Fast and robust visual object recognition in young children

Vladislav Ayzenberg[1,2], Sukran Bahar Sener[3], Kylee Novick[4], Stella F. Lourenco[4]

[1]Department of Psychology and Neuroscience, Temple University
[2]Department of Psychology, University of Pennsylvania
[3]Department of Psychology, University of Washington
[4]Department of Psychology, Emory University

Short Title:
Robust recognition in kids

Corresponding Authors:
Vladislav Ayzenberg: vayzenb@temple.edu

Abstract

By adulthood, humans rapidly identify objects from sparse visual displays and across significant disruptions to their appearance. What are the minimal conditions needed to achieve robust recognition abilities and when might these abilities develop? To answer these questions, we investigated the upper-limits of children's object recognition abilities. We found that children as young as 3 years successfully identified objects at speeds of 100 ms (both forward and backward masked) under sparse and disrupted viewing conditions. By contrast, a range computational models implemented with biologically informed properties or optimized for visual recognition did not reach child-level performance. Models only matched children if they received more object examples than children are capable of experiencing. These findings highlight the robustness of the human visual system in the absence of extensive experience and identify important developmental constraints for building biologically plausible machines.

Keywords: Development; Machine Learning; Object Classification; Shape Perception; Benchmark; Feedback

Teaser

The visual recognition abilities of preschool children rival those of state-of-the-art artificial intelligence models.

Introduction

Humans extract meaning rapidly from sparse, and often incomplete, visual information. By adulthood, participants identify objects presented as quickly as 100 ms (*1-3*) and they do so across large variations in the visible appearance of objects, such as different orientations (*4, 5*) or partial occlusion (*6, 7*). Adult participants also need very little information to infer an object's identity. For instance, they readily recognize objects when properties such as color, texture, or other internal features are removed—leaving only an object outline (*8-10*). Moreover, adults maintain high accuracy even when the contours of the object outline are distorted or deleted (*11-13*). Thus, with little visual information, the adult visual system is able to quickly and accurately identify objects. What mechanisms support such robust recognition abilities and when do these abilities develop?

Uncovering the mechanisms needed to quickly recognize objects in challenging contexts is difficult when studying only an adult sample. Indeed, when encountering objects in degraded conditions adults may rely on dedicated neural processes that can compensate for missing object information, or they may simply draw on pre-existing knowledge gained from years of visual experience with similar degraded scenarios. For instance, recurrent neural circuits in the ventral visual pathway—the primary pathway underlying visual recognition (*14*)—have been found to support recognition when objects are partially occluded or information is missing (*15-17*). In this context, the missing portions of an object are perceptually completed via feedback from higher-level areas or lateral input from other portions of the same visual area (*15-17*). Alternatively, given their breadth of experience, adults may have simply encountered the degraded object in similar conditions previously. Indeed, an extensive perceptual learning literature shows that prior exposure to similar objects or contexts improves recognition performance for objects presented in otherwise challenging conditions (*18-21*). Repeated visual exposure may even alter the response profile of neurons in the ventral pathway, leading to selectivity for commonly encountered objects or stimulus features (*22-25*). In this view, processes such as recurrence would not be needed because observers have seen similar objects previously and recognition can be accomplished using only a feedforward pass through the ventral pathway.

One approach to help identify the minimal conditions necessary to accomplish object recognition is to study developmental populations (*26-28*). Specifically, young children provide an ideal subject pool for testing mechanistic questions about object recognition because they have limited visual experience (*29, 30*) and immature visual processing pathways (*31-35*). This means that researchers can present stimuli in contexts that children are unlikely to have seen before and track children's performance in relation to the maturity of their visual system. Together, these factors may allow researchers to isolate the processes that are necessary and sufficient for object recognition even under challenging conditions. Nevertheless, little is known about when, and how, such fast and robust recognition abilities develop in childhood.

The use of developmental populations, however, can be challenging for several reasons. One challenge is that children are rarely tested in the same conditions as adults (*26, 36*), making direct comparisons between populations difficult. Specifically, object recognition in human adults, as well as adult non-human primates, is typically tested by presenting objects in a speeded task (100 to 300 ms presentation) and visually masking the stimuli (*8, 14, 37-40*). By requiring participants to respond quickly and interfering with subsequent visual processing using a mask, this design reduces participants' ability to draw on high-level inferential processes and, therefore, better isolates the visual processes related to the most recently seen stimulus (*14*). Importantly, the

development of this standardized approach has been crucial in enabling comparisons between humans and other model organisms (both biological and synthetic), thereby providing an important method by which to explore the mechanisms underlying visual perception (*21, 41*). However, because children's capacity to follow instructions or stay attentive for long experimental sessions is limited, they are rarely tested in the same challenging conditions as adults. Indeed, the lack of a standardized approach has resulted in unresolved debates regarding the robustness of children's visual recognition abilities (*42*). For instance, research with infants suggests that the processes underlying visual recognition abilities, such as global form perception (*43, 44*) and perceptual completion (*45, 46*), arise early in development (for review, see *47*). In contrast, studies with older children suggest that these same processes may not develop until adolescence (*48, 49*). However, because these studies have largely relied on either indirect measures (e.g., looking time) or especially artificial experimental procedures (e.g., aligning Gabor wavelets), they are difficult to compare across populations and may not provide a direct or accurate measure of children's abilities.

In the current study, we sought to determine the upper-bound of young children's abilities using a challenging object recognition task similar to the standardized approach typically used with adults (see Figure 1). We did this by requiring 3- to 5-year-olds to identify rapidly presented (100 - 300 ms) 2D outlines of common objects (both forward and backward masked) under different stimulus manipulations. Specifically, we included objects that had complete, undisrupted, contours, as well as objects with perturbed or deleted contours. These stimulus conditions were carefully selected to test the presence of different mechanisms in children. Complete contour objects are those that children of this age have extensive familiarity with via picture books and likely can identify via a feedforward pass through the ventral pathway (*50, 51*). This condition has the added benefit of providing a baseline for children's performance in a challenging task. By contrast, perturbed contour objects disrupt the appearance of familiar local visual features and require global form perception, and are challenging for even sophisticated computer vision models (*52-54*). Similarly, deleted contour objects are likely to be unfamiliar to children and they may require perceptual completion as well as recurrent processing within the ventral pathway (*50, 51*). As mentioned previously, these two processes—global form perception and perceptual completion—have historically been thought to develop late in childhood (*48, 49, 55-57*), but this claim has primarily been tested via indirect methods.

To further explore what kinds of processes might be necessary to accomplish robust object recognition in childhood, we also compared children to a range of deep neural network models (DNNs; see Table 1). DNNs are computational models that can be trained to accomplish various visual perception tasks, and provide researchers with a well-controlled test bed by which to evaluate theories of visual processing (*58, 59*) and their development (*60-63*). For instance, DNNs with recurrent architectures show stronger performance on challenging visual tasks compared to feedforward ones (*64, 65*) and they have internal representations that are closely aligned with the adult primate ventral pathway (*66, 67*). Other work shows that DNNs trained with naturalistic visual experience and learning objectives are sufficient to recapitulate many human-like visual biases (*68-71*) and they show human-like performance on a range of object recognition tasks (*72-75*). Like human adults, increasing the variability of DNNs' visual experience to include degraded stimuli also improves their recognition performance for objects presented in challenging conditions (*68, 76, 77*).

In the current study, we compared children to a series of biologically inspired DNNs which allowed us to explore what mechanisms support fast and robust visual recognition early in development. Specifically, we included feedforward and recurrent models with ventral-like architectures (VoneNet_FF, VoneNet_R) because they allowed us to test predictions related to object speed, neural development, and viewing condition. For instance, if rapidly presented (and masked) objects do not afford time for recurrent processing (*50*), then we may find that performance on fast stimulus durations is best matched by feedforward models, but performance on slower stimulus durations is best matched by recurrent models. Moreover, if young children have immature recurrent circuits (*42, 78*), then their recognition performance may be best matched by a DNN with a feedforward architecture across multiple durations, particularly in the deleted contour condition which was previously shown to require recurrence (*17, 64*). We also included models with different training diets to explore what kinds of visual experiences may support robust recognition. Specifically, we included DNNs trained with realistic, but 'curated', images (EcoSet, ImageNet)(*73*), images manipulated to increase their variability (Stylized EcoSet, ImageNet-Sketch)(*79*), and naturalistic videos recorded from children (6 to 32 months of age) wearing head-mounted cameras (SAYCam)(*71*). If performance on perturbed and deleted stimulus conditions depends on visual experience with similarly degraded stimuli, then models explicitly trained with objects whose contours vary (e.g., Stylized EcoSet), or models that received such experience during natural viewing (e.g., SAYCam), may perform well in these conditions. We also included models trained with an unsupervised (e.g., ResNet50_SAY, DINOv2) learning objective or one that emphasizes vision-language alignment (e.g., CVCL, CLIP). These models were included because prominent developmental hypotheses suggest that object learning occurs in an unsupervised manner in infancy (*80, 81*) and is further bootstrapped by language learning (*74, 82, 83*). Finally, to explore the performance gaps between children and current DNNs, we included models whose performance has been optimized for visual recognition via supervised (e.g., ConvNext, ViT) or unsupervised (e.g., DiNOv2) approaches. Collectively, these model comparisons were included to support the initial exploration of a broad range of developmental hypotheses. However, it is important to note that our hypotheses are neither exhaustive nor mutually exclusive with one another. Yet, by comparing children to DNNs, researchers can better understand the gaps between humans and computational models (*30, 36, 84, 85*), as well as help identify models that can be used to inform developmental theory (*86-89*).

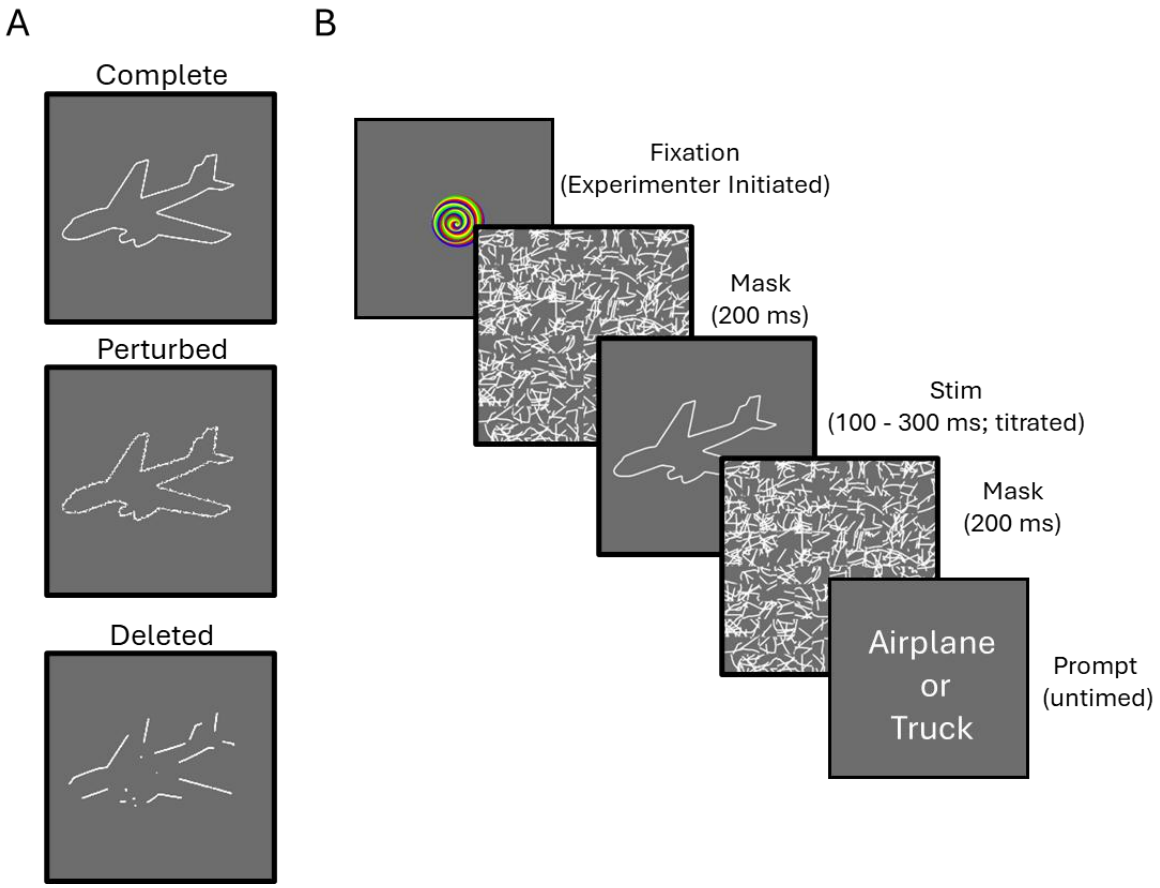


Figure 1. Stimuli and human testing procedure. (A) Children and adults were tested with object outlines that had either complete, perturbed, or deleted contours. (B) On each trial, participants were presented with an object image rapidly (100 – 300 ms duration), which was both forward and backward masked. In the prompt phase, children were asked to verbally indicate which object they saw amongst two possibilities (read by an experimenter). Adults responded by pressing an arrow key that corresponded to each object label.

Table 1. Model description. All models tested in the current study. Colored rectangles correspond to the models in Figure 4.

Name	Architecture	Learning Objective	Training Data	Dataset Description
VoneNet_FF-EcoSet	V1 + CORNet_S	Image Classification	EcoSet	Ecologically valid images in 565 categories
VoneNet_FF-EcoSet-Style	V1 + CORNet_S	Image Classification	Stylized EcoSet	Stylized ecologically valid images in 565 categories
VoneNet_R-EcoSet	V1 + CORNet_S	Image Classification	EcoSet	Ecologically valid images in 565 categories
VoneNet_R-EcoSet-Style	V1 + CORNet_S	Image Classification	Stylized EcoSet	Stylized ecologically valid images in 565 categories
ResNext50-SAY	ResNext50_32x4d	Temporal Self-Supervision	SAYCam Videos	Videos from child head-mounted cameras
CVCL	ResNext50_32x4d	Vision-Language Learning	SAYCam Videos + Text	Videos and language from child head-mounted cameras
ConvNext-ImageNet1k	ConvNext_Large	Image Classification	ImageNet 1k	1.3m curated photographs in 1k classes
ViT-ImageNet1k	ViT_B_16	Image Classification	ImageNet 1k	1.3m curated photographs in 1k classes
ViT-DINOv2	ViT_B_14	Unsupervised global image prediction	LVD-142M	142m images from the internet and other datasets
ViT-CLIP	ViT_B_16	Vision-Language Learning	CLIP data	400m image-text pairs from the internet
ResNet50-ImageNet1k	ResNet50	Image Classification	ImageNet 1k	1.3m curated images in 1k classes
ResNet50-ImageNet-Sketch	ResNet50	Image Classification	ImageNet 1k	1.3m curated images in 1k classes
ResNet50-ImageNet21k	ResNet50	Image Classification	ImageNet 21k	14m curated images in 21,000 classes
ResNet50-DINO	ResNet50	Unsupervised global image prediction	ImageNet 1k	1.3m curated images in 1k classes
ResNet50-YFCC15m	ResNet50	Vision-Language Learning	YFCC15m	15m image-text pairs from the internet
ResNet50-CLIP	ResNet50	Vision-Language Learning	CLIP data	400m image-text pairs from the internet

Results

Child performance

What are the upper limits of children's visual recognition abilities? We tested this question by presenting 3-, 4-, and 5-year-old children (N = 128) with a challenging recognition task where they were required to identify rapidly presented objects (100 to 300 ms) that were both forward and backward masked (see Figure 1 and Supplemental Video). Each child's top stimulus speed was identified using a titration procedure, wherein stimulus durations decreased by 50 ms (capped at 100 ms) for every three consecutively correct trials (see Supplemental Figure 1 for survival curve analyses). Moreover, objects could be presented with either complete, perturbed, or deleted contours in a between-subjects design. See Methods for full design details.

Overall, we found that children's performance was above chance (0.5) in all conditions, even when stimuli were presented at the fastest speed of 100 ms ($ps < .001$, $ds > 0.65$; see Figure 2), suggesting that object recognition is fast and robust from a young age.

Next, we examined whether there were differences in performance as a function of stimulus condition (complete, perturbed, or deleted), presentation time (durations: 100-300 ms), and child's age. A repeated measures ANCOVA, with age as a covariate, revealed main effects of condition, $F(2, 122) = 16.62$, $p < .001$, $\eta_p^2 = .21$, duration, $F(4, 488) = 2.64$, $p = .033$, $\eta_p^2 = .02$, and age, $F(1, 122) = 28.25$, $p < .001$, $\eta_p^2 = .19$. There were no significant 2- or 3-way interactions between the factors ($ps > .314$). Overall, participants' performance was higher at slower durations than faster ones. Moreover, post-hoc comparisons (Holm-Bonferroni corrected) revealed that performance was worse for the deleted condition than either complete or perturbed conditions ($ps < .001$, $ds > 0.66$; see Figure 2). There was no difference between complete and perturbed conditions ($p = .502$).

When age is converted into a categorical variable (3, 4, and 5 years), we found that 3-year-olds performed worse than 4- and 5-year-olds ($ps <.001$; $ds > 0.70$), but there was no difference between 4- and 5-year-olds' performance ($p = .136$, $d = 0.20$; see Figure 3). An additional comparison to adults revealed that adults performed better than children of every age and in each condition ($ps < .001$; $ds > 0.78$; see Figure 3).

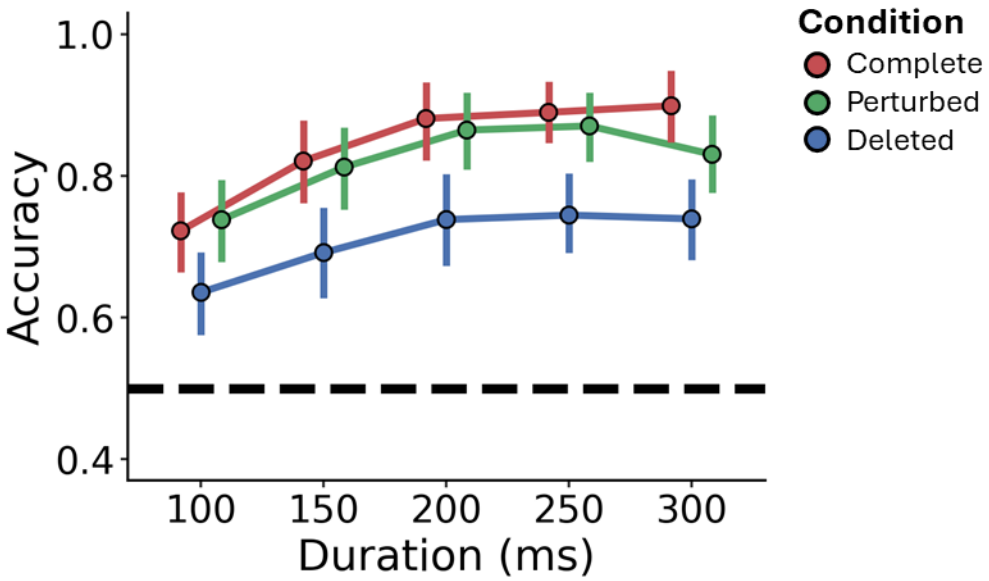


Figure 2. Children's performance in each condition. Across age, children performed above chance for each condition at each duration. Error bars depict 95 % confidence intervals. Dotted black line indicates chance performance (0.5).

To provide a more detailed understanding of the development of object recognition, we examined performance separately for each age group and condition (see Figure 3). We found that, like adults, 4- and 5-year-olds performed above chance in every condition and at all durations (*ps* < .038, *ds* > 0.63). By contrast, 3-year-olds performed above chance at all durations of the complete condition (*ps* < .007, *ds* > 1.00), but only at speeds of 200 ms or slower in the perturbed condition (*ps* < .001, *ds* > 1.27) and 250 ms or slower in the deleted condition (*ps* < .007, *ds* > 0.98). That children, especially in the youngest age group, performed worse in the deleted condition suggests that perceptual completion and recurrent processing in the ventral pathway may still be immature in preschool-aged children. Furthermore, 3-year-olds' poor performance at the fastest durations on the perturbed condition suggests that although young children have the capacity for global form perception, this capacity may not be fully mature at this age (*48*).

Altogether, these findings suggest that by 4 years of age, children rapidly extract meaning from sparse visual displays, including when there is missing information. That said, even 3-year-old children performed above chance in most conditions, despite stimuli being presented as fast as 100 ms with both forward and backward masking. Nevertheless, children's object recognition abilities are not fully mature, as even 5-year-old children performed worse than adults.

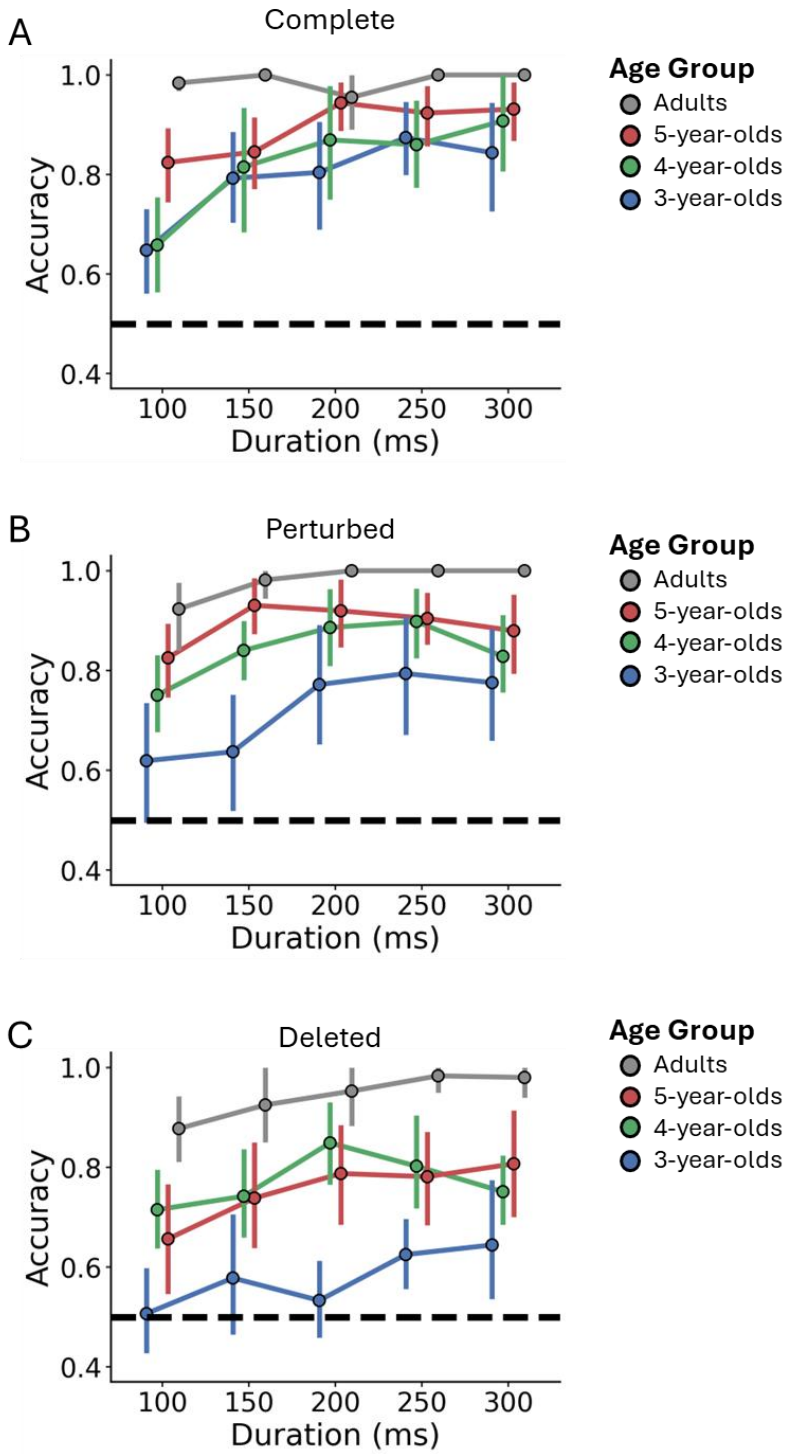


Figure 3. Performance in each condition by age group. (A) In the complete condition, participants of all ages performed above chance, even at the fastest speeds. (B) In the perturbed condition, 4- and 5-year-olds performed above chance at all speeds, whereas 3-year-olds were only above chance when durations were 200 ms and slower. (C) In the deleted condition, 4- and 5-year-olds performed above chance at all speeds, whereas 3-year-olds only performed above chance at the slowest speeds (250 ms and 300 ms). Error bars depict 95 % confidence intervals. Dotted black line indicates chance performance (0.5).

In a final analysis, we examined whether children's impressive performance in this task could be explained by low-level shape properties, such as curvature statistics (i.e., curvilinear vs. rectilinear) or shape envelope similarity (i.e., elongated vs. square-like). We computed the similarity of curvature statistics and shape envelope for each object pair, and median split the data into high and low similarity trials (see Methods). To maximize statistical power, child performance was grouped into fast (mean of 100 and 150 ms durations) and slow (mean of 200 and 250 ms durations) stimulus durations.

We did not find significant main effects of curvature or shape envelope similarity (*ps* > .139, η_p^2 = .04), nor any 2-way interactions between low-level shape similarity (curvature or shape envelope) with age group (*ps* > .085, η_p^2 < .09), stimulus condition (*ps* > .338, η_p^2 = .04), or stimulus duration (*ps* > .057, η_p^2 = .06). However, because these comparisons exhibit small-to-medium effect sizes, it is possible they would reach the threshold for statistical significance with greater power (see Figure 4). Thus, although low-level shape properties may have contributed to children's performance, they are unable to fully account for children's accuracy across stimulus durations or conditions.

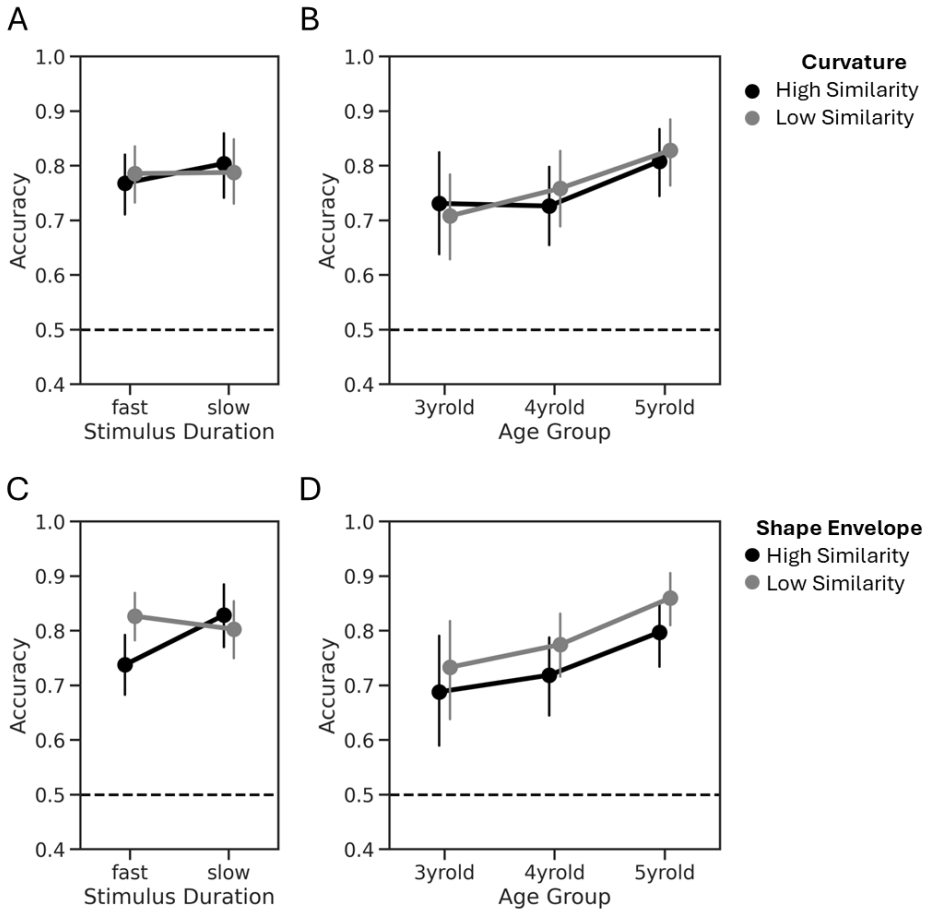


Figure 4. Influence of low-level shape features. Performance separated by (A-B) curvature and (C-D) shape envelope similarity across different (A, C) stimulus durations and (B, D) age groups. The black dotted line indicates chance performance (0.5) Error bars depict 95% confidence intervals.

Model comparisons

How does children's performance compare to biologically inspired and performance-optimized DNNs? To answer this question, we compared children (and adults) to DNN models with different types of architectures (e.g., feedforward vs. recurrent), visual experiences (e.g., curated vs. variable visual experience), and learning objectives (e.g., classification vs. vision-language alignment), as well as DNNs optimized for visual recognition tasks (see Table 1). Participants' performance in each age group (3, 4, 5 years, and adult) was split into fast (mean of 100 and 150 ms) and slow (mean of 200 and 250 ms) durations.

Overall, models performed above chance in every condition, regardless of their architecture, training experience, or learning objective (see Figure 5; Supplemental Tables 2-4). Furthermore, an analysis of model performance by layer revealed that performance was generally above chance by middle layers, and best in the top layers of the models (see Supplemental Figure 2).

Like humans, these models generally performed best on the complete condition, and worst on the deleted condition (see Supplemental Figure 3). However, whereas humans did not show a significant difference between complete and perturbed conditions, models generally performed worse on the perturbed condition (see Supplemental Figure 3). Amongst biologically inspired models, recurrent models (VoneNet_R) generally outperformed feedforward models (VoneNet_FF), which is consistent with the hypothesis that recurrence may be crucial for robust object recognition. Training with variable visual experience (i.e., Stylized EcoSet) showed a benefit, but only in the deleted condition. Models trained with videos from child head-mounted cameras (i.e., ResNext50_SAY) generally outperformed feedforward models, but not recurrent models. Interestingly, VoneNet_R performance was as strong, or stronger, than that of ConvNext-ImageNet1k and ViT-ImageNet1k, even though VoneNet models have fewer parameters (VoneNet_R: 55m params, ConvNext: 198m params; ViT: 86m params). Furthermore, ResNext50-SAY, a model trained with videos from children's perspective, matched the performance of ConvNext and ViT on the deleted condition, and even outperformed these models on the perturbed condition. These findings suggest that biologically inspired architectures and experiences may lead to improvements in the performance and efficiency of DNNs.

Next, we compared the performance of biologically inspired models to children on the basis of overlapping confidence intervals (see Supplemental Figure 3 for individual human variability). In the complete condition, recurrent models (VoneNet_R) performed as well as 3- and 4-year-old children at fast stimulus durations, but not slow stimulus durations, nor as well as 5-year-old children at any duration (see Figure 5A and Supplemental Table 2). In the perturbed condition, all biologically inspired models matched the performance of 3-year-olds at the fast durations, but not any other age group, nor 3-year-olds at the slow durations (see Figure 5B and Supplemental Table 3). In the deleted condition, all biologically inspired models, except CVCL, performed as well as children at the fast durations (see Figure 5C and Supplemental Table 4). With stylized training, VoneNet_R even surpassed 3-year-old children at the slow durations. However, these models performed worse than 4- and 5-year-olds at slow durations (Figure 5C). Nevertheless, that 3-year-old performance in the deleted condition was matched by a feedforward models (VoneNet_FF), regardless of stimulus duration, suggests that recurrent connections may still be immature at this age.

However, overall, we did not find that biologically inspired, nor performance optimized classification models (ConvNext-ImageNet1k, ViT-ImageNet1k), consistently matched children's performance across both conditions and stimulus durations. The only models to consistently match or surpass children was ViT-DINOv2, a performance-optimized vision models trained using an unsupervised learning objective, and ViT-CLIP, a vision-language model trained to associate images and text via a contrastive learning objective. Indeed, these models even matched the performance of adults in some cases (Figure 5A-C).

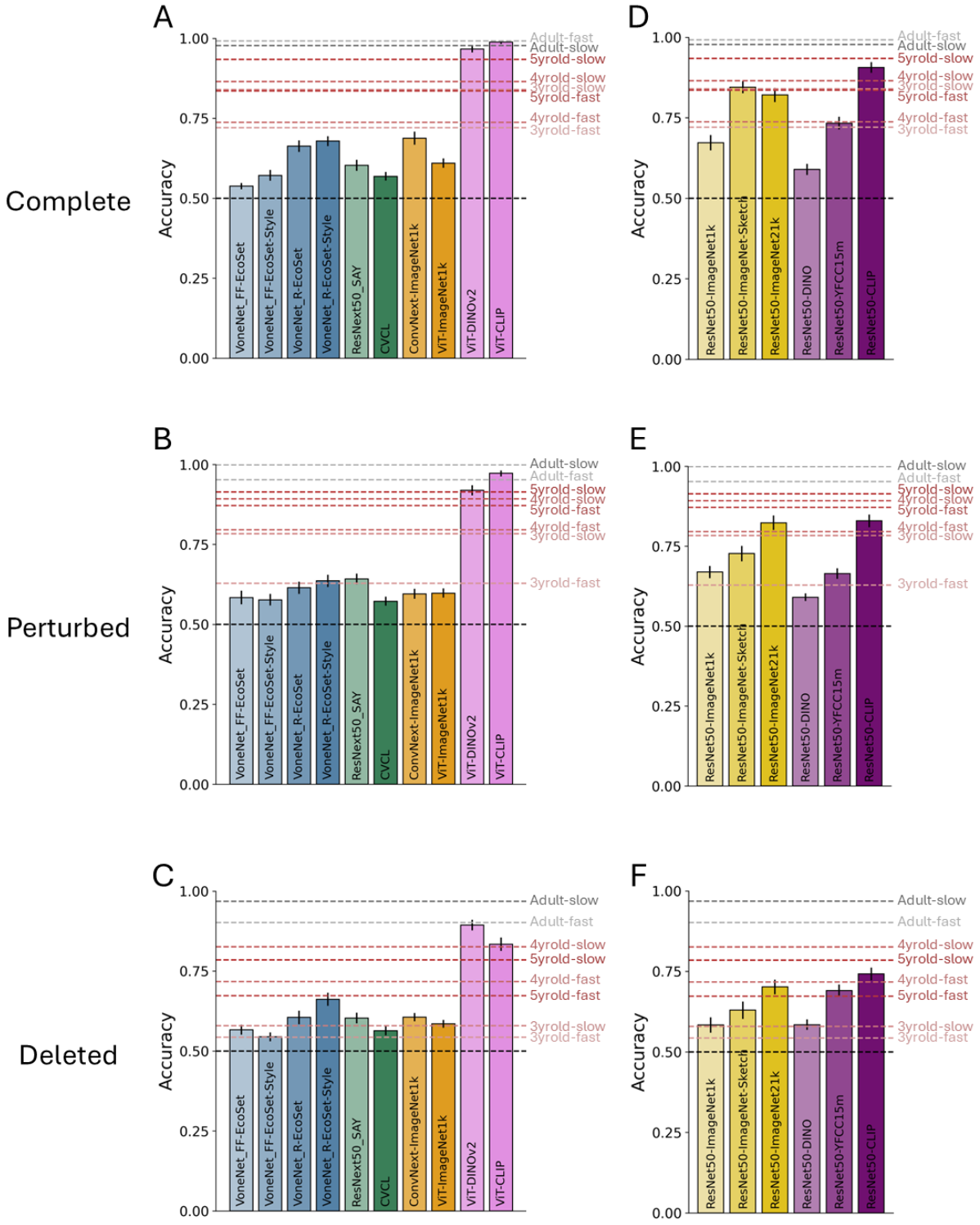


Figure 5. Model and human performance in each condition. Performance of humans and models in the (top) complete, (middle) perturbed, and (bottom) deleted contour conditions. Human data for each age (red dotted lines: children; gray dotted lines: adults) were aggregated into fast (100 ms & 150 ms) and slow (200 ms & 250 ms) stimulus durations. Humans were compared to (A-C) biologically inspired (blue: ventral-like architecture; green: trained on infant experience) and performance-optimized (orange: classification

objective; violet: unsupervised & vision-language objective) models, as well as (D-F) models selected to disambiguate between the contributions of training type, scale. and learning objective (yellow: classification objective; purple: vision-language objective). The Y-axis indicates classification accuracy. The black dotted line indicates chance performance (0.5) Error bars depict 95% confidence intervals for models. See Supplemental Figure 3 and Supplemental Tables 2-4 for variability estimates and confidence intervals for human data.

Learning Objective or Scale?

What properties of ViT-DINOv2 and ViT-CLIP account for its strong visual recognition performance? One possibility is that an unsupervised learning objective or one that pairs linguistic information with images, results in more robust visual representations. Indeed, unsupervised learning is often hypothesized to be critical for aligning the visual representations of primates and DNNs (*81, 90*). Moreover, classic developmental work has suggested that language bootstraps object recognition by guiding children's attention to diagnostic object properties (*91, 92*). However, the mediocre performance of ResNext50-SAY, an unsupervised model, and CVCL, a model trained to associate naturalistic video with linguistic information from children, suggests that neither unsupervised learning nor a vision-language objective on its own is sufficient for DNNs to match child performance.

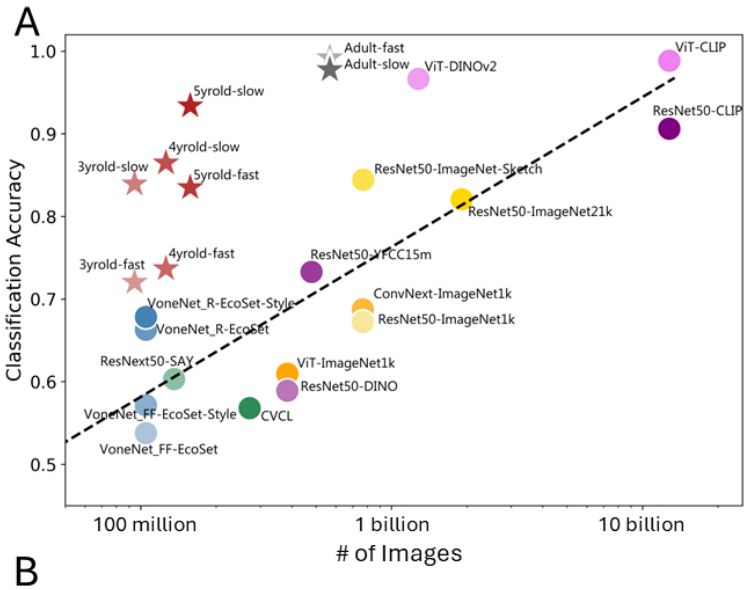
An alternative possibility is that the success of these models is attributed to their extensive object experience. Indeed, in contrast to the classification models, which were trained with ImageNet (1.3 m images) or EcoSet (1.5 m images), ViT-DINOv2 and ViT-CLIP were trained with datasets of 142 and 400 million images, respectively (see Table 1). Across multiple epochs of training, these models have been exposed to over a billion images each (ViT-DINOv2: 1.28 b; ViT-CLIP: 12.8 b). For comparison, if, from the moment of birth, a child were to see a new object every second of their life for six consecutive years, they would have seen only 189 million objects—less than half the number of object images in the CLIP dataset. Even if this estimate was increased 10-fold (i.e., 1 image every 100 ms without sleep), the oldest children in our sample would still have less than 1/6th of the total object experience of CLIP trained models. Although this estimate is not intended to be a realistic accounting of human visual experience, it serves to illustrate the scale of model training when projected onto the timeline of the human life span. Indeed, this back-of-the-envelope estimate ignores the fact that infants, conservatively, sleep an average of 14+ hours per day in the first year of life and 3- to 5-year-olds sleep at least 10 hours per day (*93, 94*).

To test whether the scale of object experience or the learning objective contributed to model performance, we tested six additional models. Specifically, we tested ResNet50 models that were pre-trained on image datasets of comparable size (ImageNet21k: 14 m images; YFFC15m: 15 m images), but with one using a standard object classification objective and dataset (ImageNet21k), and the other using a vision-language objective to align image and text embeddings (YFFC15m). We also tested a ResNet50 model trained on standard ImageNet1k using the unsupervised DINO procedure (ResNet50-DINO; 1.3 m images) (*95*). As baselines, we also included a ResNet50 model trained to classify images from the standard ImageNet1k dataset (1.3 m images) and one trained with the same procedure and dataset as CLIP (ResNet50-CLIP; 400 m images). Finally, to explore whether specific object experience with line drawings contributed to the performance of large-scale models, we also test a ResNet50 model finetuned on ImageNet sketch (ResNet50-ImageNet-sketch), a dataset comprised of line drawings (see Methods).

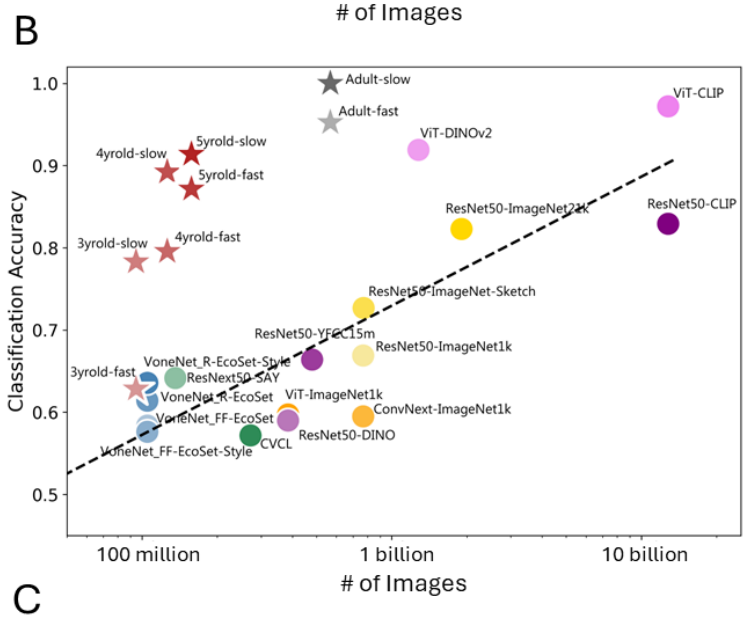
Consistent with the prior analysis, we found that ResNet50-CLIP showed the strongest performance compared to the other models, and it matched (or surpassed) children in almost every condition (see Figure 5D-F). The next best model across all conditions was ResNet50-ImageNet21k, which showed large improvement over a base ResNet50 model trained on ImageNet1k. Moreover, ResNet50-ImageNet21k outperformed both DINO and YFFC15m trained ResNet50 models (see Figure 5D-F), suggesting that training scale, and not learning objective, led to performance improvements. Finally, we found that a ResNet50 model trained on ImageNet-Sketch showed large improvements over a base ResNet50, particularly in the complete contour condition, where it even outperformed children in some cases. However, its performance was still lower than a model trained on ImageNet21k in perturbed and deleted conditions, as well as most children in the perturbed condition.

Finally, to visualize the importance of scale, we analyzed the relation between object experience and recognition performance across all models tested (Figure 6). This analysis revealed strong Spearman rank-ordered correlations between a model's classification accuracy and the total number of images it was exposed to (log-transformed) across training for every condition (Complete: *rho* = 0.83; Perturbed: *rho* = 0.78; Deleted: 0.70), which is consistent with prior work on scaling laws in DNNs (*96*). To compare this scaling pattern to humans, we plotted data from children and adults using the conservative estimate described previously (1 object every second of life without sleep). Human data were generally above and left of the model trendline, suggesting that humans develop robust visual recognition abilities with far less object experience than models. Thus, although certain kinds of experiences, such as experience with line drawings, can improve visual recognition abilities, our results suggest that scale remains the strongest driver of performance for DNNs but not humans.

Complete



Perturbed



Deleted

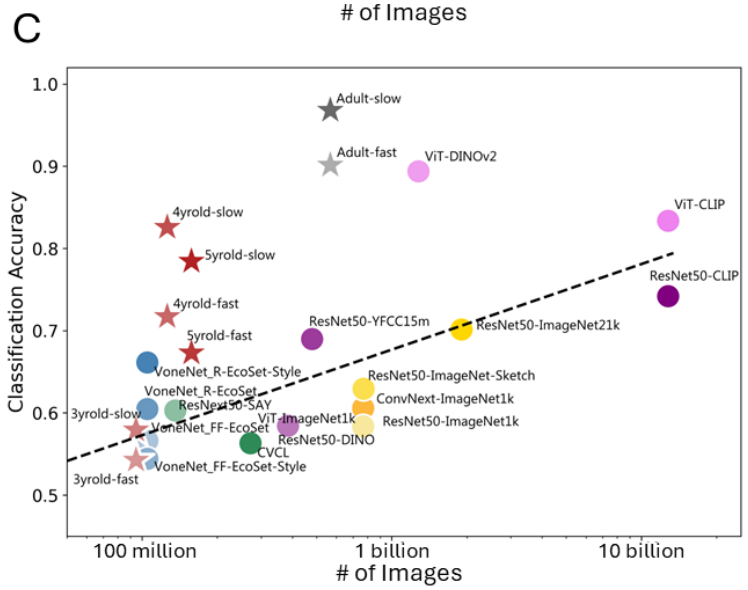


Figure 6. Recognition performance as a function of experience. Scatter plots showing the relation between classification accuracy on the y-axis for (A) complete, (B), perturbed, and (C) deleted contour conditions and total number of images models were trained with on the x-axis in log scale. Human estimates are plotted as stars and their experience is conservatively estimated as seeing one object every second of their life without sleep (see Main Text).

Discussion

In the current study, we sought to understand the development of robust object recognition in young children. Our results demonstrate that young children succeed at identifying objects from sparse visual displays, at speeds as fast as 100 ms and even when the contours are disrupted. Direct comparisons to a range of computational models revealed that children generally outperformed both biologically inspired and performance-optimized DNNs. Only by exponentially increasing a model's training data, beyond what children can realistically experience, did models match or surpass children's performance. Altogether, these findings suggest that object recognition is fast and robust from early in development, but gaps remain in our ability to approximate these processes with current computational models.

Visual recognition in young children

We found that by 4 years of age, children identified objects whose local visual features were disrupted via contour perturbations or deletions, even when presented as rapidly as 100 ms, and when stimuli were forward and backward masked. Remarkably, even 3-year-old children performed above chance, though they required somewhat slower presentation times when object contours were disrupted. These findings suggest that young children exhibit robust recognition in which they readily extract global form across variations in local features.

Our results stand in stark contrast to a collection of studies in which the findings suggested a protracted development of object recognition abilities (*42*). Specifically, these studies found that the ability to ignore local features, so as to represent global form, or to accomplish perceptual completion across disconnected contours remained difficult for children as old as 10 years of age (*48, 49, 55-57*). These findings have been mixed, however. Other studies with similar methods and age groups found the opposite, namely, that children prioritized global form over local features (*97*).

Our findings are consistent with a rich literature on younger children which suggests an earlier development of object recognition abilities. For instance, 2-year-olds preferentially use shape information to categorize objects across changes in color and texture information (*98, 99*), and with linguistic scaffolding, young children generalize object identities to sparse caricature versions of common objects (*82, 100, 101*). Furthermore, research with infants has shown that by 3 months of age, they identify novel objects from orientations not previously seen (*102-107*), and older infants are capable of categorizing novel objects across variations in contours (*43, 108-110*). Indeed, by the end of the first year of life, children can infer an object's category from only its shape silhouette (*111*), suggesting that they do not rely on color, texture, or internal local features to recognize objects. Finally, other work has shown that infants are capable of inferring the complete shape of an object even when it is partially occluded (*112-115*). However, studies with infants cannot measure speed of processing and they generally rely on indirect looking time measures to infer processes like recognition, which can be difficult to compare across different populations. Our work shows

that by at least preschool age, children are capable of identifying rapidly presented objects from sparse and disrupted visual displays.

We would not suggest, however, that children's visual recognition abilities are fully mature at preschool age. Indeed, we generally found that children's performance in all conditions and at all stimulus durations, improved with age, with all children performing worse than adults. The effect of age was most pronounced when contours were disrupted. Specifically, in the perturbed and deleted conditions, 3-year-olds performed at chance at the fastest speeds, whereas by 4 years of age, children performed above chance at all speeds. Importantly, this pattern cannot be explained by 3-year-olds' inability to conduct the task given that their performance did not differ from other children in the complete contour condition, and it was above chance at slower speeds in all conditions. Instead, given that the recognition of contour-deleted objects may require perceptual completion (*48, 49*) and recurrent processing (*17, 64*), the developmental change we observed may indicate underdeveloped recurrent connections in 3-year-old children. Moreover, 3-year-olds' performance in the fastest durations of the perturbed condition suggests that their global form perception abilities are still maturing. Yet the fact that children generally performed worse than adults on all conditions and durations suggests that there may be other developmental factors not directly related to visual recognition, such as attention or executive control, that impact children's performance. Thus, one interpretation of our results, which is consistent with the existing literature from both infants and older children, is that children have the capacity for robust object recognition from early in development, but this ability does not become fully adult-like until adolescence (*57*).

Using children as a benchmark

A major goal of the current study was to compare children to computational models to explore what mechanisms might be needed for the development of visual recognition abilities and to assess the gaps between humans and machines. Such comparisons are crucial if DNNs are to be used as tools to explore the cognitive and neural mechanisms underlying visual recognition in humans (*41, 59, 116*). Although we found that factors like a biologically inspired recurrent architecture (VoneNet-R) and naturalistic or variable experience (Stylized EcoSet, SayCam) allowed models to match, or even outperform much larger DNNs optimized for image classification (e.g., ConvNext), these models did not consistently match children's performance. Indeed, few of the hypotheses we described in the Introduction relating to performance on different stimulus durations or conditions bore out. The only exception was that the performance of 3-year-old children across all durations was well matched to a feedforward model, which provides computational evidence that recurrent circuits in humans may be immature at a young age.

The only models to consistently match, or surpass, children's performance were those trained on especially large datasets (e.g., ImageNet21k, CLIP). Could children's performance in the current study reflect prior experience with many objects? Our analyses, as well as existing research with children, suggest that this possibility is unlikely. First, as described in the results, even a conservative estimate of children's object exposure reveals that they are unable to receive as much object experience as available in the CLIP dataset until they are well into adolescence. Indeed, even if images were presented every 100 milliseconds without sleep—a timeframe shorter than the typical length of a fixation (150-300 ms)(*117*) or the latency of a saccade (~200 ms)(*118*)—it would be difficult for an 18-year-old adult to see as many images as a fully trained CLIP model. Moreover, models trained using the SAYCam dataset (ResNet50_SAY, CVCL), a dataset comprised of

head-mounted videos from children ages 6 months to 3 years (*71*), did not consistently match child performance across conditions even for the youngest ages.

However, as we discussed in the Results, these estimates of visual experience are not intended to be realistic. Indeed, human vision does not have a clear frame rate and humans often have multiple objects in view at once. Instead, these estimates are intended to contextualize model experience on the scale of the human lifespan. Whereas most model training regiments emphasize a large number of distinct object examples, the human lifespan does not afford such varied experience. Indeed, studies analyzing children's visual experience from head-mounted cameras show that children are exposed to relatively few distinct objects. For instance, in the first year of life, just 10 common objects occupies approximately 1/3 of their visual experience (*29*), with another large portion accounted for by three faces (*119, 120*), hands (*121*), and visually simple architectural features of a scene (e.g., bright ceiling light; *122*). Instead, children's experience with these few objects is extensive, such that they densely sample many views of the same few objects (*123, 124*). Furthermore, it is unlikely that children's performance in the current study can be explained by extensive experience with drawing or coloring books because children do not show the capacity to draw and color, beyond simple scribbles, until approximately 3 years of age (*125, 126*). Thus, at best, these activities occupy only a small portion of children's visual experience at the ages we tested. Moreover, our results showed that models trained using video data from children's perspective (ResNet50_SAY, CVCL) (*71*), did not result in a consistent match to child performance across conditions even for the youngest ages, further suggesting that specific kinds of prior experience did not contribute to children's performance. Like human children, SayCam trained models are presented with a continuous stream of visual information which includes multiple objects.

In contrast to children, our results suggest that DNNs are far more reliant on specific visual experiences in their training data to succeed. For instance, fine-tuning models on ImageNet-Sketch, a dataset comprised of 50,000 line drawings (*127*)—likely more than children could have experienced—led to large improvements in model performance. However, these improvements were primarily in the complete contour condition where stimuli were most similar to the training data. Even models trained on Stylized EcoSet, a training regimen that disrupts local visual features and object contours (*79*), were largely surpassed by children in both the complete and perturbed conditions. Only models trained on especially large, but opaque datasets (e.g., CLIP), matched or surpassed children, suggesting that their strong performance on our task was likely due to the presence of similar kinds of images in their training data. Indeed, CLIP trained models show catastrophic recognition failures if small aspects of their training data are lesioned (*128*). These findings add to a body of work showing that current DNNs, regardless of architecture, training, or learning objective show substantial deficits in generalizing to examples that are outside the specific distribution of their training experience (*129-131*). Altogether, our findings contribute to a growing literature that illustrates how children develop sophisticated perceptual (*30, 43, 84*), linguistic (*36, 132*), and other high-level cognitive abilities (*133-135*) from far less data than state-of-the-art machine learning models currently require (*136*).

Mechanisms of visual development

What, then, might account for the discrepancy between children and DNNs? One possibility is that we only tested recurrent models that implement local, within-layer, recurrence where information is processed by the same layer multiple times. By contrast, recurrent processing in the human

ventral pathway occurs both within and between cortical areas, a process also known as top-down feedback (*137, 138*). Indeed, top-down feedback may be particularly important for perceptual completion because it exploits the larger-receptive fields of higher-level visual areas (*17, 51, 137*). Moreover, it provides a mechanism by which prior knowledge from areas such as the pre-frontal cortex can be used to make predictions about the content of ambiguous displays (*139-141*). Thus, top-down feedback offers both a perceptual mechanism by which to fill-in missing information and a process by which to efficiently apply prior experience to the current scenario.

Incorporating continuous recurrent dynamics may also improve models' capacity to approximate the temporal dynamics of human vision. Indeed, although we used feedforward and recurrent models to examine the mechanisms needed to recognize objects at different stimulus durations, this approach is limited because the VoneNet_R uses a fixed number of recurrent loops and this type of recurrence cannot be used for the other models without altering their architecture and retraining them. Moreover, although researchers have identified methods to model temporal dynamics related to decision making (e.g., reaction time) (*142, 143*), there is currently no approach (to the best of our knowledge) for approximating stimulus duration in computational models without altering their architectures. This gap in temporal processing abilities is particularly glaring given that the spatio-temporal continuity of objects (i.e., motion) has long been identified as an important constraint for developing robust object representations (*45, 112, 144-146*).

Another possibility is that the current DNNs only approximate the ventral visual pathway, which, on its own, may not be sufficient for robust object recognition. Specifically, accumulating evidence suggests that the ventral pathway, like DNNs, is most sensitive to local visual features (*54, 147-149*), and it has difficulty representing a complete object form without input from the dorsal visual pathway and prefrontal cortex (*138, 150-155*). Dorsal visual areas develop early (*34, 156, 157*) and support the formation of global shape percepts by representing the spatial arrangement, rather than the appearance, of object parts (*153, 158*). And, as mentioned previously, pre-frontal areas may be crucial for applying prior knowledge to solve visual recognition tasks (*138*), though these areas may develop later (*159; but, see Ellis et al., 161*). Nevertheless, models that implement the processes of dorsal and pre-frontal cortices may better approximate human recognition abilities across development.

A final possibility is that the *kind* of visual experience children receive, rather than the quantity, is crucial for robust object recognition. As mentioned previously, children are exposed to relatively few objects in the first year of life, but their experiences with these objects is far more variable and extensive (*124*). Indeed, DNNs trained on children's views of objects from head-cam videos perform better than those trained on the same objects using adult viewpoints (*161*), suggesting that child experience provides better training data than that of adults'. Other work shows that building in constraints of the newborn visual system, such as infants' initially blurry and low-saturation vision (*69, 70*), improves the robustness of DNNs. And, as mentioned previously, experience with smoothly moving objects is crucial for forming robust object representations (*45, 112, 144-146*). These constraints, combined with potentially innate capacities to represent shape properties like 3D structure (*28, 103, 162, 163*), symmetry (*164, 165*), and other Gestalt cues (*45, 166*), suggest that the human visual system is well optimized to form robust object representations. Although our data suggest that naturalistic visual experience alone is insufficient for DNNs to match child performance, integrating both biologically plausible architectures and naturalistic experiences may help close the gaps between humans and machines.

Towards a developmentally inspired machine learning

In addition to incorporating biological and experiential constraints, a key condition for fostering a field of developmentally-inspired machine learning is the collection of large high-quality child datasets (*30, 84, 167, 168*). In the current study, we sought to provide a new child benchmark by which to explore the developmental mechanisms needed to recognize objects at different speeds and viewing conditions. However, in addition to comparisons of overall performance, it is also important for datasets to support multivariate analyses which can elucidate whether children and DNNs solve tasks using similar approaches. For instance, a common measure of alignment between artificial and biological organisms is representational similarity analyses (RSA), in which the correspondence between two organisms is measured as the degree to which they show correlated patterns of similarity judgments across the object pairs in a set (*169, 170*) (but, see (*171, 172*)). In the context of the current study, RSA would have allowed us to examine whether children and DNNs exhibit similar error patterns for each object target-distractor pair (e.g., confusability between airplane and truck vs. airplane and apple). However, because our study incorporated many experimental factors (5 durations × 3 conditions × 3 age-groups) and children could only conduct a small number of trials per session (30 trials), there was insufficient data for every object pair across experimental factor to allow for RSA analyses. Thus, although the approach used in the current study can shed light on what mechanisms are sufficient to achieve child-level performance, multivariate analyses provide a more precise measure for exploring whether DNNs accomplished object recognition in a human-like way.

Beyond behavior, a crucial source of insight into the mechanisms of visual recognition is neuroimaging data from children. Neural data can directly illuminate the mature and immature processes of the child visual system (*32, 173-177*). Moreover, by combining child imaging data with DNN models, we can shed light on the underlying visual representations of developing brain areas (*178-181*). However, imaging awake behaving children, particularly infants, is notoriously difficult (*182, 183*). Computational modeling offers a powerful method by which to test initial hypotheses of neural development (*63, 184*). These models can then be used to guide the design of otherwise costly and time-consuming neuroimaging experiments with children. In this view, DNNs can be 'control-reared', much like biological organisms (*63, 185*), to examine what architectures, experiences, or learning objectives lead to adult-like neural organization. Thus, we imagine a developmentally-inspired field of machine learning where insights from children are used to improve the efficiency and performance of DNNs, and inversely, DNNs are used as tools to explore difficult to test questions in children.

Materials and Methods

Participants

For human children, sample sizes and testing procedures were preregistered (https://aspredicted.org/Z6B_PF4) on the basis of pilot testing. We sought to test 135 children ages 3 through 5 years of age. Five additional children were included because they were scheduled before the target sample size was met. An additional 12 children were excluded from the analyses for not following instructions (n = 3), failing to complete the task (n = 5), technical issues (n = 3), or because of a developmental disability (n = 1). The five additional children were thus retained to compensate for those who were excluded. In total, data from 128 children (M_{age} = 4.62 years, range = 3.05 – 5.95; 64 girls and 64 boys) were analyzed. Children were randomly assigned to one of

three stimulus conditions (between-subjects: complete, perturbed, or deleted), with an equal number of children (n = 42) per condition. For comparison, we also recruited 30 adult participants (10 per condition) online. One participant did not finish the task and thus 29 adult participants were analyzed. Age and sex were not collected for adults. All participants were tested online through Children Helping Science (children) or Prolific (adults). All protocols (Children: IRB#03571; Adults:IRB#03388) were approved by the Institutional Review Board (IRB) of Emory University.

Models

For computational models, we selected a combination of models based on biologically inspired properties or performance on visual recognition tasks. See Table 1 for all models.

To examine the possible contributions of feedforward and recurrent neural circuits, we implemented a shallow feedforward and recurrent VoneNet architecture (*186*). VoneNet is a convolutional DNN whose architecture is based on the primate ventral visual pathway, with distinct layers corresponding to V1, V2, V4, and IT. What further distinguishes VoneNet from other DNNs is that layer V1 directly simulates the organization of primary visual cortex by implementing a fixed weight front-end comprised of a Gabor filter bank. The inclusion of a biologically plausible front-end was shown to improve the adversarial robustness of DNNs, as well as VoneNet's match to monkey ventral pathway (*186*). Feedforward or recurrent processing was implemented into each VoneNet architecture by manipulating the number of times a stimulus was passed through each layer on a single presentation (i.e., recurrent steps). In the feedforward architecture, each stimulus was processed only once by each layer. For the recurrent architecture, the number of recurrent steps for each layer (V2: 2, V4: 4, IT: 2) was selected on the basis of prior work (*186*).

In addition, to explore the contributions of different kinds of visual experience, each VoneNet architecture was trained on either EcoSet, an ecologically-valid stimulus set comprised of 565 basic-level categories (*73*), or stylized EcoSet, which uses style transfer techniques to randomly vary the color and texture of each EcoSet image (*79*). We specifically included stylized EcoSet because prior work found that conventional DNNs are biased to use texture information, but training on style transfer images induces a more human-like shape bias in the models (*68*). All models were trained in a supervised manner to classify images into 565 categories. Models were trained in PyTorch (*187*) using the procedure described in Dapello, Marques, Schrimpf, Geiger, Cox and DiCarlo (*186*). Thus, four biologically inspired models were trained for this study: feedforward models trained either on EcoSet or Stylized-EcoSet, and recurrent models trained on the same two datasets. The use of these models allowed us to explore two possible mechanisms underlying visual recognition: recurrent processing and variable visual experience.

We also tested two other models with biologically plausible properties. These included a ResNext50 model trained on the SAYCam dataset (*71*), a dataset consisting of naturalistic head-cam videos from three children (child: 'S', 'A', and 'Y'), using temporal self-supervision method (from here on referred to as ResNext50-SAY). Furthermore, to examine whether language may bootstrap visual learning, we tested the CVCL (Child's View for Contrastive Learning) model, a contrastive ResNext50 model that was trained to associate linguistic-utterances with co-occurring visual information from the head-cam videos of a single child from the SAYCam dataset (*74*). CVCL was included because it successfully uses biologically plausible training data and learning objective to recapitulate the visual abilities of a young child.

To estimate the gaps (or lack thereof) between child and machine perception, we also tested several performance-optimized object classification models (as of March 2024). Specifically, we tested a Vision Transformer (ViT), which was trained on ImageNet using a transformer architecture (*188*), and ConvNext, which is a state-of-the-art convolutional DNN that was optimized using the best-practices from the last decade of machine learning to maximize visual recognition performance on ImageNet (*189*). We also tested a high-performing model trained using an unsupervised learning regimen, namely ViT-DINOv2 (*190*) and a vision-language model using the ViT architecture that learned to associate 400m images and text pairs from the CLIP dataset using a contrastive learning objective (*191*).

Finally, to examine the role of learning objective and training size on performance, we also tested six versions of ResNet50. Specifically, we tested a ResNet50 trained to classify images from standard ImageNet (1000 classes, 1.4m images) (*192*) or ImageNet 21k (21k classes, 14m images) (*193*), as well as a ResNet50 architectures trained to associate text-image pairs from YFFC 15M (15m images) or the CLIP training set (400m images) (*194*), and a ResNet50 model trained on ImageNet1k using the unsupervised DINO training procedure (*95*). Finally, we tested a ResNet50 model finetuned on ImageNet-Sketch, a dataset of 50,000 line drawings (*127*). Finetuning was accomplished by taking a standard pre-trained ResNet50 model and retraining the entire model for 30 epochs using the ImageNet-Sketch dataset. To prevent catastrophic forgetting of naturalistic stimulus features, the training dataset was augmented with 50,000 randomly selected images from ImageNet1k (equal number per category).

Stimuli

Thirty unique objects (15 animate; 15 inanimate) were selected for the object recognition task created for this study. Objects were selected from the Snodgrass and Vanderwart (1980) image set by sampling age-appropriate nouns from the Peabody Picture Vocabulary Test (PPVT-4) (*195*). See Supplemental Table 1 for the list of objects used. Each object image was transformed into an outline and adapted into three stimulus conditions: complete, perturbed, and deleted contours. Stimuli with complete contours were adopted as is from Snodgrass and Vanderwart (*196*). Perturbed stimuli were created by applying the ripple distortion (amount = 200%; size = medium) to the complete contour stimuli using image editing software (Photopea; https://www.photopea.com/). The deleted contour stimuli were created by removing 50% of the contours from each complete stimulus using salience scores based on the medial axis of the object (*197*). To minimize visual discomfort during the testing session, stimulus contours were presented as white lines on a gray background (see Figure 1).

Human testing procedure

Each child was tested individually by an experimenter over Zoom using a two-alternative forced-choice (2AFC) procedure. Trials began with a colorful fixation stimulus, which remained onscreen until children attended to it (as determined by the experimenter). The stimulus display was sandwiched by forward and backward masks; the response prompt followed (Figure 1B). Masks were created by randomly overlaying object outlines across the image frame and then box scrambling the result.

During presentation of the response display, the experimenter asked the child whether they saw the target stimulus or a distractor (e.g., “Did you see an airplane or a truck?”; order randomized across trials). The distractor was always another object from the stimulus set with the same

animacy (e.g., animate target paired with animate distractor; randomly selected), ensuring that children could not rely on low-level stimulus features to identify the objects. The response phase was followed by feedback indicating whether the child responded correctly (green check) or incorrectly (gray square).

Duration of stimulus displays varied between 100-300 ms. Durations were determined via a titrated procedure. At the beginning of each session, objects were presented for 300 ms. If children correctly identified objects on three consecutive trials, then the duration decreased by 50 ms (capped at 100 ms). However, if children incorrectly identified objects for three consecutive trials, then the stimulus duration increased by 50 ms (capped at 300 ms).

Prior to testing, children were introduced to the task in a practice phase where they were required to identify slowly presented (1 s) photographs (e.g., shoes) using the same instructions as above.

Adults were tested using an identical procedure, except that testing was conducted asynchronously without an experimenter present. Adults made their response by pressing the arrow key that corresponded to either the target or distractor label.

Model testing procedures

We tested model recognition accuracy using the same stimulus set presented to children and with a comparable 2AFC procedure. In the object recognition task, children were required to determine which of two semantic labels corresponded to the target stimulus (e.g., "airplane" or "truck"). To accomplish this task, children must be able to match the stored object representation that is associated with each label (e.g., examples of previously seen airplanes) with the stimulus image they saw during the trial.

To approximate this process, we trained machine learning classifiers using the feature activations from the penultimate layer of each model on naturalistic images of each object and then tested them on each stimulus display in a pairwise fashion (e.g., train on photographs of airplanes and trucks; test on the perturbed airplane stimulus). For each stimulus category, we provided 500 images for training (15,000 images total). Training images were randomly selected from the EcoSet (*73*), ImageNet21k (*193*), and Open Images (*198*) data sets. Visual inspection of the images showed that they comprised both photographic (i.e., real) and stylized (e.g., cartoons) examples, making them comparable to a child's viewing experience. See Supplemental Figure 4 for classification performance on the images used for training.

Because there can be significant differences in the performance of a model based on which classifier is used or how many examples are used to train the classifier, we tested each model using six common classifiers (support vector machine, logistic regression, ridge regression, naïve bayes, K-nearest neighbors, and nearest centroid). We specifically included the K-nearest neighbors and nearest centroid classifier because these methods may also approximate the use of exemplar and prototype representations in humans, respectively. Furthermore, we parametrically varied the number of images used for classifier training (5, 10, 25, 50, 100, 150, 200, 250, and 300). For each stimulus, training and testing was done using a 20-fold cross validation regime (stratified k-fold). For each model, final comparisons with children were made by selecting the best performing version of the model across all classifiers and quantity of training.

For layer-wise analyses, feature activations were extracted from the final ReLU or linear readout operation of each layer block of a model. For computational efficiency, layer-wise analyses were conducted using only 150 training images and the KNN classifier.

Statistical Analyses

Preprocessing of child data. In the current task, we used a titrated procedure to determine the fastest speed at which children could identify images. However, the use of a titrated procedure meant that not all children contributed trials at every stimulus duration. Indeed, survival analyses revealed that 3-year-olds had a lower probability of reaching the fastest speeds in the perturbed and deleted conditions than 4- and 5-year-olds (but not the completed condition; see Supplemental Figure 1). However, if we only included data from children that reached each speed, then this would likely overestimate children's performance (particularly at the fastest speeds) because it would exclude every participant that might otherwise perform at chance. To address this issue, we imputed data for each missing trial using a conservative estimate that assumes children's performance for those trials would be below chance. This value was computed for participants with missing data by subtracting the standard error of the mean of their performance on existing trials from 0.5.

We chose to impute a below-chance values because imputing 0.5 (i.e., chance) would inflate the test statistic, such that, statistically, child accuracy could never fall below chance levels. That is, if the lower-bound of children's performance on any trial type is 0.5, then the lower-bound of the confidence interval would never cross the chance line. In this context, a one-sample *t*-test comparing performance to chance (0.5) would always have a positive *t*-value, because no value would ever be below 0.5. Thus, by imputing a value below chance, we strongly penalize the overall group estimate of performance for those trials and ensure that we are not overestimating the strength of children's performance. Nonetheless qualitatively similar results are found when imputed data was set to 0.5 (see Supplemental Data).

Low-level shape similarity. To examine whether child performance could be explained by low-level shape features like curvature statistics or a coarse similarity between shape envelopes we computed child recognition accuracy as a function of curvature and shape envelope similarity. The curvature statistics of every object was computed using the Normalized Contour Curvature model (*199*) and similarity was measured as the Pearson correlation between curvature values for each object pair. Shape envelopes were computed by calculating an elongation ratio for each object as the quotient between the longest and shortest axis of the object, and similarity was measured as the difference between elongation ratios for each object pair. Trials were then median split on the basis of curvature or shape envelope similarity.

Model comparison. Humans and models were compared on the basis of overlapping confidence intervals (95% CIs). Human data were split into four age groups (3-, 4-, 5-year-olds, and adults) and two speeds (fast: mean accuracies at 100 ms and 150 ms; slow: mean accuracies at 200 ms and 250 ms). Performance for models was computed as mean decoding accuracy across all pairwise comparisons of objects from the best performing version of the model across all classifiers and quantity of training examples. Confidence intervals (CIs) for humans and models were computed using a bootstrapping procedure (10,000 resamples).

References

1. K. Grill-Spector, N. Kanwisher, Visual Recognition: As Soon as You Know It Is There, You Know What It Is. *Psychol. Sci.* **16**, 152-160 (2005).
2. C. Keysers, D.-K. Xiao, P. Földiák, D. I. Perrett, The speed of sight. *J. Cognit. Neurosci.* **13**, 90-101 (2001).
3. S. Thorpe, D. Fize, C. Marlot, Speed of processing in the human visual system. *Nature* **381**, 520-522 (1996).
4. I. Biederman, M. Bar, One-shot viewpoint invariance in matching novel objects. *Vision Res.* **39**, 2885-2899 (1999).
5. I. Biederman, P. C. Gerhardstein, Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. *J. Exp. Psychol. Hum. Perception Performance* **19**, 1162-1182 (1993).
6. J. M. Brown, C. Koch, Influences of occlusion, color, and luminance on the perception of fragmented pictures. *Percept. Motor Skills* **90**, 1033-1044 (2000).
7. R. F. Murray, A. B. Sekuler, P. J. Bennett, Time course of amodal completion revealed by a shape discrimination task. *Psychonomic Bulletin & Review* **8**, 713-720 (2001).
8. I. Biederman, G. Ju, Surface versus edge-based determinants of visual recognition. *Cog. Psychol.* **20**, 38-64 (1988).
9. J. Wagemans *et al.*, Identification of everyday objects on the basis of silhouette and outline versions. *Perception* **37**, 207-244 (2008).
10. J. H. Elder, L. Velisavljević, Cue dynamics underlying rapid detection of animals in natural scenes. *J. Vision* **9**, 1-20 (2009).
11. I. Biederman, E. E. Cooper, Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cog. Psychol.* **23**, 393-419 (1991).
12. J. Wagemans *et al.*, A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bull.* **138**, 1172-1217 (2012).
13. J. Wasserstein, R. Zappulla, J. Rosen, L. Gerstman, D. Rock, In search of closure: subjective contour illusions, Gestalt completion tests, and implications. *Brain and cognition* **6**, 1-14 (1987).
14. James J. DiCarlo, D. Zoccolan, Nicole C. Rust, How Does the Brain Solve Visual Object Recognition? *Neuron* **73**, 415-434 (2012).
15. V. A. F. Lamme, P. R. Roelfsema, The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571-579 (2000).
16. P. Kok, L. J. Bains, T. van Mourik, D. G. Norris, F. P. de Lange, Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Curr. Biol.* **26**, 371-376 (2016).
17. M. E. Wokke, A. R. Vandenbroucke, H. S. Scholte, V. A. Lamme, Confuse your illusion: feedback to early visual cortex contributes to perceptual completion. *Psychol. Sci.* **24**, 63-71 (2013).
18. R. L. Goldstone, Perceptual learning. *Annu. Rev. Psychol.* **49**, 585-612 (1998).
19. Z.-L. Lu, B. A. Dosher, Current directions in visual perceptual learning. *Nat. Rev. Psyc.* **1**, 654-668 (2022).
20. L. R. Squire, J. C. Frascino, C. S. Rivera, N. C. Heyworth, B. J. He, One-trial perceptual learning in the absence of conscious remembering and independent of the medial temporal lobe. *Proc. Natl. Acad. Sci.* **118**, e2104072118 (2021).
21. M. J. Tarr, H. H. Bülthoff, Image-based object recognition in man, monkey and machine. *Cognition* **67**, 1-20 (1998).

22. N. K. Logothetis, J. Pauls, T. Poggio, Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552-563 (1995).
23. M. J. Arcaro, P. F. Schade, M. S. Livingstone, Universal Mechanisms and the Development of the Face Network: What You See Is What You Get. *Annu. Rev. Vision Sci.* **5**, 341-372 (2019).
24. S. Dehaene, L. Cohen, J. Morais, R. Kolinsky, Illiterate to literate: behavioural and cerebral changes induced by reading acquisition. *Nat. Rev. Neurosci.* **16**, 234-244 (2015).
25. K. Srihasam, J. L. Vincent, M. S. Livingstone, Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nat. Neurosci.* **17**, 1776-1783 (2014).
26. M. C. Frank, Baby steps in evaluating the capacities of large language models. *Nat. Rev. Psyc.* **2**, 451-452 (2023).
27. D. T. Benton, The elusive "Developmental Mechanism": What they are and how to study and test them. *Developmental Review* **65**, 101034 (2022).
28. J. N. Wood, Newborn chickens generate invariant object representations at the onset of visual object experience. *Proc. Natl. Acad. Sci.* **110**, 14000-14005 (2013).
29. E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, L. B. Smith, Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160055 (2017).
30. L. S. Huber, R. Geirhos, F. A. Wichmann, The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. *J. Vision* **23**, 4-4 (2023).
31. G. Golarai *et al.*, Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nat. Neurosci.* **10**, 512-522 (2007).
32. M. Nishimura, K. S. Scherf, V. Zachariou, M. J. Tarr, M. Behrmann, Size precedes view: developmental emergence of invariant object representations in lateral occipital complex. *J. Cognit. Neurosci.* **27**, 474-491 (2015).
33. K. S. Scherf, M. Behrmann, K. Humphreys, B. Luna, Visual category-selectivity for faces, places and objects emerges along different developmental trajectories. *Dev. Sci.* **10**, F15-F30 (2007).
34. V. Ayzenberg, M. C. Granovetter, S. Robert, C. Patterson, M. Behrmann, Differential functional reorganization of ventral and dorsal visual pathways following childhood hemispherectomy. *Dev. Cog. Neuro.*, 101323 (2023).
35. E. Freud, M. Behrmann, The life-span trajectory of visual perception of 3D objects. *Sci. Rep.* **7**, 1-11 (2017).
36. A. W. M. Tan *et al.*, DevBench: A multimodal developmental benchmark for language learning. *arXiv preprint arXiv:2406.10215*, (2024).
37. I. Biederman, R. J. Mezzanotte, J. C. Rabinowitz, Scene perception: Detecting and judging objects undergoing relational violations. *Cog. Psychol.* **14**, 143-177 (1982).
38. M. J. Tarr, P. Williams, W. G. Hayward, I. Gauthier, Three-dimensional object recognition is viewpoint dependent. *Nat. Neurosci.* **1**, 275-277 (1998).
39. R. Rajalingham *et al.*, Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *J. Neurosci.* **38**, 7255-7269 (2018).
40. F. A. Wichmann *et al.*, Methods and measurements to compare men against machines. *Electronic Imaging* **2017**, 36-45 (2017).
41. M. Schrimpf *et al.*, Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, (2020).
42. M. Nishimura, S. Scherf, M. Behrmann, Development of object recognition in humans. *F1000 biology reports* **1**, (2009).
43. V. Ayzenberg, S. Lourenco, Perception of an object's global shape is best described by a model of skeletal structure in human infants. *eLife* **11**, e74943 (2022).

44. D. H. Rakison, G. E. Butterworth, Infants' attention to object structure in early categorization. *Dev. Psychol.* **34**, 1310-1325 (1998).
45. E. Valenza, I. Leo, L. Gava, F. Simion, Perceptual completion in newborn human infants. *Child Dev.* **77**, 1810-1821 (2006).
46. S. P. Johnson, Development of perceptual completion in infancy. *Psychol. Sci.* **15**, 769-775 (2004).
47. V. Ayzenberg, M. Behrmann, Development of visual object recognition. *Nat. Rev. Psyc.* **3**, 73-90 (2024).
48. K. S. Scherf, M. Behrmann, R. Kimchi, B. Luna, Emergence of global shape processing continues through adolescence. *Child Dev.* **80**, 162-177 (2009).
49. I. Kovács, P. Kozma, Á. Fehér, G. Benedek, Late maturation of visual spatial integration in humans. *Proc. Natl. Acad. Sci.* **96**, 12204-12209 (1999).
50. J. Drewes, G. Goren, W. Zhu, J. H. Elder, Recurrent processing in the formation of shape percepts. *J. Neurosci.* **36**, 185-192 (2016).
51. R. von der Heydt, Figure–ground organization and the emergence of proto-objects in the visual cortex. *Front. Psychol.* **6**, (2015).
52. N. Baker, H. Lu, G. Erlikhman, P. J. Kellman, Deep convolutional networks do not classify based on global object shape. *PLoS Comp. Biol.* **14**, e1006613 (2018).
53. N. Baker, H. Lu, G. Erlikhman, P. J. Kellman, Local features and global shape information in object classification by deep convolutional neural networks. *Vision Res.* **172**, 46-61 (2020).
54. V. Ayzenberg, M. Behrmann, Does the brain's ventral visual pathway compute object shape? *Trends Cog. Sci.* **26**, 1119-1132 (2022).
55. J. Davidoff, D. Roberson, Development of Animal Recognition: A Difference between Parts and Wholes. *J. Exp. Child. Psychol.* **81**, 217-234 (2002).
56. M. Jüttner, E. Wakui, D. Petters, S. Kaur, J. Davidoff, Developmental trajectories of part-based and configural object recognition in adolescence. *Dev. Psychol.* **49**, 161-176 (2013).
57. I. Kovács, Human development of perceptual organization. *Vision Res.* **40**, 1301-1310 (2000).
58. N. Kriegeskorte, Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vision Sci.* **1**, 417-446 (2015).
59. A. Doerig *et al.*, The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431-450 (2023).
60. Y. Yermolayeva, D. H. Rakison, Connectionist modeling of developmental changes in infancy: Approaches, challenges, and contributions. *Psychol. Bull.* **140**, 224-255 (2014).
61. D. Pak, D. Lee, S. M. Wood, J. N. Wood, A newborn embodied Turing test for view-invariant object recognition. *arXiv*, (2023).
62. D. T. Benton, An associative learning account of how infants learn about causal action in animates and inanimates: A critical re-examination of 4 classic studies. *J. Exp. Psychol. General*, (2024).
63. J. N. Wood, L. Pandey, S. M. Wood, Digital twin studies for reverse engineering the origins of visual intelligence. *Annu. Rev. Vision Sci.* **10**, 145-170 (2024).
64. H. Tang *et al.*, Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci.* **115**, 8835-8840 (2018).
65. D. Linsley, J. Kim, V. Veerabadran, C. Windolf, T. Serre, Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems* **31**, (2018).
66. K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, J. J. DiCarlo, Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974-983 (2019).

67. T. C. Kietzmann *et al.*, Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci.* **116**, 21854-21863 (2019).
68. R. Geirhos *et al.*, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv*, (2018).
69. L. Vogelsang *et al.*, Potential downside of high initial visual acuity. *Proc. Natl. Acad. Sci.* **115**, 11333-11338 (2018).
70. M. Vogelsang *et al.*, Impact of early visual experience on later usage of color cues. *Science* **384**, 907-912 (2024).
71. J. Sullivan, M. Mei, A. Perfors, E. H. Wojcik, M. C. Frank, SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *PsyArxiv*, (2020).
72. T. P. O'Connell *et al.*, Approaching human 3D shape perception with neurally mappable models. *arXiv preprint arXiv:2308.11300*, (2023).
73. J. Mehrer, C. J. Spoerer, E. C. Jones, N. Kriegeskorte, T. C. Kietzmann, An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc. Natl. Acad. Sci.* **118**, e2011417118 (2021).
74. W. K. Vong, W. Wang, A. E. Orhan, B. M. Lake, Grounded language acquisition through the eyes and ears of a single child. *Science* **383**, 504-511 (2024).
75. S. Sheybani, H. Hansaria, J. Wood, L. Smith, Z. Tiganj, Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems* **36**, (2024).
76. H. Jang, D. McCormack, F. Tong, Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS Biol.* **19**, e3001418 (2021).
77. H. Jang, F. Tong, Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nat. Commun.* **15**, 1989 (2024).
78. A. Burkhalter, Development of forward and feedback connections between areas V1 and V2 of human visual cortex. *Cereb. Cortex* **3**, 476-487 (1993).
79. L. A. Gatys, A. S. Ecker, M. Bethge, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016), pp. 2414-2423.
80. R. C. O'Reilly, J. L. Russin, M. Zolfaghar, J. Rohrlich, Deep predictive learning in neocortex and pulvinar. *J. Cognit. Neurosci.* **33**, 1158-1196 (2021).
81. C. Zhuang *et al.*, Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci.* **118**, (2021).
82. L. B. Smith, Learning to recognize objects. *Psychol. Sci.* **14**, 244-250 (2003).
83. S. S. Jones, L. B. Smith, Object name learning and object perception: a deficit in late talkers. *Journal of Child Language* **32**, 223-240 (2005).
84. S. Sheybani, L. B. Smith, Z. Tiganj, S. S. Maini, A. Dendukuri, ModelVsBaby: a Developmentally Motivated Benchmark of Out-of-Distribution Object Recognition. (2024).
85. L. Zaadnoordijk, T. R. Besold, R. Cusack, The Next Big Thing (s) in Unsupervised Machine Learning: Five Lessons from Infant Learning. *arXiv*, (2020).
86. L. B. Smith, L. K. Slone, A developmental approach to machine learning? *Front. Psychol.* **8**, (2017).
87. R. Cusack, M. A. Ranzato, C. J. Charvet, Helpless infants are learning a foundation model. *Trends Cog. Sci.* **28**, 726-738 (2024).
88. L. Vogelsang, M. Vogelsang, G. Pipa, S. Diamond, P. Sinha, Butterfly effects in perceptual development: A review of the 'adaptive initial degradation' hypothesis. *Developmental Review* **71**, 101117 (2024).
89. L. Yuan, Beyond learnability: understanding human visual development with DNNs. *Trends Cog. Sci.*, (2024).
90. T. Konkle, G. A. Alvarez, A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.* **13**, 491 (2022).

91. L. B. Smith, S. S. Jones, B. Landau, Naming in young children: A dumb attentional mechanism? *Cognition* **60**, 143-171 (1996).
92. L. B. Smith, S. S. Jones, B. Landau, L. Gershkoff-Stowe, L. Samuelson, Object Name Learning Provides On-the-Job Training for Attention. *Psychol. Sci.* **13**, 13-19 (2002).
93. S. Paruthi *et al.*, Recommended Amount of Sleep for Pediatric Populations: A Consensus Statement of the American Academy of Sleep Medicine. *Journal of Clinical Sleep Medicine* **12**, 785-786 (2016).
94. in *Centers for Disease Control and Prevention*. (Centers for Disease Control and Prevention, 2024).
95. M. Caron *et al.*, in *Proceedings of the IEEE/CVF international conference on computer vision*. (2021), pp. 9650-9660.
96. Y. Bahri, E. Dyer, J. Kaplan, J. Lee, U. Sharma, Explaining neural scaling laws. *Proc. Natl. Acad. Sci.* **121**, e2311878121 (2024).
97. E. Wakui *et al.*, Earlier development of analytical than holistic object recognition in adolescence. *PLOS ONE* **8**, e61041 (2013).
98. B. Landau, L. B. Smith, S. S. Jones, The importance of shape in early lexical learning. *Cog. Dev.* **3**, 299-321 (1988).
99. S. O. Abdelrahim, M. C. Frank, in *Proc. Annu. Meet. Cog. Sci.* (2024), vol. 46.
100. A. F. Pereira, L. B. Smith, Developmental changes in visual object recognition between 18 and 24 months of age. *Dev. Sci.* **12**, 67-80 (2009).
101. L. B. Smith, From fragments to geometric shape: Changes in visual object recognition between 18 and 24 months. *Current Directions in Psychological Science* **18**, 290-294 (2009).
102. A. Slater, A. Mattock, E. Brown, J. G. Bremner, Form perception at birth: Revisited. *J. Exp. Child. Psychol.* **51**, 395-406 (1991).
103. A. Slater, V. Morison, Shape constancy and slant perception at birth. *Cog. Dev. Infancy*, 119-129 (1987).
104. A. Slater, V. Morison, D. Rose, New-born infants' perception of similarities and differences between two-and three-dimensional stimuli. *Br. J. Dev. Psychol.* **2**, 287-294 (1984).
105. C. Mash, M. E. Arterberry, M. H. Bornstein, Mechanisms of visual object tecognition in infancy: Five-month-olds generalize beyond the interpolation of familiar views. *Infancy* **12**, 31-43 (2007).
106. H. A. Ruff, Infant recognition of the invariant form of objects. *Child Dev.* **49**, 293-306 (1978).
107. K. S. Kraebel, P. C. Gerhardstein, Three-month-old infants' object recognition across changes in viewpoint using an operant learning procedure. *Infant Behav. and Dev.* **29**, 11-23 (2006).
108. D. Mareschal, P. C. Quinn, Categorization in infancy. *Trends Cog. Sci.* **5**, 443-450 (2001).
109. P. C. Quinn, A. M. Slater, E. Brown, R. A. Hayes, Developmental change in form categorization in early infancy. *Br. J. Dev. Psychol.* **19**, 207-218 (2001).
110. C. Turati, F. Simion, L. Zanon, Newborns' perceptual categorization for closed and open geometric forms. *Infancy* **4**, 309-325 (2003).
111. P. C. Quinn, P. D. Eimas, M. J. Tarr, Perceptual categorization of cat and dog silhouettes by 3- to 4-month-old infants. *J. Exp. Child. Psychol.* **79**, 78-94 (2001).
112. P. J. Kellman, E. S. Spelke, Perception of partly occluded objects in infancy. *Cog. Psychol.* **15**, 483-524 (1983).
113. P. J. Kellman, C. Yin, T. F. Shipley, A common mechanism for illusory and occluded object completion. *J. Exp. Psychol. Hum. Perception Performance* **24**, 859 (1998).
114. A. Slater *et al.*, Newborn and older infants' perception of partly occluded objects. *Infant Behav. and Dev.* **13**, 33-49 (1990).
115. S. P. Johnson, R. N. Aslin, Perception of object unity in 2-month-old infants. *Dev. Psychol.* **31**, 739 (1995).

116. J. S. Bowers *et al.*, Deep problems with neural network models of human vision. *Behav. Brain. Sci.* **46**, e385 (2023).
117. T. Tullis, B. Albert, in *Measuring the User Experience (Second Edition),* T. Tullis, B. Albert, Eds. (Morgan Kaufmann, Boston, 2013), pp. 163-186.
118. R. H. Carpenter, *Movements of the Eyes, 2nd Rev*. (Pion Limited, 1988).
119. S. Jayaraman, L. B. Smith, Faces in early visual environments are persistent not just frequent. *Vision Res.* **157**, 213-221 (2019).
120. S. Jayaraman, C. M. Fausey, L. B. Smith, The faces in infant-perspective scenes change over the first year of life. *PloS ONE* **10**, e0123780 (2015).
121. B. L. Long, G. Kachergis, K. Agrawal, M. C. Frank, A longitudinal analysis of the social information in infants' naturalistic visual experience using automated detections. *Dev. Psychol.* **58**, 2211 (2022).
122. E. M. Anderson, T. R. Candy, J. M. Gold, L. B. Smith, An edge-simplicity bias in the visual input to young infants. *Science Advances* **10**, eadj8571 (2024).
123. K. H. James, S. S. Jones, L. B. Smith, S. N. Swain, Young children's self-generated object views and object recognition. *J. Cog. Dev.* **15**, 393-401 (2014).
124. L. B. Smith, S. Jayaraman, E. Clerkin, C. Yu, The developing infant creates a curriculum for statistical learning. *Trends Cog. Sci.* **22**, 325-336 (2018).
125. C. J. Dunst, E. Gorman, Development of infant and toddler mark making and scribbling. *Centre for Early Learning Literacy Review* **2**, (2009).
126. A. Saito, M. Hayashi, H. Takeshita, T. Matsuzawa, The origin of representational drawing: a comparison of human children and chimpanzees. *Child Dev.* **85**, 2232-2246 (2014).
127. H. Wang, S. Ge, Z. Lipton, E. P. Xing, Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems* **32**, (2019).
128. P. Mayilvahanan *et al.*, In search of forgotten domain generalization. *arXiv preprint arXiv:2410.08258*, (2024).
129. S. Madan *et al.*, Benchmarking Out-of-Distribution Generalization Capabilities of DNN-based Encoding Models for the Ventral Visual Cortex. *arXiv preprint arXiv:2406.16935*, (2024).
130. S. Madan, T. Sasaki, T.-M. Li, X. Boix, H. Pfister, Small in-distribution changes in 3d perspective and lighting fool both cnns and transformers. *arXiv preprint arXiv:2106.16198* **3**, (2021).
131. R. Geirhos *et al.*, Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665-673 (2020).
132. M. C. Frank, Bridging the data gap between children and large language models. *Trends Cog. Sci.*, (2023).
133. E. Yiu *et al.*, KiVA: Kid-inspired Visual Analogies for Testing Large Multimodal Models. *arXiv preprint arXiv:2407.17773*, (2024).
134. M. K. Goddu, A. Gopnik, The development of human causal learning and reasoning. *Nat. Rev. Psyc.* **3**, 319-339 (2024).
135. A. Cherian, K.-C. Peng, S. Lohit, K. A. Smith, J. B. Tenenbaum, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2023), pp. 10834-10844.
136. A. M. Zador, A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* **10**, 3770 (2019).
137. T. S. Lee, D. Mumford, R. Romero, V. A. Lamme, The role of the primary visual cortex in higher level vision. *Vision Res.* **38**, 2429-2454 (1998).
138. M. Bar *et al.*, Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci.* **103**, 449-454 (2006).
139. C. Teufel, S. C. Dakin, P. C. Fletcher, Prior object-knowledge sharpens properties of early visual feature-detectors. *Sci. Rep.* **8**, 10853 (2018).

140. R. Hardstone *et al.*, Long-term priors influence visual perception through recruitment of long-range feedback. *Nat. Commun.* **12**, 6288 (2021).
141. M. Bar, Visual objects in context. *Nat. Rev. Neurosci.* **5**, 617-629 (2004).
142. A. Subramanian *et al.*, Benchmarking the speed–accuracy tradeoff in object recognition by humans and neural networks. *J. Vision* **25**, 4-4 (2025).
143. Y.-A. Cheng *et al.*, RTify: Aligning Deep Neural Networks with Human Behavioral Decisions. *arXiv preprint arXiv:2411.03630*, (2024).
144. J. N. Wood, S. M. Wood, The development of invariant object recognition requires visual experience with temporally smooth objects. *Cog. Sci.* **42**, 1391-1406 (2018).
145. P. J. Kellman, K. R. Short, Development of three-dimensional form perception. *J. Exp. Psychol. Hum. Perception Performance* **13**, 545 (1987).
146. E. S. Spelke, R. Kestenbaum, D. J. Simons, D. Wein, Spatiotemporal continuity, smoothness of motion and object identity in infancy. *Br. J. Dev. Psychol.* **13**, 113-142 (1995).
147. B. Wang, C. R. Ponce, Tuning landscapes of the ventral stream. *Cell Reports* **41**, (2022).
148. A. V. Jagadeesh, J. Gardner, Texture-like representation of objects in human visual cortex. *Proc. Natl. Acad. Sci.*, (2022).
149. B. Long, C.-P. Yu, T. Konkle, Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci.* **115**, E9015-E9024 (2018).
150. K. Kar, J. J. DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition. *Neuron* **109**, 164-176.e165 (2021).
151. V. Ayzenberg, C. Simmons, M. Behrmann, Temporal asymmetries and interactions between dorsal and ventral visual pathways during object recognition. *Cereb. Cortex Comms.* **4**, tgad003 (2022).
152. V. Zachariou, C. V. Nikas, Z. N. Safiullah, S. J. Gotts, L. G. Ungerleider, Spatial mechanisms within the dorsal visual pathway contribute to the configural processing of faces. *Cereb. Cortex* **27**, 4124-4138 (2017).
153. V. Romei, J. Driver, P. G. Schyns, G. Thut, Rhythmic TMS over Parietal Cortex Links Distinct Brain Frequencies to Global versus Local Visual Processing. *Curr. Biol.* **21**, 334-337 (2011).
154. E. Freud, D. C. Plaut, M. Behrmann, 'What 'is happening in the dorsal visual pathway. *Trends Cog. Sci.* **20**, 773-784 (2016).
155. O. Rose, C. R. Ponce, A concentration of visual cortex-like neurons in prefrontal cortex. *Nat. Commun.* **15**, 7002 (2024).
156. J. A. Bourne, M. G. Rosa, Hierarchical development of the primate visual cortex, as revealed by neurofilament immunoreactivity: early maturation of the middle temporal area (MT). *Cereb. Cortex* **16**, 405-414 (2006).
157. V. Ayzenberg, C. Song, M. J. Arcaro, An intrinsic hierarchical, retinotopic organization of visual pulvinar connectivity in the human neonate. *Curr. Biol.*, (2025).
158. V. Ayzenberg, M. Behrmann, The Dorsal Visual Pathway Represents Object-Centered Spatial Relations for Object Recognition. *J. Neurosci.* **42**, 4693-4710 (2022).
159. S. M. Kolk, P. Rakic, Development of prefrontal cortex. *Neuropsychopharmacology* **47**, 41-57 (2022).
160. C. T. Ellis, L. J. Skalaban, T. S. Yates, N. B. Turk-Browne, Attention recruits frontal cortex in human infants. *Proc. Natl. Acad. Sci.* **118**, e2021474118 (2021).
161. S. Bambach, D. J. Crandall, L. B. Smith, C. Yu, in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. (IEEE, 2017), pp. 290-295.
162. J. N. Wood, S. M. W. Wood, One-shot learning of view-invariant object representations in newborn chicks. *Cognition* **199**, 104192 (2020).

163. A. Slater, A. Mattock, E. Brown, Size constancy at birth: Newborn infants' responses to retinal and real size. *J. Exp. Child. Psychol.* **49**, 314-322 (1990).
164. C. Turati, F. Simion, I. Milani, C. Umiltà, Newborns' preference for faces: What is crucial? *Dev. Psychol.* **38**, 875-882 (2002).
165. V. M. Cassia, E. Valenza, F. Simion, I. Leo, Congruency as a nonspecific perceptual property contributing to newborns' face preference. *Child Dev.* **79**, 807-820 (2008).
166. V. M. Cassia, F. Simion, I. Milani, C. Umiltà, Dominance of global visual properties at birth. *J. Exp. Psychol. General* **131**, 398 (2002).
167. C. Spriet, E. Abassi, J.-R. Hochmann, L. Papeo, Visual object categorization in infancy. *Proc. Natl. Acad. Sci.* **119**, e2105866119 (2022).
168. J. E. Kiat *et al.*, Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis. *Dev. Sci.*, e13155 (2021).
169. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Sys. Neuro.* **2**, (2008).
170. M. Schrimpf *et al.*, Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, (2018).
171. C. Conwell, J. S. Prince, K. N. Kay, G. A. Alvarez, T. Konkle, A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 9383 (2024).
172. N. J. Sexton, B. C. Love, Directly interfacing brain and deep networks exposes non-hierarchical visual processing. *bioRxiv*, 2021.2006. 2028.450213 (2021).
173. T. Dekker, D. Mareschal, M. I. Sereno, M. H. Johnson, Dorsal and ventral stream activation and object recognition performance in school-age children. *NeuroImage* **57**, 659-670 (2011).
174. L. L. Emberson, S. L. Crosswhite, J. E. Richards, R. N. Aslin, The Lateral Occipital Cortex Is Selective for Object Shape, Not Texture/Color, at Six Months. *J. Neurosci.* **37**, 3698-3703 (2017).
175. X. Yan *et al.*, When Do Visual Category Representations Emerge in Infants' Brains? *Preprint at: https://www.biorxiv.org/content/10.1101/2023.05.11.539934*, 2023.2005. 2011.539934 (2023).
176. K. Grill-Spector, G. Golarai, J. Gabrieli, Developmental neuroimaging of the human ventral visual cortex. *Trends Cog. Sci.* **12**, 152-162 (2008).
177. T. S. Yates, C. T. Ellis, N. B. Turk-Browne, Emergence and organization of adult brain function throughout child development. *NeuroImage* **226**, 117606 (2021).
178. Á. T. Dineen *et al.*, The Effect of Low Acuity in Infancy on Developing Ventral Visual Stream Representations. *J. Vision* **24**, 1416-1416 (2024).
179. C. O'Doherty *et al.*, in *Cognitive Computational Neuroscience (CCN)*. (Boston, MA, 2024).
180. L. Bayet *et al.*, Temporal dynamics of visual representations in the infant brain. *Dev. Cog. Neuro.* **45**, 100860 (2020).
181. S. Xie *et al.*, Visual category representations in the infant brain. *Curr. Biol.* **32**, 5422-5432. e5426 (2022).
182. C. T. Ellis *et al.*, Re-imagining fMRI for awake behaving infants. *Nat. Commun.* **11**, 4523 (2020).
183. N. B. Turk-Browne, R. N. Aslin, Infant neuroscience: how to measure brain activity in the youngest minds. *Trends Neurosci.* **47**, 338-354 (2024).
184. N. M. Blauch, M. Behrmann, D. C. Plaut, A connectivity-constrained computational account of topographic organization in high-level visual cortex. *Proc. Natl. Acad. Sci.* **119**, (2022).
185. D. Lee, P. Gujarathi, J. N. Wood, Controlled-rearing studies of newborn chicks and deep neural networks. *arXiv*, (2021).

186. J. Dapello *et al.*, Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Advances in Neural Information Processing Systems* **33**, 13073-13087 (2020).
187. A. Paszke *et al.*, Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32**, (2019).
188. A. Dosovitskiy *et al.*, An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, (2020).
189. Z. Liu *et al.*, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (2022), pp. 11976-11986.
190. M. Oquab *et al.*, Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, (2023).
191. A. Radford *et al.*, in *International conference on machine learning*. (PMLR, 2021), pp. 8748-8763.
192. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016), pp. 770-778.
193. T. Ridnik, E. Ben-Baruch, A. Noy, L. Zelnik-Manor, Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, (2021).
194. M. Cherti *et al.*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (2023), pp. 2818-2829.
195. L. M. Dunn, D. M. Dunn, *PPVT-4: Peabody picture vocabulary test*. (Pearson Assessments, 2007).
196. J. G. Snodgrass, M. Vanderwart, A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory* **6**, 174 (1980).
197. M. Rezanejad *et al.*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2019), pp. 4116-4124.
198. A. Kuznetsova *et al.*, The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* **128**, 1956-1981 (2020).
199. A. Marantan, I. Tolkova, L. Mahadevan, Image cognition using contour curvature statistics. *Proceedings of the Royal Society A* **479**, 20220662 (2023).