Winter 2020

A Deep Learning Approach to Understanding Real-World Scene Perception in Autism

Erica Lindsey Busch
erica.l.busch.20@dartmouth.edu

A Deep Learning Approach to Understanding Real-World Scene Perception in Autism

Erica Lindsey Busch

Senior Honors Thesis in Cognitive Science, Dartmouth College

March 3, 2020

Advisors: Dr. Caroline Robertson (Dartmouth College, Department of Psychological and Brain Sciences) & Dr. Leyla Isik (Johns Hopkins University, Cognitive Science)

Contents

1. Abstract

Autism is a multifaceted neurodevelopmental condition. Around 90% of individuals with autism experience sensory sensitivities, particularly impacting visual perception. Despite this high percentage, previous studies investigating visual perception in autism impose severe limitations on our understanding. In many of these experiments, their stimuli and experimental methods are un-naturalistic and produce unreproducible and conflicting results. In this study, we investigate the nature of the real-world visual experience in autism with a cutting-edge experimental approach. First, we use virtual reality headsets with eye-trackers to measure gaze behavior while individuals freely explore real-world, everyday scenes. Then, we compare their gaze behavior to the representations within convolutional neural networks (CNNs), a class of computational models resemblant of the primate visual system. This allows us to model the stages of the visual processing hierarchy that could account for differences in visual processing between individuals with and without autism. To our knowledge, this is the first fully unbiased, data-driven approach to studying naturalistic visual behavior in autism. In brief, we found that convolutional neural networks, regardless of the task upon which they were trained, are better able to predict gaze behavior in typically developing controls than in individuals with autism. This suggests that differences in gaze behavior between the two groups are not principally driven by the semantically-meaningful features within a scene and emerge from differences earlier in visual processing.

2. Introduction

2.1 Characteristics of autism spectrum conditions

Autism spectrum condition (henceforth ASC or autism) is a complex neurodevelopmental condition affecting one out of fifty-nine individuals in the United States (CDC, 2019). Despite autism's prevalence, we know relatively little about its underlying neurobiology. The current literature on the possible neural underpinnings of autism is rife with inconsistencies, as well as the rates of diagnosis worldwide -- diagnosis rates range from 1 in 27 in Hong Kong to 1 in 3,333 in Poland (Elsabbagh et al., 2012).

On average, individuals are diagnosed with autism at age four in the U.S., although signs of autism are often visible by nine months. Sensory processing differences are particularly notable early in development and predictive of later autism diagnosis (Baranek et al., 2013; Estes et al., 2015), and thus may serve as promising early markers of the condition (Robertson & Baron-Cohen, 2017). In adults, atypical sensory perception is reported to impact over 90% of individuals with autism (Tavassoli et al., 2014), and as of 2014, sensory reactivity is included in the DSM-5 criteria for ASC. This new diagnostic criterion emphasizes the important role sensory perception and processing plays in ASC, and recent studies focus on sensory atypicalities as characteristic features of autism's neurobiology (Robertson & Baron-Cohen, 2017). Such sensory atypicalities include visual, tactile, taste, gustatory, and auditory sensitivities significantly different from control populations and populations with other clinical conditions (fragile X syndrome and developmental disabilities of mixed etiology) that appear before three years of age and persist through adulthood (Kern et al., 2006; Rogers et al., 2003).

2.2 Visual perception in autism

A common characterization of perception in autism is as emphasizing local details at the expense of the global percept, i.e. 'Seeing the trees, but not the forest' (Dakin & Frith, 2005).

In other words, they are exceptionally attentive to visual details rather than global features, which (in a subgroup of individuals with autism) is linked to savant-like drawing abilities (Baron-Cohen et al., 2009; Mottron & Belleville, 1993). Numerous studies have shown that individuals with autism are faster at detecting visual targets among distractors ('a tree within the forest'), and eye-tracker studies have shown this as an early behavioral marker of the condition in toddlers (Gliga et al., 2015; Kaldy et al., 2011; Plaisted et al., 1998). Yet, little is known about how such a detailed perceptual style manifests in real-world environments. Moreover, visual processing in autism has rarely been explored using computational approaches.

One previous study has attempted to use machine learning approaches to characterize visual processing in autism in real-world scene images. Eye-tracking studies have shown that individuals with ASC freely viewing complex, naturalistic scenes demonstrate more pixel-level saliency than controls. They also demonstrate decreased saliency for semantic-level features, as well as faces and locations considered socially-meaningful by controls (Wang et al., 2015). For example, individuals with autism viewing naturalistic scenes show preference to regions of scenes with high contrast and color instead of regions with faces, text, or other semantically informative features (Robertson & Baron-Cohen, 2017). Comparatively low responsivity to social stimuli (and nonsocial to a lesser effect) is predictive of later autism diagnosis risk for children at as young as 11 months (Baranek et al., 2013), and decreased social attention is a hallmark of ASC behavior.

These behavioral results suggest that visual perception differences in ASC originate from differences in early visual processing. Neuroimaging studies support the conclusion that atypicalities in visual perception are linked to atypical responses in primary sensory cortices. For instance, individuals with ASC have difficulty tracking the global motion of multiple objects. This could result from atypical responses in early visual cortex (Robertson et al., 2014) and primary motion area (MT) (Herrington et al., 2007; Peiker et al., 2015; Takarae et

al., 2014), as shown in numerous functional magnetic resonance imaging (fMRI) studies of motion processing. Studies have also linked motion perception abnormalities with superior temporal sulcus (STS) (Dakin & Frith, 2005), a region also implicated in perceiving and understanding social interactions (Isik et al., 2017).

2.3 Naturalistic perception

A standard measure of visual attention allocation, gaze behavior has been widely suggested as a 'behavioral marker' of autism. A key feature of these studies, though, is how little they resemble the real-world visual experience. Most of the key studies in autism vision research utilize stimuli like Gabors, line drawings, basic block patterns, and moving dots (Simmons et al., 2009; Spencer & O'Brien, 2006; among many others). One study (Wang et al., 2015) presented real-world scenes as stimuli to measure gaze behavior, but such stimuli were also presented as still images displayed on a computer screen. Most studies using naturalistic stimuli have focused on social salience or interpreting social interaction in autism, again treating autism as a condition of the social mind primarily and sensory processing secondarily (Simmons et al., 2009).

Using naturalistic stimuli in psychological studies (neuroimaging and behavioral studies alike) generates a large mass of noisy data that researchers need to sift through to draw meaningful conclusions about cognition. Computational models are powerful tools for making sense of such data. Modern artificial deep neural networks (DNNs) are biologically-inspired models that can solve some of the cognitive tasks once thought unique to humans. These models are built of layers of simple processing units (like neurons) that work in parallel and communicate in feed-forward and feed-backward projections. They are trained in order to complete a specific task, like predicting market fluctuations, natural language processing, or image recognition. This training allows a model to learn a specific set of connection weights so it can make useful predictions on unseen data (Cichy & Kaiser, 2019).

Deep neural networks were initially built as engineering tools yet quickly infiltrated other fields. A particular form of DNNs, convolutional neural networks (CNNs) have received particular attention for their outstanding performance on computer vision tasks. Since Krizhevsky et al. introduced their award-winning image recognition model in 2012, CNNs have dominated the computer vision field (Krizhevsky et al., 2012). In the following eight years, CNNs have surpassed all other computational models and even human performance at visual recognition tasks (Kriegeskorte, 2015). Beyond the computer vision field, DNNs perform remarkably in other domains, including speech recognition and translation (Sak et al., 2014) and models of brain activity.

2.4 Deep neural networks and the brain

Perhaps naturally given their biological inspiration, deep neural networks have tremendously benefitted neuroscience research. DNNs are able to predict both behavior and neural responses, informing the link between the two. CNNs, known for their visual recognition prowess, strongly resemble the brain responses in primate primary sensory cortices. Khaligh-Razavi & Kriegeskorte (2014) showed that a CNN, among a pool of 37 models, best predicted inferior temporal (IT) responses to visually-presented objects. Though several models trained on low-level features could predict early visual cortex responses, later levels of the CNN model (which are notably more representative of visuo-semantic features) exceeded all other models in predictivity of higher-order visual areas (including FFA, PPA, and LOC). Overall, when trained on large sets of labeled images, supervised CNNs best explained IT data, and specifically the later layers of the network best predicted responses in both monkey and human IT. The network's final spatially-selective layer is the most highly predictive of neural responses of all their investigated models, and each layer of the model's hierarchy increases roughly monotonically in its ability to predict IT responses (Khaligh-Razavi & Kriegeskorte, 2014).

Since 2014, many studies have supported the findings in Khaligh-Razavi and Kriegeskorte: CNNs are highly predictive of visual cortex responses, with early layers better predicting early visual cortex and later layers better predicting higher-level visual areas (Cichy & Kaiser, 2019; Isik et al., 2014; Kriegeskorte, 2015). Units in early layers of CNN have small receptive fields that are most responsive to features like edges, analogous single neurons in the early visual cortex (Hubel & Wiesel, 1962). The activity within the early layers of a CNN is more general and less task-specific, and representations increase in task-specificity in later layers. Units further along the ventral pathway have larger receptive fields and are more transformationally and translationally invariant and selective to particular shapes and semantic categories (Güçlü & van Gerven, 2015; Hung et al., 2005; Yamins et al., 2014), similar to IT cortex responses (Khaligh-Razavi & Kriegeskorte, 2014). This brain-model analogy has proven a useful framework: CNNs are now an integral feature of models that can predict visual stimuli from observed brain (BOLD) activity (Güçlü & van Gerven, 2015).

2.5 Convolutional neural networks, the brain, and gaze behavior

Unsurprisingly, CNNs are also highly predictive of visual behavioral responses. The MIT saliency benchmark presents models that are most predictive of human gaze behavior on a dataset of natural images with eye-tracking data. To date, the ten top models of the MIT saliency benchmark are all CNNs (O'Connell & Chun, 2018; MIT Saliency Benchmark). DeepGaze, a prominent saliency estimation model, utilizes prominent CNN architectures trained on object recognition. Specifically, it uses the landmark VGG architecture that revolutionized computer vision in 2015 (Simonyan & Zisserman, 2015). The newest edition, DeepGazeII can predict 87% of patterns of fixation and outperforms all other models on the saliency benchmark. Evidently, though the information represented in CNN activity is dependent upon the task the network was trained to predict, the features that a CNN learns to represent during training are transferable between tasks. This supports the notion that CNNs learn flexible feature space for an array of objectives: one can use image recognition models

to predict human gaze behavior (Kümmerer et al., 2016) and scene recognition models to predict neural responses (O'Connell & Chun, 2018).

As shown, convolutional neural networks can predict both gaze behavior and neural responses. Can they model the link between brain activity and behavior? In an elegant series of experiments, O'Connell and Chun (2018) showed a novel three-way connection between CNNs, neural responses, and gaze behavior. First, they used brain activity (BOLD responses) to directly predict gaze behavior. Then, they translated these neural responses into CNN activity patterns at specific layers of a VGG model trained to recognize scenes. These activity patterns were used to build spatial priority, or saliency, maps, indicating the most salient regions of a scene. Spatial priority maps were then used to reconstruct fixation maps in novel scenes, and they found that these CNN-reconstructed fixation maps could predict human fixation patterns. By extracting CNN's representations of scenes, they showed a novel link between visual cortex activity, CNN activity, and gaze behavior (O'Connell & Chun, 2018). Despite the exciting potential for CNNs to shed light on the real-world human visual experience, to our knowledge they have not yet been applied to 360° gaze behavior. Moreover, they have never been used to characterize visual processing in autism.

2.6 A novel approach: CNNs to understand visual processing in autism

In this study, we use convolutional neural networks to investigate gaze behavior of individuals with autism in naturalistic scenes. To do this, we designed a novel experimental approach. First, we used in-headset eye-tracking in immersive virtual reality headsets to measure spatial attention allocation in real-world, complex scenes. This affords objective insight into the day-to-day visual experiences of individuals with autism, which we compare with typically-developing controls. Second, we used CNNs to model where along the visual hierarchy scene perception and processing diverges between groups, and whether this relates to the model's training. Specifically, we utilized the hierarchical representations of scenes and objects within convolutional neural networks to directly model spatial attention

allocation across low-level visual features (e.g. color, pixel-saliency, or contrast) and high-level ones (e.g. socially or non-socially meaningful features).

Through modeling gaze behavior with CNNs, we asked a series of questions about how individuals experience the visual world. Does visual attention allocation in natural scenes vary between ASC and typically-developing controls? Are between-group differences in visual behavior predicted by different layers along the CNN hierarchy, since such layers represent increasing semanticity? Are CNN models trained for different visual recognition tasks predictive of group differences? If our models indicate group-level differences, it could inform how visual perception characterizes autism. Furthermore, this could afford researchers better insight into where along the visual hierarchy differences in visual information representations originate.

In brief, we found that CNNs overall predicted gaze behavior better for controls than for individuals with ASC. This group difference was not affected by location along the network hierarchy-- for all individuals, the later layers of the network better predicted gaze behavior than the earlier ones, and all layers better predicted controls than individuals with autism. We also found that a model pretrained on object recognition significantly predicted gaze behavior better than scene recognition in both individuals with and without autism, and it better predicted control gaze behavior than ASC. This suggests that differences in gaze behavior between the two groups are not tied to high-level representation of objects or places and begin early in visual processing.

3. Materials and Methods

3.1 Behavioral data

Behavioral data was collected from participants according to the following specifications. We will refer to this as ‘gaze data’ throughout the course of this thesis.

3.1.1 Participants.

Forty-one adults participated in this experiment (20 ASC). All participants were recruited from the local Upper Valley (NH/VT) community. Control participants (12 female; mean age 22.38 +/- 4.84 STD years) were included based upon 1) having normal or corrected vision and no colorblindness, 2) having no neurological or psychiatric conditions, and 3) having no history of epilepsy.

Twenty participants (8 female, 1 gender unspecified; mean age 23.4 +/- 7.19 STD years) had documented autism spectrum condition (ASC) diagnoses, confirmed with the Autism Diagnostic Observation Schedule Second Edition (ADOS-2) Module 4 assessment administered by a research-reliable administrator (Hus & Lord, 2014). ASC participants all had normal or corrected vision and no colorblindness. Fifty percent of ASC participants self-reported co-occurring conditions, including anxiety, depression, attention-deficit/hyperactivity disorder, and dyslexia. These co-occurrences were not controlled for or matched in the control group. An additional four participants with ASC attempted the experiment but were later excluded from analyses due to one of three reasons: task comprehension difficulty, insufficient diagnostic confirmation, or contributing fewer than 30 valid scenes.

All participants completed the Kaufman Brief Intelligence Test (Kaufman & Kaufman, 2014) and the Autism Spectrum Quotient (Baron-Cohen et al., 2001). Control participants were matched for age with participants with ASC. An additional eight individuals participated in a

pilot experiment to identify stimuli balanced for social/nonsocial salience (for more information, see section 3.1.2). Written consent was obtained from all participants in accordance with a protocol approved by the Dartmouth College Institutional Review Board.

3.1.2 Stimulus and head-mounted display.

Stimuli consisted of 360-degree "photospheres" of real-world scenes sourced from open online databases such as Flickr (flickr.com) or Youtube (youtube.com). Photospheres depicted a diverse set of indoor and outdoor settings and contained 1-3 people and non-social yet interesting objects. Pilot participant data identified a set of 60 photospheres balanced for both salient social and non-social content. Such balanced scenes were defined as ones where the top 50% of pilot participants' gaze heat is distributed across both socially meaningful features (i.e. faces or bodies) and nonsocial yet identifiable and interesting features (i.e. televisions or trees). Each photosphere was then applied to a virtual environment built in Unity version 2018.3.11f1 (unity3d.com) then integrated with a head-mounted display (Oculus Rift Development Kit 2, oculus.com, low persistence OLED screen, 2K resolution per eye, ~90 degree field of view, 75 Hz refresh).

3.1.3 Eye tracking specifications.

Two in-headset binocular eye-trackers monitored participants' gaze continuously during scene viewing (Pupil Labs version 1.9.7, 120 Hz sampling frequency, 0.6 visual degrees accuracy, 0.08 visual degrees precision, 5.7ms camera latency, 3ms processing latency). Custom scripts written in C# for Unity were used to record eye movements.

3.1.4 Experimental procedures.

During each experimental trial, participants were presented with a photosphere via the head-mounted display for twenty seconds. Each participant had the opportunity to view all 60 photospheres. The number of trials each participant actually completed varied according to participant time restraints or fatigue. On average, control participants completed 59 trials (+/-

1.8 STD trials) whereas ASC participants completed 59 trials (+/- 3.9 STD trials). Experimental trials with insufficient or low-confidence data were excluded according to the preprocessing steps (see section 3.1.6 for pre-processing details).

During viewing, participants were told to "look around each place just like you would look around a new place in real life. Pretend like you'll have to describe that place later to someone who didn't see it." Participants were given a break after every ten scenes, at which point the eye-tracker was recalibrated. Participants stood while wearing the head-mounted display and actively explored the photosphere via self-directed eye-movements and head turns. This provided an opportunity to explore the naturalistic environment from an egocentric perspective (Figure 1).

3.1.5 Practice trials and calibration routine.

The experiment had three phases: practice, calibration, and experimental trials. During the practice phase, each participant saw two scenes that were not included in analysis. They were reminded to move their heads and explore the whole scene. Practice phases ensured that participants had acclimated to virtual reality environments before beginning the experiment. Then, participants performed a 21-point calibration routine (approximately one minute) to validate eye-tracking accuracy. This calibration routine was repeated after every 10 experimental trials.

After each trial of the experimental phase, participants returned to a virtual home screen where they took a five-second break before the next trial. After leaving the home screen, participants saw a pre-trial fixation screen with a visual target at center screen. If significant gaze drift (>5 degrees visual angle) was detected at this time, the calibration routine was repeated. Re-calibration also occurred after every time a participant removed the headset.

3.1.6 Data preprocessing.

Gaze data was filtered and excluded from analysis for one of three reasons. First, we filtered for eye tracker confidence. If, for any time point in the trial, eye tracker confidence fell below 50%, we exclude that data. If we excluded more than 75% of the time points for a trial, we exclude the entire trial. Next, we filtered for adequate scene exploration. For their trial to be included in our analysis, the participant must have explored at least 60% of the scene's yaw with confident eye tracking. This ensures that the participant understood the task and actively explored the scene. Finally, we thresholded for pretrial calibration check failures. Before exploring a scene, participants fixated on a target at the scene's center and we calculated the eye tracker's drift away from their gaze so we can correct for this drift at other time points in the trial. If drift exceeded 10 degrees visual angle, we excluded that trial from analyses.

To determine fixations, we calculated the orthodromic distance and velocity between consecutive gaze points. We calculated the mean absolute deviation (MAD) in gaze position using a seven-sample sliding window of ~80ms (Voloh et al., 2019). Windows with a MAD less than 50°/s were defined as potential fixations (Peterson et al., 2016). If two group centroids were displaced by under 1° and two potential fixations occured within 150 ms, the potential fixations were concatenated. We excluded fixations that were shorter than 100ms (Peterson et al., 2016; Wass et al., 2013). This fixation routine was previously defined in Haskins et al. (in preparation).

Figure 1: Experimental design.

Participants wore immersive virtual reality headsets equipped with binocular eye trackers. On each experimental trial, participants freely explored immersive, naturalistic environments with self-generated movements (saccades and head turns).

3.2 CNN model data.

3.2.1 CNN architecture.

To model spatial attention allocation, we used VGG16, a convolutional neural network (CNN) with a deep, feed-forward architecture (Simonyan & Zisserman, 2015). This architecture consists of 18 spatially selective layers broken into five blocks of convolution operations followed by non-linear max-pooling operations. Convolution layers are made of 64 to 512, 3 by 3 filters, which slide across the activation volumes in a block with a stride of 1. The max-pooling operations use a 2 by 2 filter with a stride of two. By downsampling less relevant features and propagating forward more salient ones, the network can reduce the spatial size of its representation by half and build translational invariance in its representation (Figure 2C).

3.2.2 CNN training.

Along this processing hierarchy, the network moves from detection of low-level features (e.g. edges and contrast) to high-level, semantically meaningful concepts. A network's 'concept' is relative to the task it was trained to complete. Thus, the later layers of a network trained on scene recognition might represent concepts like roads or churches, whereas a network trained on object recognition might represent concepts like dogs or toasters.

In order to investigate the differences in task performed by participants upon exploring a novel scene, we modeled their gaze behavior in experimental trials using convolutional neural networks pretrained on two different tasks: scene recognition (Places365, (Zhou et al., 2016)) and object recognition(ImageNet, (Deng et al., 2009)). Theoretically, if individuals perform a task more similar to scene recognition in one of our experimental trials, attending to more global features of a scene, their gaze maps will resemble the top pooling-layer activation of a CNN model trained to recognize scenes. If they perform a task more focused on local details (such as the objects embedded within a scene), their gaze maps will resemble the top pooling-layer CNN map of a model trained to recognize objects. In this study, we compare models trained on two different tasks (object recognition and scene recognition) to test the local versus global attention used by individuals with and without autism when exploring novel scenes.

Both models utilize the architecture of VGG16 models with hidden layer weights optimized to categorize scenes into one of 365 categories (e.g. ice field, forest, office, ice cream shop). The network gradually builds up a representation of the input image as it attempts to categorize it utilizing relevant features it has 'learned' from training. Thus, each layer 'attends' to different identifying features of the scene, and the activations in the final layer of the network show the features most useful for classifying the scene.

The hidden layer weights learned by VGG16-Places365 are available from the Places365 GitHub (*CSAILVision/places365*, 2016/2020) as a Caffe model. We converted this Caffe

model to a Torch model (Collobert et al., 2011) using the LoadCaffe module (Zagoruyko, 2014/2020) and from Torch to Pytorch using the convert_torch_to_pytorch module (Carwin, 2017/2020). After conversions, we interfaced with the model using Pytorch (Paszke et al., 2019). The hidden layer weights learned by VGG16-ImageNet are available as a pretrained model in Pytorch (Paszke et al., 2019).

Our first model, VGG16-Places365, was trained on a scene recognition task. Our second model, VGG16-ImageNet, was trained on an object recognition task. Though they utilize the same architecture, they are trained to attend to different visual features. VGG16-Places365 attends more to global features of a scene, such as affordances and terrain. VGG16-ImageNet attends more to local features embedded within a scene, such as specific items or beings. Because these two networks are trained to perform different tasks, we can use them to quantitatively model local versus global attention during scene exploration.

Within each of our models (VGG16-Places365 and VGG16-Places365), the layers along the hierarchy can be regarded as a feature map highlighting the most salient features within a scene for that layer. We can compare feature maps from the same layer (say, pool-5 VGG16-Places365 to pool-5 VGG16-Places365) because the models share identical architectures. Visualizing the activations of these feature maps as the CNN processes an image illustrates the network building its conceptual representation of an image, from low-level to high-level features. In later analyses, we capitalize upon this gradual representation construction to investigate group differences in representation of scene features between ASC and typically-developing individuals.

3.2.3 Panoramic image feature map extraction.

A unique feature of our study is the gaze data collected from individuals freely exploring novel photospheres of real-world scenes in virtual reality. Because our scenes are spherical, though, we need to break them down into square ‘viewport’ images in order to a) account for

the equirectangular distortion of projecting a spherical scene into Cartesian space, and b) feed the network an image of its preferred input size without losing spatial resolution.

We developed a novel pipeline for extracting CNN activity maps in response to a photosphere and comparing that activation to behavioral gaze data. First, we sample 500 (x,y) points on a sphere such that points are sampled with greatest density around the equator and with decreasing density as moving away from the equator. Individual gaze behavior has been shown to be strongly biased toward the equator of images, with decreasing fixations at increasing latitudes above and below the equator (Judd et al., 2009; Sitzmann et al., 2016). Sampling accounts for equatorial bias by upsampling and downsampling our photospheres accordingly (Figure 2A&2B).

We converted these points (in radians) to Cartesian coordinates and used the Equirec2Perspec module (Fu-En.Wang, 2017/2020) to build a square viewport centered on each point. Each viewport is a 224 by 224-pixel view of the photosphere, which accounts for 90 degrees of visual angle, approximately equivalent to that of the Oculus Rift used in our experimental trials. These are large enough to capture a meaningful portion of the image contained in one field of view while avoiding equirectangular distortion (Figure 2A). The VGG16 architecture requires input volumes of size 224 by 224 x 3 (pixels x pixels x channels), but by sampling our sphere rather than simply resizing our flattened photosphere (a panoramic image of 1000 by 2000 pixels), we can feed the model an image comparable to what one participant views in one fixation within an experimental trial, without distortion. With 500 samples of 224 by 224 pixels, viewports overlapped with one another heavily (moreso around the equator, where we sample most densely). This let us average the network's activations at a given pixel when considered in numerous contexts so we could infer how the network responds to each pixel in relation to the entire photosphere.

After sampling the photosphere into viewports and propagating viewport volumes forward through the VGG16-Places365 and Object-CNN models, we extract each model's pooling

layer activations (Figure 2C) and project the viewport's activation back outward into equirectangular space using Perspec2Equirec (Fu-En.Wang, 2017/2020). We repeat this process for all 500 viewports until we have generated one CNN activity heatmap representing the activation of each desired layer in response to the whole photosphere. For the pixels in which our viewport samples overlap, within each model we average over all aggregated activation values at that pixel in order to obtain an average activation value for each pixel.

For our task-differentiation analysis (local/global), this process only involved extracting layer activations for the last pooling layer of the VGG16-Places365 and Object-CNN (directly before the softmax). For our layer analysis, the process involved extracting layer activations at each of the five pooling layers along the VGG16-Places365 hierarchy.

3.3 Comparing behavioral and model data

3.3.1 Hierarchical layer analysis

To determine which layers of a CNN trained to recognize scenes are most predictive of an individual's gaze behavior, we extracted activations from each pooling layer of VGG16-Places365 (VGG16-Places365) in response to each of our scenes. We followed our panoramic image feature map extraction pipeline (see 3.2.3) in order to build a CNN map for each of the VGG16-Places365's five pooling layers. For each of our 60 photospheres explored by participants, this yielded 5 feature maps (CNN Layer Maps), which we then smoothed using a variable-width Gaussian kernel (base filter width of 12 pixels) and z-scored across all values.

Each of our layer maps was normalized to a common scale using histogram matching (Henderson & Hayes, 2018). We averaged the five CNN Layer Maps to create a reference heatmap for each individual layer. Then, each of these maps was histogram matched to the

average reference heatmap using the MATLAB Image Processing Toolbox function *imhistmatch* (*MATLAB imhistmatch.*; *MATLAB 2019b*).

Once all CNN Layer Maps were normalized to this common scale, we used partial correlations to compare our behavioral gaze data and our model data. For each participant and each experimental trial, we z-scored and sampled their gaze map within a photosphere and the CNN Layer Maps in that scene at 500 sampling coordinates (to control for equirectangular distortion; see Figure 2B). We then correlated the gaze map with each of the five normalized CNN Layer Maps, controlling for the the scene's equator. Individual viewing behavior tends to demonstrate a strong bias toward the equator of images with decreasing fixations at latitudes above and below the equator (Judd et al., 2009; Sitzmann et al., 2017), so by partitioning the variance attributed to this equatorial bias, we controlled for individual behavioral adaptation to our task to instead focus exclusively on individual scene exploration (Groen et al., 2018) . This resulted in five correlation values for each experimental trial.

For each participant, we averaged over each experimental trial's layer correlation value to obtain an average correlation of that participant's gaze behavior with each pooling layer of the VGG16-Places365. This score indicates how predictive a given layer is of that individual's gaze behavior across scenes. After repeating this process for each participant, we averaged over all participants within each group (based on ASC diagnosis). This resulted in two sets of five correlation values: one set for how predictive each layer is of average gaze behavior for individuals with an ASC diagnosis and one set for those without an ASC diagnosis.

3.3.2 Local versus global attention analysis

To determine whether individuals use more local or global attention when exploring a novel scene, we compared gaze maps with CNNs trained to perform object recognition (VGG16-ImageNet; Object CNN) versus scene recognition (VGG16-Places365; Scene CNN). We compared individual gaze maps from each experimental trial to the top-pooling

layer activation generated by the Object CNN and the Scene CNN in response to that trial's scene. To do this, we extracted just the final max-pooling layer activation, resulting in one CNN map per model (Object-CNN Map and VGG16-Places365 Map) for each of our photospheres.

Each CNN map was normalized to a common scale using histogram matching. We averaged the Object-CNN Map and the VGG16-Places365 map to create a reference heatmap, and each of our target CNN maps was histogram matched to this average reference heatmap using the MATLAB Image Processing Toolbox function *imhistmatch* (*MATLAB imhistmatch*, n.d.; *MATLAB 2019b*, n.d.; Henderson & Hayes, 2018). By averaging the heat from the two models, we bring the two model maps into a common, middle-ground heat distribution for comparison.

After normalizing a scene's Object-CNN Map and VGG16-Places365 Map to an average heat scale, we used partial correlations to compare each model's CNN Map with participant's gaze map. Within the partial correlation, we controlled for the contribution of the opposite model's heat and for equator bias (for explanation of equator bias, see section 3.2.3). For example, when correlating an Object-CNN Map and a gaze map, we would sample both maps at each sample coordinate and then vectorize them. We would correlate these two vectors (using Pearson's correlation) while controlling for the equator and the VGG16-Places365 Map (also vectorized and sampled at the same sample coordinates) and vice-versa for the VGG16-Places365 Map.

We repeated this process for each participant and each experimental trial, then averaged the correlations over all of a participant's trials. This gave us two average scores per participant: the participant's average correlation of gaze behavior with the VGG16-Places365 and the Object-CNN. We then averaged these scores across participants within diagnosis group (ASC versus typically developing) to obtain four scores: the average correlation of gaze behavior in

individuals with ASC and without ASC with the top-level activation of the VGG16-Places365 and Object-CNN.

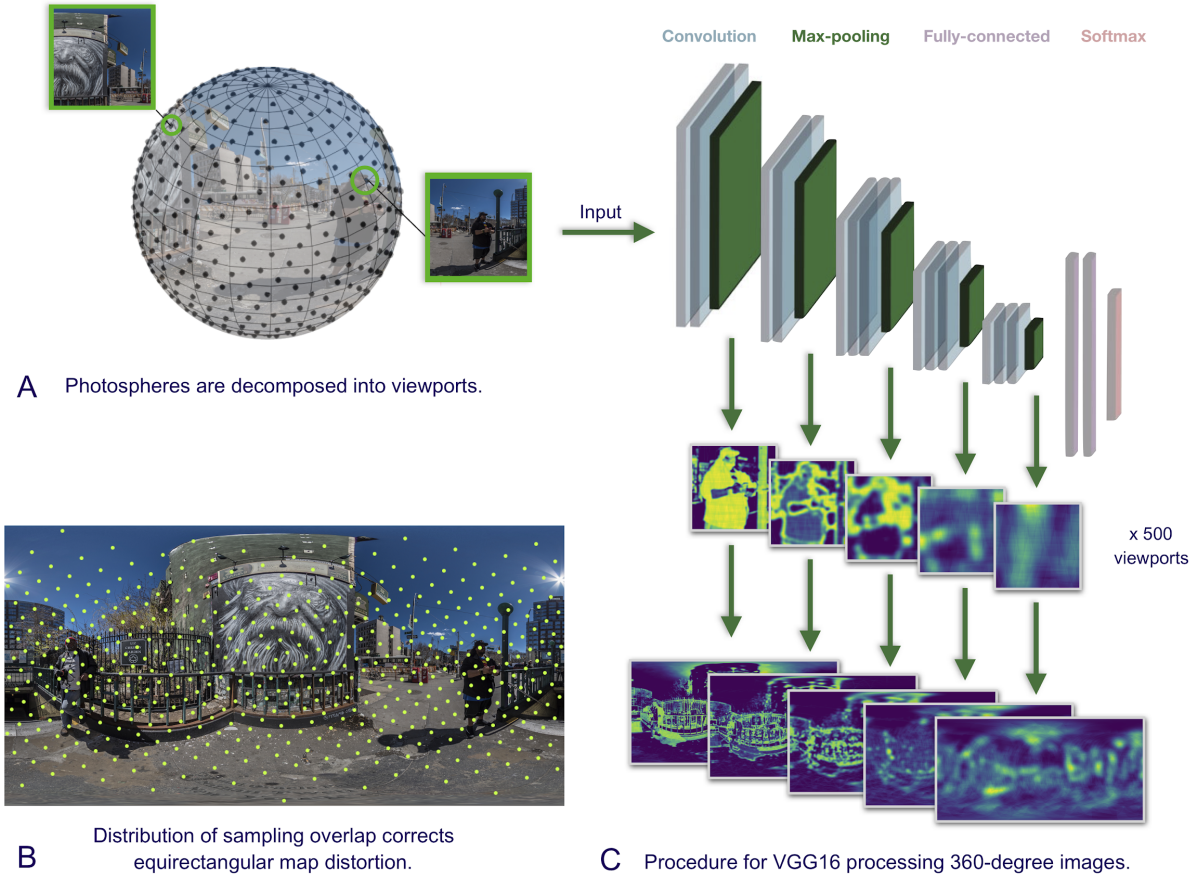


Figure 2. Photosphere-CNN processing pipeline.

A. Each scene is presented to participants as a 360-degree photosphere projected onto a virtual-reality environment in Unity. The sphere is sampled in 500 points densely around the equator and sparsely at the poles to account for the distortion of converting the photosphere into a panoramic image in Cartesian space (adapted from Haskins et al., 2020).
B. Sampling shown in equirectangular (Cartesian) space.
C. The viewport image centered around one sample point is fed as input into the CNN model (VGG16). At each max-pooling layer, we extract the network's activation in response to that viewport (adapted from O'Connell & Chun, 2018). This is repeated for all 500 viewports. These viewports are then projected back out into a panorama and aggregated in equirectangular space. Each pixel is then averaged over the number of times it was sampled.

4. Results

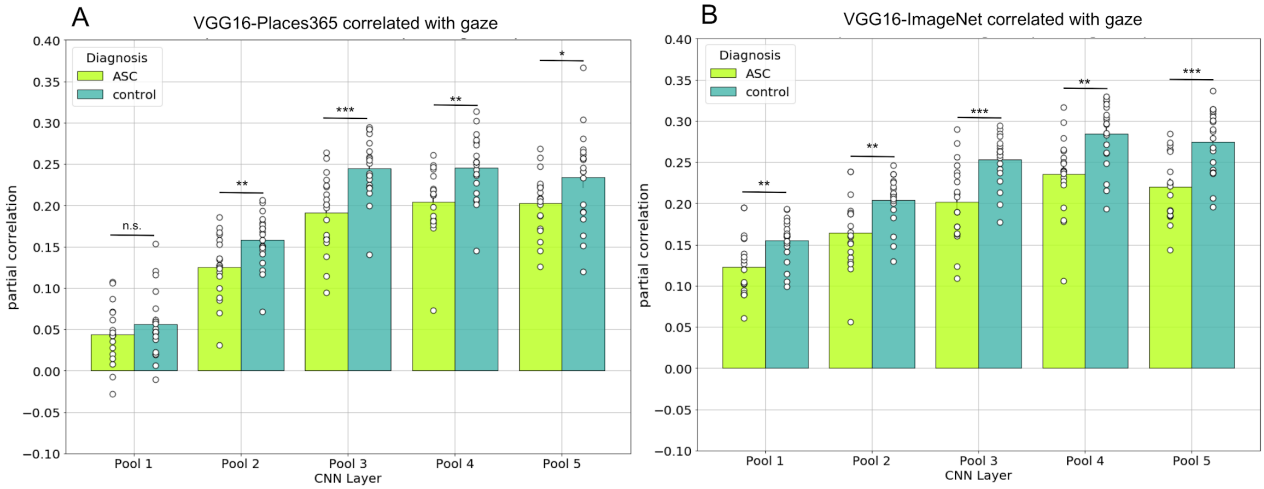
4.1 How well do the layers of a scene-trained CNN predict gaze behavior?

In this first analysis, we addressed whether different layers of a CNN processing scenes accounted for the differences in gaze behavior between individuals with and without autism. Behavioral and neuroimaging studies suggest that atypical early cortex activity could underlie the visual perceptual differences hallmark of ASC. In recent years, many computational neuroscientists have deep neural networks (particularly CNNs) as models of the primate visual hierarchy. We hypothesized that earlier CNN layers would be more predictive of gaze behavior in autism, reflecting more attention to low-level features like pixel saliency or contrast, whereas later layers (which represent more semantic information) would be more predictive of typical gaze behavior.

We used partial correlations (Groen et al., 2018) to investigate how predictive each layer of the model was of gaze behavior. This resulted in five values for each trial, for each participant, per model. VGG16-Places365 differentiated between ASC and controls significantly at pool-2, -3, -4, and -5. At each level of the VGG16-Places365, controls were consistently better predicted than individuals with autism. This pattern was consistent across the layers of VGG16-ImageNet as well, with an additional significant difference in predictivity at pool-1.

We tested the main effects of autism diagnosis (group), layer number, and model (network) with a 3 by 2 ANOVA. For both models and groups, we always find a main effect of layer, with correlations always increasing at higher layers ($F\text{-value} = 46.034$; $p < 0.0001$). We also always find a main effect of group, as individuals with autism are less predicted by both models at all layers ($F\text{-value} = 16.26$; $p < 0.0001$). We had hypothesized there would be a group by layer interaction, which was not indicated in our ANOVA (Figure 3).

To confirm that these results were not driven by a few outlier participants or scenes, we modeled individual by layer by experimental trial interactions as random effects in a linear mixed effects model. We included the same fixed effects of diagnosis and model pretraining, but added random effects of participants and trial. We found consistent patterns in our results: main effects of group and layer with no group by layer interaction (Supplemental Table S1).



C Analysis of Variance Table

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	3.29	3.2878	37.1463	1.123e-09 ***
network	1	5.94	5.9364	67.0713	2.833e-16 ***
layer	1	20.87	20.8689	235.7828	< 2.2e-16 ***
group:network	1	0.11	0.1069	1.2078	0.2718
group:layer	1	0.03	0.0266	0.3001	0.5838
network:layer	1	0	0.0005	0.0056	0.9401
group:network:layer	1	0	0.0027	0.0307	0.8608
Residuals	14681	1299.4	0.0885		

Figure 3: CNN layer analysis results.
Top: We averaged each individual's correlation with feature maps from each layer of the CNN to obtain an average correction score per individual, per layer. These scores are represented by **open circles.** We averaged across all individuals' average correlation scores within a group at each layer to obtain one group average correlation score per layer. These are represented by the colored **bars**. Significance indicated as:

*** : p < 0.05, ** : p < 0.01, *** : p < 0.001**

A. Average individual & group gaze correlations with feature maps from VGG16-Places365.
B. Average individual & group gaze correlations with feature maps from VGG16-ImageNet.
C. ANOVA table testing main effects of group, network (VGG16-Places365 versus ImageNet), and pooling layer, as well as their interactions.

4.2 Is gaze behavior in a novel scene more like an object-recognition task or a scene recognition task?

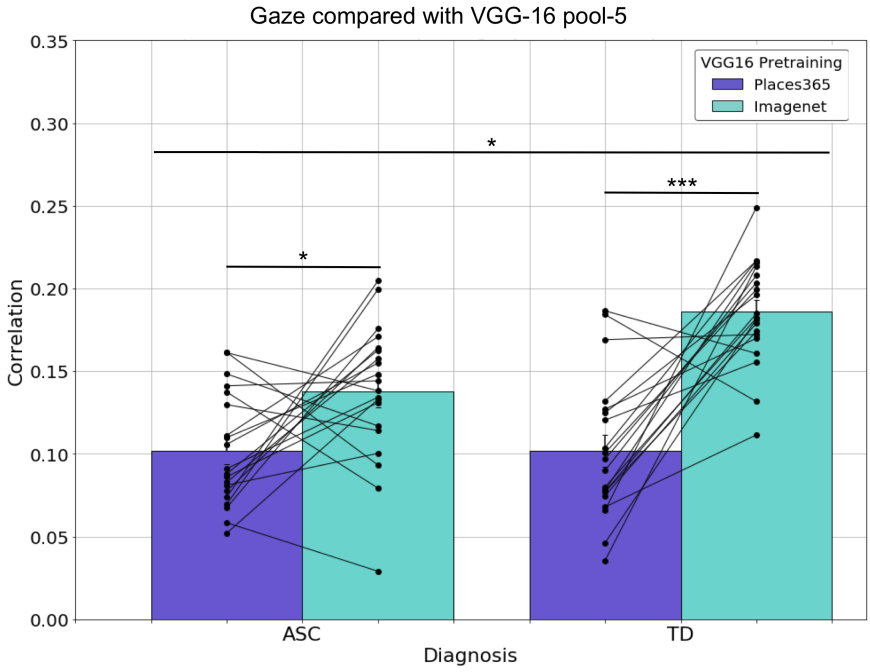
To test whether individuals exploring a novel scene allocate attention more toward local features or global features within the scene, we compared gaze maps with CNN maps of the top pooling layers across the two models: VGG16-Places365 and VGG16-ImageNet. These models have identical architectures but hidden weights optimized for scene classification and object classification respectively. By comparing a human's gaze behavior with the final pooling layer activity of CNNs completing object recognition tasks and scene recognition tasks within the same photosphere, we can identify which task is more comparable to the participant's active exploration of a scene.

We hypothesized that gaze behavior of individuals with ASC would be more similar to a high-level feature map of a network completing an object recognition task (VGG16-ImageNet) than one completing a scene recognition task (VGG16-Places365). We hypothesized that control behavior would reflect the reverse: their gaze behavior would be driven more by globally-informative features than locally-focused ones.

On average, individual gaze maps from ASC participants were significantly more correlated with the final VGG16-ImageNet map than with the final VGG16-Places365 map (t-statistic = 2.46; $p = 0.024$). However, the final VGG16-ImageNet map was also more predictive of control gaze behavior than the final VGG16-Places365 map (t-statistic = 6.12; $p < 0.0001$). It was also more predictive of control gaze behavior than it was of ASC gaze behavior (t-statistic = -4.21; $p = 0.0001$) (Figure 4A).

We modeled the effect of CNN pretraining and autism diagnosis with a two-factor ANOVA. This revealed a significant main effect of autism diagnosis ($F = 4.424$; $p < 0.05$) as well as the CNN pretraining ($F = 41.63$; $p < 0.0001$). Furthermore, we found a significant interaction of CNN pretraining and autism diagnosis ($F = 4.86$; $p < 0.05$) (Figure 4B).

To confirm that the group by network interaction was not driven by a subset of individual participants, we also built and tested a linear mixed effects model with the same fixed effects of group (ASC / typically developing) and network (VGG16-Places365/VGG16-ImageNet) and modeled individual participants and individual scenes as random effects. This revealed the same pattern of results (see Supplemental Table S2).



Analysis of Variance Table

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	0.47	0.4688	4.4243	0.03550 *
network	1	4.41	4.4119	41.6333	1.245e-10 ***
group:network	1	0.51	0.5148	4.8581	0.02758 *
Residuals	3676	389.55	0.106		

Figure 4: Local/global analysis results.
Top: We averaged each individual's experimental trial correlations with pool-5 feature maps from the two models (indicated as VGG16 pretraining) to obtain an average correlation score per individual (black points). Lines connect one individual's correlation score with both models. Bars indicate the average correlation score within a group, obtained by averaging across all of the individuals' average correlation scores within diagnostic group..
Bottom: ANOVA table modeling main effects of autism diagnosis group, VGG16 network pretraining, and significant interaction of group and network.

5. Discussion

5.1 Hierarchical layer analysis results

The first analysis in this study revealed that 360-gaze behavior in typically developing individuals is better predicted than in ASC at every level of the CNN. In both groups, higher levels better predict gaze behavior than lower levels, but within each level, controls are better predicted than individuals with autism. This effect is not driven by the high-level task performed by the model (object recognition or scene recognition) -- both groups are better predicted by VGG16-ImageNet than by VGG16-Places365, and both models are more predictive of control gaze than ASC.

We had expected that later layers would diverge in their ability to predict groups rather than earlier ones, as the model reaches its prediction at its highest layer and shows the most heat around name-abe objects or scene features. Instead, we found that feature maps from even the earliest layers of the CNN are more predictive of controls than of individuals with autism. This suggests that even early visual processing differs in autism, regardless of high-level task.

5.2 Local versus global attention analysis results

The second analysis found that gaze behavior of individuals with autism is better predicted by an object-recognition model than a scene-recognition model. We had expected this finding, as a hallmark of autistic visual perception is detail-orientation. However, we had also expected that gaze behavior of controls would be better predicted by a scene-recognition model than an object-recognition one, the reverse of our ASC prediction. We predicted this would be representative of how typically developing individuals represent a ‘global percept’ when exploring a novel 360-degree space. Instead, we found that control individuals were better predicted by an object-recognition model than a scene-recognition one, and they were even better predicted by this model than individuals with autism were.

This could indicate that the visual attention of individuals with autism is less driven by semantically-salient objects in a scene than that of controls. The top pooling layer of a CNN shows heat on what the model deems most significant for the task it is trying to complete. For our models, the top pooling layer is most concerned with name-able objects or scenes, which are semantically-meaningful. Perhaps ASC gaze behavior is characterized by comparatively less attention to semantically-salient features -- both social and non-social ones -- which would be supported by Wang et al. (2015)

Nevertheless, perhaps these findings are surprising to us because of how we conceptualized the models rather than how we conceptualized autistic gaze behavior. The local attention bias noted in the autism literature is supported by embedded figure tasks, such as the Navon task (Navon, 1977), which indicates that individuals with autism are more attentive to details than the big-picture figure, whereas controls indicate the reverse. We had predicted that in 360-degree scenes, this would translate to attention to objects within a scene, rather than attention to the global features of a scene, and the reverse for controls. Instead, we found that regardless of group, visual attention was more driven by objects within a scene than the scene itself, which suggests that these distinct models do not really target a 'local versus global attention' dichotomy in terms of the overall human visual experience.

Indeed, the feature maps produced by VGG16-Places365 show broad swaths of heat across the boundaries and terrain of a scene, whereas VGG16-ImageNet shows concentrated pockets of heat around specific features (such as a table or a book). Eye-tracking data in virtual reality does not indicate that individuals explore novel scenes in broad sweeps of the scenes' boundaries -- instead, their fixations jump between salient affordances and items (Sitzmann et al., 2017) producing pockets of heat rather than broad strokes (see supplementary figure S1). Regardless of *where* people are fixating, it seems there is an inherent bias in our analysis for gaze heat to correlate more with an object-recognition model

than with a scene-recognition one, given their heat distributions. We will address possible remedies for this in the next section.

5.3 Conclusions

This study is the first to unite eye-tracking in virtual reality and deep learning models to study the everyday visual experience of individuals with autism. Past explorations of visual processing in ASC have been driven by theory, whereas this investigation is driven by data, which affords us unbiased insight into the sensory experience. CNN feature maps are more predictive of gaze behavior in control participants than ASC participants at every location along the processing hierarchy, regardless of the type of visual recognition task the CNN is trained to perform. These models suggest that there are differences in visual processing between individuals with and without autism, beginning early in the visual hierarchy and progressing through levels that represent high-level features.

This investigation paves the way for more computational, data-driven approaches to studying the sensory experiences of individuals with autism in real-world settings. Given that sensory sensitivities impact around 90% of individuals with autism, it is essential that research focus on developing reliable, empirical methods to understand these sensitivities. Our approach also opens the door to understanding the perception of nonverbal individuals with autism, whose experiences we cannot understand from verbal reports but we can from eye-tracking data. In this report, we have only scratched the surface of what computational investigations of visual perception can teach us about autism. In the following section, we outline future extensions upon our work.

6. Future directions

Given the novelty of this study and the short time-frame in which this investigation was conducted, we propose a number of future avenues of investigation. We aim to address many of these in the coming months.

First, we would like to understand what drives the consistently lower correlation of ASC gaze behavior and CNN feature maps that is persistent across layers and networks. In our partial correlations, we attempt to control for the equatorial bias inherent in free viewing of images and scenes (Judd et al., 2009; Sitzmann et al., 2016). However, we would now like to investigate if individuals with autism tend to scan the equator of our scenes more than controls. If they do, by partialling out the equator, we are essentially handicapping our models to explain this group's gaze behavior, because we give the models less data. We also would like to understand if individuals with autism generally explore scenes less than our controls do, thus generating less gaze data for us to compare with the CNN models. In this case, the results we have found could be artifacts of our data that we could better control for in our regression models.

We would also like to understand the coherence of gaze behavior within and across diagnostic groups. Do the gaze maps of individuals with autism show less inter-subject correlation than controls do? Is one group more prone to noisy data than the other? Are there subgroups within these groups that generate more noisy data than others?

Another direction modification would be to our CNN map preprocessing pipeline. As shown in Supplementary Figure 1, feature maps from early pooling layers show sharp pixel contrast, whereas later feature maps are smoother. This has to do with the effective receptive field size considered in the model's layers, which increases by layer. However, this increasing smoothness gives a layer like pool-5 an inherent advantage in predicting gaze behavior. We propose smoothing our feature maps with a kernel that varies by layer in order to bring all

layers into equal smoothness as pool-5. Our expectation is that this would bring all layers into equal smoothness while maintaining the specific features important to that layer.

Our final proposed direction of investigation would be to use the heat from the ImageNet-generated feature maps to segment our scenes according to social and nonsocial salience. We would then like to investigate whether either group demonstrates a bias toward these different types of saliency

7. Acknowledgements

I would like to thank my advisors, Caroline Robertson and Leyla Isik, for their support and guidance throughout this project. I look forward to continuing this work in the future with both of you. I would also like to thank the Robertson Lab for their support: A.J. Haskins for her statistical prowess and R enthusiasm, Tommy Botch for co-developing the panoramic image feature map extraction pipeline, Brenda Garcia for much of the behavioral data collection, and Adam Steel for his well-timed wisdom and humor. I want to acknowledge Dartmouth's Undergraduate Research and Advising program for supporting all of my undergraduate research experiences, as well as the Neukom Institute for Computational Sciences and the Presidential Scholars Program.

8. Supplemental figures

Type III Analysis of Variance Table with Satterthwaite's method						
Factor	**Sum Sq**	**Mean Sq**	**NumDF**	**DenDF**	**F value**	**Pr(>F)**
group	0.3237	0.3237	1	14594.2	4.1016	0.04286 *
network	0.4795	0.4795	1	14597.2	6.0763	0.01371 *
layer	12.0742	12.0742	1	170.9	153.0032	< 2e-16 ***
group:network	0.0025	0.0025	1	14589.4	0.0313	0.85958
group:layer	0.0305	0.0305	1	170.8	0.3865	0.53497
network:layer	0.0025	0.0025	1	14588.3	0.0313	0.85962
group:network:layer	0.0038	0.0038	1	14586.2	0.048	0.82651

Table S1. Linear mixed effects model of CNN layer results.

We modeled the partial correlations of gaze behavior and activity at each layer of the CNN as a linear mixed effects model, whereby we treated the group, network, and layer as fixed effects but modeled the individual participants and individual scenes as random effects. This showed the same pattern of results as our traditional 3*2 ANOVA model: significant main effects of group, network, and layer with no significant interactions.

Type III Analysis of Variance Table with Satterthwaite's method						
Factor	**Sum Sq**	**Mean Sq**	**NumDF**	**DenDF**	**F value**	**Pr(>F)**
group	0.4674	0.4674	1	3628.3	5.0573	0.02458 *
network	4.1839	4.1839	1	3625.2	45.2667	1.99e-11 ***
group:network	0.5148	0.5148	1	3625.2	5.57	0.01832 *

Table S2. Linear mixed effects model of local / global attention results.

We modeled the partial correlations of gaze behavior and CNN activity as a linear mixed effects model, treating the group and network as fixed effects and modeling individual participants and individual scenes as random effects. Again, this reveals a main effect of both group and network and a significant interaction of the two, whereby the ImageNet network better predicts controls than ASC than does Places365.

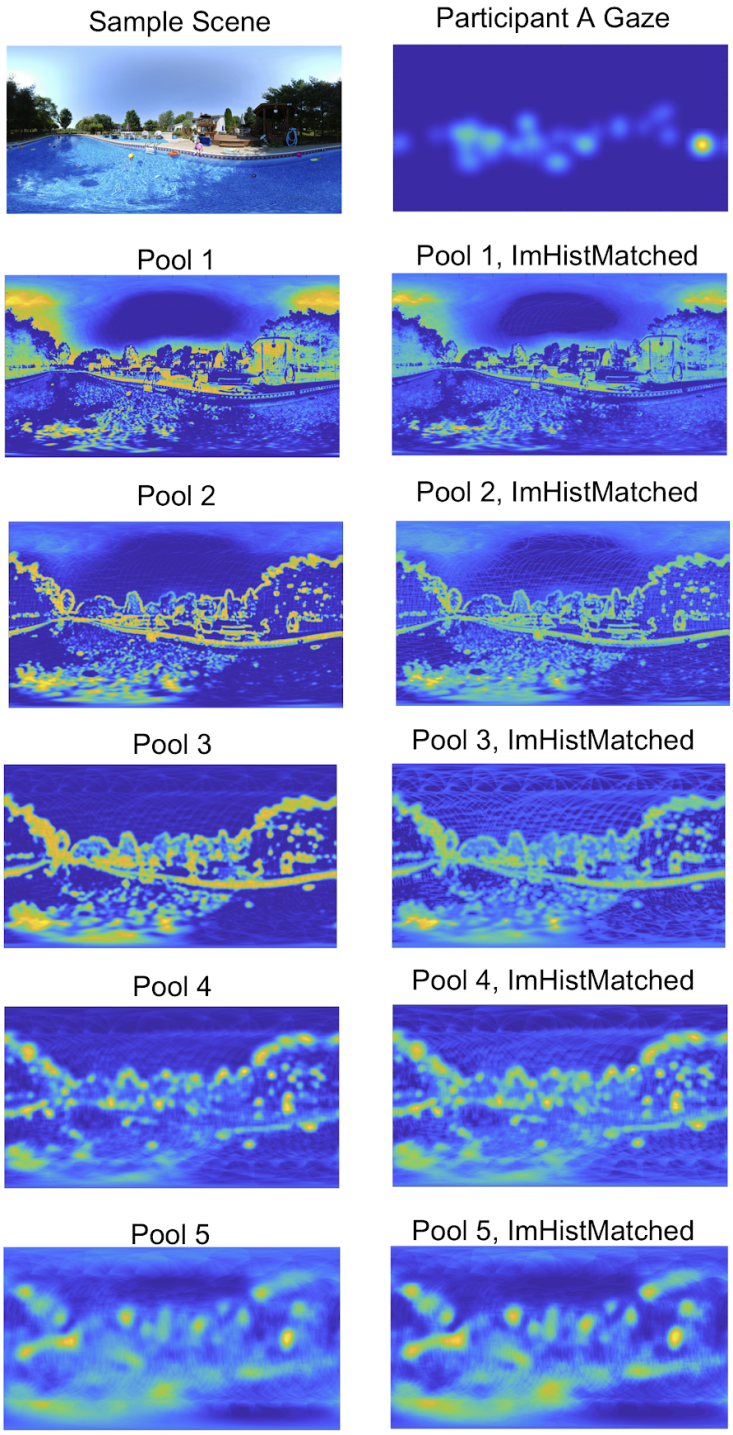
Figure S1. Example ImHistMatch of CNN layer maps to an average heat template.

Left Top: A sample stimulus.

Right Top: An example participant's gaze map for this scene.

Left column: VGG16-Places365 layer activity at each pooling layer, before preprocessing.

Right column: Layer activity after histogram normalization to a common template, which was modeled as the average of all the pooling layers in the left column.



9. References

Adjust histogram of 2-D image to match histogram of reference image—MATLAB imhistmatch. (n.d.). Retrieved February 17, 2020, from https://www.mathworks.com/help/images/ref/imhistmatch.html

Baranek, G. T., Watson, L. R., Boyd, B. A., Poe, M. D., David, F. J., & McGuire, L. (2013). Hyporesponsiveness to social and nonsocial sensory stimuli in children with autism, children with developmental delays, and typically developing children. *Development and Psychopathology*, *25*(2), 307–320. https://doi.org/10.1017/S0954579412001071

Baron-Cohen, S., Ashwin, E., Ashwin, C., Tavassoli, T., & Chakrabarti, B. (2009). Talent in autism: Hyper-systemizing, hyper-attention to detail and sensory hypersensitivity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1522), 1377–1383. https://doi.org/10.1098/rstb.2008.0337

carwin. (2020). *Clcarwin/convert_torch_to_pytorch* [Python]. https://github.com/clcarwin/convert_torch_to_pytorch (Original work published 2017)

CDC. (2019, September 3). *Data and Statistics on Autism Spectrum Disorder | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/data.html

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Collobert, R., Kavukcuoglu, K., & Farabet, C. (n.d.). *Torch7: A Matlab-like Environment for Machine Learning*. 6.

CSAILVision/places365. (2020). [Python]. MIT CSAIL Computer Vision. https://github.com/CSAILVision/places365 (Original work published 2016)

Dakin, S., & Frith, U. (2005). Vagaries of Visual Perception in Autism. *Neuron*, *48*(3), 497–507. https://doi.org/10.1016/j.neuron.2005.10.018

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (n.d.). *ImageNet: A Large-Scale Hierarchical Image Database*. 8.

Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y. S., Kauchali, S., Marcín, C., Montiel-Nava, C., Patel, V., Paula, C. S., Wang, C., Yasamy, M. T., & Fombonne, E. (2012). Global

Prevalence of Autism and Other Pervasive Developmental Disorders. *Autism Research*, *5*(3), 160–179. https://doi.org/10.1002/aur.239

Estes, A., Zwaigenbaum, L., Gu, H., St. John, T., Paterson, S., Elison, J. T., Hazlett, H., Botteron, K., Dager, S. R., Schultz, R. T., Kostopoulos, P., Evans, A., Dawson, G., Eliason, J., Alvarez, S., & Piven, J. (2015). Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life. *Journal of Neurodevelopmental Disorders*, *7*(1). https://doi.org/10.1186/s11689-015-9117-6

Fu-En.Wang. (2020). *Fuenwang/Equirec2Perspec* [Python]. https://github.com/fuenwang/Equirec2Perspec (Original work published 2017)

Gliga, T., Bedford, R., Charman, T., Johnson, M. H., Baron-Cohen, S., Bolton, P., Cheung, C., Davies, K., Liew, M., Fernandes, J., Gammer, I., Maris, H., Salomone, E., Pasco, G., Pickles, A., Ribeiro, H., & Tucker, L. (2015). Enhanced Visual Search in Infancy Predicts Emerging Autism Symptoms. *Current Biology*, *25*(13), 1727–1730. https://doi.org/10.1016/j.cub.2015.05.011

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*. https://doi.org/10.7554/eLife.32962

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Brain's Ventral Visual Pathway. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10–10. https://doi.org/10.1167/18.6.10

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106-154.2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science (New York, N.Y.)*, *310*(5749), 863–866. https://doi.org/10.1126/science.1117593

Hus, V., & Lord, C. (2014). The Autism Diagnostic Observation Schedule, Module 4: Revised Algorithm and Standardized Severity Scores. *Journal of Autism and Developmental Disorders*, *44*(8), 1996–2012. https://doi.org/10.1007/s10803-014-2080-3

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. https://doi.org/10.1073/pnas.1714471114

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*, 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

Kaldy, Z., Kraper, C., Carter, A. S., & Blaser, E. (2011). Toddlers with Autism Spectrum Disorder are more successful at visual search than typically developing toddlers. *Developmental Science*, *14*(5), 980–988. https://doi.org/10.1111/j.1467-7687.2011.01053.x

Kern, J. K., Trivedi, M. H., Garver, C. R., Grannemann, B. D., Andrews, A. A., Savla, J. S., Johnson, D. G., Mehta, J. A., & Schroeder, J. L. (2006). The pattern of sensory processing abnormalities in autism. *Autism*, *10*(5), 480–494. https://doi.org/10.1177/1362361306066564

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

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1),

417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (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

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv:1610.01563 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1610.01563

MATLAB 2019b (2019b). (n.d.). [Computer software]. The MathWorks, Inc.

Mottron, L., & Belleville, S. (1993). A Study of Perceptual Analysis in a High-Level Autistic Subject with Exceptional Graphic Abilities. *Brain and Cognition*, *23*, 279–309.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3

O'Connell, T. P., & Chun, M. M. (2018). Predicting eye movement patterns from fMRI responses to natural scenes. *Nature Communications*, *9*(1), 5159. https://doi.org/10.1038/s41467-018-07471-9

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8026–8037). Curran Associates, Inc. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Peterson, M. F., Lin, J., Zaun, I., & Kanwisher, N. (2016). Individual differences in face-looking behavior generalize from the lab to the world. *Journal of Vision*, *16*(7),

12. https://doi.org/10.1167/16.7.12

Plaisted, K., O'Riordan, M., & Baron-Cohen, S. (1998). Enhanced Visual Search for a Conjunctive Target in Autism: A Research Note. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(5), 777–783. https://doi.org/10.1017/S0021963098002613

Robertson, C. E., & Baron-Cohen, S. (2017). Sensory perception in autism. *Nature Reviews Neuroscience*, *18*(11), 671–684. https://doi.org/10.1038/nrn.2017.112

Robertson, C. E., Thomas, C., Kravitz, D. J., Wallace, G. L., Baron-Cohen, S., Martin, A., & Baker, C. I. (2014). Global motion perception deficits in autism are reflected as early as primary visual cortex. *Brain*, *137*(9), 2588–2599. https://doi.org/10.1093/brain/awu189

Rogers, S. J., Hepburn, S., & Wehner, E. (2003). Parent Reports of Sensory Symptoms in Toddlers with Autism and Those with Other Developmental Disorders. *Journal of Autism and Developmental Disorders*, *33*(6), 631–642. https://doi.org/10.1023/B:JADD.0000006000.38991.a7

Sak, H., Senior, A., & Beaufays, F. (n.d.). *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*. 5.

Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., & Pollick, F. E. (2009). Vision in autism spectrum disorders. *Vision Research*, *49*(22), 2705–2739. https://doi.org/10.1016/j.visres.2009.08.005

Simonyan, K., & Zisserman, A. (2015). *VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION*. 14.

Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., & Wetzstein, G. (2017). How do people explore virtual environments? *ArXiv:1612.04335 [Cs]*. http://arxiv.org/abs/1612.04335

Spencer, J. V., & O'Brien, J. M. D. (2006). Visual Form-Processing Deficits in Autism. *Perception*, *35*(8), 1047–1055. https://doi.org/10.1068/p5328

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., & Zhao, Q. (2015). Atypical Visual Saliency in Autism Spectrum Disorder Quantified

through Model-Based Eye Tracking. *Neuron*, *88*(3), 604–616. https://doi.org/10.1016/j.neuron.2015.09.042

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250. https://doi.org/10.3758/s13428-012-0245-6

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

YouTube. (n.d.). Retrieved February 25, 2020, from https://www.youtube.com/

Zagoruyko, S. (2020). *Szagoruyko/loadcaffe* [Protocol Buffer]. https://github.com/szagoruyko/loadcaffe (Original work published 2014)