# On the Role of Pretraining in Domain Adaptation in an Infant-Inspired Distribution Shift Task

**Deepayan Sanyal**
Department of Computer Science
Vanderbilt University
Nashville, USA
deepayan.sanyal@vanderbilt.edu

**Joel Michelson**
Department of Computer Science
Vanderbilt University
Nashville, USA
joel.p.michelson@vanderbilt.edu

**Maithilee Kunda**
School of Informatics
University of Edinburgh
Edinburgh, UK
mkunda@ed.ac.uk

## Abstract

We study a novel distribution shift inspired by infant visual experience, which involves the tradeoff between viewpoint and instance diversity. To analyze this shift, we apply domain adaptation using Joint Adaptation Networks (JAN) under varying pretraining conditions. Our results show that JAN's performance is highly sensitive to the pretraining scheme, with notable drops when semantic information about the target dataset is absent during pretraining. To investigate this dependence, we introduce a metric that measures target category separability in the pretrained feature space. Using this metric, we demonstrate a strong correlation between target separability before domain adaptation and JAN's eventual performance on the target dataset.

## 1 Introduction

Infants' visual experiences are characterized by extended bouts of experience with a small number of familiar objects (e.g., toy ducks at home), with a large number of rarer exposures to less familiar objects (e.g., real ducks at the park) [Smith et al., 2018, Herzberg et al., 2022]. This pattern of exposure to instances of a particular category yields a long-tailed distribution, where some instances (e.g. their toys/household objects) are seen very frequently, while most instances (e.g. objects they see outdoors) are seen more rarely: **(1)** The *head* of the distribution is rich in the distribution of viewpoints, i.e. *viewpoint-dominated*, while **(2)** the *tail* of the distribution is rich in the number of different category instances, i.e. *instance-dominated* . In Machine Learning (ML) parlance, this constitutes a distribution shift. We term this the **VI-Shift** (see Sec 2.1).

In our research, we use the Domain Adaptation paradigm [Ben-David et al., 2010] to study how different learning signals can help bridge this distribution shift. It is customary to use ImageNet pretraining [Long et al., 2015, French et al., 2017] when training domain adaptation models. However, ImageNet pretraining provides access to advanced semantic knowledge which developing infants might not possess. Instead, we are interested in studying how viewpoint-dominated visual experience supports learning generalizable feature representations. We find that the performance of JAN [Long et al., 2017], a popular domain adaptation model varies strongly with different pretraining conditions. This finding leads us to two questions: (1) *Why does the performance of JAN vary so strongly under*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Interpreting Cognition in Deep Learning Models.

*different pretraining conditions?* (2) *What kind of inductive biases and learning signals can be leveraged to bridge this distribution gap?* In *this* work, we focus on the **first** of these questions.

Models like JAN rely on a global alignment approach: models are trained to align the source and target domain in the feature space. However, this does not necessarily lead to category-specific alignment across the two domains. **Under what conditions does a global alignment approach lead to category alignment?** We hypothesize that the *category separability* of the target dataset after pretraining (and *before* JAN training) affects the performance of JAN. In this paper,

1. We showed that performance of JAN is dependent on the pretraining method. Interestingly, JAN severely underperforms in our *developmentally-relevant* pretraining conditions.
2. We proposed a metric to measure category separability of the target dataset after pretraining. Using this metric, we showed that, for two pretraining conditions, category separability is strongly correlated with the performance of JAN.

## 2 Methods

### 2.1 VI-Shift

As the *viewpoint-dominated* dataset, we used the Toybox dataset [Wang et al., 2018] and for the *instance-dominated* dataset, we created the category-matched IN-12 dataset.

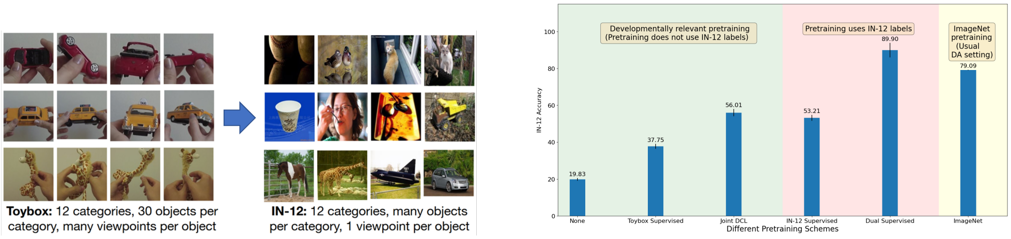
**Toybox dataset** The Toybox dataset contains short egocentric videos of objects being manipulated in different ways. The dataset contains 360 objects from 12 categories grouped into 3 super-categories: vehicles (airplanes, cars, helicopters, trucks), animals (cat, duck, giraffe, horse) and household objects (balls, cups, mugs, spoons); these categories correspond to early learned nouns among children in the US [Fenson, 2007]. For each object, the videos depict a wide variety of controlled, such as rotation, and random object manipulations (*hodgepodge*) yielding a wide range of viewpoints.

**IN-12 dataset** To create a category matched dataset for the Toybox dataset, we curated the IN-12 dataset from the ImageNet [Deng et al., 2009] and MS-COCO [Lin et al., 2014] datasets. First, we manually extracted all ImageNet classes corresponding to the 12 Toybox categories. From among these candidate classes, we select a few which describe the category at a general level (e.g. car vs police car). From these chosen classes, we randomly select 1600 images per class while ensuring that each candidate class contributed the same number of images. The entire list of the synsets are presented in Fig 5 in the Appendix. For the giraffe and helicopter categories, we extracted additional images from the MS-COCO dataset because the ImageNet synsets did not contain sufficient number of images. Fig 1a shows example images depicting this task.

### 2.2 Joint Adaptation Network (JAN)

JAN jointly minimizes a classification loss on the source dataset and alignment loss on the distribution of target features with the source features. The JAN alignment loss, called the JMMD loss, is based on Maximum Mean Discrepancy (MMD) [Gretton et al., 2012].

**Pretraining schemes for JAN** We use several different pretraining schemes, which can be grouped into 3 categories: (1) **ImageNet pretraining:** This is the default setting used in domain adaptation



(a) VI-Shift

(b) JAN performance with different pretraining

Figure 1: (a) An example of the infant-inspired VI-Shift problem, using the Toybox (left) and IN-12 (right) datasets. (b) Performance of JAN varies with different underlying pretraining schemes.

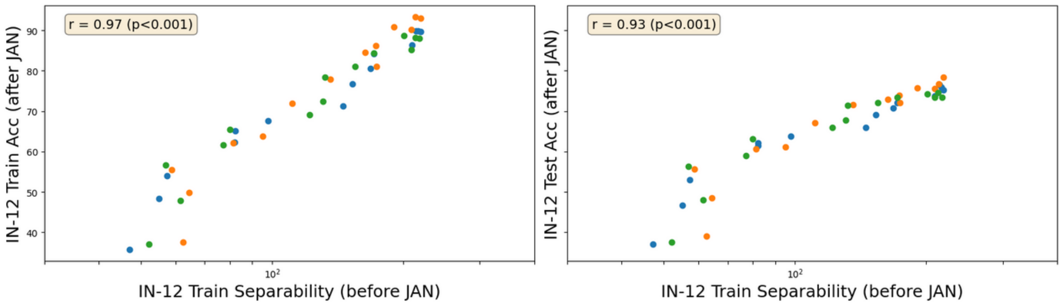


Figure 2: Scatter plot between JAN accuracy on IN-12 train/test sets and the target separability from IN-12 train set before JAN training after Dual Supervised pretraining. Text box shows the Pearson's correlation value and the p-value. There is a strong and significant correlation between *post-JAN* accuracy and *pre-JAN* separability. Different colors represent separate pretraining runs.

tasks. (2) **Pretraining using IN-12 labels:** This consists of two methods: in the first one, we train a network from scratch on IN-12 in supervised manner. In the second, we jointly train a network from scratch on both Toybox and IN-12 using supervision. (3) **Developmentally relevant pretraining:** In this setting, we do not use IN-12 labels. This group consists of no pretraining, supervised pretraining on Toybox and self-supervised pretraining on Toybox and IN-12 using Decoupled Contrastive Learning (DCL) [Yeh et al., 2022]. Further details about different pretraining schemes can be found in Section B. We use a ResNet-18 [He et al., 2016] in our experiments. More details about experiment settings and hyperparameter tuning are provided in Section C.

**JAN performance with different pretraining schemes** Figure 1b shows the performance of JAN on the VI-Shift task under different pretraining conditions. While the ImageNet-pretrained model performs strongly, the model pretrained on Toybox and IN-12 using supervised learning shows the best performance. Interestingly, only supervised training on IN-12 is insufficient for good performance. We see that all the *developmentally relevant* pretraining methods severely underperform ImageNet pretraining; even strong self-supervised methods like Decoupled Contrastive Learning performs poorly.

## 3 Relationship between target separability and JAN performance

**Earth Mover's Distance** The Earth Mover's Distance (EMD) is a metric for calculating distances between distributions and is the solution to the optimal transport problem. For two probability distributions $P$ and $Q$, it is defined as:

$$\text{EMD}(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[d(x, y)]$$

where $\Pi(P, Q)$ is the set of all distributions with marginals $P$ and $Q$.

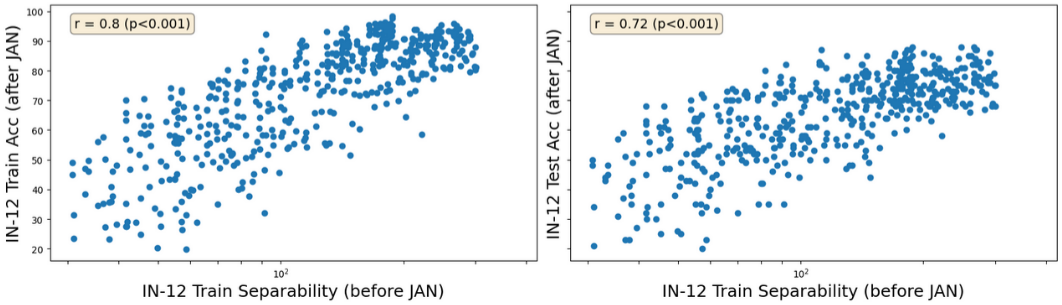


Figure 3: Scatter plot between JAN accuracy *per class* on IN-12 train/test sets and the target separability from IN-12 train set before JAN training after Dual Supervised pretraining. Text box shows the Pearson's correlation value and the p-value. There is a strong and significant correlation between *post-JAN* accuracy and *pre-JAN* separability.

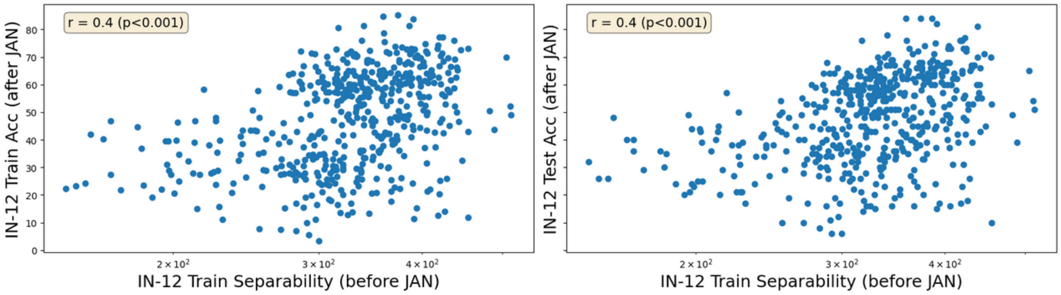


Figure 4: Scatter plot between JAN accuracy on IN-12 train/test sets and the target separability from IN-12 train set before JAN training after Joint DCL pretraining. Text box shows the Pearson's correlation value and the p-value.

**Separability** Measuring separability of a category $c$ requires aggregating the distance of $c$ from all other categories $c' \neq c$. Hence, we defined the separability $S(c)$ of a particular category $c \in C$ as:

$$S(c) = \frac{1}{|C| - 1} \sum_{c' \in C, c' \neq c} EMD(c, c')$$

**Experiment Details** To investigate the relationship between target category separability and JAN performance, we focused on the Joint DCL and Dual Supervised pretraining conditions. During pretraining, we saved the models at various stages of training. This yielded several models which display various levels of category separability. Subsequently, we trained JAN starting from each of these models. For our analysis, we investigated the dependence of JAN accuracy on the target category separability before JAN.

**Separability with Dual Supervised and DCL pretraining** Figure 2 shows the scatter plot between average accuracy on IN-12 post-JAN and the target separability on IN-12 pre-JAN for Dual Supervised pretraining; there is a strong dependence of JAN accuracy on the target separability before JAN. Figure 3 shows the plot between per-class accuracy on IN-12 post-JAN and the per-class separability on IN-12 pre-JAN for Dual Supervised pretraining. We see that the strong dependence persists even when we look at per-class accuracy, though the strength of the correlation is slightly weaker.

Figure 4 shows the scatter plot between per-class accuracy on IN-12 post-JAN and the per-class separability on IN-12 pre-JAN for Joint DCL pretraining. We find that there exists a moderate relationship between accuracy and separability, though the relationship is weaker.

## 4 Conclusion and Future Work

In this work, we introduced a metric to measure category separability in feature space and demonstrated a strong relationship between this separability and the accuracy of JAN on the target dataset. This analysis highlights the importance of understanding structural properties of the feature space prior to domain adaptation. Future work can extend this study to a broader range of distribution shifts, domain adaptation techniques, and network architectures, enabling a clearer assessment of the generality of the results. Additionally, the performance of JAN may be influenced by additional factors beyond target separability, such as the cross-domain alignment of categories; examining such factors would provide a more comprehensive account of what drives successful domain adaptation.

Returning to our motivation from infant vision, a key challenge is to design domain adaptation methods that remain effective even with weak pretraining. Insights from developmental psychology can provide valuable input in this regard. For instance, it has been argued that humans benefit from inductive biases, such as a preference for shape-based categorization, which promote the learning of generalizable representations. Incorporating such biases into neural network training could serve to build more robust and transferable features [Landau et al., 1988, Geirhos et al., 2018]. Moreover, empirical research suggests that infants leverage statistical regularities in their environment to learn strong representations [Bambach et al., 2018, Aubret et al., 2022]. Future work could investigate how analogous learning signals, derived from the statistical structure of the data, might be leveraged to improve domain adaptation under distribution shift. Pursuing this direction not only has the potential to yield stronger and more generalizable learning models but also provides a novel framework for testing hypotheses about the mechanisms of visual learning in infancy.

## References

Linda B Smith, Swapnaa Jayaraman, Elizabeth Clerkin, and Chen Yu. The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, 22(4):325–336, 2018.

Orit Herzberg, Katelyn K Fletcher, Jacob L Schatz, Karen E Adolph, and Catherine S Tamis-LeMonda. Infant exuberant object play at home: Immense amounts of time-distributed, variable practice. *Child development*, 93(1):150–164, 2022.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

Xiaohan Wang, Tengyu Ma, James Ainooson, Seunghwan Cha, Xiaotian Wang, Azhar Molla, and Maithilee Kunda. The toybox dataset of egocentric visual object transformations. *arXiv preprint arXiv:1806.06034*, 2018.

Larry Fenson. *MacArthur-Bates communicative development inventories*. Paul H. Brookes Publishing Company Baltimore, MD, 2007.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL `http://jmlr.org/papers/v13/gretton12a.html`.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European conference on computer vision*, pages 668–684. Springer, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Sven Bambach, David J Crandall, Linda B Smith, and Chen Yu. Toddler-inspired visual object learning. *Neural Information Processing Systems (NeurIPS)*, 2018.

Arthur Aubret, Markus Ernst, Céline Teulière, and Jochen Triesch. Time to augment self-supervised visual representation learning. *arXiv preprint arXiv:2207.13492*, 2022.

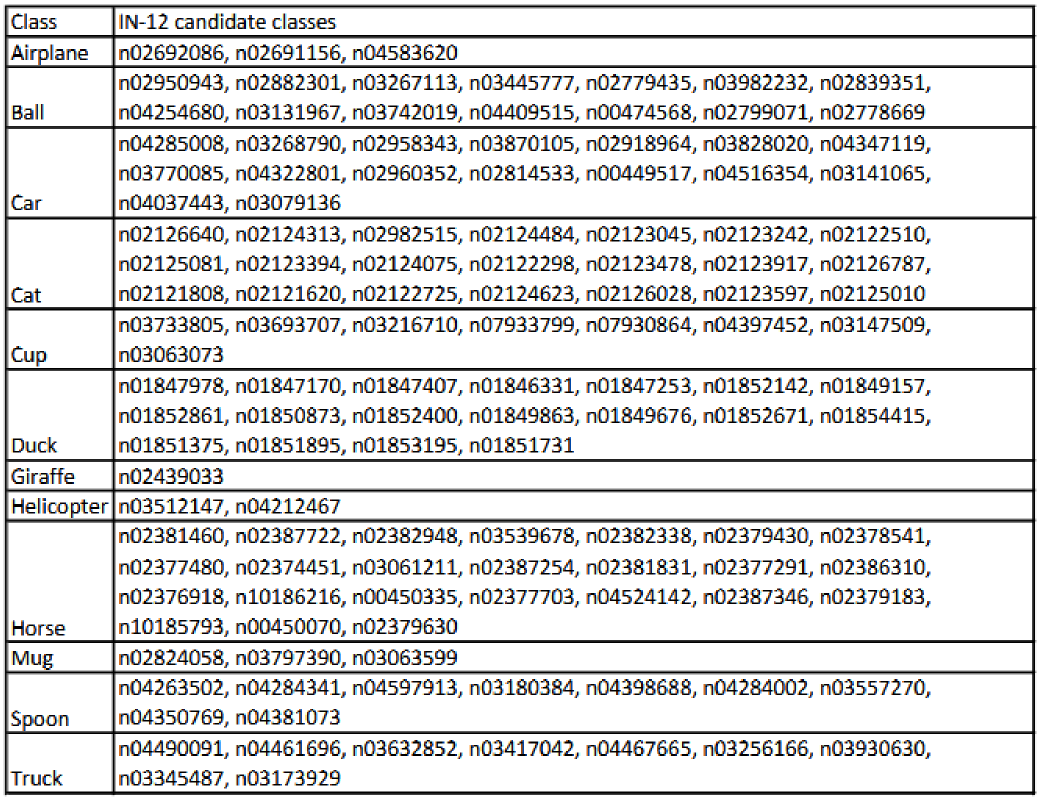
Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

## A List of ImageNet synsets used for IN-12 dataset

Fig 5 provides a list of the synsets used to compose the IN-12 dataset.



| Class | IN-12 candidate classes |
|---|---|
| Airplane | n02692086, n02691156, n04583620 |
| Ball | n02950943, n02882301, n03267113, n03445777, n02779435, n03982232, n02839351, n04254680, n03131967, n03742019, n04409515, n00474568, n02799071, n02778669 |
| Car | n04285008, n03268790, n02958343, n03870105, n02918964, n03828020, n04347119, n03770085, n04322801, n02960352, n02814533, n00449517, n04516354, n03141065, n04037443, n03079136 |
| Cat | n02126640, n02124313, n02982515, n02124484, n02123045, n02123242, n02122510, n02125081, n02123394, n02124075, n02122298, n02123478, n02123917, n02126787, n02121808, n02121620, n02122725, n02124623, n02126028, n02123597, n02125010 |
| Cup | n03733805, n03693707, n03216710, n07933799, n07930864, n04397452, n03147509, n03063073 |
| Duck | n01847978, n01847170, n01847407, n01846331, n01847253, n01852142, n01849157, n01852861, n01850873, n01852400, n01849863, n01849676, n01852671, n01854415, n01851375, n01851895, n01853195, n01851731 |
| Giraffe | n02439033 |
| Helicopter | n03512147, n04212467 |
| Horse | n02381460, n02387722, n02382948, n03539678, n02382338, n02379430, n02378541, n02377480, n02374451, n03061211, n02387254, n02381831, n02377291, n02386310, n02376918, n10186216, n00450335, n02377703, n04524142, n02387346, n02379183, n10185793, n00450070, n02379630 |
| Mug | n02824058, n03797390, n03063599 |
| Spoon | n04263502, n04284341, n04597913, n03180384, n04398688, n04284002, n03557270, n04350769, n04381073 |
| Truck | n04490091, n04461696, n03632852, n03417042, n04467665, n03256166, n03930630, n03345487, n03173929 |

Figure 5: Candidate classes from the ImageNet dataset used to create the IN-12 dataset

## B Pretraining Schemes for JAN

We use the following pretraining schemes:

(1) Random Initialization: We initialize the network with random weights and apply JAN/DANN directly.

(2) Toybox Supervised: We pretrain the network using supervised learning on Toybox.

(3) Joint DCL: We use contrastive to pretrain the network using both Toybox and IN-12 images, using the Decoupled Contrastive Loss (DCL) method [Yeh et al., 2022]. Since, we assume that labels for Toybox are available, we modify the DCL learning signal so that positive pairs for Toybox can belong to different objects from the same category.

(4) IN-12 Supervised: We pretrain the network directly on the target IN-12 dataset using supervised learning.

(5) Joint Supervised: We pretrain the network by training it jointly on both Toybox and IN-12 using supervised learning.

(6) ILSVRC pretraining: DANN/JAN training starts from a model previously trained on the ImageNet dataset. This is the default experimental setting in ML.

# C Training Details

## C.1 Pretraining Experiment Details

We use a ResNet-18 He et al. [2016] backbone in our experiments. For the pretraining methods that require training, we initialize the network with the Xavier initialization [Glorot and Bengio, 2010] and train the networks from scratch on each of the different experimental settings. We use the Adam optimizer [Kingma and Ba, 2014] for training the network. During training, we linearly increase the learning rate for the first 2 epochs of training and then decay the learning rate using a cosine decay schedule [Loshchilov and Hutter, 2016] without any restarts.

## C.2 Hyperparameter tuning details for joint DCL experiments

All models are trained for 100 epochs with an initial learning rate of 0.015 and a cosine decay schedule without restarts. Batch size was set to 256. The relative weight of the Toybox DCL loss was set to 0.25. A larger value was found to reduce the effectiveness of the IN-12 DCL signal, while a smaller weight hampered separability of Toybox clusters.

## C.3 JAN training details

For JAN, we initialize a bottleneck layer with 512 neurons. We follow the default training specifications provided in the JAN paper: the learning rate follows the schedule given by $\eta_p = 0.01(1 + 10p)^{-0.75}$, where p increases from 0 to 1 during training. The relative weight of the $l_{mmd}$ loss increases from 0 to 1 following $\lambda_p = \frac{2}{1+\exp(-10p)} - 1$. Each network is trained for 100 epochs with 100 minibatches per epoch using the SGD optimizer.