Received: 30 October 2020 | Revised: 23 June 2021 | Accepted: 1 July 2021

DOI: 10.1111/desc.13155

Developmental Science WILEY

PAPER

Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis

John E. Kiat | Steven J. Luck | Aaron G. Beckner | Taylor R. Hayes | Katherine I. Pomaranski | John M. Henderson | Lisa M. Oakes

Center for Mind & Brain and Department of Psychology, University of California, Davis, California, USA

Correspondence
Steven J. Luck, UC-Davis Center for Mind & Brain, 267 Cousteau Place, Davis, CA 95618, USA.
Email: sjluck@ucdavis.edu

John E. Kiat and Steven J. Luck have equal contribution.

Abstract

Little is known about the development of higher-level areas of visual cortex during infancy, and even less is known about how the development of visually guided behavior is related to the different levels of the cortical processing hierarchy. As a first step toward filling these gaps, we used *representational similarity analysis* (RSA) to assess links between gaze patterns and a neural network model that captures key properties of the ventral visual processing stream. We recorded the eye movements of 4- to 12-month-old infants ($N = 54$) as they viewed photographs of scenes. For each infant, we calculated the similarity of the gaze patterns for each pair of photographs. We also analyzed the images using a convolutional neural network model in which the successive layers correspond approximately to the sequence of areas along the ventral stream. For each layer of the network, we calculated the similarity of the activation patterns for each pair of photographs, which was then compared with the infant gaze data. We found that the network layers corresponding to lower-level areas of visual cortex accounted for gaze patterns better in younger infants than in older infants, whereas the network layers corresponding to higher-level areas of visual cortex accounted for gaze patterns better in older infants than in younger infants. Thus, between 4 and 12 months, gaze becomes increasingly controlled by more abstract, higher-level representations. These results also demonstrate the feasibility of using RSA to link infant gaze behavior to neural network models. A video abstract of this article can be viewed at https://youtu.be/K5mF2Rw98Is

KEYWORDS
AlexNet, attention, convolutional neural networks (CNNs), deep neural networks (DNNs), infant development, visually guided behavior

1 | INTRODUCTION

Infants rely heavily on vision to explore the environment, especially before they are able to crawl or walk. Their visual exploration—and access to new information about the visual world—depends on their ability to shift their gaze from one location to another. Thus, the development of gaze control is fundamental to infants' learning.

A growing body of research has begun to outline how the control of gaze develops across infancy. In general, the direction of gaze is controlled primarily by low-level physical salience in young infants and becomes progressively more controlled by higher-level visual information and goals over the ensuing months (Colombo, 2001; Frank et al., 2009, 2014). At the neural level, eye movements appear to be primarily controlled by the superior colliculus in newborns, with an increase

in cortical control over development (Amso & Scerif, 2015; Colombo, 2001; Johnson, 1990). However, the cerebral cortex does not develop in a monolithic manner, and little is known about how the development of individual cortical regions influences gaze control during infancy. The goal of the present study is to link the development of gaze control to the kinds of information represented in different areas of the ventral stream.

Although much is known about the development of primary visual cortex, relatively little is known about the development of higher-level visual cortical areas in human infants. This is largely due to a lack of methods for measuring functional activity in specific brain regions that are suitable for use in young infants. Event-related potentials (ERPs) have been widely used in infants (Braddick & Atkinson, 2011), but they do not have the spatial resolution necessary to differentiate between closely spaced areas of cortex. Functional near infrared spectroscopy (fNIRS) has better spatial resolution, but fNIRS has so far provided only coarse, sensor-level information about the development of higher-level visual areas (Grossmann et al., 2008; Lloyd-Fox et al., 2009). At this time, there are only a few studies of single-unit activity in higher-level visual cortex in infant macaque monkeys (e.g., Rodman et al., 1991, 1993) and only a few functional magnetic resonance imaging (fMRI) studies using complex visual stimuli in awake human infants (e.g., Biagi et al., 2015; Deen et al., 2017; Ellis et al., 2020).

In macaque monkeys, the available research indicates that the general anatomical hierarchy of the ventral cortex is present at birth (Batardière et al., 2002), and both low- and high-level areas of visual cortex are responsive within a few months after birth, with some adult-like featural specificity (Kiorpes, 2016; Rodman et al., 1991, 1993). In humans, the major visual areas can be functionally distinguished by 4–6 months after birth, but these areas do not yet exhibit the highly selective responses observed in adults (Deen et al., 2017). In addition, although considerable information about features such as orientation is available in lower-level cortical areas in the first few postnatal months, this information does not appear to control behavior in human infants. For example, behaviorally measured visual acuity in young infants is considerably worse than would be expected from the optical information available to the photoreceptors and from ERP measures of cortical activity (Candy et al., 1998; Kiorpes, 2016; Norcia et al., 1990). Thus, little is known about how the development of the various areas of visual cortex impacts visually guided behavior once cortical control of behavior has been established. The goal of the present study was therefore to take a first step toward understanding how the information represented in different areas of the ventral visual pathway predicts looking behavior in infants of different ages.

We addressed this goal by using *representational similarity analysis* (RSA) to link patterns of infant eye movements to a computational model that captures important properties of the ventral stream. RSA makes it possible to assess relationships between results obtained from widely different methods, including behavioral measures, computational models, and neural recordings. In neuroimaging research, for example, RSA is widely used to link the pattern of activation across voxels to computational models and to behavior (Cichy et al., 2014; Groen et al., 2018; Kriegeskorte et al., 2008; Wen et al., 2018). RSA solves a fundamental problem that arises in comparing results across methods, namely, that the units of measurement and the format of the empirical data are typically quite different across methods. For example, in a study assessing the processing of visual scenes, the data from a neural network model would consist of a pattern of activation across processing units for each scene, whereas fMRI data would consist of a set of activation values across voxels for each scene, and eye movement data would consist of a set of {x, y} coordinates for each scene. RSA allows us to relate these different data formats to each other by presenting the same set of inputs to each system and examining the similarity structure of the outputs of each system.

RESEARCH HIGHLIGHTS

- We used representational similarity analysis to link infant eye movements to a neural network model that captures key properties of the ventral pathway
- Lower-level layers of the network predicted the gaze patterns of younger infants better than the gaze patterns of older infants
- Higher-level layers of the network predicted the gaze patterns of older infants better than the gaze patterns of younger infants
- More generally, this study demonstrates the feasibility of using representational similarity analysis to link infant gaze behavior to computational models

The present study used RSA to link infant eye movement patterns to a convolutional neural network (CNN) model of visual scene recognition. We obtained eye tracking data from a prior study in which infants between 4 and 12 months viewed photographs of complex natural scenes (Pomaranski et al., in press). For each scene, a *fixation density map* was generated for each infant, representing how often the infant fixated each location in that scene (see examples in Figure 1a). We then computed the correlation between the fixation density maps for each pair of scenes. Each of these correlations indicates the similarity between the fixation density maps for a given pair of scenes. We then organized the pairwise correlations from 22 scenes into a 22×22 matrix, which is called a *representational similarity matrix* (RSM; see Figure 1b). A separate RSM was obtained for each infant, quantifying the *representational geometry* of the gaze patterns for that infant.

We also submitted each of the 22 scenes to a CNN (based on the AlexNet architecture; Krizhevsky et al., 2012) that was trained to classify photographs of natural scenes. Although AlexNet was originally developed as a machine vision system and not as a model of the human visual system per se, its design was inspired by the known properties of the primate ventral pathway, and the sequence of layers does a reasonable job of fitting data recorded from the sequence of areas along the ventral pathway (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Storrs et al., 2020; Wen et al., 2018; Yamins et al., 2014). The structure of the network is shown in

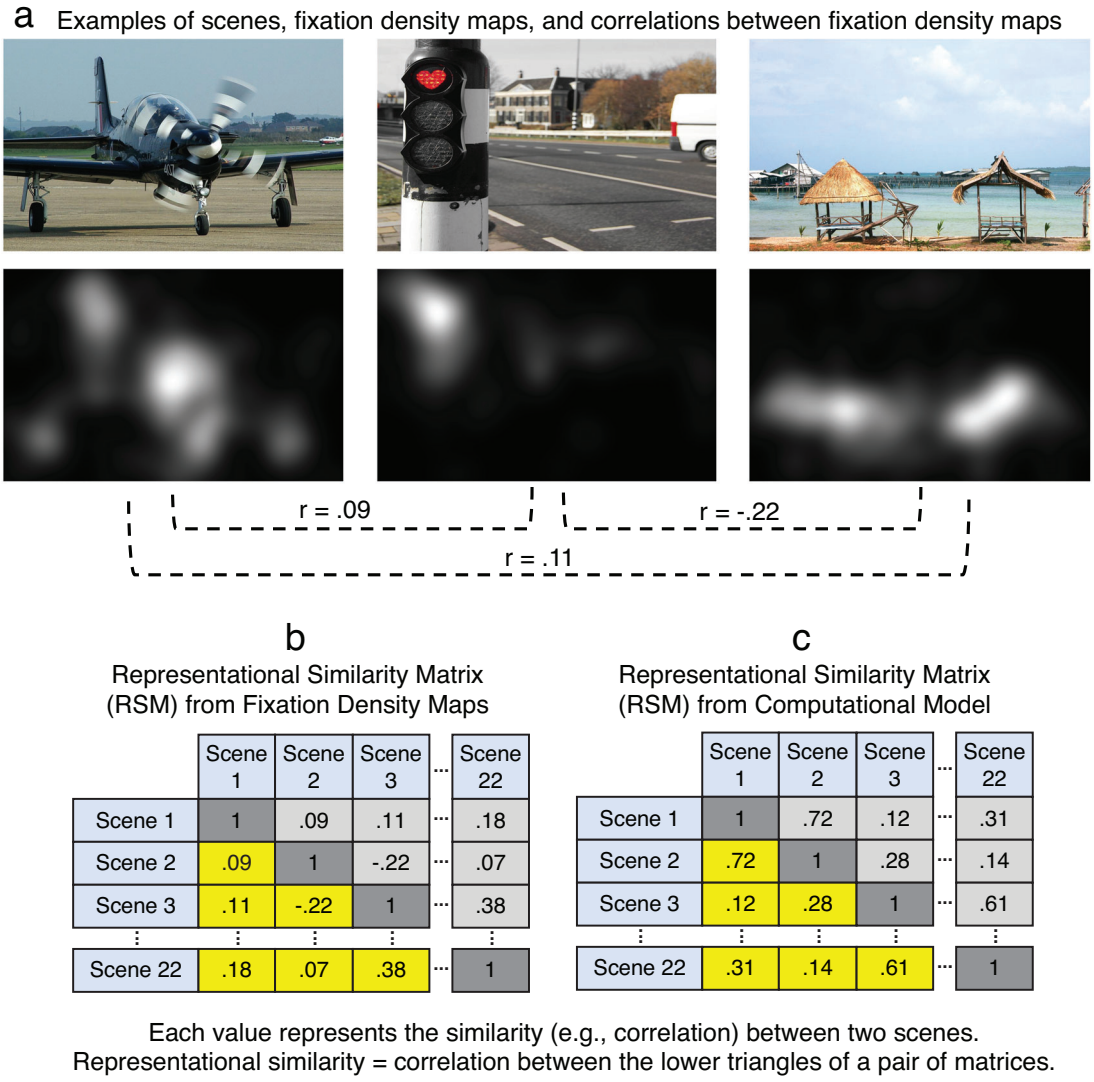

FIGURE 1 (a) Examples of three scenes used in the present study and the fixation density maps for each scene (averaged over all participants). The *r* values represent the Pearson correlation between a pair of fixation density maps. Each of these *r* values would fill one cell of a representational similarity matrix. Note, however, that single-participant fixation density maps were used for the analyses of the present study rather than average fixation density maps shown here. (b, c) Simplified examples of representational similarity matrices from the fixation density maps and from the computational model. The lower and upper triangles are mirror images, and the cells on the diagonal always have a value of one (the correlation between the data of a scene with itself). Representational similarity is typically defined as the rank order correlation between a pair of representational similarity matrices (limited to the lower triangles of the matrices, highlighted in yellow).

Figure 2, and a more detailed explanation is provided in the online supplementary materials. Passing a scene through the network produces a pattern of activation across the units in each layer. We computed the correlation between the activation patterns for each pair of scenes, separately for each spatially organized layer, yielding a separate RSM for each layer (see Figure 1c). These RSMs reflect the representational geometry of the individual layers.

To determine how the representational geometry of each layer was related to the representational geometry of the infant gaze control system—and how this relation changes over development—we computed the correlation between the eye movement RSM for each infant (as in Figure 1b) and the neural network RSM for each layer (as in Figure 1c). This approach allowed us to test the hypothesis that the representations that control shifts of gaze become more abstract over development in a manner that parallels the increasing abstraction that occurs as information flows through the ventral stream in adults. This general hypothesis leads to two specific predictions: (1) the pattern of activation in the lower layers of the network (which are analogous to lower-level areas of visual cortex) will account for the gaze patterns of younger infants better than the gaze patterns of older infants; and (2) the pattern of activation in the higher layers of the network (which are analogous to higher-level areas of visual cortex) will account for the gaze patterns of older infants better than the gaze patterns of younger infants.

If observed, these predicted patterns would confirm our general hypothesis that, over development, gaze control becomes governed

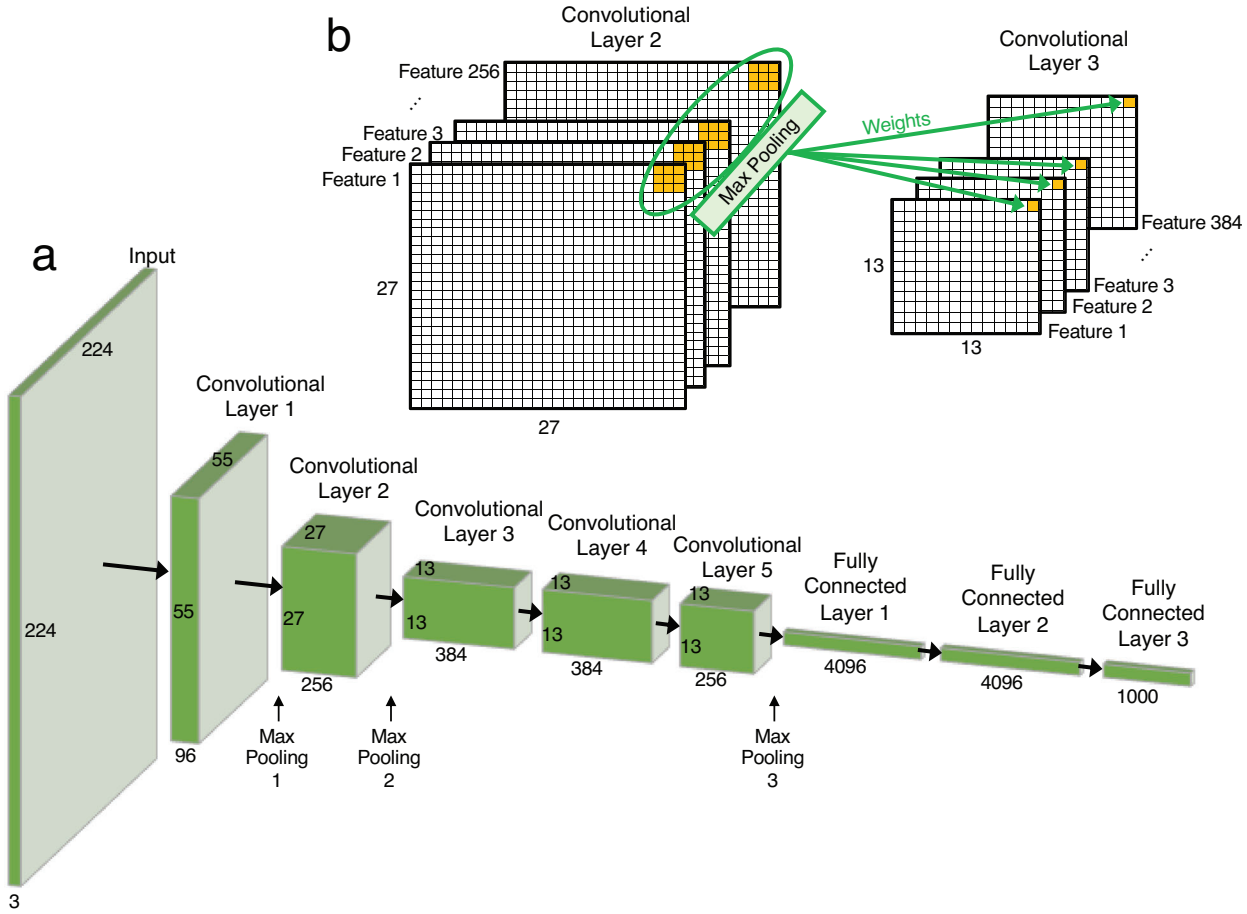

FIGURE 2 (a) General architecture of AlexNet. The input is a bitmap with three layers (red, green, and blue). Information flows forward through each successive layer, and the pattern of activation over the units in the final layers is used to guess the class of the input scene. (b) Detailed look at the flow of information between layers. Note that the output of each unit (except for the max pooling layers) is passed through a rectified linear unit, which zeros the value if it is negative.

by progressively more abstract representations like those found in higher-level areas of visual cortex. Such a pattern of results would also corroborate prior research showing that eye movement control shifts from being dominated by low-level physical stimulus features early in infancy to being dominated by higher-level information in older infants (Colombo, 2001; Frank et al., 2009, 2014). An alternative possibility is that younger infants would simply have less consistent gaze patterns than older infants, leading older infants to exhibit stronger representational similarity than younger infants for all layers of AlexNet.

It should be noted that each layer of AlexNet (except for the fully connected layers) contains three dimensions, including X and Y spatial dimensions and a Z dimension that represents the different features coded at each location of the layer. By contrast, the infant eye movement data are 2-dimensional, with X and Y spatial dimensions. RSA transforms both the 3-dimensional network representations and the 2-dimentional fixation maps into the common space of representational similarity matrices, making it possible to link them despite their different dimensionality. In addition, the mapping between the representation of space in the network (or in the brain) and the space of the eye movements may not be linear, and RSA avoids the need to assume a specific mapping function between the coordinate systems.

This study also served as a proof of principle for using RSA to link patterns of infant eye movements with computational models. RSA can be used for virtually any type of behavioral, neural, or model data as long as separate responses have been obtained for a reasonably large set of inputs (like the 22 images used in the present study). This flexibility means that RSA could be used to address a broad range of questions about the development of perception and cognition during infancy.

2 | METHODS

The data were obtained from a study that was designed for a different purpose (Pomaranski et al., in press). The specific data and analysis code used in the present study are available at https://osf.io/ehg82/.

2.1 | Participants

The final sample included 54 healthy, typically developing infants (27 boys) ranging in age from 114 to 373 days. All infants were born full term, and we excluded individuals who were at significant risk for

colorblindness (e.g., boys with a maternal grandfather who was colorblind). An additional 48 infants were tested but excluded from the analyses due to fussiness or lack of interest on the part of the infant, parental interference, calibration or tracking failure, or experimenter error. The exclusion criteria were established in Pomaranski et al., (in press). All infants were provided with a toy or t-shirt and a certificate in appreciation of their time.

The sample of infants reflected the demographics of the local community (57% were Caucasian, 22% were mixed race, and the others were African American, other, or no race reported), and the mothers were well educated (52% had earned at least a bachelor's degree).

We conducted correlations of these variables with age to determine whether there were any differences in demographic characteristics across ages. We found that infant gender ($t(54) = 0.996$, $p = 0.324$), race ($F(7,54) = 0.516$, $p = 0.818$) and parental education level ($rho(52) = -0.032$, $p = 0.821$) did not significantly vary with age.

2.2 | Stimuli

The primary stimuli were 24 images taken from the MIT Saliency Benchmark (Judd et al., 2012). Examples are shown in Figure 1a. In addition, infants were periodically shown 10-s video clips from popular children's television shows (e.g., *Sesame Street*, *Baby Einstein*) to maintain their interest.

2.3 | Apparatus

Infants' eye movements were recorded at 120 Hz using an SMI-RED N eye tracker. The eye tracker was attached to the bottom of an LCD monitor that had a camera attached to the top to record the participants' head and body position. The stimulus images completely filled the monitor (approximately 48 cm wide, 30 cm high, 1680 × 1050 resolution).

2.4 | Procedure

The protocol was reviewed and approved by the local Institutional Review Board, and parents gave informed consent. Infants sat on their parent's lap or in a highchair (with parent nearby), positioned so their eyes were approximately 60 cm from the monitor. At this viewing distance, the images were approximately 46 × 28 degrees of visual angle. We instructed parents to interact with their infant as little as possible, to remain quiet, and not to direct their infant's attention to the screen.

The session began with an automatic 5-point calibration procedure, followed by a validation. The calibration procedure was repeated until the infant's estimated gaze position was within approximately 2° of the validation locations. After calibration, parents were fitted with felt-covered sunglasses to minimize bias.

Each experimental trial began with a flashing fixation crosshair presented at the center of the screen, accompanied by an attention-grabbing sound (i.e., bells, rattle). Once fixation remained within approximately 5° of the fixation crosshair for 200 ms, the fixation crosshair was replaced with a 5-s presentation of one of the 24 images, paired with classical music. Trials were presented in blocks of four, each of which was followed by a short video clip to maintain infants' interest and to reduce fussiness. Each item was presented in only one block; the images were presented in random order within a block. An example of a trial block, with an infant's eye gaze superimposed, can be found at https://osf.io/ehg82/.

Each included infant provided data from at least 18 images (range = 18–22) of the 24 maximum images. Given variation in infants' interest across the session, not every infant saw every image. Following the prior study (Pomaranski et al., in press), we selected the images that were viewed by at least 12 infants, resulting in a final selection of 22 images for analysis. Each of these scenes was viewed by a sizable proportion of the sample (range = 51–54 infants, mean = 52.737, SD = 1.032). The entire set of 22 images is provided in Figure S1.

2.5 | Data processing

The eye tracker data were filtered into fixations using the SMI BeGaze analysis software with standard parameters for low-speed (< 200 Hz) eye-tracking. Fixations were defined as any period of gaze that was at least 80 ms in duration, with maximum dispersion of 100 pixels. The X and Y coordinates of each fixation to each image were then used to create fixation density maps, or matrices representing the number of fixations centered at each pixel location of the 1680 × 1050 image. A separate fixation density map was created for each image that a participant viewed. To account for eye tracking error and variability within a fixation period, the 1680 × 1050 pixel maps were filtered with a Gaussian kernel (full width at half maximum = 150 pixels). The maps were then normalized using MATLAB's mat2gray.m function so that all cells in the matrix contained values ranging from 0 (the least fixated pixel) to 1 (the most fixated pixel).

2.6 | AlexNet

We used a CNN based on the AlexNet architecture (Krizhevsky et al., 2012), pre-trained on the Places365 database (Zhou et al., 2018), and implemented in Caffe (Jia et al., 2014). As illustrated in Figure 2, this CNN has five convolutional layers, intermixed with three max pooling layers and followed by three fully connected layers. All convolutional layers perform a linear convolution on their inputs. A rectified linear unit (ReLU) function is applied after all convolutional and fully connected layers, which sets negative values to zero. We selected this particular CNN because it is optimized for scene classification and has been repeatedly demonstrated to correspond well with the response properties of the primate ventral stream (Cadieu et al., 2014; Groen et al., 2018 ; Lindsay, 2020). Because the fixation density maps were entirely spatial in nature, our analyses focused on the spatially organized layers of AlexNet (the convolutional and max pooling layers). We

did not analyze data from the fully connected layers because they have no spatial organization.

2.7 | Representational similarity analysis (RSA)

Representational similarity was calculated in a three-stage process. First, representational similarity matrices (RSMs) were computed from the infant fixation maps for each image, separately for each participant (as in Figure 1b). Second, RSMs were computed from the pattern of activation in AlexNet for each image (as in Figure 1c), separately for each convolutional layer (after the ReLU non-linearity, in line with the approach typically utilized in prior work linking neural networks to brain activity; Storrs et al., 2020) and max pooling layer. Finally, representational similarity was quantified as the rank order correlation between a given infant's RSM and the AlexNet RSM for a given layer. This gave us eight representational similarity estimates for each infant (one for each AlexNet layer). A linear mixed effects model was then used to assess the effects of infant age and AlexNet layer on representational similarity.

For each infant, we computed the fixation map RSM from that infant's fixation density maps (one 1680 × 1050 map for each scene). Each fixation density map was reorganized into a single vector of 1,764,000 values, and the similarity between a given pair of maps was quantified as the Pearson *r* correlation between the vectors for the two maps. This produced an RSM for each infant, in which each cell contained a correlation between the fixation density maps for a pair of scenes (as in Figure 1b).

We computed the AlexNet RSMs from the pattern of activation across all units of a given layer for each of the 22 images. For each layer, the activation pattern for each image was converted into a single vector of activation values, and the Pearson *r* correlation between each pair of vectors was computed. This yielded eight separate 22 × 22 RSMs (one for each of the eight convolutional and max pooling layers), in which each cell contained a correlation between the pattern of activation for a pair of scenes for that layer of AlexNet. RSMs were calculated separately for each infant, and the single-infant RSMs are shown in Figure S2.

Representational similarity was quantified as the Spearman *rho* rank-order correlation between an individual infant's fixation RSM and the AlexNet RSM for a given layer. We had 54 infants and 8 AlexNet layers, so this yielded 54 × 8 representational similarity values.[1] Note that parametric (Pearson *r*) correlations were appropriate when computing the individual RSMs, because the same data types (e.g., fixation density maps) were being correlated with each other at that stage of the analysis. However, rank order correlations were appropriate when quantifying the association between RSMs, because we cannot assume that the RSM for one data type (e.g., fixation density values) will be linearly related to the RSM for a different data type (e.g., AlexNet activations).

The correlation between the fixation map RSMs and the AlexNet RSMs is limited by the reliability of the fixation map data. To account for this, we estimated the *noise ceiling*, which represents the highest Spearman rho that a perfect model would be expected to achieve given the noise in the fixation maps (Nili et al., 2014). The upper bound of the noise ceiling was computed by taking the correlation between a given infant's fixation map RSM and the fixation map RSM of all subjects combined and then averaging the resulting rho values across participants. The lower bound was computed in the same way, except that a given infant's RSM was correlated with the average RSM from all the other infants.

2.8 | Statistical analyses

The statistical analyses were carried out in SAS 9.3 with PROC MIXED using custom effect size macros (Tippey & Longnecker, 2016). The representational similarity (Spearman *rho*) values served as the dependent variables in a fully-crossed random mixed effects model with infant age in days and AlexNet layer as fixed effects, with a random intercept for subjects. Age was entered into the model as a continuous variable (centered at the age of the youngest infant, 118 days). AlexNet layer was entered as a class effect rather than as a continuous variable. A Holm-Bonferroni correction was applied to adjust for multiple comparisons when each layer of AlexNet was analyzed separately, using a family-wise Type I error rate of 0.05. To assess the robustness of the estimated parameters and confidence intervals to violations of parametric model assumptions, the analyses were repeated using a non-parametric bootstrapping approach with 10,000 iterations (Appiah, 2018). All effects that were significant in the parametric analyses were significant in these non-parametric tests, so we report only the parametric results here. The results of the non-parametric tests can be found in the online supplementary materials.

3 | RESULTS

3.1 | Basic properties of the eye movements

A detailed description of the basic eye movement parameters is provided by Pomaranski et al. (in press). To summarize, there were no age-related differences in the quality of data, in the number or duration of fixations, in the saccade length, or in center bias tendencies. However, the between-infant consistency of the eye movement patterns increased with age. That is, the fixation density maps were more idiosyncratic in the younger infants than in the older infants, and the fixation density maps were both more consistent across infants and more adultlike in the older infants (see Pomaranski et al., in press, for details).

[1] Note that not all infants saw all 22 scenes. For an infant who saw *N* scenes, the RSM was an *N* × *N* matrix of the correlations between the fixation density maps for those scenes. Similarly, we extracted the corresponding model correlations to create an *N* × *N* RSM for each layer of AlexNet for use with that infant. Representational similarity was then computed using the *N* × *N* AlexNet and fixation RSMs for the scenes viewed by a given infant. Because representational similarity was computed separately for each infant, this did not create any difficulties in the analysis.

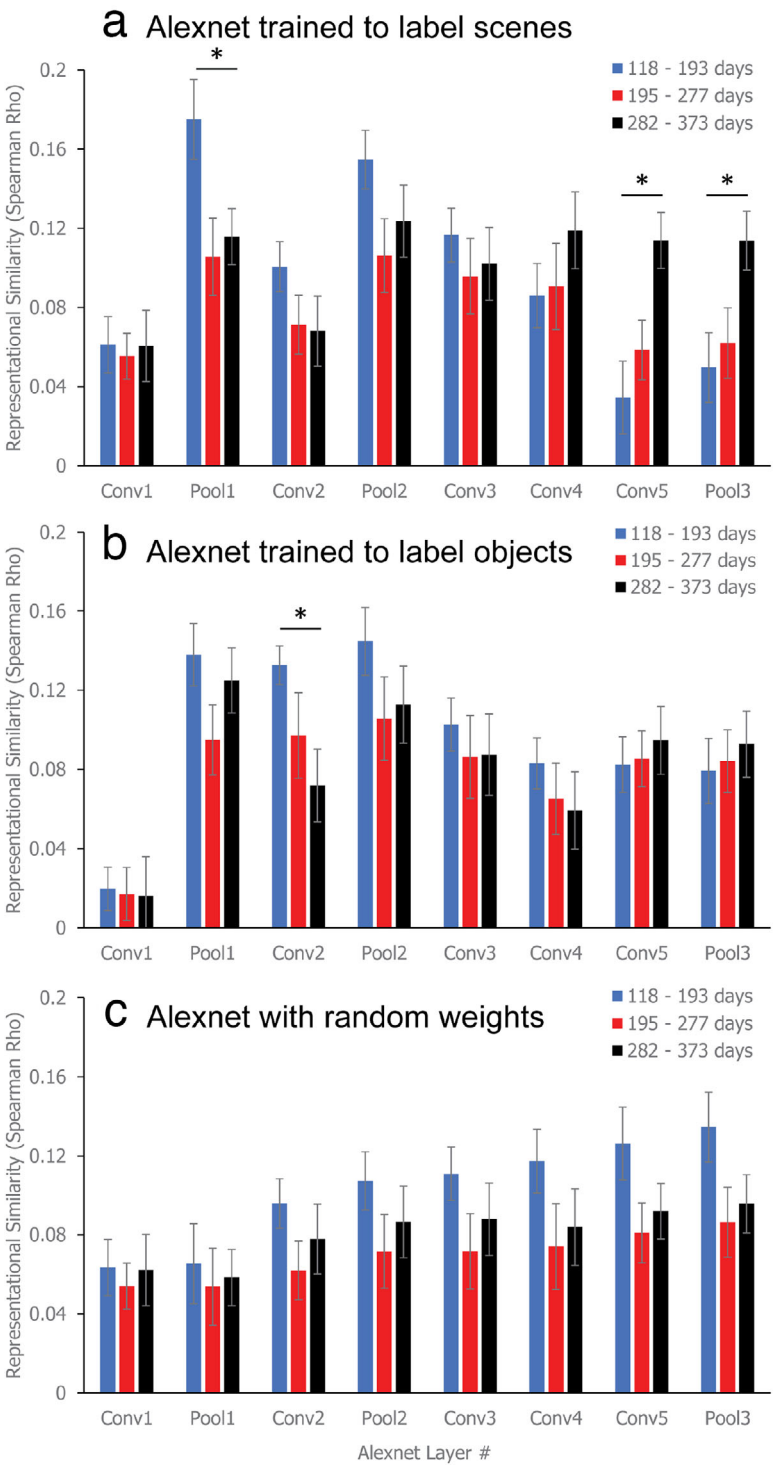

FIGURE 3 Representational similarity (in Spearman rho rank-order correlation units) between each layer of three differently trained AlexNets and the gaze patterns of infants of different age ranges. A separate rho value was calculated for each combination of participant and layer, and then the rho values were averaged across the infants within a given age range for visualization purposes. These age groups were not used in the statistical analyses, in which age was a continuous variable. Separate values are provided for versions of AlexNet that were (a) trained to classify scenes, (b) trained to classify objects, and (c) untrained, with random weights. The noise ceiling (i.e., the highest rho values that could be expected given the noise level in the gaze data) was 0.33–0.38. * = Significant effect of age ($p < 0.05$) for that layer. Error bars denote +/- 1 standard error. All bars save for those in Convolutional Layer 1 in the objects-trained network are significantly different from zero (all adjusted p's < 0.05).

3.2 | Primary representational similarity effects

For purposes of visualization, Figure 3a shows the raw representational similarity between each AlexNet layer and the fixation data, averaged over the infants in each of three equal-sized age groups (terciles: 118–193 days, 195–277 days, and 282–373 days). Note that these arbitrary age ranges were used only for visualization, and age was treated as a continuous variable in all statistical analyses. The figure shows that the mean representational similarity (Spearman rho) values, averaged across infants within an age range, were well above zero for each layer in each age range. These means were between 0.04 and 0.16, which is within the range of representational similarity values observed in many previous ERP and fMRI studies (e.g., Greene & Hansen, 2018; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Storrs et al., 2020). Moreover, these values are reasonable given the noise ceiling, which is the highest representational similarity

value that would be expected given the noise in the fixation data (lower bound = 0.33, upper bound = 0.38).

The statistical analysis showed that the predicted representational similarity level was significantly greater than zero (chance) in each layer for infants at the average age of our sample (241 days; all *p*s < 0.05 after applying the Holm-Bonferroni correction for multiple comparisons). Thus, each layer of the network could predict the gaze patterns better than chance for the average-aged infant. The average representational similarity across all layers was rho = 0.093 (SD = 0.08).

The statistical analysis indicated that the representational similarity pattern varied significantly with age. In general, representational similarity for the lower layers of the scene-trained network was greater for the younger infants than for the older infants, and representational similarity for the higher layers was greater for the older infants than for the younger infants. The mixed effects model yielded no significant main effect of age, $F(1, 52) = 0.09$, $p = 0.767$, $\omega^2 < 0.001$, consistent with the approximately equal levels of overall representational similarity across ages shown in Figure 3a. However, the analysis yielded a significant main effect of layer, $F(7, 364) = 19.07$, $p < 0.0001$, $\omega^2 = 0.240$, indicating that in general the RSMs for some AlexNet layers predicted the fixation RSMs better than others. Most importantly, the interaction between age and layer was statistically significant, $F(7, 364) = 10.83$, $p < 0.0001$, $\omega^2 = 0.131$, consistent with the observation of greater representational similarity for younger infants than for older infants in the lower layers and greater representational similarity for older infants than for younger infants in the higher layers.

Follow-up analyses examined the age effect separately for each layer, with the *p* values adjusted using the Holm-Bonferroni correction for multiple comparisons. Three layers showed a significant effect of age, and Figure 4 shows scatterplots of the correlations for these layers. In Convolutional Layer 2, representational similarity declined significantly as age increased, $t(364) = -3.01$, $p_{adjusted} = 0.020$. Conversely, in Convolutional Layer 5 and Max Pooling layer 3, representational similarity increased significantly as age increased, $t(364) = 3.67$, $p_{adjusted} = 0.002$ and $t(364) = 2.92$, $p_{adjusted} = 0.022$ respectively.

These results indicate that the lower layers of AlexNet (corresponding to lower visual cortical areas) accounted for the patterns of looking better in younger infants than in older infants, whereas the higher layers of AlexNet (corresponding to higher visual cortical areas) predicted the looking patterns better in older infants than in younger infants.

To confirm that our primary results did not reflect an undue influence of outliers, we conducted a formal outlier analysis in which the overall influence of individual cases on the log-likelihood distance of the model was assessed (Schabenberger, 2004). We identified one potential outlier case (the case with the lowest value on the Y axis for the middle panel of Figure 4). Upon exclusion of this case, the effect at Pooling Layer 3 was no longer statistically significant after correction for multiple comparisons ($p_{adjusted} = 0.110$), but there were no substantive changes in any of the other effects. In particular, the interaction between layer and age remained significant, $F(7,357 = 8.25$, $p < 0.0001$, $\omega^2 = 0.101$).

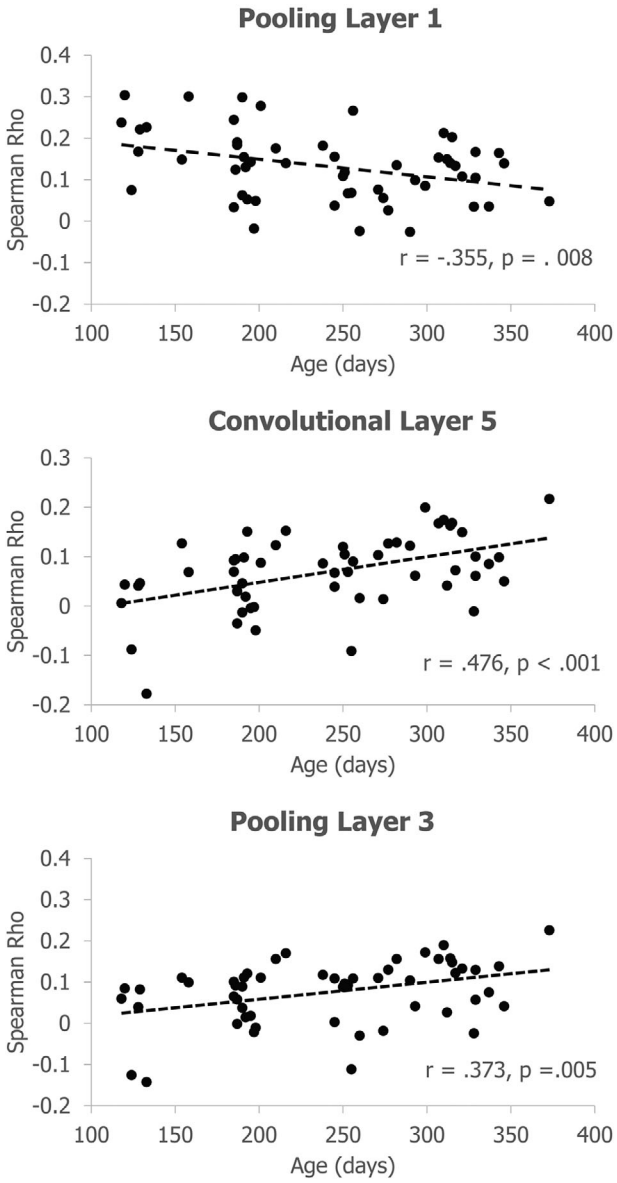

FIGURE 4 Scatterplots of representational similarity (Spearman rho) as a function of age. Each dot indicates the representational similarity between a single infant's gaze pattern and a specific layer of the scenes-trained AlexNet. Only the layers with statistically significant age effects are shown.

3.3 | Effects of network architecture

It is natural to assume that the representational similarity we observed between the infant gaze data and AlexNet depends on the specific features that AlexNet uses to represent the scenes. That is, when AlexNet is trained to classify scenes, it develops Gabor-like receptive field properties in lower layers and develops more complex shape-like representations in higher layers. However, the representational similarity we observed could also be a result of the overall architecture of AlexNet rather than the specific features that are learned during training. This overall architecture—including the hierarchical structure, the increasing receptive field sizes at higher layers, and the use of max pooling

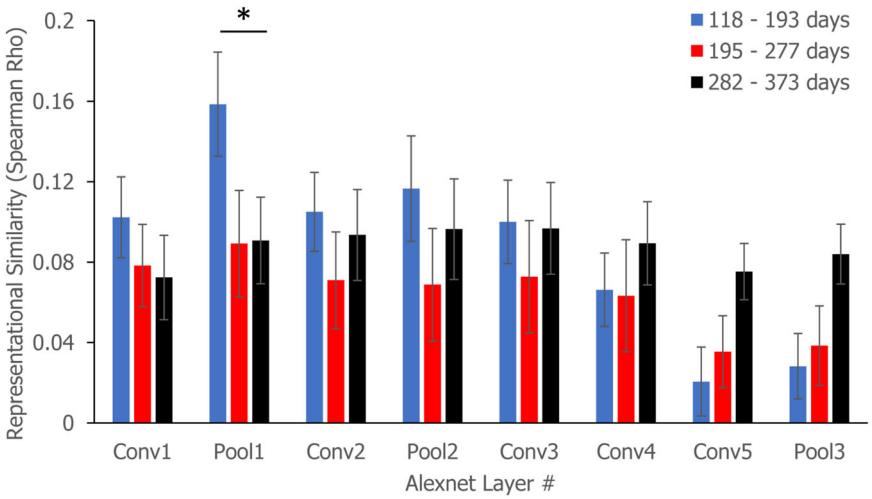

FIGURE 5 Representational similarity after collapsing across the feature channels (using the original scene-trained network trained on scenes). A separate rho value was calculated for each combination of participant and layer, and then the rho values were averaged across the infants within a given age range for visualization purposes. These age groups were not used in the statistical analyses, in which age was a continuous variable. The noise ceiling was 0.33–0.38. * = Significant effect of age ($p < 0.05$) for that layer. Error bars denote +/- 1 standard error. All individual bars are significantly different from zero (all adjusted p's < 0.05)

layers—was defined prior to training and remained constant. Although this architecture is not the result of training, it nevertheless captures some important attributes of the ventral pathway that may be important for visually guided behavior (Rosenfeld & Tsotsos, 2019). Indeed, prior investigations have shown that, in some situations, untrained networks can outperform trained networks with regard to predicting neural representations (Truzzi & Cusack, 2020).

To assess the role of this overall architecture, we asked whether we would obtain similar RSA results with an untrained version of AlexNet (for a similar approach, see Cichy et al., 2016; Herzog et al., 2020). Specifically, we passed the 22 images through an *untrained* version of AlexNet with random weights (Glorot initializer with a uniform distribution, bounds = $+/- \sqrt{\frac{6}{N_o+N_i}}$, where N_o = layer filter size × number of features, N_i = layer filter size × number of input channels). We then computed RSMs for the pattern of activation in each layer. As illustrated in Figure 3b, this random-weight version of AlexNet produced above-zero representational similarity with the infant fixation data, with significantly above-chance values for each layer at the average age (all p-values < 0.05 after adjusting for multiple comparisons). This indicates that the overall architecture of AlexNet, independent of any training, has some ability to predict infant gaze patterns.

However, the strength of the relationship between the untrained network activations and fixation behavior did not vary with infant age overall, $F(1, 52) = 0.16$, $p = 0.694$, $\omega^2 < 0.001$, or within any individual layer for infants at the average age (all adjusted p's > 0.99). Moreover, a direct comparison of the results for the trained network (Figure 3a) and the untrained network (Figure 3b) for each layer revealed that the effect of age differed significantly for Pooling Layer 1 ($p_{adjusted} = 0.034$), Pooling Layer 2 ($p_{adjusted} = 0.020$), Convolutional Layer 4 ($p_{adjusted} = 0.045$), Convolutional Layer 5 ($p_{adjusted} < 0.001$) and Pooling Layer 3 ($p_{adjusted} = 0.022$). These results indicate that the developmental pattern shown in Figure 3a was a result of the visual features that the network learned from the training procedure, and not a function of the training-independent architecture of the network.

3.4 | Effects of features

The fact that the untrained network yielded much lower representational similarity than the trained network suggests that the specific features learned by the network during training are important in explaining the gaze patterns. There are two ways that these features might have an impact. First, similarity in the specific features contained by two scenes might play an important role. For example, the similarity of the looking patterns for two scenes may depend on whether the scenes contain the same colors and shapes. An alternative possibility is that the similarity between scenes per se is not important, but the network must code the kinds of features that the visual system actually represents. That is, as long as the network represents the right kinds of features, the similarity in features between pairs of scenes may play little or no role, and the key factor may be the similarity across scenes in the locations that contain substantial visual information. For example, two scenes might elicit similar looking patterns if they both have an object in the upper left corner, even if this object is red and rectangular in one scene and blue and oval in the other.

To distinguish between these alternatives, we averaged the activation values across features at each location of a given layer, creating a 2D spatial map of averaged activations instead of a 3D map that represents the individual feature values at each location. This 2D map no longer indicates which specific features were present at a given location, but the values at each location indirectly reflect the features that can be detected by the layer. We used these 2D maps to computer the RSMs for each layer.

The results are shown in Figure 5. As in the original analysis, the main effect of age was not significant, $F(1, 52) = 0.01$, $p = 0.904$, $\omega^2 < 0.001$, indicating that the overall levels of representational similarity were similar across ages. However, the main effect of layer ($F(7, 364) = 22.03$, $p < 0.0001$, $\omega^2 = 0.448$) and the interaction between layer and age ($F(7, 364) = 11.01$ $p < 0.0001$, $\omega^2 = 0.213$) were significant for this network, indicating that the magnitude of the representational link was stronger in some layers relative to others and that the pattern varied as a function of age. This is the same qualitative pattern

of effects that was observed in the original analyses (Figure 3a). When the effect of age was analyzed separately for each layer, we found only one layer with a significant effect of age: representational similarity declined significantly with age in Pooling Layer 1, $t(364) = -2.86$, $p_{adjusted} = 0.036$. This contrasts with the original analyses, in which representational similarity also increased significantly with age in Convolutional Layer 5 and Pooling Layer 3.

We also conducted a direct comparison of this feature-collapsed network and the untrained network, including network type as a predictor variable in the statistical model. This analysis revealed that the two networks differed significantly with regard to the relationship between age and layer (network type × age × layer interaction: $F(7,780) = 9.81$, $p < 0.0001$, $\omega^2 = 0.122$). Follow-up analyses showed that the age effect was significantly greater for the feature-collapsed network than for the untrained network in Convolutional Layer 5 ($t(780) = 2.83$, $p_{adjusted} = 0.038$), and marginally significantly greater in Pooling Layer 1 ($t(780) = 2.59$, $p_{adjusted} = 0.069$).

We also conducted a direct comparison of the feature-collapsed (2D) network and the original (3D) network. This analysis revealed that the two networks did not differ significantly with regard to either the overall degree of representational similarity, $F(1,53) = 0.30$, $p = 0.585$, $\omega^2 < 0.001$, or the relationship between age and layer activation levels overall, $F(7,780) = 1.42$, $p = 0.196$, $\omega^2 = 0.003$, as well as within any individual layer (all adjusted p's > 0.99).

The finding that representational similarity was much higher for the collapsed-feature network than for the random network indicates that training of the network is important to ensure that the network is sensitive to the same kinds of features that humans detect. However, the finding that the results were similar for the collapsed-feature network and the original 3D network suggests that gaze is primarily controlled by *where* features are present more than *what* features are present, as long as a given layer is sensitive to the appropriate features. Thus, feature similarity across scenes at a given location does not appear to have a strong effect.

3.5 | Effects of training set

Although the results obtained with random weights and collapsed features indicate that training is important, they do not indicate whether the *specific* training set is important. The version of AlexNet used for the primary analyses was trained to classify complex scenes that each contained many individual objects (e.g., schools, parking lots, grocery stores), similar to the scenes viewed by the infants in this study. However, most research linking AlexNet with visual cortex has used networks that were trained to classify close-up pictures of individual objects (e.g., paper clips, brooms, grasshoppers). We therefore repeated our analyses with a version of AlexNet that was trained to classify objects of this nature, using the 1.2 million images from the ILSVRC2015 ImageNet database.

Prior evidence suggests that both scene-trained and object-trained networks have reasonable representational links to scene-related responses in early visual cortex (Blauch et al., 2019; Chang et al., 2019). Consequently, we expected to find the same effects in the lower layers for the object-trained network that we observed in the scene-trained network. However, object- and scene-trained networks appear to diverge with regard to higher-level ventral stream regions, with significantly weaker representational links for object-trained networks relative to scene-trained networks (Blauch et al., 2019). Thus, the positive age-related correlations we observed in the upper layers of the scene-trained network would not be expected for the object-trained network.

As shown in Figure 3c, this is exactly what we found. That is, we observed relatively equal levels of overall representational similarity across ages for the object-trained network, as was observed for the scene trained network. Consequently, the main effect of age on representational similarity was not significant in the object-trained network, $F(1, 52) = 0.38$, $p = 0.541$, $\omega^2 < 0.001$. However, the main effect of layer ($F(7, 364) = 12.49$, $p < 0.0001$, $\omega^2 = 0.148$) and the interaction between layer and age ($F(7, 364) = 3.63$ $p = 0.0009$, $\omega^2 = 0.034$) were significant for the object-trained network, indicating that the magnitude of the representational link was stronger in some layers relative to others and that this varied as a function of age. In particular, as shown in Figure 3c, Convolutional Layer 2 showed marginally significantly greater representational similarity to younger than to older infants ($t(364) = 2.60$, $p_{adjusted} = 0.078$), similar to Pooling Layer 1 in the scene-trained network (Figure 3a). However, whereas this age effect was reversed at the higher layers of the scene-trained network, there was no significant effect of age for any of the other layers of the object-trained network.

To directly assess the effects of training the network on objects versus scenes, we performed statistical contrasts of the age effects in the scene-trained and object-trained networks for each layer. The effect of age in the scenes-trained network was significantly more negative (i.e., a stronger representation link in younger infants) in Pooling Layer 1 ($t(780) = -3.00$, $p_{adjusted} = 0.017$) and more positive (i.e., a stronger representational link in older infants) in Convolutional Layer 4 ($t(780) = 3.09$, $p_{adjusted} = 0.014$), Convolutional Layer 5 ($t(3.98) = 3.98$, $p_{adjusted} < 0.001$), and Pooling layer 3 ($t(780) = 2.70$, $p_{adjusted} = 0.036$). These results indicate that the specific training experience of the network plays a role in the developmental pattern shown in Figure 3a, especially for the higher layers.

4 | DISCUSSION

Very little is known about how the development of gaze control in infancy is related to the types of representations found in different areas of the ventral stream. The goal of this study was to provide a first step toward filling this gap by asking how the pattern of activation in different layers of AlexNet predicts looking behavior in infants of different ages. We tested the hypothesis that gaze control relies on progressively more abstract representations over development, paralleling the increasing abstraction that occurs as information flows through the ventral stream in adults.

We obtained two key results (see Figure 3a). First, the pattern of activation in the lower layers of the network (which are analogous to

lower-level areas of visual cortex) exhibited greater representational similarity to the fixation density maps of younger relative to older infants. Second, the pattern of activation in the higher layers of the network (which are analogous to higher-level areas of visual cortex) exhibited greater representational similarity to the fixation density maps of older relative to younger infants. This is exactly the pattern that would be expected if gaze control relies on progressively more abstract representations as infants develop. This pattern cannot be explained by noisier or less consistent gaze patterns in the younger infants, which would have led to weaker representational similarity for younger than for older infants in all layers of AlexNet.

We also found that collapsing across the feature dimension produced a reduction in age-related differences in the upper layers of the network but had little to no effect on age-related differences in the lower layers. Training the network on different image classification tasks also impacted age-related differences in the upper but not the lower layers of the networks. These results indicate that *what* features are present at a given location in the higher layers impacts the pattern of looking in an age-dependent manner, whereas for the lower layers the key factor is *where* the features are located and not *what* the features are. However, the features do play some role even in the lower layers, because a completely untrained network had much lower representational similarity to the gaze patterns in all layers relative to the trained network. This presumably indicates that the network must learn what kinds of features are easily detectable by each region of the visual system.

Finally, it is interesting to note that the middle layers of the network exhibit neither greater nor weaker representational similarity as a function of age. Some evidence (Güçlü & van Gerven, 2015; Long et al., 2018) suggests that these layers represent mid-level texture and form information, which may be equally well-represented across the age ranges in our sample.

The present findings are consistent with prior research suggesting a shift over development from gaze being controlled by low-level physical features of the stimulus to gaze being controlled by higher level features. For example, several studies have shown that 3- to 4-month-old infants look first at physically salient regions of visual stimuli, and infants 6 months and older look first at regions of social significance, such as the location of a human face (Frank et al., 2009, 2014; Gliga et al., 2009; Kwon et al., 2016). In addition, when infants view photographs of naturalistic scenes, the proportion of systematic looking attributable to physical salience decreases over the first year (Pomaranski et al., in press).

Note, however, that this shift from salience-based control to higher-level sources of information in prior research is not identical to the shift observed in the present study. First, salience per se is not directly coded by the lower layers of AlexNet. Second, prior work with infants has focused mainly on faces as the higher-level stimuli (Frank et al., 2009, 2014; Kelly et al., 2019; Kwon et al., 2016), whereas the higher layers of AlexNet represent a broad range of abstract features (Wen et al., 2018). Thus, although the present findings are broadly consistent with prior work, they also provide a significant step forward by linking the development of gaze control to the types of features represented across the sequence of ventral stream areas.

The present findings are only a first step toward understanding how specific areas of the ventral stream contribute to gaze control in infants, and there are some important limits to the conclusions that can be drawn. First, our conclusions are based on a network that is broadly similar to the ventral stream rather than actual data from the ventral stream. Moreover, although this network captures key properties of the ventral stream in adults (Cadieu et al., 2014; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Storrs et al., 2020; Yamins et al., 2014), it is not known how well it matches the response properties of the ventral stream in infants. In addition, because AlexNet was trained in a supervised manner with labeled images, it is a model of representations in the mature visual system, not a model of how those representations develop. Thus, we can conclude that changes in gaze patterns over development *can be predicted by* progressively more abstract representations that are *similar* to the changes in representations that occur between lower- and higher-level areas of the adult visual stream; however, we cannot conclude that the present results reflect developmental changes in the ventral stream itself.

A second limitation is that gaze control depends on the dorsal stream and subcortical regions as well as the ventral stream, but AlexNet lacks these other regions. Thus, although we found that the specific features that are learned by this model of the ventral stream can predict age-related changes in gaze patterns, we were unable to assess the role of other brain regions. There is no doubt that these other brain regions develop across the first year, and that there are changes in the role they play in gaze control (Johnson, 1997). A more complete understanding of the control of gaze in infancy will therefore require models that include structures such as the posterior parietal cortex, the frontal eye fields, and the superior colliculus. Furthermore, it will be important for future work using this approach to investigate whether some features (e.g., faces, bodies) play a special role, as has been observed using more traditional methods (e.g., Kwon et al., 2016).

A third limitation is that RSA is inherently correlational, so the observed effects may not reflect a direct causal role of the different representational geometries captured by the different layers of AlexNet. Future research could address this limitation by constructing artificial stimuli that reflect the representations in specific layers of AlexNet (as in Bashivan et al., 2019). These synthetic stimuli could then be experimentally manipulated to see if stimuli corresponding to different layers produce different gaze patterns.

Despite these limitations, the present study makes two significant contributions to the literature. First, it provides new evidence that infants' gaze control becomes increasingly influenced by higher level visual cortical areas between 4 and 12 months. Second, it provides a proof of principle for using RSA to link the development of looking behavior in infancy to computational models of the visual system.

This approach could easily be extended to other computational models or other types of infant behavioral data. As highlighted by Wen et al. (2018), such an approach is particularly promising with regard to hypotheses regarding differences in hierarchical visual processing

when direct neural measures are difficult to obtain, as in developmental populations. The main requirement is that data must be available for a reasonably large number of distinct inputs (such as the 22 images used in the present study). RSA can also be used to link behavioral data with neural data. For example, RSA could be used to ask whether infant gaze patterns can be predicted by ERP, fNIRS, or fMRI data collected from infants or from adults. Moreover, RSA is straightforward to implement, and our analysis code is available at https://osf.io/ehg82/. Thus, RSA is a useful new analytic tool that can be used to address previously intractable questions about infant development.

ACKNOWLEDGMENT

This study was made possible by grants R01EY022525 and R01EY030127 to Lisa M. Oakes.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ETHICS STATEMENT

This study followed the United States Federal Policy for the Protection of Human Subjects and was approved by the UC-Davis Institutional Review Board.

DATA AVAILABILITY STATEMENT

The data and analysis code used in the present study are available at https://osf.io/ehg82/. The data were obtained from a previously published study (Pomaranski et al., in press), and the materials from that study are available at https://osf.io/5j4ht/.

ORCID

Steven J. Luck https://orcid.org/0000-0002-3725-1474

REFERENCES

Amso, D., & Scerif, G. (2015). The attentive brain: Insights from developmental cognitive neuroscience. *Nature Reviews. Neuroscience*, *16*(10), 606–619. https://doi.org/10.1038/nrn4025

Appiah, A. K. (2018). Bootstrap linear mixed-effects models using SAS® procedures. *MWSUG 2018*, *Paper HS-118*, 1–17.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*, 6439. https://doi.org/10.1126/science.aav9436

Batardière, A., Barone, P., Knoblauch, K., Giroud, P., Berland, M., Dumas, A.-M., & Kennedy, H. (2002). Early specification of the hierarchical organization of visual cortical areas in the macaque monkey. *Cerebral Cortex*, *12*(5), 453–465. https://doi.org/10.1093/cercor/12.5.453

Biagi, L., Crespi, S. A., Tosetti, M., & Morrone, M. C. (2015). BOLD response selective to flow-motion in very young infants. *PLOS Biology*, *13*(9), e1002260. https://doi.org/10.1371/journal.pbio.1002260

Blauch, N. M., Peres, F. D. A. B., Farooqui, J., Zar, A. C., Plaut, D., & Behrmann, M. (2019). Assessing the similarity of cortical object and scene representations through cross-validated voxel encoding models. *Journal of Vision*, *19*(10), 188d–188d. https://doi.org/10.1167/19.10.188d

Braddick, O., & Atkinson, J. (2011). Development of human visual function. *Vision Research*, *51*(13), 1588–1609. https://doi.org/10.1016/j.visres.2011.02.018

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Candy, T. R., Crowell, J. A., & Banks, M. S. (1998). Optical, receptoral, and retinal constraints on foveal and peripheral vision in the human neonate. *Vision Research*, *38*(24), 3857–3870. https://doi.org/10.1016/S0042-6989(98)00080-7

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*, 49. https://doi.org/10.1038/s41597-019-0052-3

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755. https://doi.org/10.1038/srep27755

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Colombo, J. (2001). The development of visual attention in infancy. *Annual Review of Psychology*, *52*, 337–367. https://doi.org/10.1146/annurev.psych.52.1.337

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., Kanwisher, N., & Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications*, *8*, 13995. https://doi.org/10.1038/ncomms13995

Ellis, C. T., Skalaban, L. J., Yates, T. S., Bejjanki, V. R., Córdova, N. I., & Turk-Browne, N. B. (2020). Re-imagining fMRI for awake behaving infants. *Nature Communications*, *11*, 4523. https://doi.org/10.1038/s41467-020-18286-y

Frank, M. C., Amso, D., & Johnson, S. P. (2014). Visual search and attention to faces in early infancy. *Journal of Experimental Child Psychology*, *118*. 13–26. https://doi.org/10.1016/j.jecp.2013.08.012

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*(2), 160–170. https://doi.org/10.1016/j.cognition.2008.11.010

Gliga, T., Elsabbagh, M., Andravizou, A., & Johnson, M. (2009). Faces Attract Infants' Attention in Complex Displays. *Infancy*, *14*(5), 550–562. https://doi.org/10.1080/15250000903144199

Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Computational Biology*, *14*(7), e1006327. https://doi.org/10.1371/journal.pcbi.1006327

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *ELife*, *7*, e32962. https://doi.org/10.7554/eLife.32962

Grossmann, T., Johnson, M. H., Lloyd-Fox, S., Blasi, A., Deligianni, F., Elwell, C., & Csibra, G. (2008). Early cortical specialization for face-to-face communication in human infants. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1653), 2803–2811. https://doi.org/10.1098/rspb.2008.0986

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Herzog, S., Tetzlaff, C., & Wörgötter, F. (2020). Evolving artificial neural networks with feedback. *Neural Networks*, *123*, 153–162. https://doi.org/10.1016/j.neunet.2019.12.004

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *Proceedings of the 22nd ACM International Conference on Multimedia*, 675–678. https://doi.org/10.1145/2647868.2654889

Johnson, M. H. (1990). Cortical maturation and the development of visual attention in early infancy. *Journal of Cognitive Neuroscience*, *2*(2), 81–95. https://doi.org/10.1162/jocn.1990.2.2.81

Johnson, M. H. (1997). *Developmental cognitive neuroscience: An introduction*. Blackwell Publishers, Inc.

Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. *MIT Computer Science and Artificial Intelligence Laboratory Technical Reports*. https://dspace.mit.edu/handle/1721.1/68590

Kelly, D. J., Duarte, S., Meary, D., Bindemann, M., & Pascalis, O. (2019). Infants rapidly detect human faces in complex naturalistic visual scenes. *Developmental Science*, *22*(6), e12829. https://doi.org/10.1111/desc.12829

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kiorpes, L. (2016). The puzzle of visual development: Behavior and neural limits. *Journal of Neuroscience*, *36*(45), 11384–11393. https://doi.org/10.1523/JNEUROSCI.2937-16.2016

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In. Pereira F, Burges, C. J. C., . Bottou L, & Weinberger K. Q (eds.) *Advances in Neural Information Processing Systems*, *25*, (pp. 1097–1105). Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Kwon, M.-K., Setoodhenia, M., Baek, J., Luck, S. J., & Oakes, L. M. (2016). The development of visual search in infancy: Attention to faces versus physical salience. *Developmental Psychology*, *52*, 537–555.

Lindsay, G. W. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, *14*. https://doi.org/10.3389/fncom.2020.00029

Lloyd-Fox, S., Blasi, A., Volein, A., Everdell, N., Elwell, C. E., & Johnson, M. H. (2009). Social perception in infancy: A Near Infrared Spectroscopy Study. *Child Development*, *80*(4), 986–999. https://doi.org/10.1111/j.1467-8624.2009.01312.x

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024. https://doi.org/10.1073/pnas.1719616115

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *Plos Computational Biology*, *10*(4), e1003553.

Norcia, A. M., Tyler, C. W., & Hamer, R. D. (1990). Development of contrast sensitivity in the human infant. *Vision Research*, *30*(10), 1475–1486. https://doi.org/10.1016/0042-6989(90)90028-J

Pomaranski, K. I., Hayes, T. R., Kwon, M. K., Henderson, J. M., & Oakes, L. M. (in press). Developmental changes in natural scene viewing in infancy. *Developmental Psychology*.

Rodman, H. R., Scalaidhe, S. P., & Gross, C. G. (1993). Response properties of neurons in temporal cortical visual areas of infant monkeys. *Journal of Neurophysiology*, *70*(3), 1115–1136. https://doi.org/10.1152/jn.1993.70.3.1115

Rodman, H. R., Skelly, J. P., & Gross, C. G. (1991). Stimulus selectivity and state dependence of activity in inferior temporal cortex of infant monkeys. *Proceedings of the National Academy of Sciences*, *88*(17), 7572–7575. https://doi.org/10.1073/pnas.88.17.7572

Rosenfeld, A., & Tsotsos, J. K. (2019). Intriguing Properties of Randomly Weighted Networks: Generalizing While Learning Next to Nothing *16th Conference on Computer and Robot Vision (CRV)*, 9–16. https://doi.org/10.1109/CRV.2019.00010

Schabenberger, O. (2004). *Mixed model influence diagnostics*. SAS Institute Inc.

Storrs, K. R., Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2020). Noise ceiling on the crossvalidated performance of reweighted models of representational dissimilarity: Addendum to Khaligh-Razavi & Kriegeskorte (2014). *BioRxiv*. https://doi.org/10.1101/2020.03.23.003046

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human IT well, after training and fitting. *BioRxiv*. https://doi.org/10.1101/2020.05.07.082743

Tippey, K. . G., & Longnecker, M. T. (2016). An ad hoc method for computing pseudoeffect size for mixed models. *Proceedings of South Central SAS Users Group Forum*. http://www.scsug.org/wp-content/uploads/2016/11/Ad-Hoc-Method-for-Computing-Effect-Size-for-Mixed-Models_PROCEEDINGS-UPDATE-1.pdf

Truzzi, A., & Cusack, R. (2020). Convolutional neural networks as a model of visual activity in the brain: Greater contribution of architecture than learned weights. *Bridging AI and Cognitive Science*. ICLR 2020. La Jolla, CA: International Conference on Learning Representations.

Wen, H., Shi, J., Chen, W., & Liu, Z. (2018). Deep Residual Network Predicts Cortical Representation and Organization of Visual Features for Rapid Categorization. *Scientific Reports*, *8*(1), 3752. https://doi.org/10.1038/s41598-018-22160-9

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kiat, J. E., Luck, S. J. , Beckner, A. G., Hayes, T. R., Pomaranski, K. I., Henderson, J. M., & Oakes, L. M. (2021). Linking Patterns of Infant Eye Movements to a Neural Network Model of the Ventral Stream Using Representational Similarity Analysis. *Developmental Science*, e13155. https://doi.org/10.1111/desc.13155