RESEARCH ARTICLE

Perception of an object's global shape is best described by a model of skeletal structure in human infants

Vladislav Ayzenberg[1]*, Stella Lourenco[2]

[1]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, United States; [2]Department of Psychology, Emory University, Atlanta, United States

Abstract Categorization of everyday objects requires that humans form representations of shape that are tolerant to variations among exemplars. Yet, how such invariant shape representations develop remains poorly understood. By comparing human infants (6–12 months; N=82) to computational models of vision using comparable procedures, we shed light on the origins and mechanisms underlying object perception. Following habituation to a never-before-seen object, infants classified other novel objects across variations in their component parts. Comparisons to several computational models of vision, including models of high-level and low-level vision, revealed that infants' performance was best described by a model of shape based on the skeletal structure. Interestingly, infants outperformed a range of artificial neural network models, selected for their massive object experience and biological plausibility, under the same conditions. Altogether, these findings suggest that robust representations of shape can be formed with little language or object experience by relying on the perceptually invariant skeletal structure.

Editor's evaluation

This well-conducted study uses relatively large sample sizes, comprehensive statistical testing, and state-of-the-art modeling to provide novel evidence that human infants generalize shape from single examples on the basis of the "shape skeleton", a structural description of the part structure of the shape. It will be of interest to researchers working on object shape processing and on the development of visual perception.

***For correspondence:** vayzenb@cmu.edu

Competing interest: The authors declare that no competing interests exist.

Funding: See page 15

Received: 22 October 2021
Accepted: 09 May 2022
Published: 25 May 2022

Reviewing Editor: Marius V Peelen, Radboud University, Netherlands

Introduction

The appearance of objects within a single category can vary greatly. For instance, despite the shared category of dog, the exemplars (i.e. different dogs) may have different snouts, tails, and/or torsos. Despite this variability, humans readily categorize never-before-seen dog breeds as members of the same basic-level category. How do we do this? Although there is widespread agreement that shape information is crucial for object categorization (***Biederman, 1995***; ***Mervis and Rosch, 1981***), it remains unclear how humans come to form global representations of shape that are tolerant to variations among exemplars.

It has been suggested that global shape information becomes crucial for object categorization in early childhood because linguistic experience, particularly the learning of object labels (e.g. 'dog'), draws children's attention to global shape as a diagnostic cue—inducing a so-called 'shape bias' (***Landau et al., 1998***; ***Smith et al., 1996***; ***Smith et al., 2002***). Indeed, labels may bootstrap object recognition abilities more generally, such that even prelinguistic children are better at individuating and categorizing objects when verbal labels are provided (***Ferry et al., 2010***; ***Xu et al., 2005***). The

advantage of labeled object experience is particularly evident in supervised artificial neural networks (ANNs), which have begun to match the object recognition abilities and neural representations of human adults (***Krizhevsky et al., 2017***; ***Rajalingham et al., 2018***; ***Schrimpf et al., 2018***). These models learn the diagnostic properties of objects following training with millions of labeled naturalistic images. With appropriate experience, these models may even develop a shape bias that supports object categorization, at least when generalizing across variations in color or texture (***Ritter et al., 2017***; ***Tartaglini et al., 2022***). Thus, global representations of shape may develop with labeled object experience that highlights the diagnostic properties of objects.

An alternative possibility, however, is that rather than labeled experience, humans develop global shape representations by relying on (non-linguistic) invariant perceptual properties inherent to the object (***Biederman, 1987***; ***Feldman, 1997***; ***Rakison and Butterworth, 1998***; ***Sloutsky, 2003***). One such property is known as the shape skeleton—a quantitative model that describes an object's global shape via a series of internal symmetry axes (***Blum, 1967***; ***Feldman and Singh, 2006***). These axes define the topological arrangement of object parts, making models of skeletal structure tolerant to local variations in shape typical of basic-level exemplars (***Ayzenberg et al., 2019a***; ***Wilder et al., 2011***). From this perspective, extensive experience with objects and linguistic labels may not be necessary to form global shape representations, and, instead, one might rely on the shape skeleton (***Feldman, 1997***). However, the contributions of labeled experience and the shape skeleton are difficult to examine because, by adulthood, humans have had massive amounts of labeled object experience, making the source of their shape representations ambiguous. Here we tested whether human infants (who have little linguistic or object experience) represent global object shape according to a shape skeleton.

Object representations in infancy have most often been tested using visual attention procedures. In these experiments, infants are typically habituated to stimuli that share a common visual dimension (e.g. shape), but vary according to other dimensions (e.g. color). Infants are then tested with objects that are either familiar (e.g. similar in shape to the habituated object) or novel (i.e. different in shape to the habituated object). If infants learn the relevant dimension during the habituation phase, then their looking times are longer for the novel object compared to the familiar one. Habituation paradigms provide an informative window into infants' object representations because they reveal what properties infants learned during the habituation phase and subsequently generalized to the test phase. Using this approach, researchers have shown that newborns can already discriminate between simple 2D shapes (***Slater et al., 1983***) and display shape constancy, such that they recognize a shape from a novel orientation (***Slater and Morison, 1985***). By 6 months of age, infants' shape representations are also robust to variations among category exemplars, such that they can categorize objects using only the stimulus' shape silhouette (***Quinn et al., 1993***; ***Quinn et al., 2001a***), as well as extend category membership to objects with varying local contours, but the same global shape (***Quinn et al., 2002***; ***Quinn et al., 2001b***; ***Turati et al., 2003***). However, the mechanisms underlying global shape representation remain unclear. Indeed, because infants in these studies were habituated to multiple (often familiar) objects, it is unclear whether shape representations in these studies were learned from the statistics of the habituation period (***Oakes and Spalding, 1997***; ***Younger, 1990***) or, rather, are an invariant perceptual property infants extract from objects more generally.

In the current study, we used a habituation paradigm to examine how 6–12 months old infants represent the global shape of objects. In particular, we tested whether infants classified never-before-seen objects by comparing the similarity between their shape skeletons. Importantly, we habituated infants to only a single object so as to measure their pre-existing shape representations, rather than ones they may have learned over the course of habituation. We chose to test 6–12-month-olds because the visual experience at this age is dominated by just a few common objects (~10 objects; ***Clerkin et al., 2017***) and they have relatively little linguistic understanding (~7 words; ***Bergelson and Swingley, 2012***). Moreover, we elucidate the mechanisms that support object perception in infancy by comparing infants to a range of computational models. Of particular interest was a flux-based medial axis algorithm that computes a shape skeleton from images (***Rezanejad and Siddiqi, 2013***). If infants represent objects using the shape skeleton, then their classification judgments should best match that of a Skeletal model.

In addition, we included four ANNs, known for their success on object recognition tasks but which do not represent the shape skeleton. These ANNs included two ResNet models, one trained on

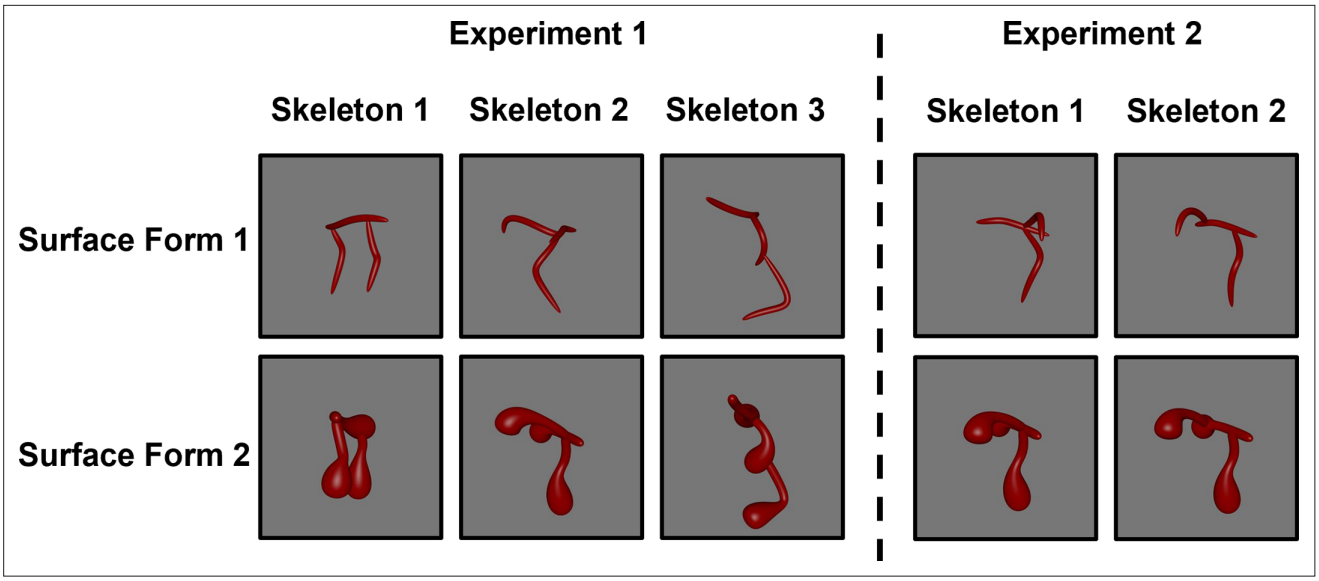

Figure 1. Screen shots of the stimuli used in Experiment 1 (left) and Experiment 2 (right). Objects were presented as rotating videos during habituation and test phases.

ImageNet (ResNet-IN), a convolutional ANN frequently used in object recognition tasks, and a variation of ResNet trained on Stylized-ImageNet (ResNet-SIN), an image set that leads models to develop a shape bias, at least in relation to color and textures cues (***Geirhos et al., 2018***). Two other ANNs were CorNet-S, a top-performing model of object recognition behavior and neural processing in primates, as measured by the brain-score benchmark (***Schrimpf et al., 2018***), and ResNext-SAY, a model trained with an unsupervised learning algorithm on first-person videos from infants (***Orhan et al., 2020***). All of these ANNs were included because they exhibit varying degrees of biological plausibility in terms of neural organization or visual experience (***Russakovsky et al., 2015***; ***Schrimpf et al., 2018***). If infant performance is best matched by ANNs, then this would suggest that global shape representations might be learned as a diagnostic cue following extensive object experience. Importantly, because none of these models represent the shape skeleton, they make for an excellent contrast to the Skeletal model. Finally, we also included a model of pixel similarity, and FlowNet, a model of optic flow (***Ilg et al., 2017***) in order to assess the extent to which shape representations may be supported by lower-level visual properties like image similarity (***Kiat et al., 2022***; ***Xie et al., 2021***) or motion trajectory (***Kellman, 1984***; ***Kellman and Short, 1987***; ***Wood and Wood, 2018***). Altogether, these comparisons provided a novel approach to understanding object perception in human infants.

Because the strength of any object classification task depends on the degree of dissimilarity between training (i.e. habituation) and test objects, infants were tested with objects that mimicked the variability of basic-level category exemplars (***Figure 1***). Within-category objects comprised objects with the same skeletons, but visually distinct component parts. Variation in the component parts was generated by manipulating the objects' surface forms, thereby changing both the image-level features and the non-accidental properties (NAPs; ***Biederman, 1987***), without altering the skeleton. Between-category objects comprised different skeletons and surface forms. If infants are capable of classifying objects vis-à-vis a shape skeleton, then they should look longer at objects with different skeletons compared to those with the same skeleton, even though both objects differ from the habituated object in surface form. Importantly, we tested whether within-category and between-category objects were equally discriminable by infants to ensure that any differences in looking time were not related to infants' ability to differentiate surface forms. Moreover, if infants rely on objects' shape skeletons to determine similarity, then their classification performance should be best matched by the Skeletal model, rather than ANNs, or other models of vision. Thus, the current study provides critical insight regarding the nature of shape representations at an age when experience with labeled objects is minimal, and, crucially, provides a novel benchmark by which to evaluate the biological plausibility of computational models of vision.

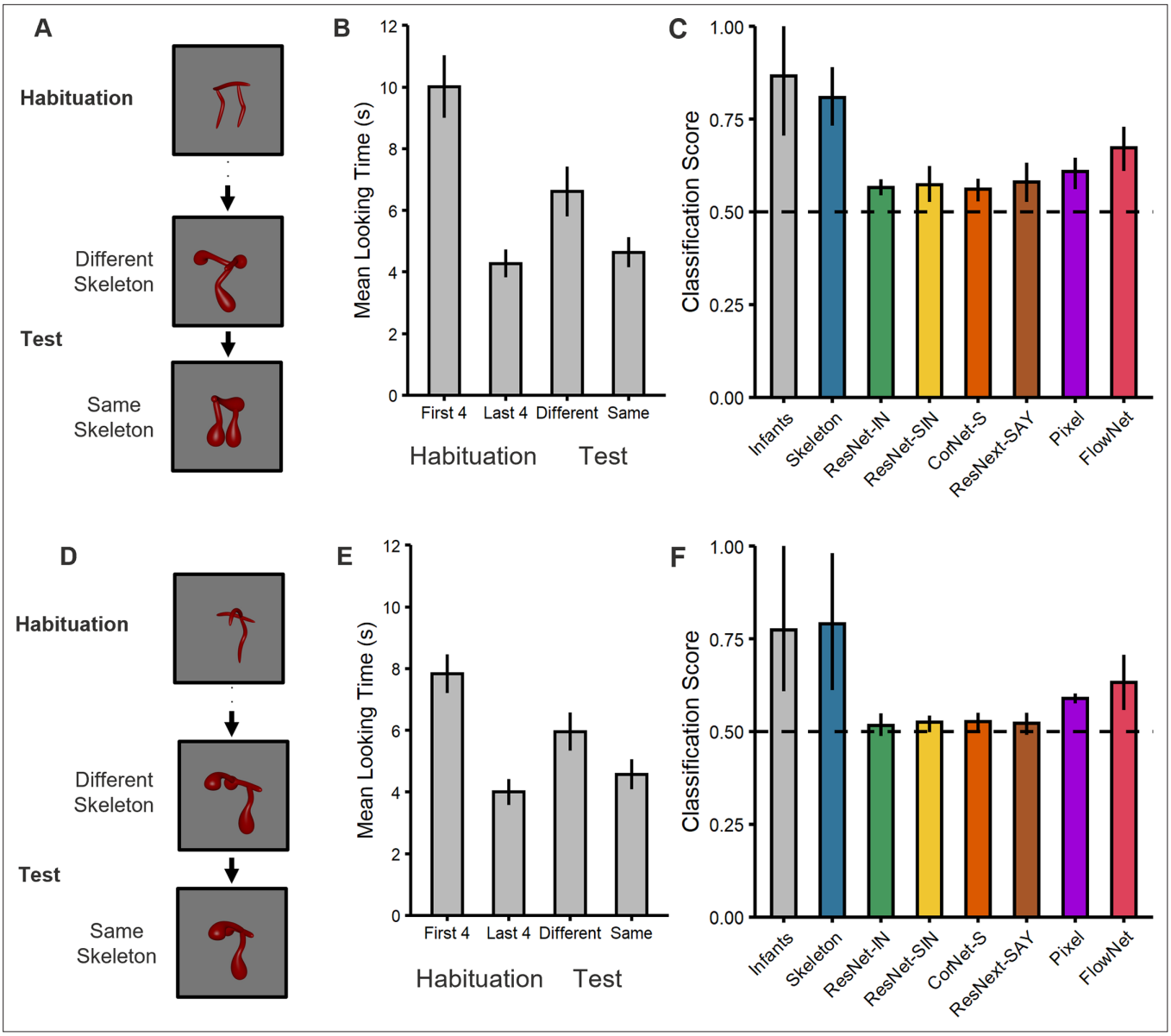

Figure 2. Experimental design and results for (top) Experiment 1 and (bottom) Experiment 2. (**A, D**) Illustration of the experimental procedure administered to infants and the computational models in (**A**) Experiment 1 and (**D**) Experiment 2. Infants and models were habituated to one object and then tested with objects that consisted of either the same or different shape skeleton. Both types of test objects (counterbalanced order) differed in their surface forms from the habituation object. (**B, E**) Mean looking times for (**B**) Experiment 1 and (**E**) Experiment 2. For the habituation phase, results are shown for the first four and last four trials. For the test phase, results are shown for the two types of test objects (i.e. same and different skeletons; 3 test trials each). Error bars represent SE. (**C, F**) Classification performance for infants and models for (**C**) Experiment 1 and (**F**) Experiment 2. Error bars represent bootstrapped confidence intervals, and the dashed line represents chance performance.

Results

Infants' looking times were analyzed using two-sided paired sample *t*-tests (α=0.05) and standard measures of effect size (Cohen's *d*). Furthermore, to ensure that sample size decisions did not unduly influence the results, we also conducted non-parametric (Binomial tests) and Bayesian analyses. A Bayes factor (BF_{10}) was computed for each analysis using a standard Cauchy prior (*d*=0.707). A BF_{10} greater than 1 is evidence that two distributions are different from one another, whereas a BF_{10} less than 1 is evidence that two distributions are similar to one another (***Rouder et al., 2009***).

How do infants represent shape?

A comparison of the two types of test trials in Experiment 1 (***Figure 2A***) revealed that, across the test phase, infants looked longer at the object with the different skeleton compared to the one with the matching skeleton (t[33]=3.04, p=0.005, d=0.52, 95% CI [0.16, 0.88], BF_{10}=8.42; ***Figure 2B***), with the majority of infants showing this effect (25/34 infants, p=0.009). In addition, a comparison between the end of habituation (mean of last 4 trials) and looking times across the test phase revealed that dishabituation only occurred for the object with the different skeleton (t[33]=3.36, p=0.002, d=0.58, 95% CI [0.21, 0.94], BF_{10}=17.47; 25/34 infants, p=0.009), not the object with the matching skeleton (t[33]=1.00, p=0.325, d=0.17, 95% CI [–0.16, 0.51], BF_{10}=0.29; ***Figure 2B***). Thus, infants treated objects with matching skeletons as more similar to one another than objects with different skeletons.

However, one might ask whether infants were simply unable to differentiate between surface forms, leading to greater looking at the object with a different skeleton. To test this possibility, we compared infants' looking times on the first test trial following habituation to the last trial during habituation. Because the first test trial immediately follows habituation, this comparison allows for a direct measure of perceptual discriminability between habituation and test objects when the memory demands are minimal and in the absence of carry-over effects between test objects. This analysis revealed that infants were indeed capable of discriminating objects on the basis of surface form alone. That is, infants dishabituated to the object with the same skeleton but different surface form, t(15) = 3.76, p=0.002, d=0.94, 95% CI [0.34, 1.52], BF_{10}=22.23, with the majority of infants showing this effect (14/16, p=0.004). Moreover, and crucially, infants' looking times on the first test trial did not differ for either test object (8.31 s vs. 9.18 s; t[32]=0.50, p=0.624, d=.17, 95% CI [–0.51, 0.84], BF_{10}=0.36). These findings demonstrate that not only could infants differentiate between surface forms, but also, that the two types of test objects were matched for discriminability relative to the habituation object. These findings argue against a pure discrimination account and, instead, support the interpretation that infants classified objects on the basis of skeletal similarity.

But did infants actually rely on the object's shape skeleton to judge similarity or some other visual representation? To explore this possibility, we compared infants' classification behavior to a flux-based Skeletal model, a range of ANNs (ResNet-IN, ResNet-SIN, and CorNet-S, ResNext-Say), and models of image similarity (Pixel) and motion (FlowNet). If infants relied on the shape skeleton to classify objects, then their performance would be best matched by the Skeletal model, rather than the others.

Models were tested with the same stimuli presented to infants using a procedure comparable to the habituation paradigm. More specifically, because habituation/dishabituation can be conceived as a measure of alignment between the stimulus and the infant's internal representation (***Mareschal et al., 2000***; ***Westermann and Mareschal, 2004***), we tested models by feeding their outputs into an autoencoder and measuring the error signal across habituation and test phases (see Methods). Like habituation paradigms, the error signal of an autoencoder reflects the degree of alignment between the internal representation of the model and the input stimulus (for review, see ***Yermolayeva and Rakison, 2014***). Unlike conventional classifiers, which often require multiple labeled contrasting examples (e.g. Support Vector Machines), autoencoders allow for measuring a model's performance following exposure to just one exemplar, as with infants in the current study. Moreover, like infant learning during habituation, the learned representation of an autoencoder reflects the entire habituation video, rather than the representation of individual frames. Most importantly, unlike other techniques, autoencoders can be tested using the same habituation and test criteria as infants (see Methods). For comparison, performance for both infants and models was converted into a classification score (see Methods) and significance was assessed using bootstrapped confidence intervals (5000 iterations).

These analyses revealed that all models performed above chance (0.50; see ***Figure 2C***). However, and importantly, infant performance was best matched to the Skeletal model. Both infants and the Skeletal model performed significantly better than all other models, except FlowNet. That infants outperformed the ANNs suggests that extensive object experience may not be necessary to develop robust shape representations. Likewise, that infants outperformed the Pixel model suggests that infant performance is not explained by the low-level visual similarity between objects. However, the success of FlowNet does suggest that motion information may contribute to representations of shape. Nevertheless, FlowNet's performance did not differ from that of any ANN or the Pixel model, leaving its exact role in object classification unclear. Altogether, these results suggest that infants' performance is

most closely aligned with the Skeletal model and, thus, that infants classified objects, at least in part, by relying on the similarity between objects' shape skeletons.

Can infant performance be explained by another representation of global shape?

An alternative explanation for the findings from the first experiment is that, rather than the shape skeleton, infants classified objects using another representation of global shape—namely, the coarse spatial relations among object parts (***Biederman and Gerhardstein, 1993***; ***Hummel and Stankiewicz, 1996***). Whereas a Skeletal model provides a continuous, quantitative description of part relations, a model based on coarse spatial relations describes part relations in qualitative terms (e.g. two parts below a third vs. two parts on either side of a third). In Experiment 1, test objects with different skeletons also consisted of part relations that could be considered qualitatively different from that of the habituated object, making it unclear whether infants relied on coarse spatial relations instead of the shape skeletons.

To address this possibility, in Experiment 2, coarse spatial relations were held constant between habituation and test phases (i.e. two parts on either side of a third; ***Figures 1 and 2D***). Thus, if infants relied on coarse spatial relations for object classification, then they would fail to dishabituate to the test object that differed in the shape skeleton (but not coarse spatial relations). However, we found that infants continued to look longer at the object with the different skeleton compared to the one with the matching skeleton (t[47]=2.60, p=0.012, d=0.38, 95% CI [0.08, 0.67], BF_{10}=3.18; ***Figure 2E***), with the majority of infants showing this effect (33/48 infants, p=0.013). Moreover, infants only dishabituated during the test phase to the object with the different skeleton (t[47]=3.63, p<0.001, d=0.52, 95% CI [0.22, 0.82], BF_{10}=39.77; 36/48 infants, p<0.001), not the one with the matching skeleton (t[47]=1.42, p=0.163, d=0.16, 95% CI [–0.08, 0.49], BF_{10}=0.40), as in Experiment 1. The results from this experiment rule out the possibility that infants classified objects on the basis of their coarse spatial relations, rather than their shape skeletons.

Importantly, as in the previous experiment, we also compared infants' looking times on the first test trial to the last trial during habituation to ensure that infants could distinguish surface forms and that the two test objects were equally discriminable. We found that infants dishabituated to the object with the same shape skeleton but different surface form, t(22)=3.51, p=0.002, d=0.73, 95% CI [0.26, 1.19], BF_{10}=19.71, with the majority of infants showing this effect (17/23, p=0.035), suggesting that discrimination was possible on the basis of surface form alone. Moreover, infants' looking times on the first test trial did not differ for the two types of test objects (6.87 s vs. 10.83 s; t[45]=1.47, p=0.149, d=0.43, 95% CI [–0.15, 1.00], BF_{10}=0.69), confirming comparable discriminability. Thus, as in Experiment 1, these findings suggest that although infants were capable of discriminating both types of test objects, they nevertheless treated objects with the same skeletons as more similar to one another.

To ensure that the effects in Experiment 2 were not unduly influenced by the larger sample size, we computed bootstrapped CIs on a smaller sample. For each bootstrap procedure (10,000 iterations), we calculated Cohen's d on data that were resampled (without replacement) to match the sample size of Experiment 1 (n=34). We found that infants dishabituated to the test object with a different skeleton, 95% CI [0.22, 0.87], but not the same skeleton, 95%CI [–0.13, 0.61]. Infants also looked longer at the test object with a different skeleton than the test object with the same skeleton, 95% CI [0.20, 0.61], confirming the robustness of these effects regardless of sample size. Finally, analyses of the first test trial confirmed that infants discriminated between surface forms when objects had the same skeletons, 95% CI [0.22, 0.81,] and that looking times did not significantly differ between the two types of test trials, 95% CI [–0.15, 0.49].

To examine whether infants' performance was best described by a Skeletal model we, again, compared infants to a model that represents the shape skeleton, as well as to models that do not. These analyses revealed that the Skeletal, Pixel, and FlowNet models all performed above chance (***Figure 2F***), but infant performance was most closely matched to the Skeletal model. None of the ANNs performed above chance. As in Experiment 1, both infants and the Skeletal model significantly outperformed the ANNs and the Pixel model, but not FlowNet. These findings again suggest that infants' judgments reflect the use of skeletal structure and that there may be a role for motion information in this ability.

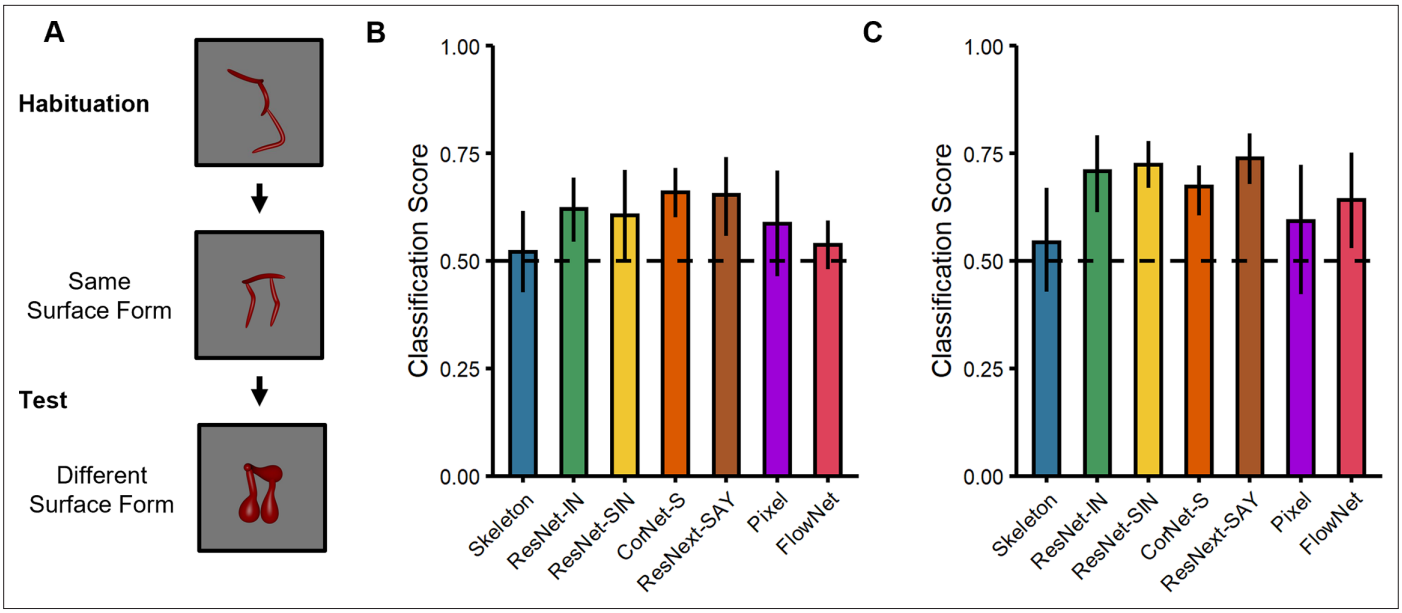

Figure 3. Experimental design and results for the surface form classification task used with the computational models. (**A**) Illustration of the experimental procedure administered to models. (**B–C**) Classification performance of models on stimuli from (**B**) Experiment 1 and (**C**) Experiment 2. Error bars represent bootstrapped confidence intervals and dashed lines represent chance performance.

Model classification using local shape properties

Given the competitive performance of ANNs on other object recognition tasks, and their match to adult performance in other studies (***Schrimpf et al., 2018***), one might wonder why ANNs performed so poorly on our task. One possibility is that ANNs rely on local shape properties (e.g. surface form), which were irrelevant to our task (***Baker et al., 2018***; ***Baker et al., 2020***). To test this possibility, we examined whether ANNs were capable of classifying objects using local shape properties, namely, the surface forms of the objects. For comparison, we also tested the other models' performance on this task (Skeleton, Pixel, and FlowNet).

Using a comparable procedure to the previous experiments, all models were tested with objects in which the surface forms either matched or differed from the habituated object; both test objects differed in their shape skeletons (***Figure 3A***). For objects from Experiment 1 (***Figure 1***), the analyses revealed that ResNet-IN, CorNet-S, and ResNext-SAY performed above chance (Ms = 0.62–0.66, 95% CIs [0.53–0.59, 0.70–0.76]; ***Figure 3B***). By contrast, none of the other models (Skeleton, ResNet-SIN, Pixel, FlowNet) performed differently from chance (Ms = 0.52–0.61, 95% CIs [0.40–0.48, 0.60–0.72]; ***Figure 3B***) when classifying objects by surface form (i.e. across shape skeletons). For objects from Experiment 2 (***Figure 1***), all models performed above chance (Ms = 0.67–0.74, 95% CIs [0.51–0.67, 0.72–0.80]), except Skeletal and Pixel models (Ms = 0.54–0.59, 95% CIs [0.40–0.41, 0.68–0.73]; ***Figure 3C***). Altogether, these findings demonstrate that most ANNs are capable of classifying objects using local shape properties. These findings are consistent with other research showing sensitivity to local shape properties in ANNs (***Baker et al., 2018***).

General discussion

Although it is well known that shape is crucial for object categorization, much speculation remains about how representations of global shape are formed in development. Here, we demonstrate that infants represent global shape by extracting a skeletal structure. With only one exemplar as a reference, infants classified objects by their shape skeletons across variations in local properties that altered component parts of the objects, including when coarse spatial relations could not be used as a diagnostic cue. Moreover, a Skeletal model provided the closest match to infants' performance, further suggesting that infants relied on the shape skeleton to determine object similarity. Based on these findings, we would argue that the formation of robust shape representations in humans is largely rooted in sensitivity to the shape skeleton, an invariant representation of global shape.

The role of other visual properties in object perception

It is important to acknowledge that our results also suggest that object classification can be accomplished on the basis of image-level similarity. In particular, when generalizing across the surface form, a Pixel model performed above chance. Does this mean that infants' performance can be explained by low-level image similarity between habituation and test objects rather than the shape skeleton? We would suggest not. First, we found that infants discriminated both test objects from the habituation object equally, which argues against the possibility that infants' performance was strictly based on differences in image-level similarity. Second, both infants and the Skeletal model outperformed the Pixel model in both experiments, further arguing against an account based on image-level similarity alone. Nevertheless, we would not argue that image similarity plays no role in object perception, particularly in infants. Indeed, recent studies comparing the visual representations of infants and computational models reveal that low-level visual similarity explains more variance in infants' behavioral and neural responses than the upper layers of ANNs (***Kiat et al., 2022***; ***Xie et al., 2021***). Moreover, recent studies suggest that object categorization in infancy may be supported by the representations of the early visual cortex (V1-V3), rather than the higher-level ventral cortex, as in adults (***Spriet et al., 2022***).

Like image-level representations, shape skeletons, themselves might be an emergent property of early visual regions. Specifically, several studies have suggested that skeletal representations may emerge in area V3 as a consequence of between-layer recurrent interactions among border-ownership cells in area V2 and grouping cells in area V3 (***Ardila et al., 2012***; ***Craft et al., 2007***). Interestingly, representations of skeletal structures in V3 are invariant across changes in component parts, which lends support to their role in categorization (***Ayzenberg et al., 2022***; ***Lescroart and Biederman, 2013***). Other research on the anatomical and functional organization of V3 in primates shows evidence of functional maturity shortly after birth (***Arcaro and Livingstone, 2017***; ***Ellis et al., 2020***; ***Wiesel and Hubel, 1974***), further raising the possibility that skeletal structure is represented in this area with little object experience or any language.

In the present study, we also found that FlowNet, a model of optic flow, appeared comparable to infants and the Skeletal model at classifying objects when they differed in surface form. This finding highlights the role of motion in the formation of invariant object representations (***Lee et al., 2021***; ***Ostrovsky et al., 2009***). It is known that infants are better at recognizing objects from novel viewpoints when they are first familiarized with moving objects rather than static images (***Kellman, 1984***; ***Kellman and Shipley, 1991***; ***Kellman and Short, 1987***). Moreover, controlled-rearing studies with chicks have found that viewpoint-invariant object recognition occurs only if the chicks are raised in environments in which objects exhibit smooth, continuous motion (***Wood and Wood, 2018***). Indeed, several studies suggest that motion information may initially bootstrap infants' ability to extract 3D shape structure (***Kellman and Arterberry, 2006***; ***Ostrovsky et al., 2009***; ***Wood and Wood, 2018***). Thus, motion information may work in concert with shape skeletons to support object perception in infancy. Yet, it is important to note that classifying objects on our task was also possible without relying on motion cues (***Quinn et al., 2001a***; ***Quinn et al., 2001b***). Indeed, the Skeletal model, which does not incorporate any motion information, was more closely aligned to the performance of human infants than was FlowNet.

In contrast to infants and the other models, ANNs were not capable of classifying objects across variations in surface form. However, they were capable of classifying objects by their surface form (across variation in shape skeleton), ruling out alternative explanations for their poor performance in the first task based on idiosyncrasies of the stimuli or testing procedures. Moreover, this finding suggests that, regardless of the specific architecture or training type, conventional ANNs rely on qualitatively different mechanisms than do humans to represent objects. This class of models may be especially sensitive to local object properties, making them susceptible to spurious changes in the objects' contours (***Baker et al., 2018***; ***Baker et al., 2020***). By contrast, representations of skeletal structure in humans are particularly robust to perturbations (***Ayzenberg et al., 2019a***; ***Feldman and Singh, 2006***) and, thus, may be especially well-suited to basic-level categorization. Although conventional ANNs (e.g. ResNet-SIN) can generalize across variations in color and texture (***Geirhos et al., 2018***; ***Tartaglini et al., 2022***), fundamental changes to ANNs' architectures and/or training may be needed before they can achieve human-like categorization on the basis of global shape.

Despite our claim that infants relied on skeletal similarity to classify objects, we would not suggest that it is the only information represented by humans. Indeed, by adulthood, humans also use visual properties such as texture (***Jagadeesh and Gardner, 2022***; ***Long et al., 2018***) and local contours (***Davitt et al., 2014***), as well as inferential processes such as abstract rules (***Ons and Wagemans, 2012***; ***Rouder and Ratcliff, 2016***) and the object's generative history (***Fleming and Schmidt, 2019***; ***Spröte et al., 2016***) to reason about objects. Properties such as texture and local features may even override shape skeletons in certain contexts, such as when identifying objects in the periphery (***Gant et al., 2021***) or during subordinate-level categorization (***Davitt et al., 2014***; ***Tarr and Bülthoff, 1998***). We suggest that the shape skeleton may be uniquely suited to the basic level of categorization, wherein objects have similar global shapes but vary in their component parts (***Biederman, 1995***; ***Rosch et al., 1976***). However, more research will be needed to understand the extent to which children may also make use of other properties when categorizing objects.

Implications for one-shot categorization

A key aspect of our design was that infants were habituated to a single, novel object. We did this to better understand the pre-existing visual representations infants rely on for object perception, rather than the ones they may learn over the course of habituation. Using this approach, we found that infants reliably classified objects on the basis of the shape skeleton. Might this result also suggest that infants are capable of one-shot categorization using the shape skeleton? One-shot categorization is the process of learning novel categories following exposure to just one exemplar (***Lake and Piantadosi, 2019Lake et al., 2011***; ***Morgenstern et al., 2019***; ***Shepard, 1987***). We cannot answer this question definitively, given that visual attention paradigms make it difficult to distinguish between category learning per se and judgments of visual similarity. However, it is intriguing to consider whether the shape skeleton may support rapid object learning at an age when linguistic and object experience is minimal.

How well do our results align with the one-shot categorization literature from older children and adults? On the one hand, our results are consistent with studies showing that one-shot categorization of objects by older children and adults involves identifying invariant visual properties of the objects, namely shape (***Biederman and Bar, 1999***; ***Feldman, 1997***; ***Landau et al., 1988***). Moreover, research using simple visual objects (e.g. handwritten characters) suggests that a skeleton-like compositional structure is central to one-shot categorization (***Lake et al., 2011***; ***Lake et al., 2015***). On the other hand, our results would suggest that such categorization is possible at a much earlier age than has previously been suggested. Indeed, the extant research with both children (***Landau et al., 1998***; ***Smith et al., 2002***; ***Xu and Kushnir, 2013***) and adults (***Coutanche and Thompson-Schill, 2014***; ***Rule and Riesenhuber, 2020***) has been taken as evidence that such categorization abilities are accomplished by leveraging extensive category and linguistic knowledge. Thus, our results are consistent with the hypothesis that one-shot category learning may be possible early in development by relying on a perceptually invariant skeletal structure.

However, we would not suggest that object experience is irrelevant in infancy. As mentioned previously, experience with moving objects may play a role in supporting the extraction of a 3D shape structure. Moreover, although infants' visual experience is dominated by a small number of objects, this experience includes a large volume of viewpoints for each object (***Clerkin et al., 2017***). Interestingly, infants' chosen views of objects are biased toward planar orientations (***James et al., 2014***; ***Slone et al., 2019***), which may serve to highlight the shape skeleton, though it is unclear whether such a bias is a consequence or a cause of skeletal extraction. Moreover, certain visual experiences, such as low visual acuity at birth, may be particularly important for highlighting global shape information (***Cassia et al., 2002***; ***Ostrovsky et al., 2009***). Indeed, ANNs trained with a blurry-to-clear visual regimen showed improved performance for categories where global information is important, such as faces (***Jang and Tong, 2021***; ***Vogelsang et al., 2018***).

Altogether, our work suggests that infants form robust global representations of shape by extracting the object's shape skeleton. With limited language and object experience, infants generalized across variations in local shape properties to classify objects—a feat not matched by conventional ANNs. Nevertheless, by comparing infants' performance to existing computational models of vision, the present study provides unique insight into humans' representations of shape and their capacity for categorization, and may serve to inform future computational models of human vision.

Materials and methods

Participants

A total of 92 full-term infants participated in this study. Ten infants were excluded (Experiment 1: 5 for fussiness and 1 because of equipment failure; Experiment 2: 3 for fussiness and 1 because of equipment failure). The final sample included 34 infants in Experiment 1 (M=9.53 months, range = 6.47–12.20 months; 18 females) and 48 infants in Experiment 2 (M=9.12 months, range = 6.17–12.00 months; 20 females). Each infant was tested only once. All families gave informed consent according to a protocol approved by the Emory University Institutional Review Board (IRB) under the project 'Spatial Origins' (Study Number IRB0003452).

The sample size for Experiment 1 was determined using a priori power analysis with a hypothesized medium effect size (d=0.50; 1 - β>.8). For Experiment 2, we hypothesized that objects with the same coarse spatial relations would be more difficult to discriminate because the shape skeletons were more similar to one another (compared to Experiment 1), leading to an attenuated effect. Accordingly, to retain adequate power, we tested 14 more infants than in Experiment 1, the exact number of which was determined according to a fully counterbalanced design. Importantly, we nevertheless find the same results in Experiment 2 when the sample size is subsampled (n=34) to match that of Experiment 1.

Stimuli

For Experiment 1, six videos of 3D novel objects were rendered from the stimulus set created by ***Ayzenberg and Lourenco, 2019b***; ***Figure 1***. The objects were comprised of three distinct skeletons selected (from a set of 30) on the basis of their skeletal similarity. The skeletal similarity was calculated in 3D, object-centered, space as the mean Euclidean distance between each point on one skeleton and the closest point on the second skeleton following maximal alignment. A k-means cluster analysis (k=3) was used to select three distinct skeletons, one from each cluster (***Figure 1***). We ensured that the three skeletons were matched for discriminability by analyzing participants' discrimination judgments using data from ***Ayzenberg and Lourenco, 2019b***. Adult participants (n=42) were shown images of two objects (side-by-side) with either the same or different skeletons (same surface form). Participants were instructed to decide whether the two images showed the same or different objects. A repeated-measures ANOVA, with skeleton pair as the within-subject factor, revealed that the three skeletons used in Experiment 1 did not significantly differ in their discriminability, $F(2, 64)=0.11$, p=0.898.

For Experiment 2, we selected one object from Experiment 1 whose skeleton could be altered without changing the coarse spatial relations. We altered the object's skeleton by moving one segment 50% down the length of the central segment (***Figure 1***).

Each object was also rendered with two different surface forms, which changed the component parts and image-level properties of the object without altering its skeleton (***Figure 1***). The selection of these surface forms was based on adult participants' data from the study of ***Ayzenberg and Lourenco, 2019b***. In a match-to-sample task, participants (n=39) were shown one object (sample) placed centrally above two choice objects. One of the choice objects matched the sample's skeleton, but not the surface form, and the other choice object matched the sample's surface form, but not the skeleton. Participants were instructed to decide which of the two choice objects was most likely to be from the same category as the sample object. Participants performed worst at categorizing objects by their skeleton when surface form 1 was paired with surface form 2, M=0.58, compared to the other surface forms (Ms = 0.61–0.78). Thus, by choosing the surface forms that presented adult participants with the greatest conflict, we provided infants with an especially strong test of generalization on the basis of the skeletal structure.

In a separate set of analyses, we tested whether the surface forms were comprised of qualitatively different component parts by having participants rate each surface form on the degree to which it exhibited a specific non-accidental property (NAP). During a training phase, adult participants (n=34) were taught four NAPs (drawn from ***Amir et al., 2012***). They then rated the degree to which each surface form exhibited a particular NAP. The four NAPs were: (1) *taper*, defined as whether the thickness of an object part was reduced towards the end; (2) *positive curvature*, defined as whether an object part bulged outwards; (3) *negative curvature*, defined as the degree to which an object part caved inwards; and (4) *convergence to vertex*, defined as whether an object part ended in a point. Prior to the statistical analyses, we ensured that all participants in the sample exhibited

reliable performance (αs>0.7). A repeated-measures ANOVA, with NAP as the within-subject factor and surface form as the between-subject factor, revealed a significant main effect of surface form, $F(1, 66)=64.00$, $p<0.001$, such that surface forms corresponded to different NAPs. Thus, because objects between habituation and test phases consisted of different NAPs, it could not be used as a diagnostic cue for categorizing the different test objects.

We also tested whether objects with different surface forms, but the same skeleton, had significantly different image-level properties, in order to ensure that both objects presented to infants during the test phase differed from the habituation object in these properties. Each object video was converted into a sequence of images (30 frames/s; 300 frames total), which were analyzed with the Gabor-jet model (***Margalit et al., 2016***). This model overlays a 12 × 12 grid of Gabor filters (5 scales × 8 orientations) on each image. The image is convolved with each filter, and the magnitude and phase of the filtered image is stored as a feature vector. Paired *t*-tests were used to compare the feature vectors from each frame of one video to the corresponding frames of the second video. To provide an estimate of the image-level difference across the entire video, the resulting p-values from each *t*-test were then averaged across frames. This analysis revealed that objects with different surface forms (but same skeleton) had significantly different image-level properties ($p=0.002$), whereas objects with different skeletons (but same surface forms) did not ($p=0.090$). In other words, the surface forms used in the present study were actually more variable than the shape skeletons with respect to image-level properties.

Finally, we ensured that surface forms were matched in discriminability to the selected skeletons. Adult participants ($n=41$) conducted a surface form discrimination task, wherein they were shown images of two objects (side-by-side) which consisted of either the same or different surface forms (same skeleton). Participants were instructed to decide whether the two images showed the same or different objects. Participants discriminated between surface forms 1 and 2 significantly better than would be predicted by chance, $M=0.86$, $t(40) = 8.95$, $p<.001$, and importantly, discrimination accuracy between surface forms did not differ from discrimination accuracy between skeletons, $t(80) = 0.02$, $p=.981$.

Procedure for testing infants

Each infant sat on their caregiver's lap approximately 60 cm from a 22-inch computer monitor (1920×1080 px). Caregivers were instructed to keep their eyes closed and to refrain from interacting with the infant during the study session. The experiment was controlled by a custom program written in Visual Basic (Microsoft) and gaze data were recorded with an EyeLink 1000 plus eye tracker recording at 500 Hz (SR-Research). Prior to the start of the experiment, the eye tracker was calibrated for each infant using a 5-point calibration routine. Looking times were coded as any fixation falling within the screen for at least 500 ms. Any trial failing to meet this criterion was not analyzed (2.03% of trials in Experiment 1; 2.04% of trials in Experiment 2).

The experiment consisted of a habituation phase, in which infants were presented with one object, followed by a test phase where classification was tested using objects with matching and different skeletons. Both test objects differed from the habituated object in their surface form (see ***Figure 2A and D***). Each trial began with an attention-getting stimulus, which remained onscreen until infants fixated on it for 2 s. On each trial, infants were then shown a video of a single object rotating back-and-forth across 60° (12° per second). Each video remained on screen for 60 s or until infants looked away for 2 s. Videos were used instead of static images to maintain infants' attention.

Each infant was habituated to an object with one of three possible skeletons in Experiment 1 and with one of two possible skeletons in Experiment 2 (see ***Figure 1***), with half of the infants habituated to each surface form in each experiment. Infants met the habituation criterion when the average looking time in the preceding four trials was less than 50% of the average looking time in the first four trials. Test trials were presented after infants had habituated or following 24 habituation trials, whichever came first. All infants were presented with a total of six test trials, alternating between objects with the same or different skeletons (3 test trials of each type). The type of first test trial (same or different skeleton) was counterbalanced across infants.

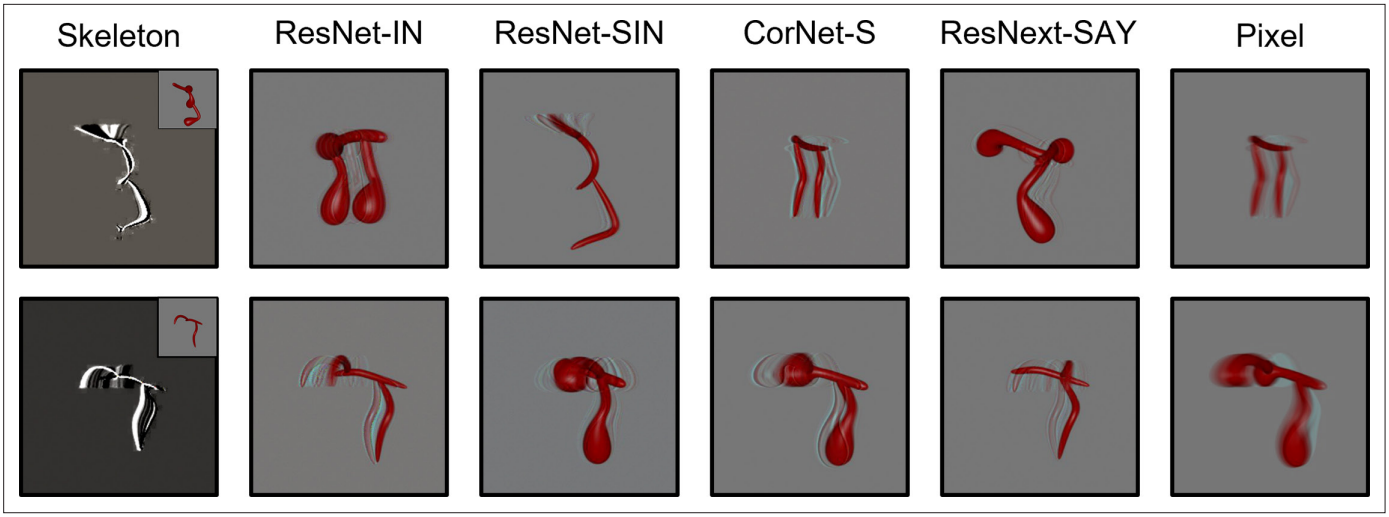

Figure 4. Examples of autoencoder reconstructions using objects from Experiment 1 (top) and Experiment 2 (bottom) for all models except FlowNet. FlowNet reconstructions are not possible because it requires multiple frames as input. For the Skeletal model, the inset displays the original input image. Each reconstruction was created by feeding a random frame from the habituation object video to each model immediately following its habituation to said video.

Model descriptions

For our Skeletal model, we used a flux-based medial axis algorithm (***Dimitrov et al., 2003***; ***Rezanejad and Siddiqi, 2013***) which computes a 'pruned' skeletal structure tolerant to local contour variations (***Feldman and Singh, 2006***). A pruned Skeletal model was selected for its biological plausibility in describing human shape judgments (***Ayzenberg et al., 2019a***; ***Feldman et al., 2013***; ***Wilder et al., 2011***; ***Wilder et al., 2019***).

The four ANNs tested in the present study were: ResNet-IN, ResNet-SIN, CorNet-S, and ResNext-SAY. ResNet-IN and ResNet-SIN are 50-layer residual networks (***He et al., 2016***) chosen for their strong performance on object recognition tasks. ResNet-IN was trained on ImageNet (***Russakovsky et al., 2015***), a dataset consisting of high-quality naturalistic photographs, whereas ResNet-SIN was trained on Stylized-ImageNet, a variation on the conventional ImageNet dataset that decorrelates color and texture information from object images using style transfer techniques (***Geirhos et al., 2018***; ***Huang and Belongie, 2017***). The third model, CorNet-S, is a 5-layer recurrent network explicitly designed to mimic the organization and functional profile of the primate ventral stream (***Kubilius et al., 2019***). It was chosen because it is a biologically plausible model of primate object recognition behavior and neural processing, as measured by the brain-score benchmark (***Schrimpf et al., 2018***). Finally, ResNext-SAY uses an updated version of the ResNet architecture and was designed to approximate the visual processing abilities of an infant (***Orhan et al., 2020***). It was trained using a self-supervised temporal classification method on the SAYCam dataset (***Sullivan et al., 2020***), a large, longitudinal dataset recorded from three infants' first-person perspectives.

To test whether infant-looking behaviors could be accounted for by low-level image properties, we also tested a model of pixel similarity and FlowNet, a model of motion flow (***Dosovitskiy et al., 2015***; ***Ilg et al., 2017***). For the Pixel model, the raw image frame was passed into the evaluation pipeline and classification was based on this information alone. FlowNet estimates the motion energy between adjacent video frames and is able to successfully use motion information to segment objects from the background as well as support action recognition in videos. Here we used an iteration of FlowNet known as FlowNet2-S, which was pre-trained on the MPI-Sintel and 'flying chairs' datasets (***Butler et al., 2012***; ***Mayer et al., 2016***).

Model analyses

Models were evaluated on the same stimulus sets presented to infants and tested using methods similar to the habituation/dishabituation procedure. One way to conceive of this procedure is as a measure of alignment between the stimulus and the infant's internal representation of the stimulus.

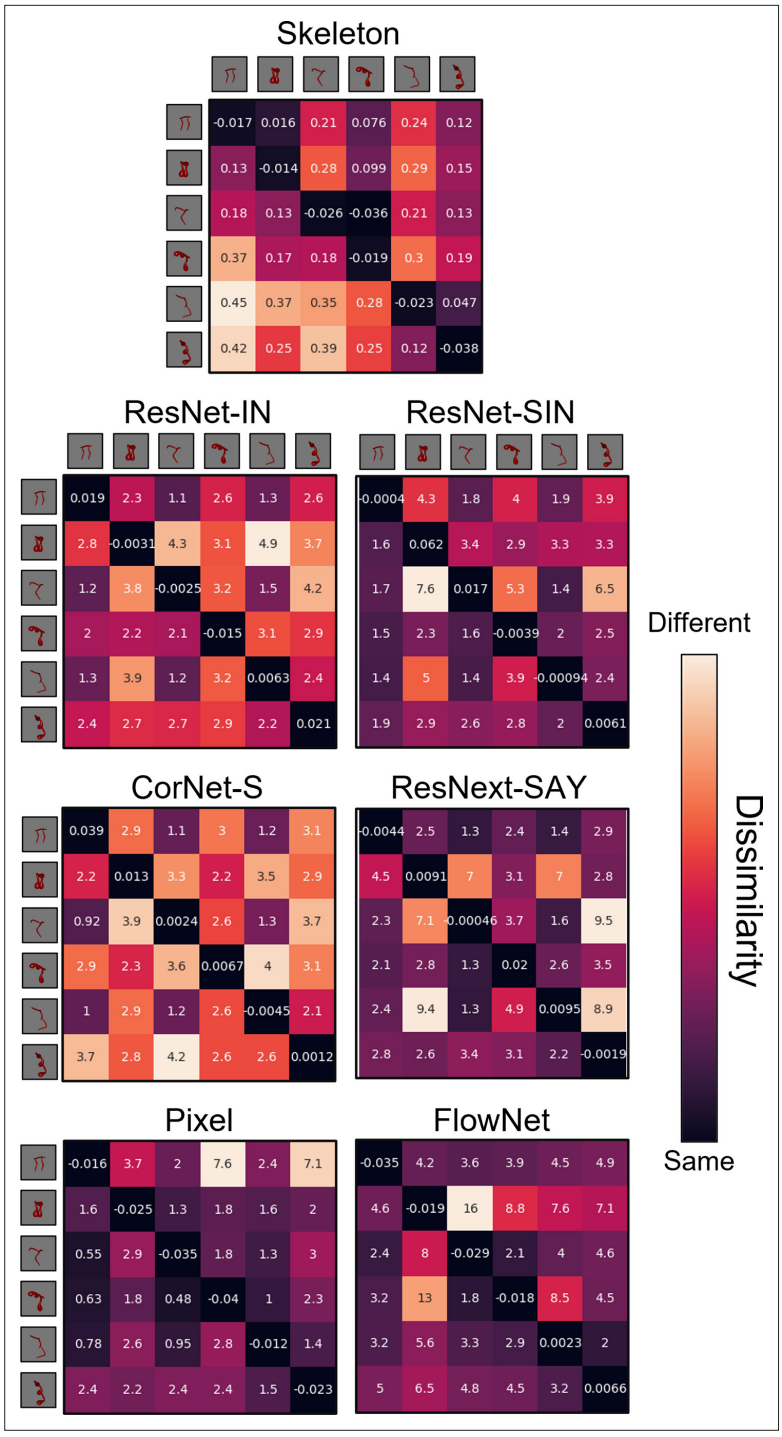

Figure 5. Dissimilarity matrices for each computational model in Experiment 1. Dissimilarity for each object pair was calculated as the error from an autoencoder following habituation to one object and testing on a second object. Internal values of each cell in the matrix indicate the error between habituation and test objects. Error-values are normalized to the end of habituation. Dissimilarity matrices are asymmetrical because the error value changes depending on which object the model was habituated to. The object adjacent to each row is the habituation object, and the object adjacent to each column is the test object.

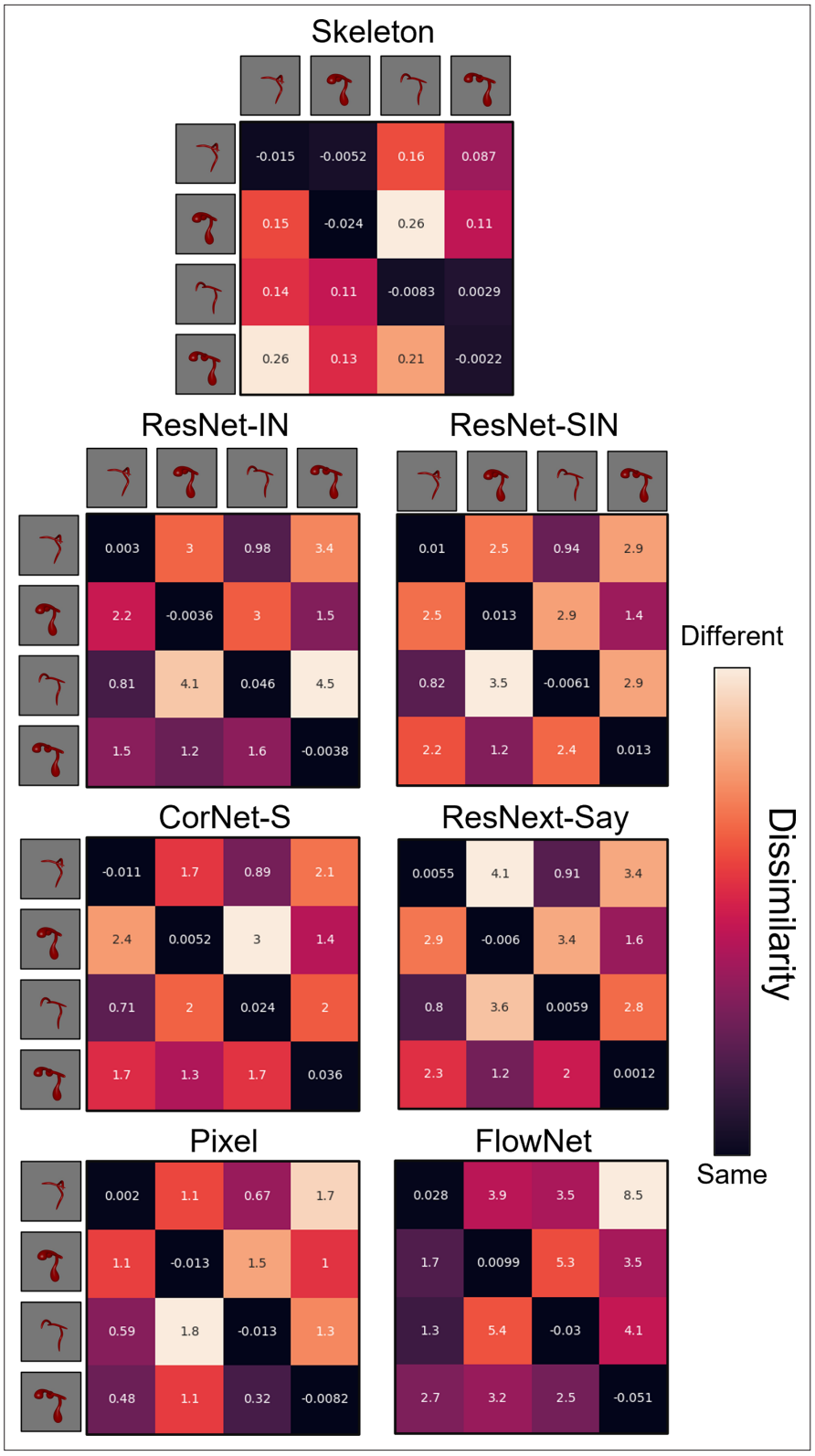

Figure 6. Dissimilarity matrices for each computational model in Experiment 2. Dissimilarity for each object pair was calculated as the error from an autoencoder following habituation to one object, and testing on a second object. Internal values of each cell in the matrix indicate the error between habituation and test objects. Error-values are normalized to the end of habituation. Dissimilarity matrices are asymmetrical because the error value changes depending on which object the model was habituated to. The object adjacent to each row is the habituation object, and the object adjacent to each column is the test object.

Infants will continue looking at a display for as long as they perceive a mismatch, or 'error', between the stimulus and their representation. To approximate this process, we converted each model into an autoencoder, which was 'habituated' and tested using the same criteria as infants (***Mareschal et al., 2000***; ***Westermann and Mareschal, 2004***). An autoencoder is an unsupervised learning model that attempts to recreate the input stimulus using a lower-dimensional set of features than the input. Like infants, the error signal from the output layer of an autoencoder remains high when there is a mismatch between the input stimulus and the internal representation. Each model was converted into an autoencoder by passing the outputs of the model to a single transposed convolutional decoding layer. For ANNs, outputs were extracted from the penultimate layer of the model, 'AvgPool' (with ReLu) for each frame of the video. For the Skeletal model, each frame of the video was first binarized and the skeleton extracted. The resulting skeletal image was blurred using a 3-pixel Gaussian kernel and passed into the decoder. For FlowNet, image representations of the flow fields were generated for each pair of adjacent frames before being passed into the decoder. No image preprocessing was conducted for the Pixel model. To match the output dimensions of each ANN, each image from the Skeletal, FlowNet, or Pixel model was passed through a single convolutional feature extraction layer with Max and Average pooling, before being sent to the decoding layer.

During the habituation phase, models were shown repeated presentations (epochs) of an object. Habituation was accomplished by training the models on each video using the Adam optimizer and a mean squared error loss function. For the ANNs, the weights of the decoding layer were updated during habituation. The weights of the pretrained models were frozen. For the Skeletal, FlowNet, and Pixel models, the weights of both the output and decoding layers were updated to support efficient feature extraction. The Skeletal, FlowNet, or Pixel model backbone was not altered. Models were said to have habituated once the average error signal in the last four epochs was below 50% of the average error in the first four epochs. The Skeletal model and ANNs met the habituation criteria within 8 epochs. Pixel and FlowNet models met the habituation criteria within 22 and 43 trials, respectively. The Skeletal, ANN, and Pixel models showed good reconstruction of the habituated objects (see ***Figure 4***). Reconstructions using FlowNet were not possible because multiple frames were used as input. At test, models were presented with objects that had the same/different skeletons or the same/different surface forms as the habituated object, and the error signal was recorded. See ***Figures 5 and 6*** for the error signal between all object pairs.

Classification score

For comparison, the performance levels of infants and computational models were converted to a classification score. Organisms' responses (i.e. looking time/error signal) in the test phase were first normalized to the end of the habituation phase (last 4 trials/epochs) by taking the difference between the two. The response to the novel object was then converted into a proportion by dividing it by the combined response to the novel and familiar test object. For both the models and infants, a classification score of 0.50 reflects chance performance.

Additional information

Funding

Funder	Grant reference number	Author
National Institutes of Health	T32 HD071845	Vladislav Ayzenberg

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Vladislav Ayzenberg, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing - original draft, Writing - review and editing; Stella Lourenco, Conceptualization, Methodology, Project administration, Resources, Supervision, Writing - original draft, Writing - review and editing

Author ORCIDs
Vladislav Ayzenberg http://orcid.org/0000-0003-2739-3935
Stella Lourenco http://orcid.org/0000-0003-3070-7122

Ethics
Human subjects: All families gave informed consent according to a protocol approved by the Emory University Institutional Review Board (IRB) under the project 'Spatial Origins' (Study Number IRB0003452).

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.74943.sa1
Author response https://doi.org/10.7554/eLife.74943.sa2

Additional files

Supplementary files
- Transparent reporting form

Data availability
All stimuli and data are available at: https://osf.io/4vswu/.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Ayzenberg V, Lourenco SF	2022	Perception of an object's global shape is best described by a model of skeletal structure in human infants	https://doi.org/10.17605/OSF.IO/4VSWU	Open Science Framework, 10.17605/OSF.IO/4VSWU

References

Amir O, Biederman I, Hayworth KJ. 2012. Sensitivity to nonaccidental properties across various shape dimensions. *Vision Research* **62**:35–43. DOI: https://doi.org/10.1016/j.visres.2012.03.020, PMID: 22491056

Arcaro MJ, Livingstone MS. 2017. A hierarchical, retinotopic proto-organization of the primate visual system at birth. *eLife* **6**:e26196. DOI: https://doi.org/10.7554/eLife.26196, PMID: 28671063

Ardila D, Mihalas S, von der Heydt R, Niebur E. 2012. 46th Annual Conference on Information Sciences and Systems (CISS. *Princeton* **1**:1–4. DOI: https://doi.org/10.1109/CISS.2012.6310946

Ayzenberg V, Chen Y, Yousif SR, Lourenco SF. 2019a. Skeletal representations of shape in human vision: Evidence for a pruned medial axis model. *Journal of Vision* **19**:1–21. DOI: https://doi.org/10.1167/19.6.6, PMID: 31173631

Ayzenberg V, Lourenco SF. 2019b. Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific Reports* **9**:1–13. DOI: https://doi.org/10.1038/s41598-019-45268-y, PMID: 31249321

Ayzenberg V, Kamps FS, Dilks DD, Lourenco SF. 2022. Skeletal representations of shape in the human visual cortex. *Neuropsychologia* **164**:108092. DOI: https://doi.org/10.1016/j.neuropsychologia.2021.108092, PMID: 34801519

Baker N, Lu H, Erlikhman G, Kellman PJ. 2018. Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology* **14**:e1006613. DOI: https://doi.org/10.1371/journal.pcbi.1006613, PMID: 30532273

Baker N, Lu H, Erlikhman G, Kellman PJ. 2020. Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research* **172**:46–61. DOI: https://doi.org/10.1016/j.visres.2020.04.003, PMID: 32413803

Bergelson E, Swingley D. 2012. At 6-9 months, human infants know the meanings of many common nouns. *PNAS* **109**:3253–3258. DOI: https://doi.org/10.1073/pnas.1113380109, PMID: 22331874

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological Review* **94**:115–147. DOI: https://doi.org/10.1037/0033-295X.94.2.115, PMID: 3575582

Biederman I, Gerhardstein PC. 1993. Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology. Human Perception and Performance* **19**:1162–1182. DOI: https://doi.org/10.1037//0096-1523.19.6.1162, PMID: 8294886

Biederman, I. (1995). Visual object recognition (Vol. 2): MIT press Cambridge, MA, USA.

Biederman I, Bar M. 1999. One-shot viewpoint invariance in matching novel objects. *Vision Research* **39**:2885–2899. DOI: https://doi.org/10.1016/s0042-6989(98)00309-5, PMID: 10492817
Blum H. 1967. A transformation for extracting descriptors of shape. Wathen-Dunn W (Ed). *Models for the Perception of Speech and Visual Form*. Cambridge, MA: MIT Press. p. 362–380.
Butler DJ, Wulff J, Stanley GB, Black MJ. 2012. A Naturalistic Open Source Movie for Optical Flow Evaluation. *Berlin, Heidelberg* **7577**:611–625. DOI: https://doi.org/10.1007/978-3-642-33783-3_44
Cassia VM, Simion F, Milani I, Umiltà C. 2002. Dominance of global visual properties at birth. *Journal of Experimental Psychology. General* **131**:398–411. DOI: https://doi.org/10.1037/0096-3445.131.3.398, PMID: 12214754
Clerkin EM, Hart E, Rehg JM, Yu C, Smith LB. 2017. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **372**:20160055. DOI: https://doi.org/10.1098/rstb.2016.0055, PMID: 27872373
Coutanche MN, Thompson-Schill SL. 2014. Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology. General* **143**:2296–2303. DOI: https://doi.org/10.1037/xge0000020, PMID: 25222265
Craft E, Schütze H, Niebur E, von der Heydt R. 2007. A neural model of figure-ground organization. *Journal of Neurophysiology* **97**:4310–4326. DOI: https://doi.org/10.1152/jn.00203.2007, PMID: 17442769
Davitt LI, Cristino F, Wong ACN, Leek EC. 2014. Shape information mediating basic- and subordinate-level object recognition revealed by analyses of eye movements. *Journal of Experimental Psychology. Human Perception and Performance* **40**:451–456. DOI: https://doi.org/10.1037/a0034983, PMID: 24364701
Dimitrov P, Damon JN, Siddiqi K. 2003. Flux invariants for shape. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. .
Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, Smagt PVD, Cremers D, Brox T. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. IEEE International Conference on Computer Vision. Santiago. DOI: https://doi.org/10.1109/ICCV.2015.316, PMID: 26540673
Ellis CT, Yates TS, Skalaban LJ, Bejjanki VR, Arcaro MJ, Turk-Browne NB. 2020. Retinotopic Organization of Visual Cortex in Human Infants. *Neuroscience* **1**:437. DOI: https://doi.org/10.1101/2020.12.01.407437
Feldman J. 1997. The Structure of Perceptual Categories. *Journal of Mathematical Psychology* **41**:145–170. DOI: https://doi.org/10.1006/jmps.1997.1154, PMID: 9237918
Feldman J, Singh M. 2006. Bayesian estimation of the shape skeleton. *PNAS* **103**:18014–18019. DOI: https://doi.org/10.1073/pnas.0608811103, PMID: 17101989
Feldman J, Singh M, Briscoe E, Froyen V, Kim S, Wilder J. 2013. An Integrated Bayesian Approach to Shape Representation and Perceptual Organization. Dickinson SJ, Pizlo Z (Eds). *Shape Perception in Human and Computer Vision: An Interdisciplinary Perspective*. London: Springer. p. 55–70.
Ferry AL, Hespos SJ, Waxman SR. 2010. Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Development* **81**:472–479. DOI: https://doi.org/10.1111/j.1467-8624.2009.01408.x, PMID: 20438453
Fleming RW, Schmidt F. 2019. Getting "fumpered": Classifying objects by what has been done to them. *Journal of Vision* **15**, 15. DOI: https://doi.org/10.1167/19.4.15
Gant JM, Banburski A, Deza A. 2021. Evaluating the Adversarial Robustness of a Foveated Texture Transform Module in a CNN. NeurIPS 2021 Workshop SVRHM. 1–12.
Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. 2018. ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. [arXiv]. DOI: https://doi.org/10.48550/arXiv.1811.12231
He K, Zhang X, Ren S, Sun J. 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. DOI: https://doi.org/10.1109/CVPR.2016.90
Huang X, Belongie S. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. 2017 IEEE International Conference on Computer Vision (ICCV). Venice. DOI: https://doi.org/10.1109/ICCV.2017.167
Hummel JE, Stankiewicz BJ. 1996. Categorical relations in shape perception. *Spatial Vision* **10**:201–236. DOI: https://doi.org/10.1163/156856896x00141, PMID: 9061832
Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T. 2017. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI. DOI: https://doi.org/10.1109/CVPR.2017.179
Jagadeesh AV, Gardner JL. 2022. Texture-like Representation of Objects in Human Visual Cortex. *Neuroscience* **1**:849. DOI: https://doi.org/10.1101/2022.01.04.474849
James KH, Jones SS, Smith LB, Swain SN. 2014. Young Children's Self-Generated Object Views and Object Recognition. *Journal of Cognition and Development* **15**:393–401. DOI: https://doi.org/10.1080/15248372.2012.749481, PMID: 25368545
Jang H, Tong F. 2021. Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of Vision* **21**:e6. DOI: https://doi.org/10.1167/jov.21.12.6, PMID: 34767621
Kellman PJ. 1984. Perception of three-dimensional form by human infants. *Perception & Psychophysics* **36**:353–358. DOI: https://doi.org/10.3758/BF03202789
Kellman PJ, Short KR. 1987. Development of three-dimensional form perception. *Journal of Experimental Psychology. Human Perception and Performance* **13**:545–557. DOI: https://doi.org/10.1037//0096-1523.13.4.545, PMID: 2965746

Kellman PJ, Shipley TF. 1991. A theory of visual interpolation in object perception. *Cognitive Psychology* **23**:141–221. DOI: https://doi.org/10.1016/0010-0285(91)90009-d, PMID: 2055000

Kellman PJ, Arterberry ME. 2006. Infant Visual Perception. Kuhn D, Siegler RS, Damon W, Lerner RM (Eds). *Handbook of Child Psychology: Cognition, Perception, and Language*. John Wiley & Sons Inc. p. 109–160.

Kiat JE, Luck SJ, Beckner AG, Hayes TR, Pomaranski KI, Henderson JM, Oakes LM. 2022. Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis. *Developmental Science* **25**:e13155. DOI: https://doi.org/10.1111/desc.13155, PMID: 34240787

Krizhevsky A, Sutskever I, Hinton GE. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60**:84–90. DOI: https://doi.org/10.1145/3065386

Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, Majaj N, Schmidt K. 2019. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. Advances in Neural Information Processing Systems 32 (NeurIPS 2019. .

Lake B, Salakhutdinov R, Gross J, Tenenbaum J. 2011. One shot learning of simple visual concepts. Proceedings of the annual meeting of the cognitive science society. .

Lake BM, Salakhutdinov R, Tenenbaum JB. 2015. Human-level concept learning through probabilistic program induction. *Science (New York, N.Y.)* **350**:1332–1338. DOI: https://doi.org/10.1126/science.aab3050, PMID: 26659050

Lake BM, Piantadosi ST. 2019. People Infer Recursive Visual Concepts from Just a Few Examples. *Computational Brain & Behavior* **3**:54–65. DOI: https://doi.org/10.1007/s42113-019-00053-y

Landau B, Smith LB, Jones SS. 1988. The importance of shape in early lexical learning. *Cognitive Development* **3**:299–321. DOI: https://doi.org/10.1016/0885-2014(88)90014-7

Landau B, Smith L, Jones S. 1998. Object perception and object naming in early development. *Trends in Cognitive Sciences* **2**:19–24. DOI: https://doi.org/10.1016/s1364-6613(97)01111-x, PMID: 21244958

Lee D, Gujarathi P, Wood JN. 2021. Controlled-Rearing Studies of Newborn Chicks and Deep Neural Networks. [arXiv]. DOI: https://doi.org/10.48550/arXiv.2112.06106

Lescroart MD, Biederman I. 2013. Cortical representation of medial axis structure. *Cerebral Cortex (New York, N.Y* **23**:629–637. DOI: https://doi.org/10.1093/cercor/bhs046, PMID: 22387761

Long B, Yu CP, Konkle T. 2018. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *PNAS* **115**:E9015–E9024. DOI: https://doi.org/10.1073/pnas.1719616115, PMID: 30171168

Mareschal D, French RM, Quinn PC. 2000. A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology* **36**:635–645. DOI: https://doi.org/10.1037/0012-1649.36.5.635, PMID: 10976603

Margalit E, Biederman I, Herald SB, Yue X, von der Malsburg C. 2016. An applet for the Gabor similarity scaling of the differences between complex stimuli. *Attention, Perception & Psychophysics* **78**:2298–2306. DOI: https://doi.org/10.3758/s13414-016-1191-7, PMID: 27557818

Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T. 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR. Las Vegas, NV, USA. DOI: https://doi.org/10.1109/CVPR.2016.438

Mervis CB, Rosch E. 1981. Categorization of Natural Objects. *Annual Review of Psychology* **32**:89–115. DOI: https://doi.org/10.1146/annurev.ps.32.020181.000513

Morgenstern Y, Schmidt F, Fleming RW. 2019. One-shot categorization of novel object classes in humans. *Vision Research* **165**:98–108. DOI: https://doi.org/10.1016/j.visres.2019.09.005, PMID: 31707254

Oakes LM, Spalding TL. 1997. The role of exemplar distribution in infants' differentiation of categories. *Infant Behavior and Development* **20**:457–475. DOI: https://doi.org/10.1016/S0163-6383(97)90036-9

Ons B, Wagemans J. 2012. Generalization of Visual Shapes by Flexible and Simple Rules. *Seeing and Perceiving* **25**:237–261. DOI: https://doi.org/10.1163/187847511X571519, PMID: 21771394

Orhan EA, Gupta PV, Lake BM. 2020. Self-Supervised Learning through the Eyes of a Child. [arXiv]. DOI: https://doi.org/10.48550/arXiv.2007.16189

Ostrovsky Y, Meyers E, Ganesh S, Mathur U, Sinha P. 2009. Visual parsing after recovery from blindness. *Psychological Science* **20**:1484–1491. DOI: https://doi.org/10.1111/j.1467-9280.2009.02471.x, PMID: 19891751

Quinn PC, Eimas PD, Rosenkrantz SL. 1993. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception* **22**:463–475. DOI: https://doi.org/10.1068/p220463, PMID: 8378134

Quinn PC, Eimas PD, Tarr MJ. 2001a. Perceptual categorization of cat and dog silhouettes by 3- to 4-month-old infants. *Journal of Experimental Child Psychology* **79**:78–94. DOI: https://doi.org/10.1006/jecp.2000.2609, PMID: 11292312

Quinn PC, Slater AM, Brown E, Hayes RA. 2001b. Developmental change in form categorization in early infancy. *British Journal of Developmental Psychology* **19**:207–218. DOI: https://doi.org/10.1348/026151001166038

Quinn PC, Bhatt RS, Brush D, Grimes A, Sharpnack H. 2002. Development of form similarity as a Gestalt grouping principle in infancy. *Psychological Science* **13**:320–328. DOI: https://doi.org/10.1111/1467-9280.00459, PMID: 12137134

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. 2018. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience* **38**:7255–7269. DOI: https://doi.org/10.1523/JNEUROSCI.0388-18.2018, PMID: 30006365

Rakison DH, Butterworth GE. 1998. Infants' attention to object structure in early categorization. *Developmental Psychology* **34**:1310–1325. DOI: https://doi.org/10.1037//0012-1649.34.6.1310, PMID: 9823514
Rezanejad M, Siddiqi K. 2013. Flux graphs for 2D shape analysis. *Advances in Computer Vision and Pattern Recognition*. Springer. p. 41–54. DOI: https://doi.org/10.1007/978-1-4471-5195-1_3
Ritter S, Barrett DG, Santoro A, Botvinick MM. 2017. Cognitive psychology for deep neural networks: A shape bias case study. Proceedings of the 34 th International Conference on Machine Learning. .
Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. 1976. Basic objects in natural categories. *Cognitive Psychology* **8**:382–439. DOI: https://doi.org/10.1016/0010-0285(76)90013-X
Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* **16**:225–237. DOI: https://doi.org/10.3758/PBR.16.2.225, PMID: 19293088
Rouder JN, Ratcliff R. 2016. Comparing Exemplar- and Rule-Based Theories of Categorization. *Current Directions in Psychological Science* **15**:9–13. DOI: https://doi.org/10.1111/j.0963-7214.2006.00397.x
Rule JS, Riesenhuber M. 2020. Leveraging Prior Concept Learning Improves Generalization From Few Examples in Computational Models of Human Object Recognition. *Frontiers in Computational Neuroscience* **14**:586671. DOI: https://doi.org/10.3389/fncom.2020.586671, PMID: 33510629
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115**:211–252. DOI: https://doi.org/10.1007/s11263-015-0816-y
Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, Kar K, Bashivan P, Prescott-Roy J, Geiger F, Schmidt K, Yamins DLK, DiCarlo JJ. 2018. Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like? *Neuroscience* **1**:e7. DOI: https://doi.org/10.1101/407007
Shepard RN. 1987. Toward a universal law of generalization for psychological science. *Science (New York, N.Y.)* **237**:1317–1323. DOI: https://doi.org/10.1126/science.3629243, PMID: 3629243
Slater A, Morison V, Rose D. 1983. Perception of shape by the new-born baby. *British Journal of Developmental Psychology* **1**:135–142. DOI: https://doi.org/10.1111/j.2044-835X.1983.tb00551.x
Slater A, Morison V. 1985. Shape constancy and slant perception at birth. *Perception* **14**:337–344. DOI: https://doi.org/10.1068/p140337, PMID: 4088795
Slone LK, Smith LB, Yu C. 2019. Self-generated variability in object images predicts vocabulary growth. *Developmental Science* **22**:e12816. DOI: https://doi.org/10.1111/desc.12816, PMID: 30770597
Sloutsky VM. 2003. The role of similarity in the development of categorization. *Trends in Cognitive Sciences* **7**:246–251. DOI: https://doi.org/10.1016/s1364-6613(03)00109-8, PMID: 12804690
Smith LB, Jones SS, Landau B. 1996. Naming in young children: A dumb attentional mechanism? *Cognition* **60**:143–171. DOI: https://doi.org/10.1016/0010-0277(96)00709-3, PMID: 8811743
Smith LB, Jones SS, Landau B, Gershkoff-Stowe L, Samuelson L. 2002. Object name learning provides on-the-job training for attention. *Psychological Science* **13**:13–19. DOI: https://doi.org/10.1111/1467-9280.00403, PMID: 11892773
Spriet C, Abassi E, Hochmann JR, Papeo L. 2022. Visual object categorization in infancy. *PNAS* **119**:e2105866119. DOI: https://doi.org/10.1073/pnas.2105866119, PMID: 35169072
Spröte P, Schmidt F, Fleming RW. 2016. Visual perception of shape altered by inferred causal history. *Scientific Reports* **6**:1–11. DOI: https://doi.org/10.1038/srep36245, PMID: 27824094
Sullivan J, Mei M, Perfors A, Wojcik EH, Frank MC. 2020. SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded from the Infant's Perspective. *PsyArXiv*. DOI: https://doi.org/10.31234/osf.io/fy8zx
Tarr MJ, Bülthoff HH. 1998. Image-based object recognition in man, monkey and machine. *Cognition* **67**:1–20. DOI: https://doi.org/10.1016/s0010-0277(98)00026-2, PMID: 9735534
Tartaglini AR, Vong WK, Lake BM. 2022. A Developmentally-Inspired Examination of Shape versus Texture Bias in Machines. [arXiv]. DOI: https://doi.org/10.48550/arXiv.2202.08340
Turati C, Simion F, Zanon L. 2003. Newborns' Perceptual Categorization for Closed and Open Geometric Forms. *Infancy* **4**:309–325. DOI: https://doi.org/10.1207/S15327078IN0403_01
Vogelsang L, Gilad-Gutnick S, Ehrenberg E, Yonas A, Diamond S, Held R, Sinha P. 2018. Potential downside of high initial visual acuity. *PNAS* **115**:11333–11338. DOI: https://doi.org/10.1073/pnas.1800901115, PMID: 30322940
Westermann G, Mareschal D. 2004. From Parts to Wholes: Mechanisms of Development in Infant Visual Object Processing. *Infancy* **5**:131–151. DOI: https://doi.org/10.1207/s15327078in0502_2, PMID: 33401785
Wiesel TN, Hubel DH. 1974. Ordered arrangement of orientation columns in monkeys lacking visual experience. *The Journal of Comparative Neurology* **158**:307–318. DOI: https://doi.org/10.1002/cne.901580306, PMID: 4215829
Wilder J, Feldman J, Singh M. 2011. Superordinate shape classification using natural shape statistics. *Cognition* **119**:325–340. DOI: https://doi.org/10.1016/j.cognition.2011.01.009, PMID: 21440250
Wilder J, Rezanejad M, Dickinson S, Siddiqi K, Jepson A, Walther DB. 2019. Local contour symmetry facilitates scene categorization. *Cognition* **182**:307–317. DOI: https://doi.org/10.1016/j.cognition.2018.09.014, PMID: 30415132
Wood JN, Wood SMW. 2018. The Development of Invariant Object Recognition Requires Visual Experience With Temporally Smooth Objects. *Cognitive Science* **42**:1391–1406. DOI: https://doi.org/10.1111/cogs.12595, PMID: 29537108
Xie S, Hoehl S, Moeskops M, Kayhan E, Kliesch C, Turtleton B, Köster M, Cichy RM. 2021. Visual Category Representations in the Infant Brain. [bioRxiv]. DOI: https://doi.org/10.1101/2021.11.03.466293

Xu F, Cote M, Baker A. 2005. Labeling guides object individuation in 12-month-old infants. *Psychological Science* **16**:372–377. DOI: https://doi.org/10.1111/j.0956-7976.2005.01543.x, PMID: 15869696

Xu F, Kushnir T. 2013. Infants Are Rational Constructivist Learners. *Current Directions in Psychological Science* **22**:28–32. DOI: https://doi.org/10.1177/0963721412469396

Yermolayeva Y, Rakison DH. 2014. Connectionist modeling of developmental changes in infancy: approaches, challenges, and contributions. *Psychological Bulletin* **140**:224–255. DOI: https://doi.org/10.1037/a0032150, PMID: 23477448

Younger B. 1990. Infants' detection of correlations among feature categories. *Child Development* **61**:614–620 PMID: 2364738.,