.67	72.64 (0.42)	75.93 (0.38)
3.33	75.69 (0.52)	73.87 (0.51)

Table 3: Comparison of model performance using the *SimCLR + Transform* model in the Fixed and Range settings as the gap between frames is varied from 0 to 3.33 seconds. This table uses images from the Toybox dataset extracted at 3fps. We report mean and std over 2 runs.

notations for the intermediate frames. Further, for this set of experiments, we used only the rotation videos. This allows us to avoid the randomness from the hodgepodge video and study the effect of viewpoint variation in a more structured and regular manner. We increase the gap parameter from 0s to 3.33s in steps of 0.67s. The other settings remain similar to the 1fps experiments. Table 3 shows our results in this setting. The first thing we note is that, because the total amount of training data increases close to 3-fold, the accuracy increases in both the *Self* and *Transform* settings. This is consistent with previous results in the machine learning literature showing that more data is generally beneficial. Secondly, we also note that the performance in both settings remains competitive even when the gap between frames is reduced to 0.67s. This demonstrates that the learning signal remains robust even when the gap between frames is reduced to 0.67s. With the Toybox videos, this gap corresponds to an average angular distance of 12° between viewpoints. These results suggest that during object manipulation, it is possible to leverage even small variations in viewing angles to learn good visual representations.

Experiment 3

In the previous experiments, we have seen that representations learned using self-supervision are beneficial for category learning on the Toybox dataset. In the final set of ex-

Model	Dataset				
	Cifar-10	Cifar-100	CORe50	ALOI	IN-12
SimCLR + Self	60.99 (0.76)	26.67 (0.66)	28.91 (0.55)	79.91 (0.16)	49.49 (0.25)
SimCLR + Transform	63.86 (0.11)	34.44 (0.13)	38.96 (0.62)	95.07 (0.12)	60.37 (1.79)
SimCLR + Object	63.11 (0.42)	34.22 (0.11)	36.35 (0.51)	95.47 (0.07)	60.16 (0.58)
SimCLR + Class	60.35 (1.51)	32.01 (0.22)	39.75 (0.05)	90.62 (0.09)	60.83 (1.33)

Table 4: Performance of the models trained with different learning signals on various transfer experiments

periments, we examine how these representations generalize to other kinds of classification tasks. Do the benefits we see by equating different physical views of objects in classifying Toybox images transfer to other datasets as well? To accommodate a variety of classification tasks, we use several downstream tasks to measure transfer performance. The phenomenon of machine learning methods developing bias towards their training dataset is well-documented (Torralba & Efros, 2011). Our aim in this set of experiments is to show that the benefit from using the learning signal is not limited only to the Toybox dataset, but extends to other datasets as well. We will refrain from providing a detailed description of the datasets, but will point out some aspects of the datasets which we find relevant for this paper.

In the computer vision community, use of large-scale datasets is mainstream. These datasets function as good test data to evaluate the generality of models. To include these kinds of datasets, we choose the CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009). While the CIFAR-10 dataset has some classes overlapping with the Toybox dataset, the CIFAR-100 dataset has classes of natural scenes and a much larger variety of classes than the Toybox dataset. While these internet-based datasets have a large number of instances for each class, there is usually only one image of each instance and it has been shown that the images in the dataset have a skewed distribution over viewpoints due to cameraman bias. To include more datasets where evaluation is done over multiple viewpoints, we include an object classification task on the CORe50 (Lomonaco & Maltoni, 2017) dataset and an instance classification task on the ALOI (Geusebroek et al., 2005) dataset. Finally, we examine if the representations learned from the Toybox dataset are transferable to real-world instances of the same categories. For this, we have curated the IN-12 dataset using images from the popular ImageNet (Deng et al., 2009) and MS-COCO (Lin et al., 2014) datasets for the Toybox classes. Specifically, we identify classes in the ImageNet dataset which overlap with the Toybox classes and randomly sample from each of these candidate classes to select 1700 images for each Toybox category. From these 1700 images, we use 1600 images per class for training and 100 images per class for testing the network.

Results

Table 4 shows our results for the transfer learning experiments. We see that the *Transform* model performs better than the *Self* model and is competitive with the *Object* models on

all the transfer tasks. The improvement in performance is strong for the datasets with multiple viewpoints (CORe50 and ALOI), thus showing that learning from multi-view egocentric experience of object manipulation benefits downstream performance for other multi-view datasets as well. The relative jump in performance is highest for CIFAR-100, thus demonstrating the general strength of the learned representations even for classification tasks where the image classes are vastly different. Looking at how the representations learned from the Toybox images transfer to real-world images from the same categories (IN-12 dataset), we find that similar trends hold in this case as well. It is interesting that even in these transfer conditions, the *Class* models generally perform worse than the *Transform* models, though it performs slightly better for the CORe50 dataset.

Conclusion and Discussion

We have considered the problem of learning from the visual experience of embodied object manipulation and proposed a mechanism by which good representations which support image classification can be learned. We do this by utilizing a learning signal which minimizes the representational distance between different physical views of the same object. Through our experiments, we showed that this signal enables learning good representations which support categorization. We further showed that this signal is robust to the magnitude of difference between the viewpoint-pair which generate the learning signal. Finally, we demonstrated that the generality of learning with this signal by showing that the learned model can transfer non-trivially to a diverse classification tasks.

Our work leads to several important questions that will be addressed in future work: 1) While our work shows the effectiveness of the learning signal for downstream classification tasks, research has shown that similar algorithms can lead to relevant information being lost in the model (Xiao et al., 2021). In order to understand the development of robust human vision that can perform diverse visual tasks, further research looking at the interaction between learning signals and the efficacy of the learned representations at different tasks needs to be done. 2) Our approach requires the use of strong image augmentations. This is likely due to the fact that CNNs can learn to use color histograms as a shortcut (Chen et al., 2020) during the self-supervised training and this problem is especially acute in the case of exemplar-based datasets like the Toybox dataset. Further research needs to be done to understand how the human visual system avoids such issues.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful and constructive comments.

References

- Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2018). Toddler-inspired visual object learning. *Neural Information Processing Systems (NeurIPS)*.
- Baumgartner, H. A., & Oakes, L. M. (2011). Infants' developing sensitivity to object function: Attention to features and feature correlations. *Journal of Cognition and Development, 12*(3), 275-298. Retrieved from <https://doi.org/10.1080/15248372.2010.542217> doi: 10.1080/15248372.2010.542217
- Bushnell, E. W., & Boudreau, J. P. (1993). Motor development and the mind: The potential role of motor abilities as a determinant of aspects of perceptual development. *Child development, 64*(4), 1005-1021.
- Casey, K., Elliott, M., Mickiewicz, E., Silva Mandujano, A., Shorter, K., Duquette, M., ... Casillas, M. (2022). Sticks, leaves, buckets, and bowls: Distributional patterns of children's at-home object handling in two subsistence societies. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning, 2020*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255).
- Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision* (pp. 1422-1430).
- Fenson, L., et al. (2007). *Macarthur-bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD.
- Geusebroek, J.-M., Burghouts, G. J., & Smeulders, A. W. (2005). The amsterdam library of object images. *International Journal of Computer Vision, 61*, 103-112.
- Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... others (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems, 33*, 21271-21284.
- Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (cvpr'06)* (Vol. 2, pp. 1735-1742).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- Herzberg, O., Fletcher, K. K., Schatz, J. L., Adolph, K. E., & Tamis-LeMonda, C. S. (2022). Infant exuberant object play at home: Immense amounts of time-distributed, variable practice. *Child development, 93*(1), 150-164.
- Jang, H., & Tong, F. (2021). Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of vision, 21*(12), 6-6.
- Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications, 13*(1), 491.
- Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324. doi: 10.1109/5.726791
- Li, N., & DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron, 67*(6), 1062-1075.
- Li, N., & DiCarlo, J. J. (2012). Neuronal learning of invariant object representation in the ventral visual stream is not dependent on reward. *Journal of Neuroscience, 32*(19), 6611-6620.
- Libertus, K., & Needham, A. (2010). Teach to reach: The effects of active vs. passive reaching experiences on action and perception. *Vision Research, 50*(24), 2750-2757. (Perception and Action: Part I) doi: <https://doi.org/10.1016/j.visres.2010.09.001>
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755).
- Lomonaco, V., & Maltoni, D. (2017). Core50: a new dataset and benchmark for continuous object recognition. In *Conference on robot learning* (pp. 17-26).
- Misra, I., & Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6707-6717).
- Needham, A. (2000). Improvements in object exploration skills may facilitate the development of object segregation in early infancy. *Journal of Cognition and Development, 1*(2), 131-156. doi: 10.1207/S15327647JCD010201
- Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision* (pp. 69-84).
- Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural*

- Information Processing Systems*, 33, 9960–9971.
- Qian, R., Meng, T., Gong, B., Yang, M.-H., Wang, H., Belongie, S., & Cui, Y. (2021). Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6964–6974).
- Rakison, D. H., & Krogh, L. (2012). Does causal action facilitate causal perception in infants younger than 6 months of age? *Developmental science*, 15(1), 43–53.
- Ruff, H. A. (1982). Role of manipulation in infants' responses to invariant properties of objects. *Developmental Psychology*, 18(5), 682.
- Schwarzer, G., Freitag, C., Buckel, R., & Lofruth, A. (2013). Crawling is associated with mental rotation ability by 9-month-old infants. *Infancy*, 18(3), 432–441. doi: <https://doi.org/10.1111/j.1532-7078.2012.00132.x>
- Slone, L. K., Smith, L. B., & Yu, C. (2019). Self-generated variability in object images predicts vocabulary growth. *Developmental science*, 22(6), e12816.
- Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in psychology*, 2124.
- Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: motor skill acquisition facilitates three-dimensional object completion. *Developmental psychology*, 46(1), 129.
- Sprekeler, H., Michaelis, C., & Wiskott, L. (2007). Slowness: An objective for spike-timing-dependent plasticity? *PLoS Computational Biology*, 3(6), e112.
- Stojanov, S., Mishra, S., Thai, N. A., Dhanda, N., Humayun, A., Yu, C., ... Rehg, J. M. (2019). Incremental object learning from contiguous views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8777–8786).
- Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021, 05). SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective. *Open Mind*, 5, 20–29.
- Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Cvpr 2011* (pp. 1521–1528).
- Tschannen, M., Djolonga, J., Ritter, M., Mahendran, A., Houlsby, N., Gelly, S., & Lucic, M. (2020). Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13806–13815).
- Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, 115(44), 11333–11338.
- Wang, J., Jiao, J., & Liu, Y.-H. (2020). Self-supervised video representation learning by pace prediction. In *Computer vision—eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xvii 16* (pp. 504–521).
- Wang, X., & Gupta, A. (2015). Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision* (pp. 2794–2802).
- Wang, X., Ma, T., Ainooson, J., Cha, S., Wang, X., Molla, A., & Kunda, M. (2018). The toybox dataset of egocentric visual object transformations. *arXiv preprint arXiv:1806.06034*.
- Wood, J. N., & Wood, S. M. (2018). The development of invariant object recognition requires visual experience with temporally smooth objects. *Cognitive Science*, 42(4), 1391–1406.
- Xiao, T., Wang, X., Efros, A. A., & Darrell, T. (2021). What should not be contrastive in contrastive learning. In *International conference on learning representations, 2021*.
- Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision* (pp. 649–666).