Cognition 245 (2024) 105690

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

Spatial relation categorization in infants and deep neural networks

Guy Davidson [a,*], A. Emin Orhan [a], Brenden M. Lake [a,b]

[a] Center for Data Science, New York University, United States of America
[b] Department of Psychology, New York University, United States of America

ARTICLE INFO

Keywords:
Spatial relation categorization
Cognitive development
Neural Networks
Connectionist models
Pretrained computer vision models

ABSTRACT

Spatial relations, such as above, below, between, and containment, are important mediators in children's understanding of the world (Piaget, 1954). The development of these relational categories in infancy has been extensively studied (Quinn, 2003) yet little is known about their computational underpinnings. Using developmental tests, we examine the extent to which deep neural networks, pretrained on a standard vision benchmark or egocentric video captured from one baby's perspective, form categorical representations for visual stimuli depicting relations. Notably, the networks did not receive any explicit training on relations. We then analyze whether these networks recover similar patterns to ones identified in development, such as reproducing the relative difficulty of categorizing different spatial relations and different stimulus abstractions. We find that the networks we evaluate tend to recover many of the patterns observed with the simpler relations of "above versus below" or "between versus outside", but struggle to match developmental findings related to "containment". We identify factors in the choice of model architecture, pretraining data, and experimental design that contribute to the extent the networks match developmental patterns, and highlight experimental predictions made by our modeling results. Our results open the door to modeling infants' earliest categorization abilities with modern machine learning tools and demonstrate the utility and productivity of this approach.

1. Introduction

Our understanding of the visual world around us is mediated by spatial relations, as they help distinguish individual objects and combine them in order to understand visual scenes (Johnson, 2010; Piaget, 1954). A breadth of experimental work explored how infants form categories and make categorical judgments (Bomba & Siqueland, 1983; Eimas & Quinn, 1994; Younger & Cohen, 1985) and specifically how infants form category representations for spatial relations (Casasola & Cohen, 2002; Casasola et al., 2003; Quinn, 1994; Quinn et al., 1999). Despite the importance of relations, little computational work has examined how infants could learn to categorize spatial relations, and why some categories are acquired before others over the course of development.

Our goal in this article is not to build a bespoke model of spatial relation categorization, for example by fitting models to developmental data, or by training models to categorize between different relations. Instead, we identify several key findings in the development of relation learning, translate their experimental paradigms to tasks suited for modern deep neural networks, and investigate whether *absent any explicit relational training*, models can categorize between spatial relations such as "above versus below" or "between versus outside". Fig. 1 summarizes our approach. We find that the networks we study are capable of making such categorizations, albeit with substantial variation by the relation examined, the data on which models were trained, and other experimental factors. Given this success, we then evaluate whether the performance of models tracks with the developmental findings that motivated this work—that is, to what extent do the relations that infants acquire later in development also challenge models more. By examining these correspondences, we hope to explore potential computational mechanisms underlying the developmental findings as well as pathways for modern neural networks to be utilized and further developed as models of cognitive development.

Our work builds on neural network modeling traditions in both cognitive development and machine learning. Related prior work in cognitive development has used neural network models to study aspects of infant categorization (French et al., 2004; Mareschal et al., 2000) and spatial language (Regier, 1995), and other computer visions approaches to identify visual relations and capture development patterns (Ullman et al., 2019), usually with models tailored to the particular problems of interest. Here we pursue a different approach: we study the extent

* Corresponding author.
E-mail address: guy.davidson@nyu.edu (G. Davidson).

https://doi.org/10.1016/j.cognition.2023.105690
Received 22 April 2023; Received in revised form 6 December 2023; Accepted 7 December 2023
Available online 8 February 2024

Fig. 1. Spatial Relation Categorization in Infants and Deep Neural Networks. Left: after being familiarized with stimuli depicting a particular relation ("familiarization"), infants find novel stimuli depicting the same relation ("same relation") less surprising than stimuli depicting a different relation ("other relation") as measured by looking times (Quinn, 2003). Right: to evaluate neural networks using a similar paradigm, we present three stimuli to a model, extract a vector embedding for each stimulus, and examine whether the "familiarization" stimulus embedding is more similar to the "same relation" stimulus embedding or to the "other relation" stimulus embedding.

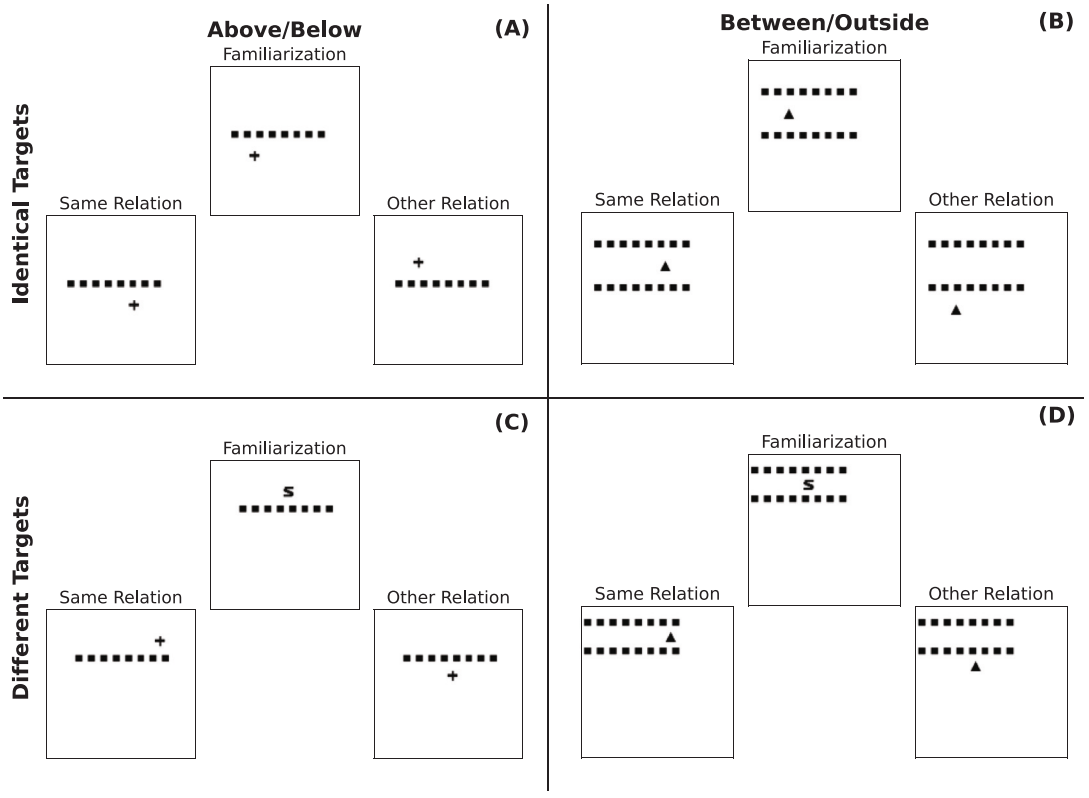

Fig. 2. Example Stimuli. In each triplet, the central stimulus is the familiarization example, the left one is the same relation test, and the right one is the other relation test. (A): *above/below* with identical target objects. (B): *between* with identical target objects. (C): *above/below* with identical target objects. (D): *between* different target objects.

to which powerful, general-purpose computer vision neural networks can account for phenomena of interest without being trained to do so. We focus our attention on convolutional neural networks, a class of computer vision architectures that have proven to be useful models in the cognitive sciences. For instance, Lindsay (2021) reviews their use as models of the visual system, and Battleday et al. (2021) study the extension of these models from the visual system toward higher-level cognitive capacities such as judgments of similarity and categorization. Machine learning work on relation learning using deep neural networks tends to focus on bespoke architectures, such as Relation Networks (Santoro et al., 2017) or PrediNet (Shanahan et al., 2019), and see Battaglia et al. (2018) for a review. Other recent work focuses on graph-based networks (Baldassarre et al., 2020) or on learning to generate images with particular relations (Liu et al., 2021). In comparison, our contribution is to evaluate the relatively generic neural network architectures, pretrained on two sources of realistic image data, on their representation of simple spatial relations without explicit training or architectural modifications.

We examine the extent to which models replicate several findings on the development of infant relation categorization, focusing on the relations "above versus below" and "between versus outside" (see Quinn (2003) for a review) and the "containment" relation (see review by Casasola, 2008). In a series of studies (Quinn, 1994, 2002, 2004; Quinn et al., 2003, 1996, 1999), Quinn and colleagues use similar methodologies to establish several patterns regarding the development of relational categories. Using stimuli similar to the one in Fig. 2,

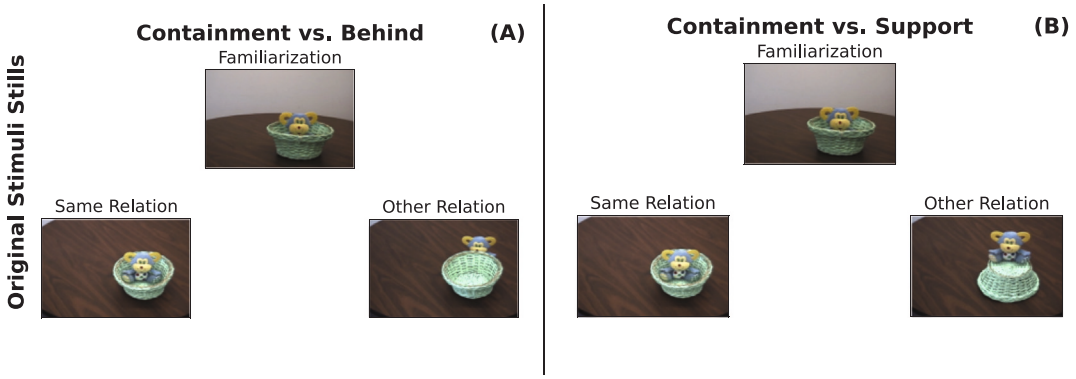

Fig. 3. Example stimuli from Casasola et al. (2003). We present the stimuli in a similar triplet form to the one used in Fig. 2. In each triplet, the central stimulus is the familiarization example, the left one is the same relation test, and the right one is the other relation test. We depict the final frames of the stimuli videos presented to infants in Casasola et al. (2003), reproduced from Casasola (2008, Figure 1). (A): comparing a test probe depicting the *containment* relation to a test probe depicting the *behind* relation. (B): comparing a test probe depicting the *containment* relation to a test probe depicting the *support* relation.

babies were familiarized with several stimuli of the type appearing in the middle of each triplet. The infants were then shown the two test stimuli, one depicting the same relation and one depicting the opposite relation. To establish the existence of a category representation, the studies measured the amount of time spent looking at the stimulus depicting the opposite relation, divided by the total looking time at both test stimuli. The higher this percentage is, the stronger a novelty preference (Fantz, 1964) the infant displays, and the more evidence it provides for a categorical representation of the familiarized relation.

Quinn (2003) surveys two primary findings. The first finding is that, by 3-4 months of age, infants can categorize "above versus below" (or "left versus right", Quinn, 2004), although they fail to categorize "between" (Fig. 2; (A) and (B)). By 6–7 months, infants can also categorize "between". In a representative experiment, Quinn (1994) familiarized infants with several stimuli, all containing a dot either above or below a horizontal bar (Fig. 2; Familiarization). After familiarization, infants were presented with a novel category preference test, finding that infants look longer at a stimulus with the dot on the other side of the bar (Fig. 2; Other relation) compared to a new location on the same side (Fig. 2; Same relation).

The second finding is that infants categorize spatial relations depicting specific objects before categorizing the same relations composed of varying objects. Quinn et al. (2003, 1996) replicate the previous experiments except that the target object varies between familiarization and test (Fig. 2; (C) and (D)). In both cases, changing the target object requires the infants to be older to show the same novelty preference—from 3–4 months to 6–7 months for above versus below, and from 6–7 months to 9–10 months for between versus outside.

A second line of work by Casasola and Cohen (2002) and Casasola et al. (2003) studies the emergence of the containment relation. In a representative experiment, Casasola et al. (2003, Experiment 2) examine infants' category for the *containment* relation (one object placed inside another object). The authors familiarized the infants with a video clip depicting the *containment* relation—one object being picked up and placed inside another one (whose final frame is represented in Fig. 3(A/B), "Familiarization"). Casasola et al. (2003) then tested the infants using three different test probes, all filmed from a different camera angle. The first probe also depicted a *containment* relation (Fig. 3(A/B), "Same Relation"). The second probe showed an object being picked up and placed *behind* another object (Fig. 3(A), "Other Relation" under "Containment vs. Behind"). The third and final test probe presented a *support* relation, with the object being picked up and placed on top of another one (Fig. 3(B), "Other Relation"). They find that even when controlling for the degree of object occlusion in their stimuli, infants reliably find the test probes depicting the *containment* relation as most similar to the familiarization probes, as measured by looking times. This is taken as evidence that the infants constructed a category representation of the *containment* relation.

In Experiments 1–4, we evaluate a collection of pretrained, large-scale computer vision neural network models on tasks inspired by the various developmental experiments and findings surveyed. We view the pretraining as a proxy for prior visual experience, and compare the experience gained from egocentric video capture based on one baby's experiences (SAYCam, Sullivan et al., 2020) to experience from a popular computer vision benchmark (ImageNet, Russakovsky et al., 2015), neither of which explicitly requires relational categorization:

1. In Experiment 1, we find that the models succeed in capturing developmental findings surveyed by Quinn (2003) using the *above/below* and *between* relations. We also compare the different model architectures and pretraining approaches and find that representations from the models trained on developmentally-realistic data appear to promote relational information more than the alternatives we evaluate.
2. In Experiment 2, we flip the relations on their side and evaluate the models on the relations *left/right* and *sideways between*. We find that our initial set of models fails to replicate the developmental findings of interest, and identify model training choices that explain the deviation and enable recovering the initial findings.
3. In Experiment 3, we examine the extent to which the networks' representations of these relations are sufficiently abstract to handle different types of stimuli. We do so by generating more complex three-dimensional scenes that more closely resemble real relation scenarios. We find success in replicating all findings of interest from Experiment 1, demonstrating that the relational representations are abstract enough to generalize to a substantially different class of visual stimuli.
4. In Experiment 4, we find that the networks we evaluate struggle to recover the relevant empirical patterns with the *containment*, *behind*, and *support* relations as described by Casasola et al. (2003). We explore these results to examine the extent to which the models we evaluate embed information about these more complex relations and discover that the information is still present and linearly decodable even when the models struggle on the task using a generic similarity metric.

We find that the pretrained visual representations are sufficient for the categorical perception of simple relations (*above/below* and *between*). Moreover, these representations are sufficiently abstract for handling both 2D and 3D stimuli. In the case of the more complex relations of *containment*, *behind*, and *support*, the embeddings contained sufficient information to linearly decode the relation with very high accuracies, even when representational similarity was not driven by the spatial

Table 1
Methodology comparison between the developmental literature we model and our formulation.

	Infant experiments	Our methodology
Participants	Infants, ranging from 3–4 months old (Quinn, 1994) to 9–10 months old (Quinn et al., 2003)	Deep neural network models, either pretrained on a computer vision dataset or randomly initialized and untrained.
Stimuli	Exemplars depicting a target object in a particular spatial relation (e.g. above, below, between, contained in) with respect to one or two reference objects.	Stimuli representing the same spatial categories, either rendered to abstract 2D representation (Experiments 1 & 2) or a richer 3D rendering (Experiments 3 & 4).
Familiarization examples	A small number (often four) of exemplars depicting a particular spatial relation, with the target object moved around a small radius maintaining the same relation.	A single rendered image of an object in a particular spatial relation, other than in Appendix B.3.2, where we use four familiarization examples instead.
Category representation	Formed by the infant, either before or during the lab study.	Embedding (latent representation) of the familiarization stimuli, extracted from the model without explicit relational training.
Test Probes	Two stimuli, one of the target object in a different position in the same relation, and one of the target object in a similar position in the opposite relation.	Functionally identical stimuli rendered using our methods.
Categorization measure	Relative looking times at the two test probes—longer looking times at opposite relation imply different category (e.g., Westermann & Mareschal, 2014).	Similarity between the embedding of the familiarization stimulus and the embeddings of test probes—higher similarity of same relation test probes implies the same category.
Relative difficulty of categories	Age range at which infants first display a categorical response for different conditions.	Relative difference in overall accuracy levels across different conditions.

relation. We conclude by attempting to identify useful methodological aspects to support future work and highlight current gaps and open questions.

2. General methodology

In translating the experimental setup used for infants to a task suitable for deep neural networks, we must contend with the fact that, unlike infants, neural networks do not get bored, and do not show a preference for novel stimuli (Fantz, 1964). The developmental results we seek to model (e.g. Casasola et al., 2003; Quinn, 1994) follow a novelty-preference methodology, which interprets longer looking times as evidence of an unfamiliar or novel stimulus (Fagan, 1970; Slater, 1995). That is, a preference for novel stimuli is based on "the infant comparing the currently available stimulus with a remembered stimulus" (Oakes, 2010), or more specifically, comparing the internal representations of the stimuli. Following this logic, when infants are simultaneously shown a pair of test stimuli, they will tend, all else equal, to look longer at the test item that is more different compared to the remembered familiarization stimuli.

In the model simulations, we compare the experimental stimuli using pre-trained deep neural networks, which have no specific training for relations, to examine the extent to which they can explain the behavioral findings. For the networks, the similarity of two visual stimuli is computed via their vector representations in a high-dimensional feature space (e.g., Jozwik et al., 2017). This feature space was shaped through pre-training by natural scene statistics in the environment of a developing child pre-training on egocentric recording from one young child; SAYCam or the goal of performing object recognition on a common computer vision benchmark (pre-training on ImageNet). We then evaluate whether the stimuli preferred by infants (as being more novel), on average, are more distinct from the familiarization stimuli for the networks. See comparison summary in Table 1, and additional details below.

Similarity and classification accuracy. We use cosine similarity[1] between the embeddings of the familiarization stimulus and the two possible test stimuli. We use classification accuracy, measured across many procedurally-generated variations of the stimuli, as an assessment of categorical perception (Goldstone & Hendrickson, 2010): that is, we consider a trial to be accurately classified when the model embeds the two congruent images (depicting the same relation) more similarly than the two incongruent images (depicting different relations), where the incongruent pair acts as a perceptual lure that matches in other non-categorical dimension(s).

Relative difficulty of categories. As a further step, assuming there is a categorical response, we examine whether capacities demonstrated by infants earlier in development are also easier for the networks we evaluate, which is an assumption we make in order to compare the developmental phenomena to model performance. For instance, given that Quinn (1994) demonstrated that infants acquire category representations for "above or below" earlier in development than for "between or outside", then we would examine whether the model is more capable in the *above/below* condition compared to the *between/outside* condition. We do not seek to map between models and infants at various ages or establish a correspondence between changes in accuracy and changes in age. Instead, we assume that relational categories that are more consistently captured in the model's embeddings, that is, are somehow more salient to these generic visual learners, would also be easier for infants to acquire. We examine this hypothesis throughout the four experiments detailed below.

Adaption during familiarization. To account for infants' adaption during familiarization, we do not adapt the representations of the large-scale pre-trained neural network models; instead, we operationalize adaptation as storing a representation of the familiarization stimuli (either via a prototype of a set of exemplars; Appendix B.3.2). Alternatively, to more explicitly model the familiarization process as in past work (French et al., 2004; Mareschal et al., 2000), a second neural network (e.g., an autoencoder) could be trained from scratch on the feature representation of familiarization stimuli, and then used as a means to measuring the perceived differences with the test stimuli. We did not pursue this approach here, instead opting for parameter-free comparison in the high-level feature space, although future work could add more detailed habituation/dishabituation modeling.

Neural networks with pre-trained weights. We explore a range of model configurations that are representative of current computer vision approaches, though far from exhaustive. We use three model architectures: two convolutional ones, MobileNetV2 (Sandler et al., 2018, see Experiment 1a for additional details) and ResNeXt (Xie et al., 2017,

[1] Given two embeddings vectors $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^D$ in a D-dimensional embedding space, their cosine similarity is defined as $S_{cos}(\mathbf{e}_1, \mathbf{e}_2) = \frac{\mathbf{e}_1 \cdot \mathbf{e}_2}{\|\mathbf{e}_1\| \|\mathbf{e}_2\|}$, that is, the angle between the feature vectors of two stimuli. Note that the cosine metric is often used to compare vector similarity, but it is far from the only such metric. In Appendix B.3.3 we explore using the Euclidean distance instead of the cosine similarity metric and find qualitatively similar results.

Experiment 1a), and one Vision Transformer, ViT-B/14 (Dosovitskiy et al., 2021, Experiment 1b). We use the two aforementioned datasets, ImageNet (Russakovsky et al., 2015, Experiment 1a) trained with supervised classification, and SAYCam (Sullivan et al., 2020, Experiment 1a) trained with temporal classification (Orhan et al., 2020, Experiment 1a). These datasets differ in their collection method (ImageNet: photographed images available on the internet; SAYCam: captures from head-mounted cameras on infants), and in several other important dimensions that follow from that, such as image quality, perspective, and the subject matter—imposing a limit on the strength and degree of control of the comparison between them. That said, to offer a closer control on the training methods, we also compare models trained on both datasets with the DINO algorithm (Caron et al., 2021, Experiment 1b). We deviate from the basic setup and set of models where necessary to provide a richer understanding of our results. Where appropriate, we compare the role of data augmentations (Experiment 2b), examine changes to our setup (Appendix B.3.3 studying an alternative distance metric, Appendix B.9 studying pre-pooling embeddings), and in one case, examine a linear decoding paradigm to further study a failure mode (Appendix B.10). To the extent our results identify any of these variations perform better or worse in our tasks, we do not consider as a demonstration they are better computer vision models at large; rather, we consider them a better match to developmental patterns and thus a more promising basis for future developmental modeling.

Neural networks with random weights. In addition to the pretrained models, we also evaluate randomly initialized versions of the same model architectures. The untrained models allow us to observe whether or not the inductive biases conveyed by the architecture alone are sufficient to embed objects in the same relation more similarly, as, e.g., Saxe et al. (2011) offer evidence that particular convolutional neural network architectures can offer effective embeddings even absent any training. In Experiments 1–3, we attempt to generate our triplets in a perceptually matched fashion (e.g., the target object in both test probes is approximately equidistant from the target object in the familiarization stimulus), and so we expect to find randomly initialized model accuracy to be at chance or slightly above. In Experiment 4, however, our stimuli are no longer perceptually matched, as we attempt to generate stimuli similar to Casasola et al. (2003). In that case, we hope the randomly initialized model accuracy levels offer a measure of baseline similarity between the familiarization stimuli and the test probes. That is, if the untrained models find one condition substantially easier or harder than chance, we should calibrate our understanding of the performance reached by our trained models accordingly.

3. Experiment 1: Classifying *above/below* and *between/outside* from 2D stimuli

We begin by studying the extent to which large-scale, pretrained computer vision models recover the two developmental findings reviewed by Quinn (2003). The first is that infants acquire a categorical representation for "above or below" earlier in development than for "between or outside", and the second is that infants acquire categorical representations for specific objects earlier in development than abstract categorical representations for varying objects.

3.1. Experiment 1a: Initial findings

In our first experiment, we use pretrained models to examine (a) whether the representations produced by these models capture the spatial relations, and (b) to what extent they recover the developmental findings of interest. This experiment varies several factors: computer vision architecture, pretraining dataset, and stimulus rendering details.

3.1.1. Methods

Model Architectures. We evaluate two computer vision architectures, to validate any findings we discover are not unique to a specific model and examine whether more performant architectures also fare better on our developmental comparison.

- MobileNetV2: This model aims to offer competitive performance with fairly limited computational resources (using only 3.4M parameters), offering an efficient trade-off between compute resources required and performance attained (Sandler et al., 2018).
- ResNeXt: This model is considered a highly capable computer vision backbone for various tasks (Xie et al., 2017). We use the ResNeXt-50 variety of this architecture, which uses 23M parameters.

As these models use substantially different parameter counts, the comparison between them provides evidence of the extent to which our results depend on model size. We visualize the ResNeXt architecture and where we extract our vector embeddings from in Fig. 4, and see Appendix A.1 for additional details.

Pretraining. We test the embeddings created by randomly initialized models (as outlined above) and compare them to models trained on two other datasets. One dataset and training approach reflects common practice in computer vision, while the other offers a closer comparison to a developing child:

- ImageNet: A landmark computer vision dataset, offering 1.2M images in 1000 object classes (Russakovsky et al., 2015). ImageNet does not correspond to an infant's natural experience but it is commonly used for general computer vision pretraining, offering a useful comparison. The ImageNet models were pretrained using the standard classification task as described in the torchvision documentation.[2]
- SAYCam: This dataset consists of longitudinal headcam videos from a small number of babies (Sullivan et al., 2020). This offers the opportunity to train vision models on a subset of the experience a child receives in development, albeit ranging to older ages than the infants studied in the experiments modeled. We utilize a pretrained network from Orhan et al. (2020) trained with temporal classification, a self-supervised learning algorithm inspired by psychologically plausible mechanisms. Temporal classification only makes use of the temporal ordering of data to supervise the learning process. We use models trained on a single child's footage (child S), approximately two hours per week while the child was between 6–30 months old, a total of 221 h.

Stimulus Generation. We synthesize custom stimuli to probe the model in this task (Fig. 2). We sample location(s) for the reference object(s) and then place the target objects relative to them. Similarly to Quinn (1994, 1996, 1999), we place the target object in one relation relative to the reference object in the familiarization example, and then place it in a different location in the same relation (first test probe) or in the other relation (second test probe). The target objects in the test probes are both equidistant from the target object in the familiarization probe, controlling for any effect of distance on the representational similarity. We examine triplets where the target object matches between the familiarization and probe stimuli ("identical targets"; Quinn, 1994; Fig. 2; (A)) and (B) and triplets where the probe stimuli use a different target object ("different targets"; Quinn et al., 1996; Fig. 2; (C) and (D)). We explore a few ways to render the reference and target objects, detailed in Appendix A.1. We render these stimuli to 224 × 224 pixel images.

Methods Summary. We evaluate models from two architectures, either randomly initialized or pretrained on one of two visual datasets, on two relations (*above/below* and *between*), using stimuli rendered with three different approaches. For each relation and rendering method, we sample 1024 triplets (identical for all models) and report the average accuracy for each model and pretraining setting—how often are the

[2] https://pytorch.org/vision/stable/models.html.

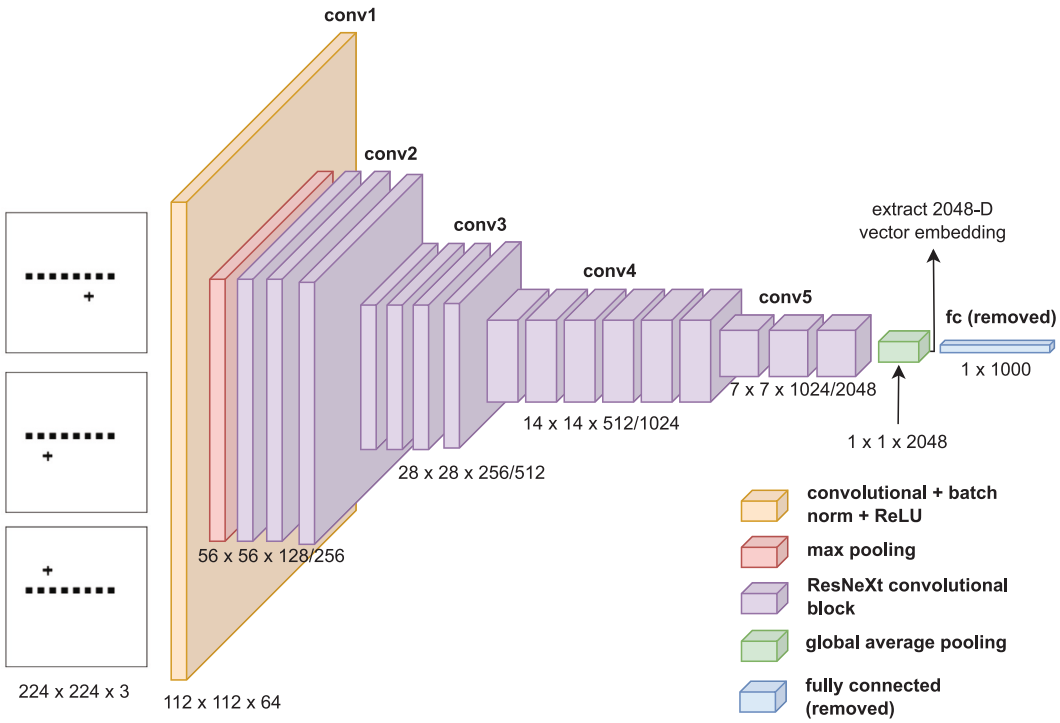

Fig. 4. ResNeXt model diagram. We pass 224 x 224 images into the model, which begins with a convolutional block (orange) and a pooling layer (red), and then proceeds to ResNeXt convolutional blocks (purple) operating on increasingly smaller representations of the input images (see Appendix A.1 for further details). We extract our vector embedding, which with this architecture has 2048 entries, after the global average pooling layer (in green). In a standard classification setting this embedding would be classified using a fully connected layer (blue), which we remove from the models we evaluate.

embeddings for the congruent pair of stimuli more similar (using cosine similarity) than the embeddings for the incongruent pair. For every set of results, we compute a mean accuracy and standard error of the mean (SEM) over the 1024 triplets, and below we report different aggregations of these mean accuracy measurements across experimental conditions of interest. We omit drawing error bars as the averaged SEMs all fall below 2% accuracy.

3.1.2. Results

A summary of the results is shown in Fig. 5 and Table 2. Without pretraining, the models performed near chance, with levels of accuracy ranging from 0.47 to 0.58. This suggests that inductive biases conferred by the architecture alone are insufficient for representing relations (see the results marked by an 'X' in Fig. 5), and confirms our stimuli are roughly perceptually matched between the different relations. Therefore, we focus our analysis on the trained models. We aggregate across the different stimulus generation approaches (subsection A.1) as qualitative results are consistent between them (Appendix Fig. B.3). Across both pretraining datasets (Fig. 5, circles for SAYCam and squares for ImageNet) and model architectures (green for MobileNetV2, orange for ResNeXt), models tended to represent the same relation test probes more similarly to the habituation stimuli than the different relation probes. This is seen in the consistent above-chance levels of accuracy, which vary by model and experimental condition, but range between roughly 60% and almost 100%. Given that we find that the networks appear to represent these stimuli in a manner reflecting relational categories, we can examine to what extent the models reflect the findings reviewed by Quinn (2003).

Using both architectures and training datasets, we recover both developmental phenomena of interest. Analogously to infants acquiring the *above/below* relation earlier in development, we found consistently higher levels of accuracy for each model and dataset in the *above/below* relation compared to the *between* relation (compare left-side results to right-side results in Fig. 5(A), or examine the 'By relation' column in Table 2). We also observed slightly higher levels of accuracy in the conditions using the same target objects across all three stimuli than the conditions using different targets in the test stimuli, corresponding to infants acquiring category representations with identical target objects before acquiring them with varying targets (compare left-side results to right-side results in Fig. 5(B), or examine the 'By targets' column in Table 2). We ran several additional controls to more closely match the above/below and between/outside conditions (e.g., such that each condition uses two horizontal bars), and to vary the number of habituation stimuli. The results were remarkably consistent across these factors (see Appendix B.3 for details).

To better understand how the embeddings learned by these models solve this task, we synthesize a set of test stimuli by tiling a target object across a canvas with respect to a fixed reference object (Fig. 6, right). We produce such sets of stimuli with the Quinn-like stimulus generation approach and embed these with the models we evaluated in Experiment 1a. To visualize, we perform unsupervised dimensionality reduction using PCA[3] (using $n = 32$ principal components), and further reduce dimensionality to 2D (X and Y coordinates) using t-SNE (van der Maaten & Hinton, 2008) with the cosine distance metric. We color each marker (representing a single stimulus, Fig. 6, right) by the vertical position of the target object in the stimulus (Fig. 6, left). Unsurprisingly, we find no structure in the 2-D representations of the untrained model embeddings. Models trained on ImageNet show a separation between stimuli whose target object was above the bar (shades of red) and stimuli whose target object was below the bar (shades of blue). Models trained on SAYCam show a much stronger separation between these categories. Stimuli rendered with our other two stimulus generation approaches replicate these results (Appendix Figs. B.9, B.10).

We repeat this embedding visualization procedure with synthesized stimuli with two reference objects that match our "between" relation

[3] As suggested by the scikit-learn documentation when applying t-SNE to high-dimensional data, we first applied PCA to denoise and accelerate the pairwise distances computed in t-SNE.

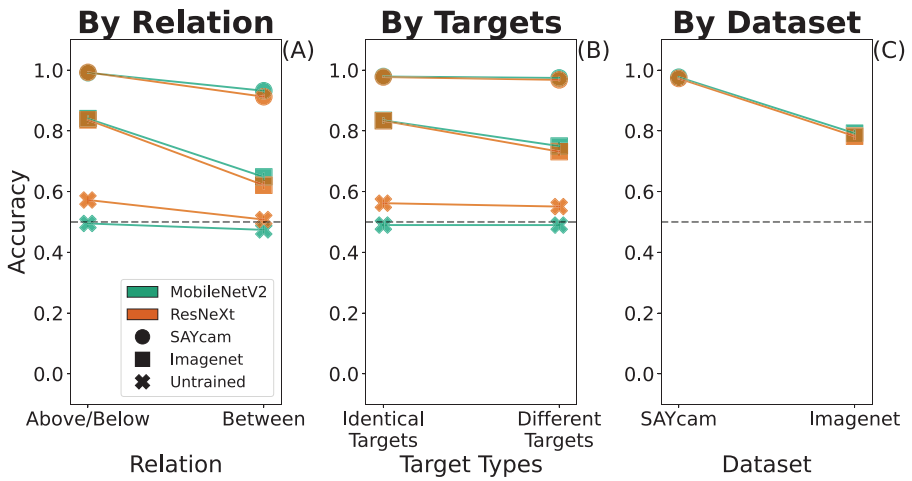

Fig. 5. Models examined represent relational categories and recover developmental phenomena. All three figures reflect the same set of experimental results, aggregated by different conditions of interest: (A): in a comparison between relations—*above/below* (left) versus *between* (right)—accuracy is higher in the *above/below* relation (B): in a comparison between target types—identical target objects (left) versus different target objects (right)—accuracy is higher when using identical target objects. (C): in a comparison between pre-training datasets—self-supervised SAYCam (left) versus supervised ImageNet (right)—accuracy is higher when using the SAYCam dataset. The color reflects model architecture, and the marker the training method. The dashed line indicates chance accuracy (50%).

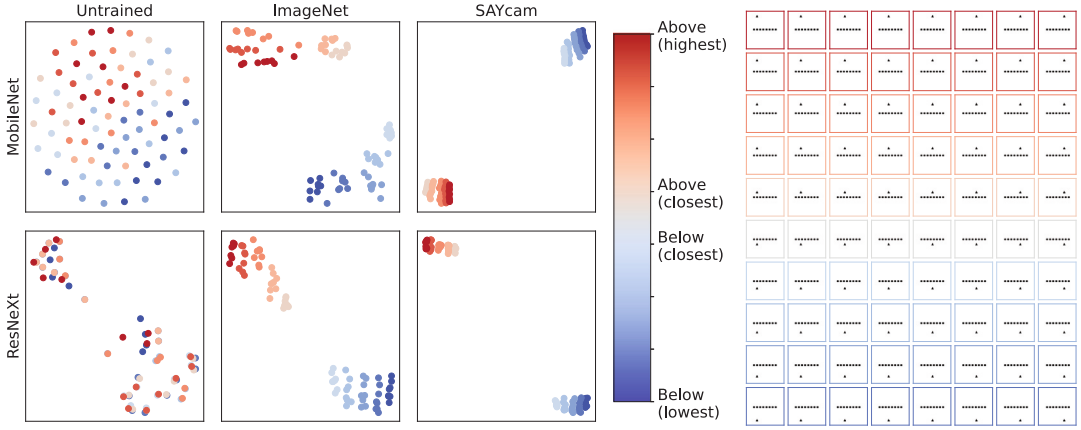

Fig. 6. Categorical perception of *above/below* in our model embeddings. We synthesize a set of controlled stimuli (right) by varying the position of a target object in relation to a fixed reference object. We embed these stimuli with the networks and reduce dimensionality to 2-D (see text for details). Each stimulus is colored by the vertical position of the target object (see the color bar). We find that while there is little structure in the untrained model embeddings, both the ImageNet-trained models and the SAYCam-trained ones produce embeddings preserving the relational structure. Rows: model architectures (top: MobileNetV2, bottom: ResNeXt). Columns: model training methods (left: untrained, middle: supervised training on ImageNet, right: self-supervised training on SAYCam).

Table 2
Summary of experiment 1a/b results. We report the mean levels of accuracy for each combination of training data, model architecture, relation, and target type variation. The right-most columns report the mean difference across the two manipulations corresponding to the developmental phenomena of interest: the "By relation" column offers the mean drop in accuracy from the *above/below* condition to the *between* one, and the "By targets" column offers the mean drop in accuracy from the "identical targets" condition to the "different targets" one. In both cases, the mean change found is congruent with the developmental phenomena examined—we observe higher accuracies in the conditions infants acquire earlier in development. Margins represent the standard errors of the mean.

Relation			*Above/Below*		*Between/Outside*		Mean change in accuracy	
Experiment	Training	Model	Identical targets	Different	Identical targets	Different	By relation	By targets
1a	Untrained	MobileNetV2	0.50 ± 0.02	0.49 ± 0.02	0.47 ± 0.02	0.47 ± 0.02	0.02	0
		ResNeXt	0.58 ± 0.02	0.56 ± 0.02	0.51 ± 0.02	0.51 ± 0.02	0.06	0.01
1a	ImageNet	MobileNetV2	0.88 ± 0.01	0.80 ± 0.01	0.68 ± 0.01	0.61 ± 0.01	0.19	0.08
		ResNeXt	0.89 ± 0.01	0.78 ± 0.01	0.66 ± 0.01	0.58 ± 0.02	0.22	0.09
1a	SAYCam(S)	MobileNetV2	0.99 ± 0.00	0.99 ± 0.00	0.93 ± 0.01	0.93 ± 0.01	0.06	0
		ResNeXt	1.00 ± 0.00	0.99 ± 0.00	0.92 ± 0.01	0.91 ± 0.01	0.08	0.01
1b	DINO-ImageNet	ResNeXt	0.93 ± 0.01	0.82 ± 0.01	0.64 ± 0.01	0.57 ± 0.02	0.27	0.09
		ViT-B/14	0.92 ± 0.01	0.77 ± 0.01	0.76 ± 0.01	0.59 ± 0.02	0.17	0.16
1b	DINO-SAYCam(S)	ResNeXt	0.99 ± 0.00	0.98 ± 0.00	0.78 ± 0.01	0.76 ± 0.01	0.22	0.02
		ViT-B/14	0.97 ± 0.00	0.89 ± 0.01	0.91 ± 0.01	0.73 ± 0.01	0.11	0.13
Mean difference							**0.14**	**0.06**

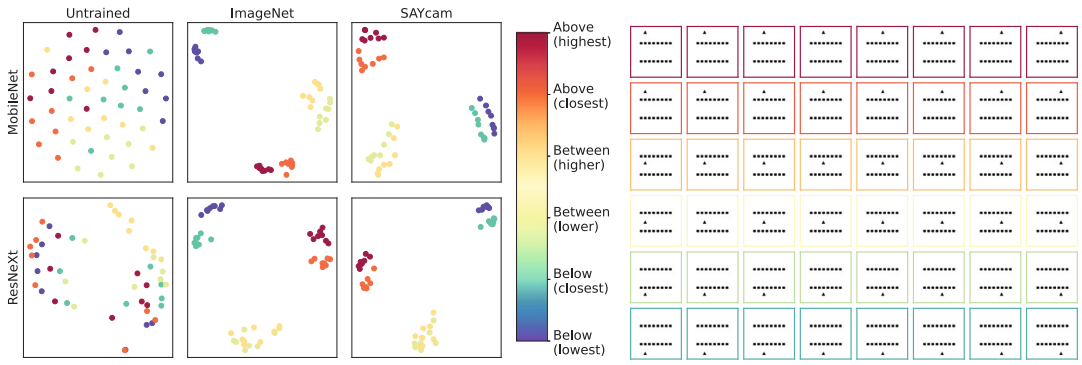

Fig. 7. Categorical perception of *between/outside* in our model embeddings. We synthesize a set of controlled stimuli (right) by varying the position of a target object in relation to two fixed reference objects. We embed these stimuli with our ResNeXt models and reduce dimensionality to 2-D (see text for details). Each stimulus is colored by the vertical position of the target object (see the color bar). As in Fig. 6, we find clustering preserving the relational information in the trained model embeddings. Rows: model architectures (top: MobileNet, bottom: ResNeXt). Columns: model training methods (left: untrained, middle: supervised training on ImageNet, right: self-supervised training on SAYCam).

stimuli (Fig. 7, right). We once again observe no structure in the embeddings produced by the untrained models (Fig. 7, left). The models trained on ImageNet show three separate clusters: above both reference objects (red and orange), between the two reference objects (light green and light orange), and below both reference objects (blue and green). The SAYCam-trained models show even tighter clustering, indicating stronger similarities within each group and more pronounced differences between the groups. Alternative stimuli renderings replicate these results as well (Appendix Figs. B.11, B.12).

We observe that our SAYCam-trained models, which acquire their perceptual features from the visual experience of young children, outperformed the ImageNet-trained models, which acquire their perceptual features from categorizing objects curated using a web search. We see this effect both quantitatively, in the higher accuracy reached by these models (Fig. 5), and qualitatively, in the tightness of the embedding clusters visualized (Figs. 6, 7). Although these results are consistent with an exciting intuitive conclusion ("models trained on infants' visual experience develop stronger relational features"), our results are confounded by the various differences between the datasets (collection method, subject matter, image quality, etc.). Additionally, the models we evaluated in this experiment were trained using different approaches, each appropriate to a particular dataset. The ImageNet model was trained using supervised learning to label objects according to their category, whereas our SAYCam models were trained in a self-supervised fashion using a temporal classification approach that does not require object labels. We deconfound the role of the training method in the next experiment.

3.2. Experiment 1b: Improved model and dataset controls

In this experiment, we introduce a third model architecture and train two of the architectures on the previous datasets using the same training method. Fixing the training algorithm allows us a controlled comparison of the effect of each dataset. We leave all other aspects of Experiment 1a unchanged.

3.2.1. Methods

Model Architectures. We evaluate one of the models from the previous experiment and add another prevalent computer vision architecture:

- ResNeXt: identical to the architecture we evaluated in Experiment 1a (Xie et al., 2017).
- ViT-B/14: We add the Vision Transformer model (Dosovitskiy et al., 2021) as another model architecture. This category of models applies the Transformer architecture (Vaswani et al., 2017) to images by extracting individual image patches, flattening each patch to a vector, embedding each vector independently using a small linear model, and passing a sequence of the vector embeddings representing the image into a series of Transformer blocks. We use the ViT-B/14 variation of the model, which uses the "Base" model size offered by Dosovitskiy et al. (2021) with a 14 x 14 patch size.

Beyond its overall recent success in a variety of computer vision tasks, we add this architecture as the Transformer self-attention architecture might offer a stronger inductive bias to relational representation than the convolutional neural networks we compared in Experiment 1a.

Pretraining. In this experiment, we study models trained using the DINO algorithm (Caron et al., 2021). DINO is a self-supervised learning algorithm that does not rely on labels, allowing us to use it with both of our datasets (whereas ImageNet contains a label for every image, SAYCam does not). DINO relies on generating multiple views of each input image through data augmentations, and learning representations that are similar between different views of the same image, but different for views of different images. We direct the reader to Caron et al. (2021) for further details.

3.2.2. Results

A summary of the results is presented in Fig. 8 and Table 2. We continue to successfully recover the two developmental phenomena of interest. Accuracy in the *above/below* relation is consistently higher than accuracy in the *between* relation, and accuracy when using the same target objects is consistently higher than accuracy when using different target objects. We also replicate the training dataset pattern from Experiment 1a—the models trained on SAYCam reliably reach higher levels of accuracy than the models trained on ImageNet ($t(7) = 12.797, p < 0.0001$, paired two-tailed). As this experiment controls for the training algorithm used, we see additional evidence that training on a child's egocentric visual experience yields a representation with more pronounced relation-based similarity than training on assorted object images.

3.3. Experiment 1 discussion

We find that large-scale, pretrained computer vision models successfully replicate a variety of developmental patterns in infant relation categorization. Across a variety of model architectures, training approaches, and control conditions, we observe that:

1. Absent any explicit relational training, embeddings extracted from the networks we study consistently display higher similarity for stimuli representing the same spatial relation, suggesting that broad visual expertise is sufficient to induce sensitivity to some relational categories.

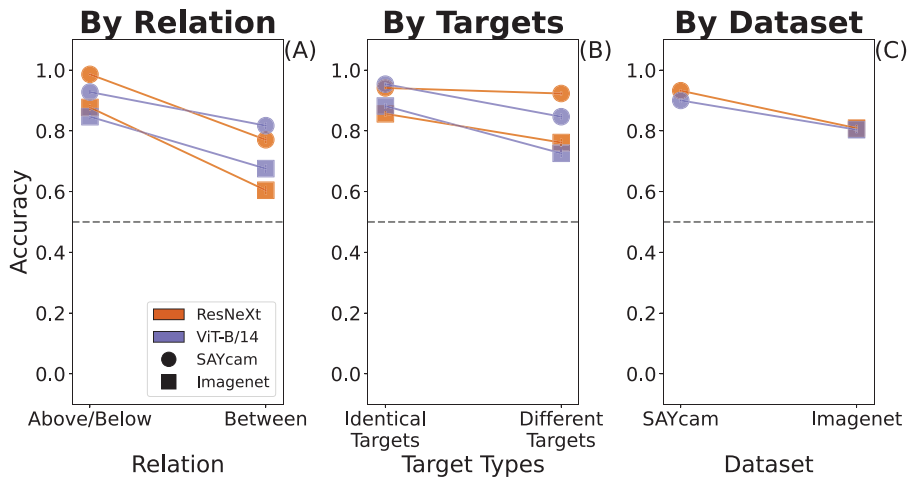

Fig. 8. DINO-trained models continue to recover developmental phenomena of interest. All three figures reflect the same set of experimental results, aggregated by different conditions of interest: (A): in comparison between relations—*above/below* (left) versus *between* (right)—accuracy is higher in the *above/below* condition. (B): in comparison between target types—identical target objects (left) versus different target objects (right)—accuracy is higher in the identical target objects condition. (C): in comparison between pre-training datasets—SAYCam (left) versus ImageNet (right)—accuracy is higher for the models trained with DINO on the SAYCam dataset. Color indicates model architecture and the marker type indicates the training dataset. The dashed line indicates chance accuracy (50%).

2. Consistent with Quinn (1994) and Quinn et al. (1999), who found that infants acquire category representations for *above/below* earlier in development than *between*, our pretrained models display higher levels of accuracy on the *above/below* relation than on the *between* relation.
3. Consistent with Quinn et al. (2003, 1996), who found that infants acquire category representations for consistent target objects earlier in development than for varying objects, our pretrained models display higher levels of accuracy when target objects remain identical ("identical targets") than when target objects vary ("different targets").

We observe no meaningful difference between the two convolutional neural networks explored in Experiment 1a, neither in overall performance nor in their ability to model the developmental phenomena. We do find that the networks trained on the developmentally relevant visual experience of SAYCam outperform models trained on the generic object recognition data in ImageNet. We find this to be true both when models were trained using different approaches that match each dataset (Experiment 1a) and when trained using an identical approach that could be applied to both datasets (Experiment 1b). However, these datasets differ in more than just their developmental relevance—ImageNet is a collection of photographs collected from the internet, while SAYCam is infant head-mounted camera data—which makes the comparison between them limited if potentially promising for future work. Although it is plausible that training models on naturalistic visual experience could increase their utility as cognitive models, we view our evidence as preliminary. One potential piece of supporting evidence: In concurrent work, Orhan and Lake (in press) find that models trained with visual data from child S in SAYCam perform at around 70% of ImageNet-trained models across a diverse range of downstream evaluations with real-world stimuli. With these findings in hand, we proceed to study another set of relations examined by Quinn and colleagues.

4. Experiment 2: Classifying *left/right* and *between/outside (sideways)* from 2D stimuli

Quinn (2004) followed up on the "above or below" experiments of Quinn (1994), and demonstrated two distinct phenomena. The first is that if the "above or below" stimuli are rotated by 90 degrees to become a "left or right" category distinction, 3–4 month-old infants continue to demonstrate a categorical preference, preferring test stimuli with the target on a novel side of the bar. Conversely, when the reference object was rotated at an angle of 45°, 3–4 month-old infants show no preference to objects placed on a novel side of this diagonal reference object, unlike both previous examples. Fig. 9 shows example stimuli with the reference objects rotated by 90 degrees, where *left/right* replaces *above/below* and the sideways between relation replaces the previous between one. Other than the angle at which the stimuli are rendered, all other experimental details remain identical to experiments 1a and 1b.

4.1. Experiment 2a: Evaluating models on the flipped relations

4.1.1. Methods

Model Architectures. We use the architectures from Experiments 1a and 1b: MobiletNetV2, ResNeXt, and ViT-B/14.

Pretraining. We use the pretraining approaches explored in Experiments 1a and 1b: randomly initialized models, supervised pretraining on ImageNet, self-supervised temporal classification on SAYCam, and self-supervised training with DINO on both ImgeNet and SAYCam.

Stimulus Generation. We generate stimuli identically to Experiments 1a and 1b and rotate the images by 90 degrees.

4.1.2. Results

Fig. 10 depicts results on the "left/right" and "between (sideways)" relations with the networks from Experiment 1a (panels (A) and (B)) and with DINO models from Experiment 1b (panels (C) and (D)). One configuration of models, ResNeXt models trained with DINO on SAYCam, performs well on the "left/right" relation (an abnormality we currently have no explanation for). The remaining models perform at chance or below on both relations, a substantial accuracy drop from the initial relations we examined. This represents a drastic qualitative deviation from the developmental results we model, where infants showed no meaningful change in the degree to which they construct a category representation for a relation contingent on whether it was presented vertically or horizontally. To attempt to isolate the cause of this effect, and identify potential conditions under which the networks recover the developmental findings on these relations, we train several additional models in the next experiment.

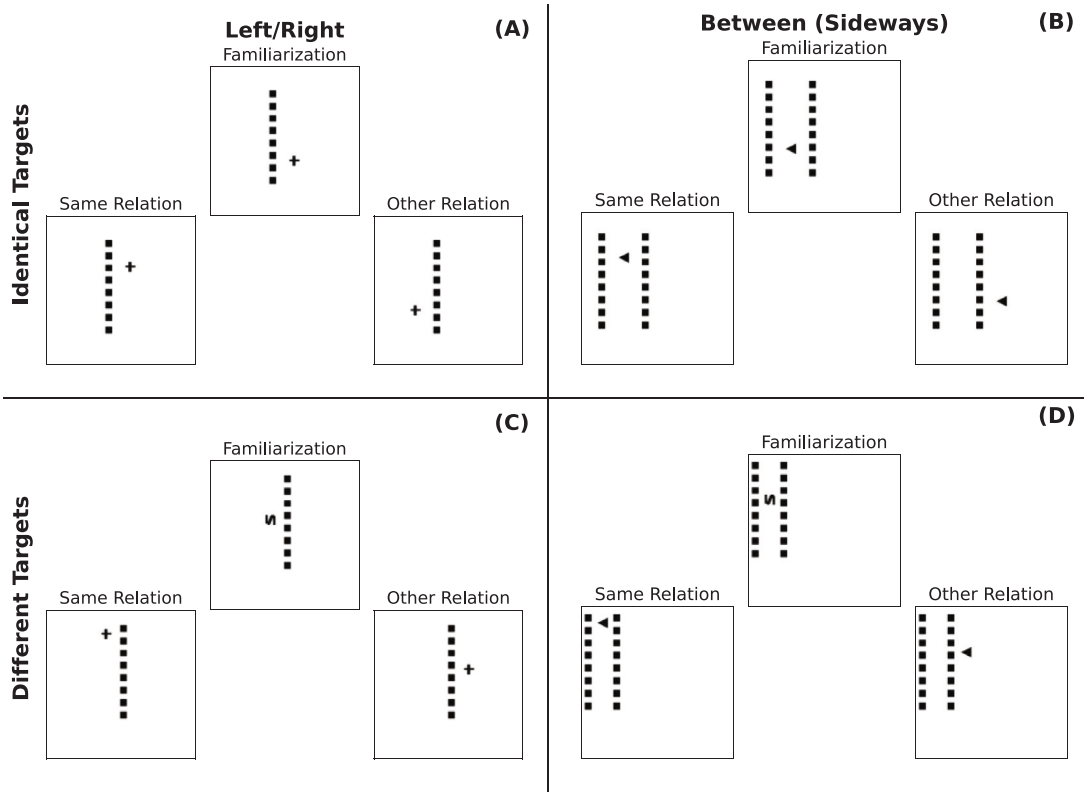

Fig. 9. Example stimuli rotated 90°. In each triplet, the central stimulus is the familiarization example, the left one is the same relation test, and the right one is the other relation test. (A): *left/right* with identical target objects. (B): *between (sideways)* with identical target objects. (C): *left/right* with identical target objects. (D): *between (sideways)* different target objects.

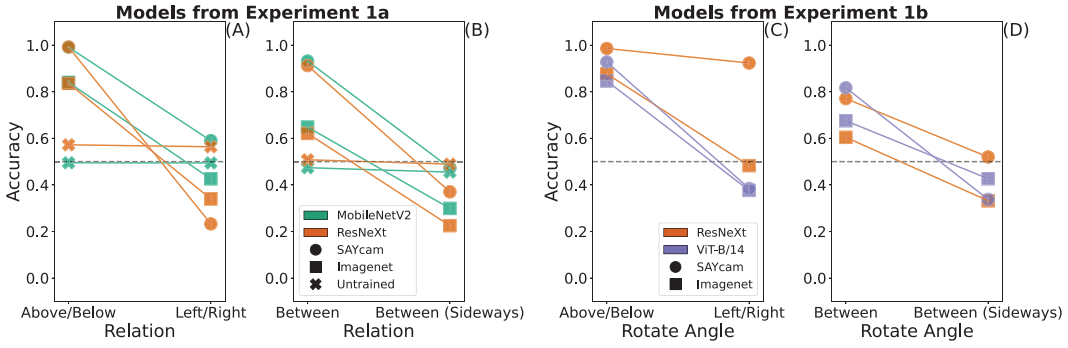

Fig. 10. Most models evaluated perform below chance on the sideways-presented relations. Models evaluated in Experiment 1a show a substantial accuracy drop from the *above/below* to the *left/right* relation (panel A) and from the *between* to the *between (sideways)* relation (panel B), other than untrained models which are unaffected. Models evaluated in Experiment 1b show similar patterns (panels C and D), other than the ResNeXt models trained with DINO on the SAYCam dataset (we offer no explanation for this aberration). Colors indicate model architecture, and marker types indicate the training dataset. The dashed lines indicate chance accuracy (50%).

4.2. Experiment 2b: Evaluating the effect of flipping data augmentations

Data augmentation refers to a set of techniques to modify the input data to a deep neural network as it is being trained, in an attempt to enable the network to learn representations that generalize and transfer better from limited amounts of data (Shorten & Khoshgoftaar, 2019). Horizontal axis flipping (across the vertical axis) is among the most common data augmentations for naturalistic image data. It is predicated on the natural symmetry across this axis (the mirror images of most objects are semantically similar or equivalent to their originals), and is trivial to implement. We hypothesize that it is this data augmentation that causes the effect we observe. By training models with horizontal flipping, we encourage the networks to represent images with a target object to the left of a reference and images with a target object to the right of a reference similarly to each other, and perhaps more similarly than to other images depicting the same relation.

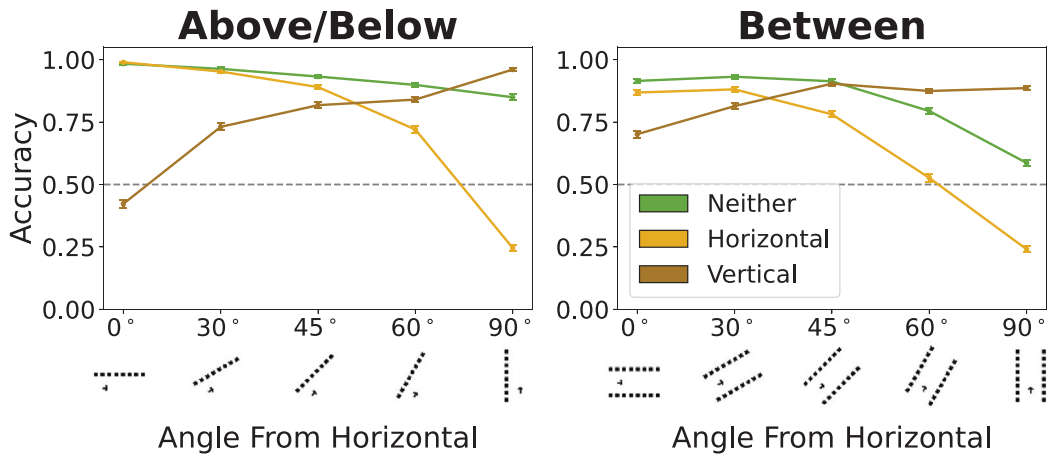

Fig. 11. The presence and type of augmentation explains (most of) the change by stimulus angle. Models trained with neither flipping augmentations (green) show a slight degradation from 0° to 90°. Models trained with the standard horizontal flipping augmentation (yellow) show a drastic degradation, matching the results shown in Fig. 10. Models trained with a non-standard vertical flipping augmentation show the opposite trend, improving gradually in accuracy from 0° to 90°. These patterns hold to varying extents in both relations.

4.2.1. Methods

Model Architecture. We perform this experiment with the ResNeXt architecture used in Experiments 1a and 1b.[4]

Pretraining. We train the models for this experiment on the SAY-Cam dataset (Sullivan et al., 2020) using temporal classification (Orhan et al., 2020), as in Experiment 1a. We train three variants on this, manipulating only the types of flips performed in data augmentation:

Neither: a model trained without any flipping as part of its data augmentation suite.

Horizontal: a model trained with horizontal flipping as part of its data augmentation suite (identical to the baseline ResNeXt-SAYCam model in Experiment 1a).

Vertical: a model trained with vertical flipping as part of its data augmentation suite.

All other data augmentations (such as color jittering, random blurring, or random cropping) were identical between these models. We also note that these augmentations were only active during model pretraining. They were not active during any of the evaluations we report.

Stimulus Generation. We use the same approach to generating stimuli as is detailed in Experiment 1a, with the exception of our rotation procedure, which is detailed in Appendix A.2 We rotate stimuli at angles of $30°, 45°, 60°, 90°, 120°, 150°$, and $180°$ counter-clockwise from the horizontal. When plotting the results, we group by the effective angle from the horizontal, e.g. as rotating at an angle of 120° is equivalent to 60° above the horizontal, we group the results for 60° and 120° under 60°. We render a collection of these rotated stimuli and the effect that each type of flipping would have on them, with one reference object (Appendix Fig. B.13) and with two reference objects (Appendix Fig. B.14).

4.2.2. Results

A summary of the results for the three flipping model variants, evaluated across the various stimulus rotation angles, is shown in Fig. 11. We found that the model with horizontal augmentations only (plotted in yellow) recovered the results from previous experiments, with high levels of accuracy at 0° (compare to Fig. 5), low ones at 90° (compare to Fig. 10), and gradual degradation in the intermediate angles. The other two flipping models provide comparison cases to demonstrate the causal effect of the standard horizontal flipping. The model without any augmentations (plotted in green) showed a much more mild yet consistent degradation in accuracy from 0° to 90°. We take this gradation to be the extent to which the model learns to favor horizontal symmetry absent any data augmentation, only from the natural statistics of the training data learned by the models (Geisler, 2008). Conversely, the model trained with vertical flipping (plotted in brown) depicted the opposite effect. Its accuracy was lowest at 0°, and as the stimuli are rendered closer to vertical, its accuracy gradually improved, peaking at 90°. Qualitatively, we find that the model trained with neither flipping directions recovers the developmental phenomenon from Quinn (2004), where discriminating *above/below* is equally easy as discriminating *left/right*. However, none of the networks recover the other finding from Quinn (2004), that infants show an inability to discriminate between objects with respect to a diagonal reference object. Regardless of what sort of flipping was applied, the three models evaluated in Experiment 2b all reached fairly high levels of accuracy at 45°. The model trained without any flipping also shows a substantial preference to *between* stimuli presented horizontally (0°) compared to vertically (90°). Quinn et al. (2003) studied both conditions and while comparing them was not the authors' explicit purpose, they report results separately. Infants in both Experiments 2 and 4 in Quinn et al. (2003) showed slightly stronger novelty preferences to stimuli presented horizontally (Tables 1 and 4 respectively in Quinn et al., 2003). However, the differences displayed by the infants are notably smaller in magnitude than the differences displayed by our model, and there is no developmental evidence that infants reliably categorize the horizontal condition before the vertical.

4.3. Experiment 2 discussion

We evaluate the existing models from Experiment 1 and specially-trained models with custom data augmentations on stimuli rotated to various angles, and find that:

1. Unlike Quinn (2004, experiment 1), who found that infants distinguish "left or right" at a similar age to "above or below", many of the networks we study have a strong preference for stimuli depicting vertical relations. We discover this is an artifact of the data augmentation often used to train these models, and show that models without data augmentation display a weaker preference.
2. Unlike Quinn (2004, experiment 3), who found that infants fail to distinguish between objects on opposite sides of a diagonal reference object (presented at an angle of 45 degrees), the networks we examine consistently categorize relations presented

[4] As this experiment required training models in unique conditions, and produced clear results, we opted not to repeat it with other architectures from Experiment 1.

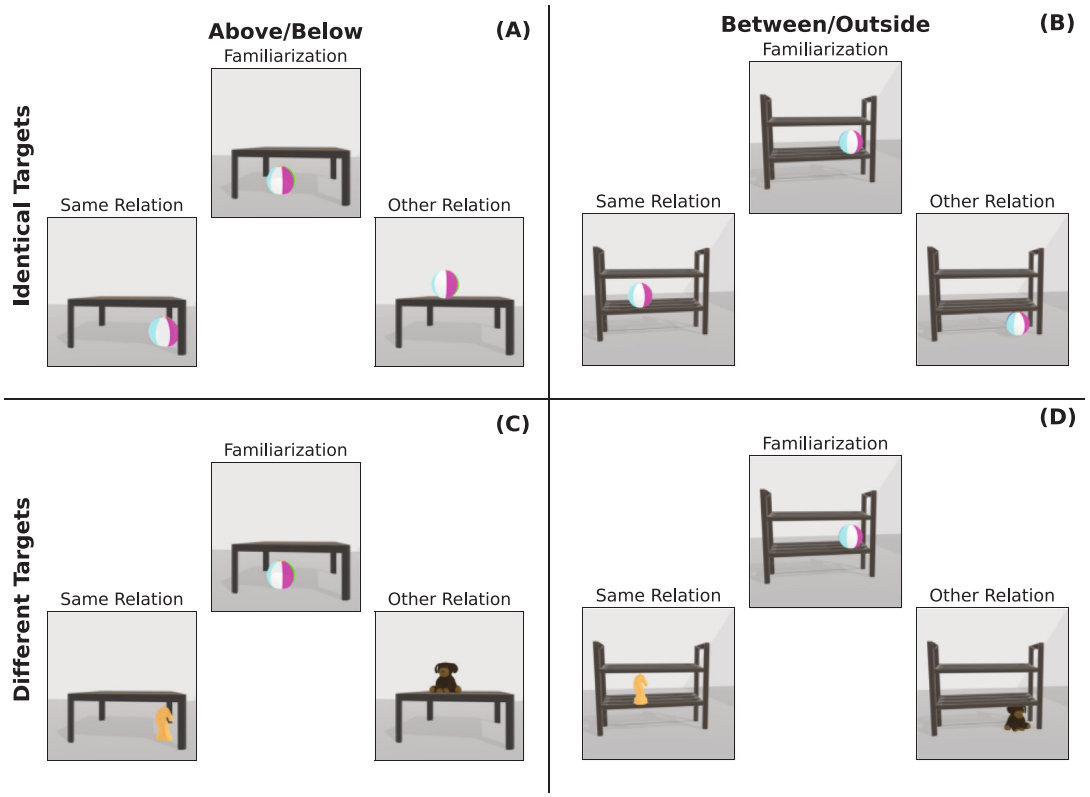

Fig. 12. Experiment 2 Example Stimuli. In each triplet, the central stimulus is the familiarization example, the left one is the same relation test, and the right one is the other relation test. (A): *above/below* with identical target objects. (B): *between* with identical target objects. (C): *above/below* with identical target objects. (D): *between* different target objects.

at this angle. This holds both relative to one reference object (*above/below* at 45 degrees) as well as relative to two reference objects (*between/outside* at 45 degrees).

Although our results from Experiment 1 broadly suggest that pretrained computer vision neural networks can model important aspects of infant relation categorization, our findings suggest that some care and caution are required in the choice of model and training setup. Models trained with horizontal flipping, a data augmentation approach designed to mimic the horizontal reflection invariance many real-world objects display, struggled once the task evaluated was in direct contrast to the augmentation. Although we did not examine this in other contexts, it is not out of the question that other augmentations, such as color jittering, image solarization, or manipulations of image brightness or contrast might be in conflict with evaluating the development of visual perception.

Item (2) above summarizes a discrepancy from the developmental results with stimuli presented at an angle of 45°. The infants evaluated in Experiment 3 of Quinn (2004) were 3–4 month-olds, the youngest age bracket evaluated across the experiments surveyed. Although infants at that age successfully appeared to develop a categorical response in the "above or below" condition, they failed to do so in the "between" condition, while 6–7 month-old infants were able to. The networks we evaluate show comparable levels of accuracy for *above/below* at an angle of 45° and "between" at an angle of 0° (compare the accuracies for these angles in Fig. 11). This suggests that to the extent the levels of accuracy displayed by the networks track the developmental difficulty of these relations, we would predict that 6–7 month-old infants should be able to form category representations for "object on either side of a diagonal bar". We leave it to future work to experimentally examine this prediction.

5. Experiment 3: Classifying *above/below* and *between/outside* from 3D-rendered stimuli

In Experiment 1, we found that pretrained computer vision models appear to categorically represent spatial relations when evaluated with abstract stimuli that resemble developmental experiments. To study how generalizable our findings are, in this experiment we follow a similar methodology to Experiment 1 although with a different, more complex approach to stimulus rendering. Experiment 1 employed simple 2D renderings, either closely matching the stimuli Quinn showed infants (Fig. 2) or in alternative control conditions (Appendix Figs. B.1, B.2). In this experiment, we evaluate models on 3D renders of scenes instantiating the same spatial relations Fig. 12. These stimuli are more similar to the images used to train the networks we examine, and therefore allow evaluation of the extent to which model embeddings organize by categorical representations in more realistic data.

We begin our examination of the more realistic stimuli by reproducing our results from Experiments 1a and 1b, using the same models in similar conditions:

5.1. Methods

Model Architectures. We evaluate the three model architectures evaluated in Experiment 1: MobileNetV2, ResNeXt, and ViT-B/14.

Pretraining. We compare the same pretraining approaches from Experiments 1a and 1b. Experiment 1a: randomly initialized and untrained models, supervised classification pretraining on ImageNet, and self-supervised temporal classification on SAYCam. Experiment 1b: self-supervised pertaining using the DINO algorithm on both the ImageNet and SAYCam datasets.

Stimulus Generation. We render stimuli using Blender (Blender Online Community, 2018) (see Fig. 12) following a similar procedure to the one described in Experiment 1a. We refer the reader to Appendix

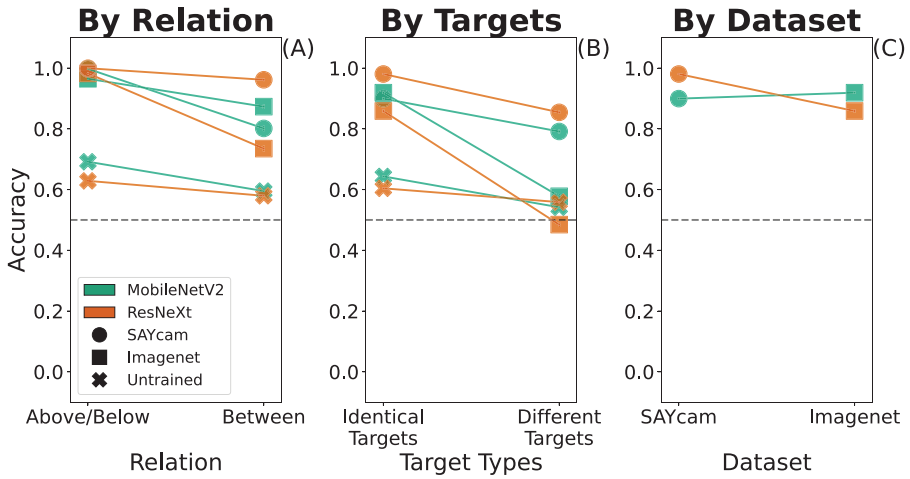

Fig. 13. Models from experiment 1a recover the same patterns with 3D-rendered stimuli. Both in a comparison between relations (panel (A)) and identical or different targets (panel (B)), the networks continue to recover the same developmental patterns with the more complex stimuli (compare this to Fig. 5). The effect of the choice of the training dataset is not evident with these stimuli and is inconsistent across models. Color indicates the model architecture and marker type indicates the training method. The dashed line indicates chance accuracy (50%).

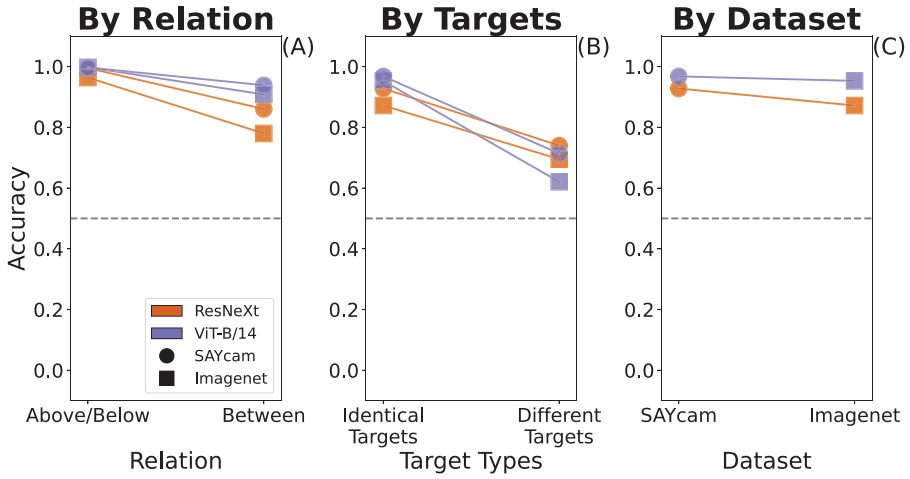

Fig. 14. DINO-trained models continue to recover developmental phenomena of interest. As in Fig. 13, the networks continue to recover the phenomena studied in Experiment 1, both when comparing by relation (panel (A)) and when comparing by identical or different target objects (panel (B)). Color indicates the model architecture and the marker type indicates the training dataset. The dashed line indicates chance accuracy (50%).

A.3 for complete details. As in Experiment 1, we examine triplets where the target object matches between the familiarization and test stimuli ("identical targets"), and ones where the target object varies in the two test stimuli ("different targets"). We render scenes of both relations (*above/below* and *between/outside*) using eight different target objects: a beach ball, a chess knight, a Lego piece, a toy pineapple, a ping-pong paddle, a toy robot, a rubber duck, and a stuffed animal (see Appendix Fig. B.15 for examples). We generate 256 unique scenes, each with all eight target objects, resulting in 2048 total triplets for each of the two relations.

Methods Summary. We evaluate the same set of models and pre-training approaches we used in Experiment 1, on the same two relations (*above/below* and *between*). We render stimuli that are richer than those used in Experiment 1 to examine how sensitive the models are to relational information in more visually complex stimuli. Our stimuli in Experiment 2 use one of eight different target objects. For each relation, we sample 2048 triplets and report the average accuracy for each model and pretraining approach—how often the embeddings for the congruent pair of stimuli are more similar (using cosine similarity) than the embeddings for the incongruent pair.

5.2. Results

A summary of our findings using 3D stimuli is shown in Fig. 13, which parallels our previous findings using abstract stimuli (Fig. 5). We again find the randomly initialized and untrained models around chance accuracy (results marked with an 'X'), though at somewhat higher accuracy levels than in Experiment 1, suggesting the stimuli in this experiment (which are more realistic and richer) are not as perceptually controlled as simple stimuli of Experiment 1. We omit the randomly initialized models from further discussion. We find that these stimuli tend to make the task harder for the networks we examine—most trained models have lower accuracies in this experiment than in the previous one. For instance, of the models evaluated in Experiment 1a, the MobileNetV2 models reached 3%–10% lower accuracy levels in this experiment, and the ResNeXt models were 9%–12% lower. The DINO-trained models show a deviation in this pattern—the DINO ResNeXt models had an accuracy roughly 20% lower in this experiment, while the ViT-B/14 models had an accuracy that was 2%–5% higher in this experiment compared to the previous one. However, even with the relative difficulty, we find the same pattern of results seen in the developmental experiments. The neural networks attain higher accuracy on the *above/below* task than on the *between* task, analogously

to infants acquiring category representations for this relation earlier in development. The networks we evaluate consistently reach higher accuracy on the "identical targets" condition compared to the "different targets" one, reflecting the same pattern in infants. We also continue to observe a higher accuracy in models trained on the SAYCam dataset, although this effect is abated.

Fig. 14 mirrors Fig. 8, depicting results from our DINO-trained models. The same developmental phenomena previously outlined continue to present themselves: *above/below* is easier than *between*, identical targets are easier than different ones, and SAYCam-trained models (slightly, yet consistently) outperform ImageNet-trained ones ($t(7) = 4.109, p < 0.005$, paired and two-sided).

5.3. Discussion

We reproduce the developmental phenomena explored in Experiment 1 with 3D-rendered stimuli. The use of richer, 3D-rendered stimuli allows us to conclude that the networks' ability to represent relational categories, and their ability to mirror findings from the developmental literature, generalizes to other types of stimuli and is not an artifact of using the simplistic setting of Experiment 1. Unlike the abstract stimuli explored in Experiment 1, the stimuli explored in this experiment would violate physical common sense if rotated by 90°, adding a potential confound, and so we opted against evaluating a version of Experiment 2 with rotated versions of these stimuli. Given the discovery that we can reproduce findings on spatially simple relations such as "above or below" or "between or outside" with more realistic stimuli, we ask: can we reproduce developmental patterns with more complex relations as well?

6. Experiment 4: Classifying *containment*, *behind*, and *support* from 3D-rendered stimuli

Following the success of replicating Experiment 1's results with 3D rendered stimuli in Experiment 3, in this experiment we use similarly rendered stimuli to examine more spatially complex relations. Casasola et al. (2003, Experiment 2[5]) studied whether 6 months old infants can categorize scenes based on whether or not they show a containment relation. Infants were habituated to a short video depicting a hand placing an object inside a container, that is, in a containment relation (Fig. 15(A/B), "Familiarization"). Infants were then tested with the familiarization video and with three novel probes. The test probes varied along two key dimensions: relation (whether or not they also depicted a containment scene) and occlusion (what fraction of the target object is visible at the end of the video). The first test probe (Fig. 15(A/B), "Same relation") showed the same event filmed from a higher camera angle: this produces the same relation but with novel occlusion—the higher camera angle causes much more of the object to be visible. The second test probe (Fig. 15(A), "Other relation") uses the same high angle but places the object behind the container. This results in similar occlusion to the familiarization video but with the object placed in a novel relation with respect to the container. Finally, the third probe (Fig. 15(B), "Other relation") offered both a novel relation and occlusion: filmed from the same angle, this test showed an object being placed on top of an upside-down container, in a support relation. This final test probe serves as a control condition with mismatches on both relation and occlusion. As expected, Casasola et al. (2003, Figure 2) found the lowest test-time looking times when showing the familiarization stimulus a second time. The 'containment' test probes (with novel degrees of occlusion) were found to elicit significantly shorter looking times than the 'behind' stimuli, which depict a novel relation with similar degrees of occlusion to the familiarization stimulus. On this basis, Casasola et al. (2003) conclude (and see also Casasola (2008) for a review) that 6-month-old infants successfully form a category representation for the containment relation. In this experiment, we will evaluate whether the models we tested in Experiments 1 and 2 can replicate these patterns.

A key methodological difference between our experimental setup and the one used by Casasola et al. (2003) is our use of still images, rather than videos. We motivate this decision from two perspectives. First, to be able to compare to our previous results in Experiments 1 and 2, we wished to use the same models, and as these models are trained on single images,[6] rather than videos, we opted to adapt the task. Second, models trained for image classification or self-supervised image-level tasks are more widely available than models trained for video classification. To the extent we hope this work can serve as methodological inspiration for studying other developmental phenomena with pretrained models, we wished to examine whether translating video stimuli to representative still images is sufficient to recover developmental findings. We generate our stimuli to match the terminal frames of the videos Casasola et al. (2003) used (Fig. 15, top). As in Experiments 1 and 2, we generate stimuli triplets to compare the similarity between an embedding of a single familiarization and the embeddings of two test probes. We visualize our two comparison cases in the bottom half of Fig. 15. In both cases, we use a familiarization stimulus showing a containment event from a low angle, similar to the familiarization event used by Casasola et al. (2003). We also use the same type of same relation test stimulus, depicting a containment event from a higher angle. In one condition, "Containment vs. Behind", we compare to a stimulus depicting the target object behind the container, rendered from the same higher angle (Fig. 15, left). In the other condition, "Containment vs. Support", we compare to a stimulus rendering the target object supported by the container, with the container flipped upside-down, also rendered from the same higher angle (Fig. 15, right).

6.1. Methods

Model Architectures. We evaluate the same three architectures from Experiments 1 and 2: MobileNetV2, ResNeXt, and ViT-B/14.

Pretraining. We use the same pretraining approaches from the previous experiments: randomly initialized models serving as a control, supervised pretraining on ImageNet, self-supervised pretraining on SAYCam, and models trained using self-supervised DINO on both the SAYCam and ImageNet datasets.

Stimulus Generation. As in Experiment 3, we use Blender (Blender Online Community, 2018) to render stimuli. Our stimuli use four different containers, and eight different target objects, identical to the ones used in Experiment 3. For additional details, see Appendix A.4. We render four images for each stimulus (see Fig. 15(C/D) for examples split into triplets). The first is the familiarization stimulus, with the target object in the containment relation and a lower camera angle. We then raise the camera to a higher angle and render three test stimuli. The first is the *containment* test stimulus, where we render the same scene as in the familiarization stimulus from the new camera angle, matching the familiarization relation but differing in occlusion. The second is the *behind* test stimulus, where we move the target object behind the container, creating similar occlusion between the container and test object to the familiarization stimulus, but in a different spatial relation. The third is the *support* test stimulus, where we flip the container upside-down, and place the target object on top of it. This offers a stimulus mismatched in both dimensions (relation and occlusion) to the familiarization stimulus. We render 128 unique scenes by sampling

[5] We skip the perceptually mismatched stimuli examined in Experiment 1 by Casasola et al. (2003) and proceed directly to the better-controlled stimuli the authors used in Experiment 2.

[6] With the exception of the models trained using Temporal Classification with the SAYCam dataset, which are trained using the temporal ordering of short video clips.

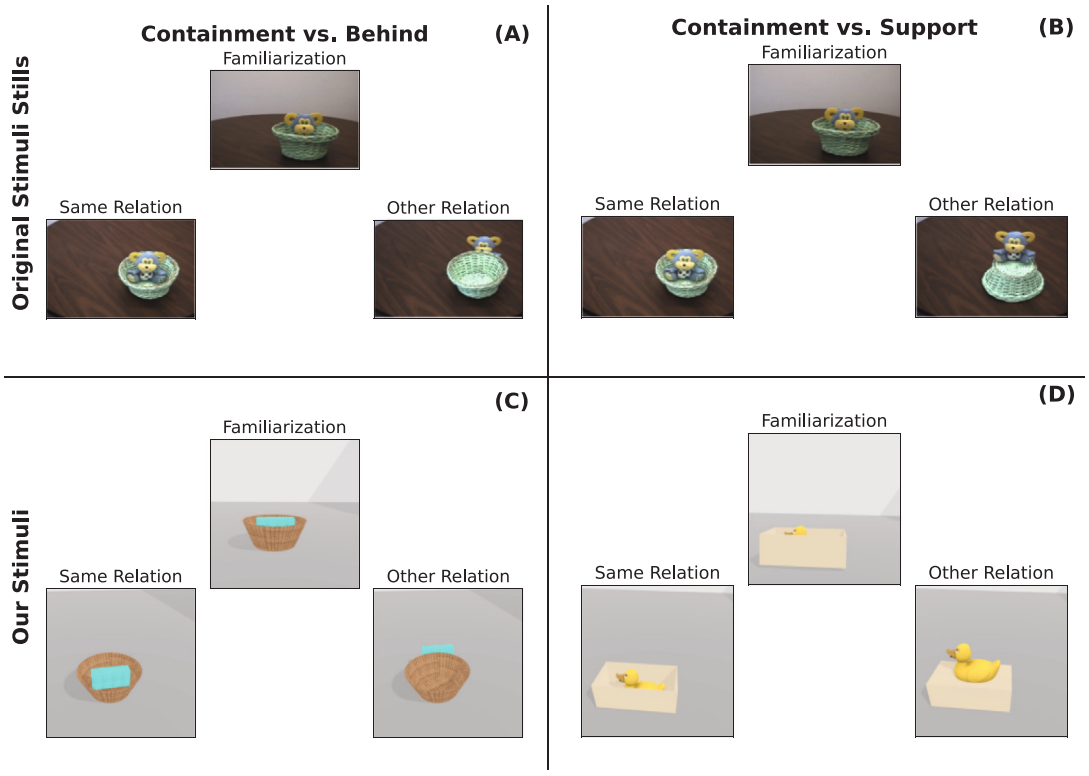

Fig. 15. Experiment 4 Example Stimuli. In each triplet, the central stimulus is the familiarization example, the left one is the same relation test, and the right one is the other relation test. Top: the final frames of the stimuli video presented to infants, reproduced from Casasola (2008, Figure 1). Bottom: our rendering of matching stimuli. Left: comparing a test probe depicting the *containment* relation to a test probe depicting the *behind* relation. Rendered using the wooden basket container and Lego target object. Right: comparing a test probe depicting the *containment* relation to a test probe depicting the *support* relation. Rendered using the shorter cardboard box container and rubber duck target object.

different camera parameters (see Appendix A.4 for full details), each with all 32 combinations of the four containers and eight target objects, resulting in a total of 4096 unique sets of stimuli. See Appendix Figs. B.16 to B.19 for visualizations of scenes with all objects, and Appendix Fig. B.25 for visualization of camera parameter variations for a single object combination.

Methods Summary. We evaluate the same collection of models pre-trained on the same datasets as in Experiments 1 and 2. We compare accuracies in two primary conditions. In the first, "Containment vs. Behind", we use a low-angle *containment* scene as our familiarization stimulus, and use test probes depicting a *containment* scene rendered from a higher angle, and a *behind* scene rendered from the same higher angle. In the second, "Containment vs. Support", we use the same familiarization stimulus and first test probe and replace the *behind* test probe with a *support* scene rendered from the same higher angle. In both conditions, we report the average accuracy over the 4096 sets of stimuli—how often the embeddings for the pair of stimuli depicting a containment relation are more similar (using cosine similarity) than the embeddings for the familiarization stimulus and the incongruent test probe.

6.2. Results

We compare the networks' levels of accuracy between the "Containment vs. Behind" condition and the "Containment vs. Support" condition in Fig. 16. We find that across various training datasets and model architectures, all of the networks reached higher levels of accuracy when comparing two *containment* scenes to a *support* scene. This by itself did not surprise us, as this is the easier foil relation, which does not match the degree of occlusion in the familiarization stimuli (compare the right-hand triplets in Fig. 15 to the left-hand ones). In the better-matched condition of "Containment vs. Behind", the networks peaked around chance accuracy. That is, most models we evaluated

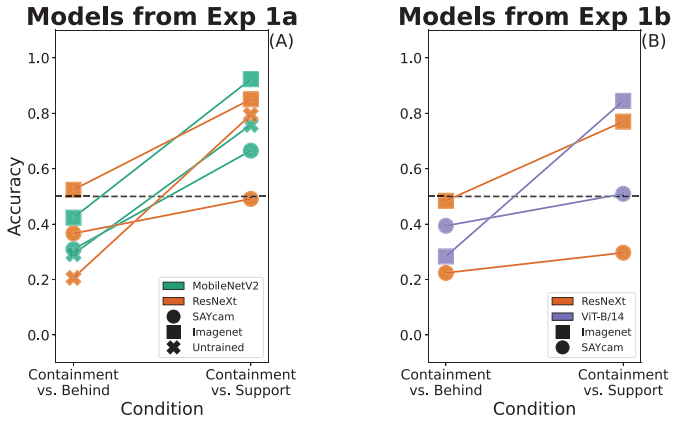

Fig. 16. Models prefer the *containment* test probe over the *support* one, but not over the *behind* one. We compare the average accuracy over triplets where the foil test probe depicts the *behind* relation ("Containment vs. Behind", left data points) to the average accuracy over triplets where the foil test probe depicts the *support* relation ("Containment vs. Support", right data points). The networks, including the untrained ones, are fairly consistent with their preferences. The *containment* test probe is often judged as more similar to the familiarization stimulus when paired with a *support* foil, but not when paired with a *behind* foil. This holds with both the models discussed in Experiment 1a (panel (A)) and the DINO-trained models introduced in Experiment 1b (panel (B)). The color indicates the model architecture and the marker type indicates the training method. The dashed line indicates chance accuracy (50%).

systematically found the *behind* test probes (which match in occlusion, but not in relation) more similar to the familiarization containment stimuli than the *containment* test probes. This is in direct opposition to the patterns infants demonstrated in Casasola et al. (2003), who found the *containment* test stimuli to be the least surprising ones. In another

reversal from Experiment 1b, we found that the models trained on ImageNet (both using DINO and in a supervised fashion) outperformed the models trained on SAYCam (compare the circle markers to the squares in Fig. 16). A final unexpected result was a consistent preference present in the randomly initialized model, reaching substantially lower accuracies in the "Containment vs. Behind" condition than in the "Containment vs. Support" one. To see if this is an anomaly, we replicated our initial randomly initialized model with nine additional ones. We demonstrate in Appendix Fig. B.20 that while the degree of this preference varies, all of our randomly initialized models (across both the MobileNetV2 and ResNeXt architectures) replicated this pattern.

We hypothesized that one potential culprit in the models' failure in the "Containment vs. Behind" condition might be the pooling operations that precede our embedding extraction. In the MobileNetV2 and ResNeXt architectures (but not in the ViT-B/14 one), the final two-dimensional representation of each input image is pooled in order to create an embedding vector. The pooling mechanism struck us as potentially related to the failure as it collapses much of the spatial information, which might leave more remaining information in the degree of occlusion (which roughly corresponds to how many pixels of the target object are visible) than in the spatial relation. To examine this hypothesis, we repeat our similarity judgments, using embeddings extracted before the pooling operation. We find that the networks consistently reached higher accuracy when similarity was compared using embeddings extracted before pooling (see Appendix B.9 for the complete details, Appendix Fig. B.22 for a summary of the results, and Appendix Table 3 for the complete results). Excluding the untrained models, the networks reach a mean accuracy of 0.798 pre-pooling in the "Containment vs. Behind" condition compared to a mean accuracy of 0.389 post-pooling. Similarly, in the "Containment vs. Support" condition, the networks reach a mean accuracy of 0.933 pre-pooling compared to a mean accuracy of 0.666 post-pooling. This supports a hypothesis that the relational information is more prominent in the pre-pooling embeddings, compared to the post-pooling embeddings, at least as measured by the cosine similarity. It appears sensible that the pooling operation, which collapses across spatial locations in an attempt to extract a location-invariant representation of the stimulus, reduces the degree of spatial and relational information preserved.

Finally, we explore the extent to which the relational information remains present in the final embeddings, by performing a series of linear decoding experiments. We opt for linear decoding as a standard approach to examine whether a particular property is represented in an embedding, and as one that requires training the minimal number of parameters (as many as the embedding size times the number of categories decoded). In Appendix B.10, we detail how we split our dataset into training and test sets and evaluated the ability to decode the relation present in an image using a single linear layer, leaving the rest of the model fixed. We find a consistent ability to decode which relation is present from our trained models (see Appendix Fig. B.24), suggesting the relational information remains present in the embeddings even if it is no longer a primary contributor to representational similarity.

6.3. Discussion

In this experiment, we examine the extent to which our success in replicating phenomena in categorizing simpler visual relations (above or below, between or outside) transfers to more complex relations (containment, behind, support). Unlike in previous experiments, where models, for the most part, mirrored developmental phenomena, here we fail to recover the main phenomenon of interest. We use containment stimuli rendered from a low angle as our familiarization examples, and three types of test probes: *containment* stimuli rendered from a higher angle (matching on the relation, but not object occlusion), *behind* stimuli (matching on occlusion, but not on the relation), and *support* stimuli (matching on neither relation nor occlusion). We discover that our models repeatedly embed the *behind* test probes more similarly to the familiarization stimuli than the *containment* test probes. This is inconsistent with the findings outlined in Casasola et al. (2003), where infants measured lower looking times to the *containment* test over the *behind* one. When tasking our models to compare the *containment* test probe to the easier *support* one, our models do substantially better. Therefore, we hypothesize that similarity in the model embeddings is driven first by lower-level perceptual features, and second by higher-level relational ones.[7] The positive results in Experiment 3 offer evidence that the negative results observed in this experiment are not solely attributable to the 3D rendering approach we used for stimuli in this experiment, suggesting the discrepancy arises from other factors. In Appendix B.9, we demonstrate that pre-pooling embeddings show similarity driven by relations, and in Appendix B.10, we validate that relational information is maintained in the final embeddings, as we successfully decode with very high levels of accuracy.

The model-to-infant comparison in this experiment is less faithful than in Experiments 1 and 2, as the infants in Casasola et al. (2003) watched short video clips depicting the object being placed in the specified relation to the container, rather than making judgments based on still images. We are not aware of any work investigating the extent to which infants make similar relational judgments of the containment relation from still images. To the extent Experiments 1 and 2 demonstrated that pretrained neural network models can successfully recover patterns in infant relation categorization, we would hypothesize that infants would be less consistent in judging stimuli by relational similarity if only offered still images. We note that although the models developed by Ullman et al. (2019) successfully categorize still images of relations, they do so after being trained on video stimuli. Their findings imply that deep neural network models trained on videos could potentially offer a closer match to the findings outlined by Casasola et al. (2003), and we leave that for future work to examine.

7. General discussion

We investigate the capacity of various large-scale pretrained computer vision neural network models to replicate findings regarding the development of relation categorization. We first find that without explicit relational training, the trained models we evaluate learn embeddings that tend to represent stimuli depicting the same relation more similarly to each other than stimuli representing different relations. We then successfully recover most patterns of interest relating to how infants process relations such as "above or below" and "between or outside" (Quinn, 2003). We observe that the neural networks we study show similar difficulty gradations to the infants: the networks reach higher accuracy levels on the *above/below* relation than on the *between/outside* one, mirroring infants' ability to form categorical relations for the former earlier in development than for the latter. Infants also respond categorically to stimuli depicting identical target objects earlier in development than to stimuli using different target objects; likewise, the models we evaluate have higher accuracy levels when using identical target objects than when using different ones. We encounter these patterns both with 2D stimuli closely resembling the developmental ones (Experiment 1) and with rendered 3D stimuli that capture the same relations in different visual formats (Experiment 3). However, when evaluating the same pretrained neural networks on the *containment* relation (Casasola et al., 2003), we find (in Experiment 4) that the models appear to organize their embedding primarily by object visibility and secondarily by relation, even when relational information is present. This is evident in the models' consistently lower levels

[7] Casasola and Cohen (2002) and Casasola et al. (2003, Experiment 1) raised the concern that perhaps infants also make similarity judgments according to such lower-level features, and assuaged this concern in Casasola et al. (2003, Experiment 2).

of accuracy when probed with a foil that is matched on occlusion (depicting the *behind* relation), compared to when probed with a foil that is mismatched on occlusion (with the *support* relation).

We find that shortcomings in the networks' abilities to replicate developmental patterns, and the variation between models, can help highlight methodological nuances meaningful for future work. In Experiment 2a, we find that many models are consistent with infants' patterns when a relation is presented vertically (e.g. "above or below"), but drastically inconsistent with the same patterns when a relation is presented horizontally (e.g. "left or right"). To explain this inconsistency, we evaluate (in Experiment 2b) the effect of image flipping, a particular form of data augmentation often used in pretraining computer vision models, and discover that the use of image flipping explains the observed deviation. We also find that on the visually simpler relations of "above or below" and "between", models trained on the egocentric, developmentally realistic SAYCam dataset (Sullivan et al., 2020) outperform models trained on ImageNet, using both simpler stimuli (Experiment 1b) and more complex, rendered stimuli (Experiment 3). We mostly observe the opposite pattern on the *containment* relation stimuli evaluated in Experiment 4 (with the exception of the pre-pooling results discussed in Appendix B.9). These results can enrich previous literature on the role of developmentally plausible data for computer vision (Smith & Slone, 2017), although it is worth noting that the training data between ImageNet and SAYCam differs across a number of dimensions (how data was collected, image quality, perspective, and subject matter). Furthermore, our mixed results in Experiment 4 suggest the benefits are not universal across all benchmarks and evaluations. We do not view this evidence as suggesting choices that perform better on versions of our task make for better computer vision engineering artifacts; instead, we consider them to offer better models of the developmental phenomena we study, and hope they inspire further investigation in this direction. In a supplemental experiment (Appendix C), we identify that for neural networks to recover similar patterns from symbolic inputs, they should flexibly allow comparing between multiple objects, as only the architectures that cannot (the MLP and RelationNet) struggled to learn a relation entirely.

One contribution our work makes is to evaluate models trained on SAYCam (Sullivan et al., 2020), the best available proxy for a child's visual experience, using developmental behavioral paradigms. Prior work has used this dataset to train models (such as some of the pretrained models from Orhan et al., 2020 we use in this work), or to evaluate such models on cognitive biases (Tartaglini et al., 2022). Forthcoming work by Vong et al. (in press) uses this dataset to evaluate grounded language acquisition through cross-situation word learning, and work from Orhan and Lake (in press) studies the representations that models learn from this dataset absent strong inductive biases. We hope that this line of work can serve as inspiration for future work in computational developmental psychology and computer vision. From the perspective of developmental psychology, we are excited about the ability to evaluate suggested computational mechanisms in the context of rich, large-scale data—beyond, for instance, fitting models to choice data in an attempt to compare them, we can now train models on proxies of perceptual inputs and examine emergent phenomena of these models. This serves to test ideas about the role of naturalistic and egocentric data in the development of artificial vision systems, particularly in comparison to the development of human vision (Smith & Slone, 2017), though ImageNet and SAYCam differ across a number of dimensions beyond the type of data. Future work should pursue detailed and controlled comparisons to better account for the roles of these different factors.

Our approach also differs from most work on learning relations with deep neural networks. Prior work (Santoro et al., 2017; Shanahan et al., 2019) focused on developing and evaluating custom architectures for relation learning. We show that the embeddings learned by pretrained models with no explicit relational bias allow judging similarity based on relation, and we leave it to future work to further study how much relational information is decodable from these embeddings, and to more explicitly compare to architectures tailored to relational learning. Future work could also examine the role of language in relation learning, building on the experimental results reviewed by Casasola (2008) and the modeling work of Westermann and Mareschal (2014) with modern multi-modal neural networks such as CLIP (Radford et al., 2021) or text-based visual classification methods (Jaini et al., 2023).

One limitation of our methodology arises from the use of relatively generic neural network architectures as our models of infant relation categorization. While neural networks have a long history as cognitive models generally (Rumelhart et al., 1987) and as models of the visual system specifically (Lindsay, 2021), there are questions regarding the inferences these models enable (Saxe et al., 2020). For instance, Bowers et al. (2022) discuss mechanistic concerns in the use of neural networks to model human vision, primarily that networks use different features than people, such as over-reliance on texture and local features. As outlined above, we find the models we evaluate mostly capable of accounting for the developmental phenomena we study. However, we evaluate the similarity between static representations generated by the networks (see General Methodology), while infant categorization in the familiarization-test paradigm can be a more dynamic process (Althaus et al., 2020; Schöner & Thelen, 2006). Future work could explore models that capture additional process-level details, such as by adapting the autoencoder models of French et al. (2004) or self-organizing maps of Althaus et al. (2020) to operate over the types of feature embeddings we use here, extracted from modern computer vision networks.

When infants discriminate between categories in a laboratory study, it is often unclear whether these abilities reflect top-down processing of categories acquired outside the lab, or bottom-up processing of categories developed during the familiarization phase (Thelen and Smith, 1994; Murphy, 2002, ch. 9; French et al., 2004, Newcombe et al., 2005). Our supplemental experiment in Appendix C directly examines the learnability of relational categories using a supervised learning paradigm on datasets ranging from as few as 8 examples to a few thousand data points. We view this as analogous to the first possibility, of learning relation concepts from numerous varied examples, as infants might acquire these categories over an extended period outside the lab. The pretrained computer vision models used in Experiments 1–4 do not separate between the top-down and bottom-up hypotheses. The pretraining process guides the model in acquiring useful perceptual features to represent its inputs, which may also serve to promote relational similarity in the models' embeddings. These models may also acquire a more abstract latent concept of the different relations—as we cannot rule this possibility out, we cannot adjudicate between top-down processing of prior categories and bottom-up processing of categories developed in familiarization. We consistently find that the networks we evaluate perform better on the single-reference object relations (e.g. *above/below*) compared to the relations requiring reasoning with respect to multiple objects (e.g. *between*), including the models trained explicitly to classify relations in Appendix C. In infants, this is explained by a developmental transition from encoding location with respect to single landmarks, to encoding local spatial frameworks with respect to multiple objects (Huttenlocher & Newcombe, 1984; Quinn, 2003). Although we do not explicitly test this hypothesis, in terms of representations formed by the networks, we find that adding additional reference objects to *above/below* stimuli does not make the task harder (Appendix Fig. B.5 in Appendix B.3). This suggests the difficulty of the task might be more related to the number of objects required to make a relational categorization, rather than merely the number of objects present in a scene.

Finally, our work allows us to make an experimental prediction and raise a source of uncertainty. In Experiment 2b, we discovered that the pretrained models we evaluated reach high levels of accuracy when stimuli are presented at a 45° angle, unlike the infants evaluated by Quinn (2004, Experiment 3). The high levels of accuracy reached

by the models make the prediction that slightly older infants (e.g. 6-7-months-old) than those evaluated by Quinn (2004) would demonstrate evidence for a category representation for an object on either side of a diagonal line. We also note a lack of experimental evidence (to the best of our awareness) for whether or not infants construct category representations for the containment relation from static stimuli. Both experimental work (Casasola & Cohen, 2002; Casasola et al., 2003) and computational models (Ullman et al., 2019) rely on dynamic video stimuli. Further experimental work could demonstrate at what stage of development a categorical response to still image stimuli depicting the containment relation is acquired, which would shed light on the discrepancy between our findings and existing experimental results.

CRediT authorship contribution statement

Guy Davidson: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **A. Emin Orhan:** Resources, Software. **Brenden M. Lake:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank Wai Keen Vong for helpful ideas and thoughtful discussions at several points in time over this project's various stages. We appreciate the feedback we received from various members of the Human and Machine Learning Lab at NYU during lab meetings. We would like to thank Alexa Tartaglini, Pat Little, Reuben Feinman, and Wai Keen Vong for their comments on the manuscript. Finally, we would like to thank Kenna Heller for her work on Fig. 1.

This work was supported by the DARPA Machine Common Sense program and National Science Foundation Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science.

Appendix A–C. Supplementary methods and results

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105690.

References

Althaus, N., Gliozzi, V., Mayor, J., & Plunkett, K. (2020). Infant categorization as a dynamic process linked to memory. *Royal Society Open Science*, *7*.

Baldassarre, F., Smith, K., Sullivan, J., & Azizpour, H. (2020). Explanation-based weakly-supervised learning of visual relations with graph networks. In *LNCS: vol. 12373, Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 612–630). Springer Science and Business Media Deutschland GmbH.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., ... Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2021). From convolutional neural networks to models of higher-level cognition (and back again). *Annals of the New York Academy of Sciences*, *1505*, 55–78.

Blender Online Community (2018). *Blender - a 3D modelling and rendering package*. Amsterdam: Blender Foundation, Stichting Blender Foundation.

Bomba, P. C., & Siqueland, E. R. (1983). The nature and structure of infant form categories. *Journal of Experimental Child Psychology*, *35*, 294–328.

Bowers, J. S., Malhotra, G., Dujmovic, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 9650–9660).

Casasola, M. (2008). The development of infants' spatial categories. *Current Directions in Psychological Science*, *17*(1), 21–25.

Casasola, M., & Cohen, L. B. (2002). Infant categorization of containment, support and tight-fit spatial relationships. *Developmental Science*, *5*(2), 247–264.

Casasola, M., Cohen, L. B., & Chiarello, E. (2003). Six-month-old infants' categorization of containment spatial relations. *Child Development*, *74*(3), 679–693.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Eimas, P. D., & Quinn, P. C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, *65*, 903–917.

Fagan, J. F. (1970). Memory in the infant. *Journal of Experimental Child Psychology*, *9*, 217–226.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668–670.

French, R. M., Mermillod, M., Mareschal, D., & Quinn, P. C. (2004). The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: Simulations and data. *Journal of Experimental Psychology: General*, *133*(3), 382–397.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *WIREs Cognitive Science*, *1*(1), 69–78.

Huttenlocher, J., & Newcombe, N. S. (1984). The child's representation of information about location. In C. Sophian (Ed.), *Origin of cognitive skills* (pp. 81–111). Hillsdale, NJ: Erlbaum.

Jaini, P., Clark, K., & Geirhos, R. (2023). Intriguing properties of generative classifiers.

Johnson, S. P. (2010). How infants learn about the visual world. *Cognitive Science*, *34*(7), 1158–1184.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*, 2017–2031.

Liu, N., Li, S., Du, Y., Tenenbaum, J. B., & Torralba, A. (2021). Learning to compose visual relations. *Advances in Neural Information Processing Systems*, *28*, 23166–23178.

Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, *36*(5), 635–645.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Newcombe, N. S., Sluzenski, J., & Huttenlocher, J. (2005). Preexisting knowledge versus on-line learning: What do young infants really know about spatial location? *Psychological Science*, *16*(3), 222–227.

Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development*, *11*, 268.

Orhan, A. E., Gupta, V. V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In *NeurIPS 2020*. arXiv.

Orhan, A. E., & Lake, B. M. (in press). Learning high-level visual representations from a child's perspective without strong inductive biases. *Nature Machine Intelligence*.

Piaget, J. (1954). *The construction of reality in the child*. New York, NY: Basic Books.

Quinn, P. C. (1994). The categorization of above and below spatial relations by young infants. *Child Development*, *65*(1), 58–69.

Quinn, P. C. (2002). Category representation in young infants. *Current Directions in Psychological Science*, *11*(2), 66–70.

Quinn, P. C. (2003). Concepts are not just for objects: Categorization of spatial relation information by infants. In D. H. Rakison, & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion*. Oxford University Press.

Quinn, P. C. (2004). Spatial representation by young infants: Categorization of spatial relations or sensitivity to a crossing primitive? *Memory and Cognition*, *32*(5), 852–861.

Quinn, P. C., Adams, A., Kennedy, E., Shettler, L., & Wasnik, A. (2003). Development of an abstract category representation for the spatial relation between in 6- to 10-month-old infants. *Developmental Psychology*, *39*(1), 151–163.

Quinn, P. C., Cummins, M., Kase, J., Erin, M., & Weissman, S. (1996). Development of categorical representations for above and below spatial relations in 3- to 7-month-old infants. *Developmental Psychology*, *32*(5), 942–950.

Quinn, P. C., Norris, C. M., Pasko, R. N., Schmader, T. M., & Mash, C. (1999). Formation of a categorical representation for the spatial relation between by 6- to 7-month-old infants. *Visual Cognition*, *6*(5), 569–585.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, *139*, 8748–8763.

Regier, T. (1995). A model of the human capacity for categorizing spatial relations. *Cognitive Linguistics*, *6*(1), 63–88.

Rumelhart, D. E., McClelland, J. L., & Group, P. R. (1987). *Parallel distributed processing*. The MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. In *IJCV*.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *CVPR 2018*. IEEE Computer Society.

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T., & London, D. (2017). A simple neural network module for relational reasoning. In *NeurIPS 2017*.

Saxe, A. M., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., & Ng, A. Y. (2011). On random weights and unsupervised feature learning. In *Proceedings of the 28th international conference on international conference on machine learning* ICML '11, (pp. 1089–1096). Madison, WI, USA: Omnipress.

Saxe, A., Nelli, S., & Summerfield, C. (2020). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67.

Schöner, G., & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review*, *113*, 273–299.

Shanahan, M., Nikiforou, K., Creswell, A., Kaplanis, C., Barrett, D., & Garnelo, M. (2019). An explicitly relational neural network architecture.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*, 1–48.

Slater, A. (1995). Visual perception and memory at birth. In *Advances in infancy research, Vol. 9* (pp. 107–162).

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology*, *8*, Article 296143.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. (2020). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective.

Tartaglini, A. R., Vong, W. K., & Lake, B. M. (2022). A developmentally-inspired examination of shape versus texture bias in machines. In *Proceedings of the 44th annual meeting of the cognitive science society: Cognitive diversity, CogSci 2022* (pp. 1284–1290). The Cognitive Science Society.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Ullman, S., Dorfman, N., & Harari, D. (2019). A model for discovering 'containment' relations. *Cognition*, *183*, 67.

van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-SNE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *NeurIPS 2017*. Long Beach, CA, USA.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (in press). Grounded language acquisition through the eyes and ears of a single child. *Science*.

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *369*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *CVPR 2017* (pp. 5987–5995).

Younger, B. A., & Cohen, L. B. (1985). How infants form categories. In *Psychology of learning and motivation - Advances in research and theory, Vol. 19* (pp. 211–247). Academic Press.