View-Invariant Object Category Learning, Recognition, and Search: How Spatial and Object Attention are Coordinated Using Surface-Based Attentional Shrouds

Arash Fazl[1], Stephen Grossberg, Ennio Mingolla

Department of Cognitive and Neural Systems
Center for Adaptive Systems
and
Center of Excellence for Learning in Education, Science, and Technology
Boston University
677 Beacon Street,
Boston, MA 02215

Running Title: Spatial and Object Attention in Category Learning

Technical Report CAS/CNS-TR-07-011

April, 2007

Revised: April, 2008

Cognitive Psychology, in press

All correspondence should be addressed to

Professor Stephen Grossberg
Department of Cognitive and Neural Systems
Boston University
677 Beacon St.,
Boston, MA 02215
Phone: 617-353-7858
Fax: 617-353-7755
Email: steve@cns.bu.edu

[1] A.F. was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-01-1-0397), the National Science Foundation (NSF SBE-0354378), and the Office of Naval Research (ONR N00014-01-1-0624). S.G. was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-01-1-0397), the National Science Foundation (NSF SBE-0354378), and the Office of Naval Research (ONR N00014-01-1-0624). E.M. was supported in part by the National Science Foundation (NSF SBE-0354378) and the Office of Naval Research (ONR N00014-01-1-0624).

ABSTRACT

How does the brain learn to recognize an object from multiple viewpoints while scanning a scene with eye movements? How does the brain avoid the problem of erroneously classifying parts of different objects together? How are attention and eye movements intelligently coordinated to facilitate object learning? A neural model provides a unified mechanistic explanation of how spatial and object attention work together to search a scene and learn what is in it. The ARTSCAN model predicts how an object's surface representation generates a form-fitting distribution of spatial attention, or "attentional shroud." All surface representations dynamically compete for spatial attention to form a shroud. The winning shroud persists during active scanning of the object. The shroud maintains sustained activity of an emerging view-invariant category representation while multiple view-specific category representations are learned and are linked through associative learning to the view-invariant object category. The shroud also helps to restrict scanning eye movements to salient features on the attended object. Object attention plays a role in controlling and stabilizing the learning of view-specific object categories. Spatial attention hereby coordinates the deployment of object attention during object category learning. Shroud collapse releases a reset signal that inhibits the active view-invariant category in the What cortical processing stream. Then a new shroud, corresponding to a different object, forms in the Where cortical processing stream, and search using attention shifts and eye movements continues to learn new objects throughout a scene. The model mechanistically clarifies basic properties of attention shifts (engage, move, disengage) and inhibition of return. It simulates human reaction time data about object-based spatial attention shifts, and learns with 98.1% accuracy and a compression of 430 on a letter database whose letters vary in size, position, and orientation. The model provides a powerful framework for unifying many data about spatial and object attention, and their interactions during perception, cognition, and action.

Keywords: category learning, view-based learning, object recognition, spatial attention, object attention, parietal cortex, inferotemporal cortex, saccadic eye movements, attentional shroud, Adaptive Resonance Theory, surface perception, V2, V3A, V4, PPC, LIP, basal ganglia.

1. Introduction. What is an object? How can we learn what an object is without any external supervision? In particular, how does the brain learn to recognize a complex object from multiple viewpoints? Consider what happens when we first look at an object that is not instantly recognizable. We make scanning eye movements, directing our foveas around to a variety of points of interest, or views, on the object. The object's retinal representations of these views are greatly distorted by cortical magnification in cortical area V1 (Figure 1). The brain somehow combines several such distorted views into an object recognition category that is invariant to where we happen to be gazing at the moment. Future encounters with the same object can therefore lead to fast recognition no matter what view we happen to see**.**

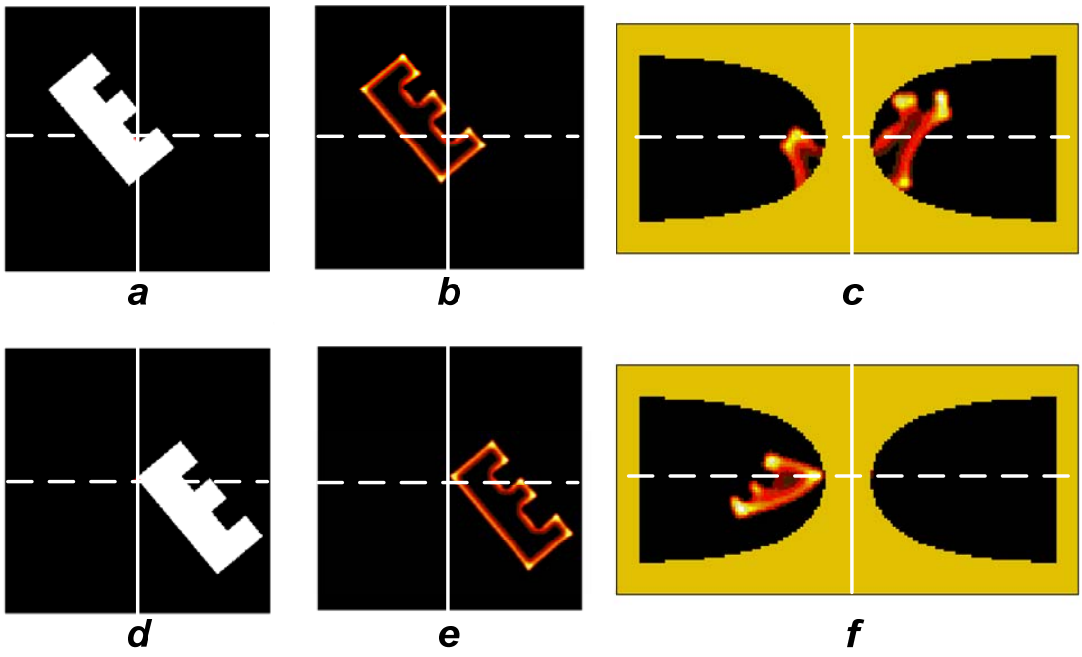

Figure 1: Visual input distortion due to cortical magnification. The activity generated in the primary visual cortex by a foveation (view) of an object depends on the position of fixation point on the object. Each saccade greatly distorts this map. (a) and (d): Images cast on the retina when the eye looks at different positions of the same tilted letter E. The center of the gaze is indicated by the intersection of the solid vertical and dashed horizontal lines. (b) and (e): Processing of the corresponding images of (a) and (d) by center-surround operators enhances contrast along edges, particularly at corners. (c) and (f): Simulated cortical magnification using the logarithmic-polar transformation (see text for details). (c) corresponds to the boundary images in (b) whereas (f) corresponds to those in (e). A gaze that is centered at the middle of the letter E, as in (a), activates peri-foveal areas of both hemispheres, whereas gazing at the top left corner of the letter E, as in (d), activates the left hemisphere only. For clarity, unlike human brain topography, the cortical representations are flipped upside down and foveal ends are juxtaposed.

Accumulating evidence supports the hypothesis that the brain learns about individual views of an object, coded by "view-tuned units." As this happens through time, neurons that respond to

different views of the same object learn to activate the same neuronal population, creating a "view-invariant unit." In other words, the brain learns to link multiple view-specific categories of an object to a view-invariant categorical representation of the object (Bradski & Grossberg, 1995; Bulthoff & Edelman, 1992; Bulthoff, Edelman & Tarr, 1995; Carpenter & Ross, 1993; Logothetis, Pauls, Bulthoff & Poggio, 1994; Riesenhuber & Poggio, 2000; Seibert & Waxman, 1992).

Many view-based models have focused on changes in retinal patterns that occur when a three-dimensional (3D) object rotates about its object-centered axis with respect to a fixed observer. However, as noted above, complex objects are often actively explored with saccadic eye movements. For example, in studying a face, eye movements may focus the eyes, nose, mouth, hair, ears, and other distinctive features. When we consider how eye movements help us to learn about an object, a fundamental *view-to-object binding problem* must be confronted.

How does the brain know that the views that are foveated on successive saccades belong to the same object? How does the brain avoid the problem of erroneously learning to classify parts of different objects together? Two identical eye movements may focus the eyes on two views of a single object, or on views of two different objects (Figure 2). Only views of the same object should be linked through learning to the same view-invariant object category. How does the brain know which views belong to the same object, even before it has learned a view-invariant category that can represent the object as a whole? How does the brain do this without an external teacher; that is, under the unsupervised learning conditions that are the norm during many object learning experiences *in vivo*?

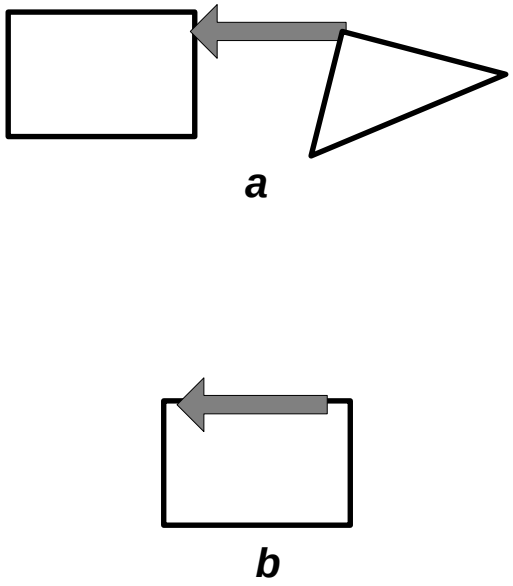

Figure 2: Saccades between and within objects. (a) The arrow indicates a saccade that moves the fovea from one object to another. (b) An example of a saccade with the same displacement as the one in (a), but which moves the fovea to another location on the same object. Learning multiple views of the same object should happen in case (b) and not in case (a), because the image cast on the retina before and after saccade in (a) do not belong to the same object.

We hypothesize that this is achieved through the coordinated use of spatial and object attention. Many studies of spatial attention have focused on its spatial distribution and how it influences

visual perception. Here we predict that spatial attention plays a crucial role in controlling view-invariant object category learning. In particular, several authors have reported that the distribution of spatial attention can configure itself to fit an object's form. Form-fitting spatial attention is sometimes called an *attentional shroud* (Tyler & Kontsevich, 1995). We explain how an object's pre-attentively formed surface representation can induce such a form-fitting attentional shroud. Moreover, while this attentional shroud remains active, we predict that it accomplishes two things:

First, it ensures that eye movements tend to end at locations on the object's surface, thereby enabling views of the same object to be sequentially explored.

Second, it keeps the emerging view-invariant object category active while different views of the object are learned and associated with it.

Thus, the brain avoids what would otherwise seem to be an intractable infinite regress: If the brain does not already know what the object is, then how can it, without external guidance, prevent views from several objects from being associated? Our proposal is that the *pre-attentively formed surface representation of the object* provides the object-sensitive substrate that prevents this from happening, even before the brain has learned knowledge about the object. This hypothesis is consistent with the burgeoning psychophysical literature showing that 3D boundaries and surfaces are the units of pre-attentive visual perception (Elder & Zucker, 1993; Grossberg, 1987a, 1987b, 1994; Grossberg & Mingolla, 1987; He & Nakayama, 1995; Paradiso & Nakayama, 1991; Raizada & Grossberg, 2003; Rogers-Ramachandran & Ramachandran, 1998) and that attention selects these units for recognition (Kahneman & Henik, 1981; LaBerge, 1995).

This proposed solution can be stated more formally as a temporally-coordinated cooperation between the brain's What and Where cortical processing streams: The Where stream maintains an attentional shroud whose spatial coordinates mark the surface locations of a current "object of interest," whose identity has yet to be determined in the What stream. As each view-specific category is learned by the What stream, it focuses object attention via a learned top-down expectation on the critical features that will be used to recognize that view and its variations in the future. When the first such view-specific category is learned, it also activates a cell population at a higher cortical level that will become the view-invariant object category.

Suppose that the eyes or the object move sufficiently to expose a new view whose critical features are significantly different from the critical features that are used to recognize the first view. Then the first view category is reset, or inhibited. This happens due to the mismatch of its learned top-down expectation, or prototype of attended critical features, with the newly incoming view information. This top-down prototype focuses object attention on the incoming visual information. Object attention hereby helps to control which view-specific categories are learned by determining when the currently active view-specific category should be reset, and a new view-specific category should be activated. However, the view-invariant object category should *not* be reset every time a view-specific category is reset, or else it can never become view-invariant. This is what the attentional shroud accomplishes: It inhibits a tonically-active reset signal that would otherwise shut off the view-invariant category when each view-based category is reset. As the eyes foveate a sequence of object views through time, they trigger learning of a sequence of view-specific categories, and each of them is associatively linked through learning with the still-active view-invariant category.

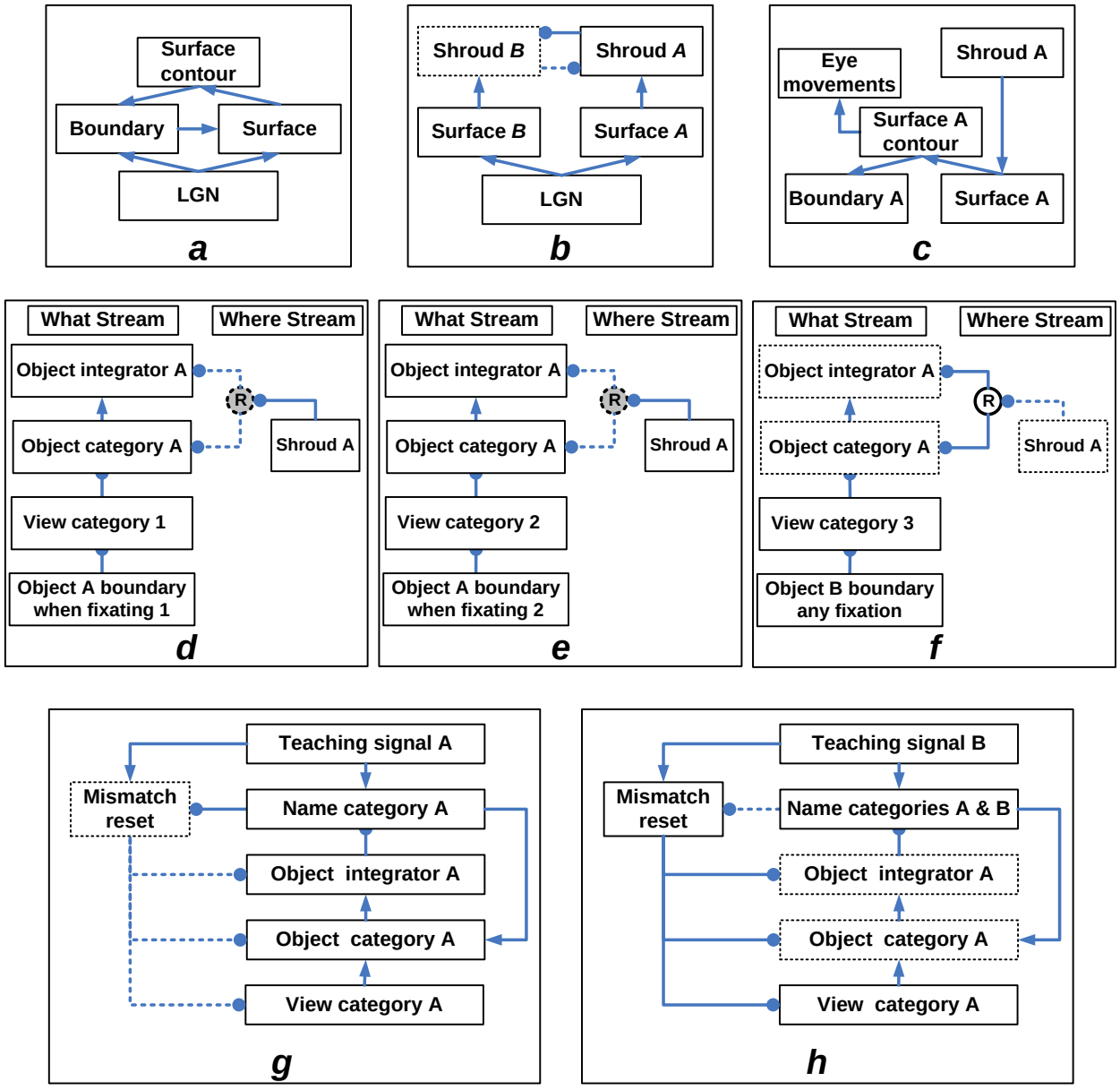

Figure 3: Schematic of ARTSCAN operations. (a) to (c): Where stream operations. (d) to (f): Unsupervised learning in ARTSCAN. (g) and (h): Supervised learning in ARTSCAN. (a) Pre-attentive boundary/surface interaction. The visual image represented in the LGN input is processed by two cortical streams: boundaries and surfaces. Wherever there is a closed boundary on the boundary map, a surface will form as the result of gated diffusion on the surface map. These completed surfaces will in turn up-regulate their corresponding boundaries through feedback via surface contours. (b) Shroud formation. If there are more than one surface present, the competition between their representations on the spatial attention map results in a winner, called the attentional shroud. The coordinate transform between the retinotopic surface map and the head-centric spatial attention map in a gain field is not shown in these simplified diagrams, but is discussed in the section titled "Retinotopic Surfaces and Head-centric Shrouds." (c) Attentional shroud effect on boundaries. The attentional shroud enhances its corresponding surfaces through feedback.

The circuit in (c) conveys this effect to surface contour and boundary maps. Surface contour feedback to the eye movement map increases the activity of all of the hotspots on an *attended* object, making them possible winners as the next saccade target. (d) The eyes fixate point 1 on object *A,* while the shroud has formed around that object. The feedback discussed in (a) to (c) has already down-regulated any other object boundary activities. This boundary activation excites view category 1, object category neuron *A* and its corresponding object integrator neuron. (e) If the eyes move to fixation point 2 on the same object, the new object boundary map activity might activate a different view category neuron 2, but it will activate the same object category and integrator *A*, because the attentional shroud is still active around the object *A* and inhibits the category reset neuron shown as R. (f) If the attentional shroud collapses around object *A*, the eyes can look at a different object and another view category neuron will get active. Collapse of the attentional shroud disinhibits the category reset neurons which inhibits all neurons in the two object layers, so these view category neurons will *not* get associated with object category neuron *A* anymore. (g) If ARTSCAN receives the name of the object it is visiting, e.g. object *A*, by a teaching signal *A*, it will associate it with the active object category and integrator neurons at that time. The activated name category neuron also inhibits the mismatch reset neuron. (h) Incorrect recall of object *B*'s name. In the same scenario as in (g), if the bottom-up input from object boundaries eventually excites name category *A,* but the teaching signal activates name *B*, both name category neurons *A* and *B* get activated and, due to shunting normalization, the activity of both will decrease to below a threshold such that none of them can inhibit the mismatch reset neuron anymore. This allows the teaching signal to activate the mismatch reset neuron and inhibit both object layers and stop learning. This also increases the vigilance parameters in the view category layer.

When the eyes move off an object, its attentional shroud collapses in the Where stream, thereby disinhibiting the reset mechanism that shuts off the view-invariant category in the What stream. When the eyes look at a different object, its shroud can form in the Where stream and a new view category can be learned that can, in turn, activate the cells that will become the view-invariant category in the What stream. Figure 3 summarizes the main operations of the model, which we call the ARTSCAN model because it shows how object category learning mechanisms of Adaptive Resonance Theory, or ART (Carpenter & Grossberg, 1987, 1993; Carpenter, Grossberg & Reynolds, 1991; Grossberg, 1999a, 2003), can be regulated during active SCANning by saccadic eye movements. These results have been reported in preliminary form in Fazl, Grossberg, and Mingolla (2004, 2005, 2006)**.**

The above discussion of attentional shroud formation concerns *exogenously* activated spatial attention: the bottom-up signals from a pre-attentively activated surface representation can compete for spatial attention to form a form-fitting shroud. *Endogenously* activated spatial attention can also activate a form-fitting shroud, even if it initially activates a much smaller region, say with a Gaussian *spotlight* of attention (Posner, 1980). The ARTSCAN model provides a mechanistic explanation of how the concept of an attentional spotlight can be reconciled with concepts about object-fitting attentional shrouds. Thus, if such a volitionally activated Gaussian sends top-down signals topographically to the surface representation, then the surface representation can use filling-in to spread the attentional input within the entire surface that is surrounded by the object's pre-attentively formed boundary. Then this enhanced surface

activity can activate spatial attention bottom-up throughout the region of the surface and again compete to create a form-fitting shroud (Figure 4). One consequence of this combination of bottom-up visual input and top-down endogenous attentional input to a surface representation is that the effective contrast of the attended surface may increase, as has been observed in the recorded responses of V4 neurons (Reynolds, Pasternak & Desimone, 2000).

Two other points will also be made now to bridge between our mechanistic descriptions of attention and more qualitative descriptions of attention in the experimental literature. These distinctions may also help to resolve controversies between space-based and object-based concepts of visual attention (Egeth & Yantis, 1997; Egly, Driver & Rafal, 1994; Serences, Schwarzbach, Courtney, Golay & Yantis, 2004; Yantis & Serences, 2003). Prior approaches to object-based attention, including the CODE theory of visual attention (LaBerge, 1995; LaBerge & Brown, 1989; Logan, 1996), use a gradient of spatial attention around visual features to explain object-based attention. Objects are words and features are individual letters in their experiments and models. ARTSCAN, on the other hand, uses boundary-surface interactions that induce figure-ground segregation to suggest how spatial and object attention may be coordinated for purposes of object learning, and how the brain can deal with deal with large and overlapping objects.

First, the word "object attention" is often used in a way that does not sharply differentiate different underlying neural mechanisms. At least three different neural mechanisms can control attention in a manner that is object-related: (1) an attentional shroud can fit an object's *surface* shape (Baylis & Driver, 2001; Cavanagh, Labianca & Thornton, 2001; Moore & Fulton, 2005; Tyler & Kontsevich, 1995); (2) attention can flow along an object *boundary* (Roelfsema, Lamme & Spekreijse, 1998; Scholte, Spekreijse & Roelfsema, 2001) and (3) attention can select the critical feature pattern, or prototype, that characterizes a learned object category (Blaser, Pylyshyn & Holcombe, 2000; Carpenter & Grossberg, 1987; Cavanagh, Labianca & Thornton, 2001; Duncan, 1984; Grossberg, 1980b; Kahneman, Treisman & Gibbs, 1992; O'Craven, Downing & Kanwisher, 1999). Only the third type of attention is formed entirely within the What cortical stream. The other two types involve What-Where inter-stream interactions, and thereby clarify how object and spatial attention interact.

Second, spatial attention need not form a shroud around only one object. A large literature clarifies that spatial attention can form over more than one object (Downing, 1988; Eriksen & Yeh, 1985; LaBerge & Brown, 1989; McMains & Somers, 2005; Pylyshyn & Storm, 1988; Yantis, 1992). How this happens will be the subject of a future study. Its possible utility is intelligently scanning a scene for several different target objects is noted in Section 9.

The ARTSCAN model and relevant data are described in Sections 2-7. Section 8 shows that the model can simulate reaction times (RTs) in human data about object-based spatial attention shifts. Reaction times are faster when responding to the non-cued end of an attended object than to a location outside the object, and slower engagement of attention to a new object occurs if attention has to be first disengaged from another object. Section 8 also describes how the model learns a letter database whose letters vary in size, position, and orientation. Finally, Section 9 discusses how the ARTSCAN model compares with other attention models in the literature.

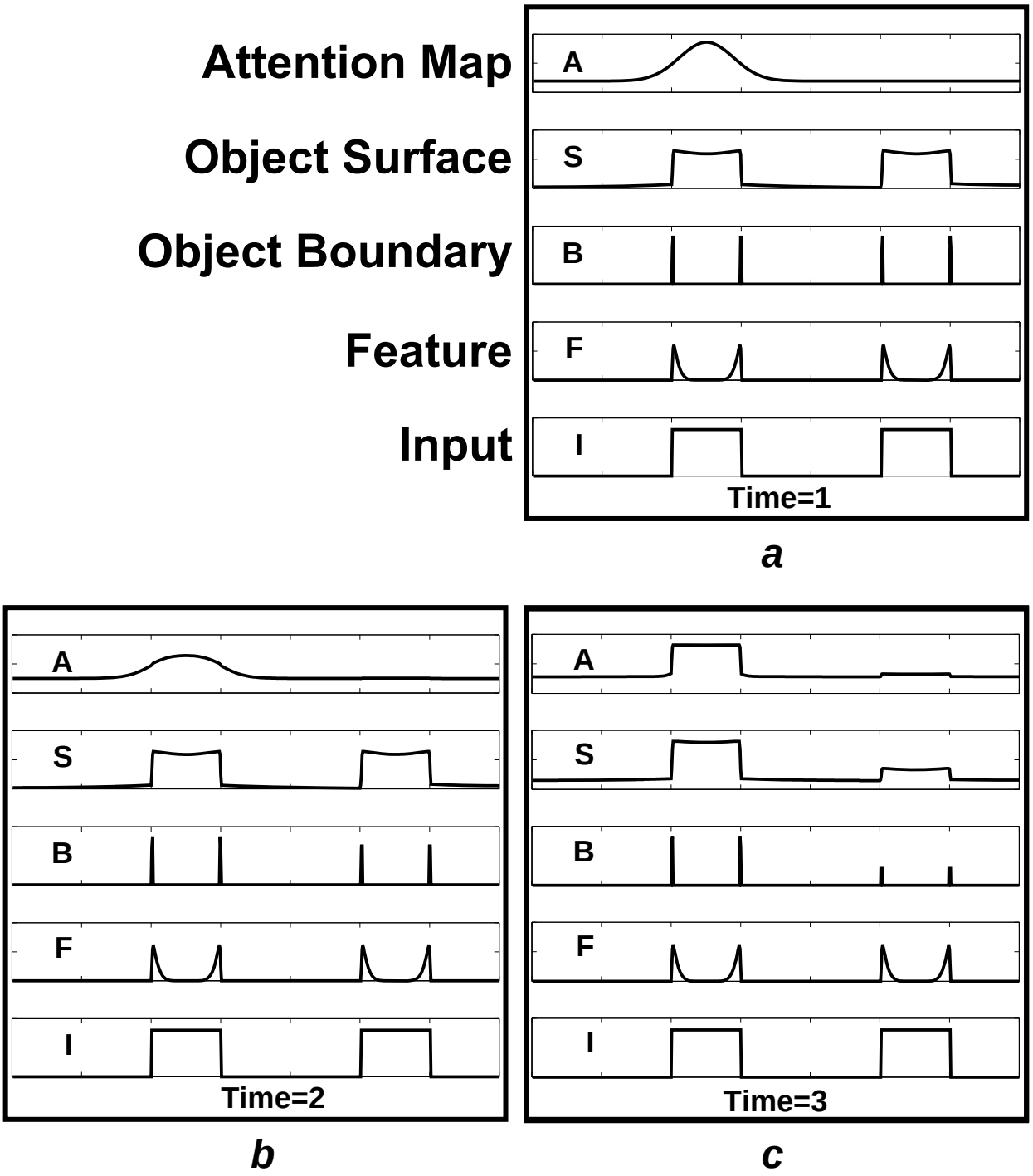

Figure 4: Schematic representation of interaction between surface filling-in and spotlight of attention through time. (a), (b), and (c) are consecutive instances in the dynamics of the Where stream. (a) Initially, two one-dimensional inputs, "I" at the bottom of (a), give rise to two sets of features and object boundaries, "F" and "B", respectively, in (a), by contrast enhancement. The features start diffusing and are bounded inside the boundaries and re-create the surfaces in the object surface map "S" in (a) by a *filling-in* process . Note that these surfaces have equal activity at this initial time. Suppose that at this same time on the attention map "T", a Gaussian spot-light of attention exists around the

location corresponding to the left surface and that there is feedback between the attention and the surface maps. Note that parts of the spotlight are located outside the boundaries of the surface. (b) The same 1D maps as in (a), but after some time has elapsed. Notice that the Gaussian spotlight of attention is morphing toward the shape of surface. (c) The same maps in a later instance of time. Now the activity in the attention map has totally gained the shape of its corresponding surface; i.e., it is form-fitting like a shroud. Those parts of the shroud that initially fell outside the boundaries of the surface have diffused away and do not exist any more. The corresponding surface on the left on the surface map has also gained more activity in the mentioned positive feedback loop. The other *unattended* surface is suppressed.

2. View-Invariant Object Recognition with Eye Movements and Cortical Magnification. Two main ideas have been proposed regarding how biological visual systems achieve *object constancy*, or the recognition of objects seen in a variety of views: "object-centered" or "structural description" theories propose that objects are represented as descriptions of parts and their spatial arrangements in a three-dimensional coordinate system that is centered on the object itself (Cooper, Biederman & Hummel, 1992; Marr & Nishihara, 1978). One problem with such theories is that one can easily find objects that do not seem to conform to any pre-determined set of parts. According to "view-based" object recognition theories, objects are represented as collections of view-specific representations, leading to recognition performance that is a function of previously seen object views (Bulthoff & Edelman, 1992; Edelman & Poggio, 1991; Tarr & Bulthoff, 1995, 1998; Tarr, Williams, Hayward & Gauthier, 1998). As noted above, ARTSCAN employs a view-based object learning and recognition strategy.

View-based models propose that there exist view-dependent neurons whose activity is tuned to a certain view of an object. ARTSCAN proposes how a view-sensitive *category* can be learned, which can be activated by a similar set of object views. Such view-sensitive category neurons in Figures 3d-f respond to a limited range of object transformation, and their response diminishes as that object rotates, translates or expands beyond that preferred range. Retinal image changes caused by saccadic eye movements can cause response changes in such view-sensitive category neurons. If several such view-sensitive neurons can all learn how to send excitatory signals to a shared neuron, or neuron population, the resulting *view-invariant category (object category)* neuron can then tolerate the total range of transformations spanned by the union of the view-sensitive neurons, and therefore keeps responding no matter which of those views of the object is observed; see the object category neurons in Figures 3d-h. Object integrator neurons in ARTSCAN integrate the activity coming from each of the object category activations. The integrator neurons are sensitive to the number of activations of the corresponding object category through time, and thereby accumulate evidence in favor of that category's interpretation of incoming signals.

Most view-based models have neglected two important facts about object learning: the transformation from retinal to cortical representation is space-variant (Daniel & Whitteridge, 1961;Drasdo, 1977;Schwartz, 1977), and the eyes move actively to explore the world. The model of Riesenhuber and Poggio (1999), for example, uses a spatially homogenous representation of the scene, ignoring the huge distortion that cortical magnification introduces into the real retinal image (Figure 1). These intrinsic variations in the input to the cortex, created by the combination of eye movements and cortical magnification on the same stationary object, are often stronger than the extrinsic variations due to rigid transformations of the object itself. Some models do

treat cortical magnification (Baloch & Waxman, 1991; Bradski & Grossberg, 1995; Seibert & Waxman, 1992), but do so *statically,* by foveating on an object's center-of-mass. Because the foveal parts of the retina have much higher resolution, primates typically scan an object with two or more eye movements in order to utilize the fovea's high resolution to better discriminate salient object features.

The ARTSCAN model predicts how the brain has evolved to deal with the challenges raised by multiple foveations and cortical magnification in a view-based object recognition system. It associates different view-sensitive categories due to eye movements with the same view-invariant object category. Object category neurons hereby learn to tolerate both *extrinsic* rigid object transformations and *intrinsic* variations due to saccades on the same object.

3. Unsupervised Learning of a View-Invariant Category as the Eyes Move. An even more basic issue concerns the *view-to-object binding problem* that was briefly mentioned above: In a scene filled with different objects, the eye movements can land the fovea on any object, and not necessarily on the object just observed (Figure 2). How does an object neuron know that a new cortical activation pattern belongs to the same object that it was learning, so keeps learning it, whereas another cortical activation pattern does not, so stops learning it? Failure of the learning/recognition system to detect this difference results in erroneously associating views of different objects with the same object category neuron and poor performance (see Results). We show that by combining perceptual boundary and surface representations, spatial and object attention, category learning, and eye movement mechanisms, the object recognition areas of the brain can correctly learn which view-dependent categories belong to the same view-invariant categories object neurons, even under unsupervised learning conditions.

How does this happen? Figure 3a notes that perceptual boundaries and surfaces can form pre-attentively, automatically, and in parallel within the visual cortical interblob and blob streams (see Section 4). An object's surface representation, in turn, activates an attentional shroud whose shape fits around the object's form (Figure 3b). The shroud inhibits a *reset* cell population that is otherwise tonically active and nonspecifically inhibits the object category and integrator levels (Figure 3d). Just so long as the shroud remains active, it inhibits the reset cells and enables the object category to remain active and to be associated with multiple object view categories (Figure 3e). When the shroud collapses, for reasons that are explained below, the reset signal inhibits the active object category and integrator neurons (Figure 3f), thus stopping view-invariant learning before views of other objects can be erroneous bound to the previous object. Formation and collapse of the attentional shroud in the Where stream hereby automatically parses into different object categories the stream of visual information hitting the object recognition areas in the What stream as eye movements occur. Previous models have attributed this reset mechanism to the occurrence of sufficiently large saccades (Baloch & Waxman, 1991). ARTSCAN mechanisms explain how such reset occurs only for those saccades that move the fovea away form an attended object. Even large saccades that explore the same object while its attentional shroud remains active do not produce reset.

All of the What stream processes explained so far operate under unsupervised learning conditions: They can learn a view-invariant object category even without an external teacher. Much biological recognition learning goes on under unsupervised conditions, without being taught to name the learned categories, notably in children. Creatures in computer animations, or Greebles (Gauthier & Tarr, 1997), are good examples; we learn to *recognize* the creatures from different angles, yet we might not be able to name them. A biologically relevant learning mechanism needs to be able to function during both unsupervised and supervised learning

conditions, including situations when unsupervised learning trials are mixed with supervised learning trials in unpredictable ways. We use an Adaptive Resonance Theory, or ART, category learning and recognition model to realize this property (Carpenter & Grossberg, 1987, 1993; Carpenter, Grossberg, Markuzon, Reynolds & Rosen, 1992; Grossberg, 1980a, 1999b). In order to learn and recall a category's name during supervised learning trials, the model incorporates teaching signals that can activate name categories, to which the object categories can be linked by means of associative learning (Figure 3g).

After some supervised training, the model guesses the name of the object that it is attentively observing. If a teaching signal is provided on such a guessing trial, two scenarios might happen: Either the teaching signal *matches* the model's prediction and confirms it, in which case learning occurs (Figure 3g), or there is a *mismatch reset*, which resets the active object category and integrator neurons, thereby stopping learning in the previously active category, while generating a memory search, or bout of hypothesis testing, aimed at discovering and learning a better object category with which to classify the object (Figure 3h). The ARTSCAN model hereby posits that at least two reset mechanisms are involved in the learning of view-invariant object categories: Collapse of a spatial attentional shroud in the Where stream (Figure 3f), and mismatch of an object name prediction with an externally supplied name in the What stream (Figure 3h).

Various data suggest that view-sensitive categories, view-invariant object categories, and category naming neurons are part of the What, or ventral object recognition stream in the mammalian visual cortex (Ashby & Ell, 2001; Booth & Rolls, 1998; Haxby*, et al.*, 1991; Lueschow, Miller & Desimone, 1994; Rolls, Judge & Sanghera, 1977; Vuilleumier, Henson, Driver & Dolan, 2002). In particular, Vuilleumier *et al.* (2002) showed a gradual posterior/anterior progression from view-variant to view-invariant areas, with category names represented in the inferior frontal lobes.

Given that an attentional shroud can prevent object category reset, we need to consider how shrouds are formed and collapse through time.

4. Boundary and Surface Representations control Shrouds and Scans. Much perceptual and neurobiological evidence support the 1984 prediction of Grossberg and his colleagues that the units of pre-attentive visual perception are boundaries and surfaces (Cohen & Grossberg, 1984; Grossberg, 1984; Grossberg & Mingolla, 1985a, 1985b); see Figures 3a-c. The model that embodies this prediction is often called the BCS/FCS model, for Boundary Contour System and Feature Contour System. Grossberg generalized this hypothesis to predict in 1987 that 3D boundaries and surfaces are the units of 3D vision and figure-ground perception. This prediction is part of the FACADE (Form-And-Color-And-DEpth) model of 3D vision and figure-ground separation, that has been used to explain and predict a wide range of perceptual and neurobiological data; see Grosserg (1994, 2003) and Raizada and Grossberg (2003) for reviews. These results about how boundaries and surfaces contribute to object perception were derived during the same period when data concerning the allocation of object attention began to appear (e.g., Duncan, 1984; Hochberg & Peterson, 1987; Kahneman & Henik, 1981).

Perceptual boundaries are predicted to form in the (LGN Parvo)-(V1 Interblob)-(V2 Interstripe)-V4 cortical stream, while perceptual surfaces are predicted to form in the (LGN Parvo)-(V1 Blob)-(V2 Thin Stripe)-V4 stream. Moreover, perceptual boundaries and surfaces can form pre-attentively, and do so as part of the process of separating figures from their backgrounds in depth. Supportive psychophysical data (Beardslee & Wertheimer, 1958; Driver & Baylis, 1996; Rubin, 1921), fMRI data (Kourtzi & Kanwisher, 2001), and electrophysiological

data (Baylis & Driver, 2001) have shown that whether an edge is assigned to a figure or to a background is an important factor for attracting attention, activating object recognition areas, and remembering it later. Baylis and Driver (2001) also argued that figure-ground separation happens prior to attentive selection of an object, and that it is an obligatory mechanism yoked to bottom-up image luminance that does not need a top-down, possibly attentive, influence to operate.

The ARTSCAN model simplifies surface processing, and thus the model's computational load, by eliminating mechanisms for boundary completion, because the stimuli in the present simulations are planar geometric shapes with no missing or hidden boundaries or depth cues. 3D boundary completion mechanisms can seamlessly be added to ARTSCAN when they are required to recognize more complex environments. The surface process that we use is schematized in Figure 3a. As in the BCS/FCS model (Grossberg & Todorović, (1988), the LGN activates boundaries and surfaces in parallel. Then the boundaries gate the filling-in of surface lightness and color via boundary-to-surface signals. Figure 3a also contains a surface-to-boundary feedback pathway (via a surface contour process). This process was introduced in the FACADE model to ensure perceptual *consistency*: it explains how, even though boundaries and surfaces form according to *complementary* computational rules, they give rise to a *consistent* visual percept (Grossberg, 1994, 2003). Surface-to-boundary feedback assures consistency by confirming and strengthening the boundaries that lead to successful filling-in of surfaces, and inhibiting boundaries that do not. The FACADE model predicts that the surface-to-boundary feedback mechanism also plays a key role in 3D figure-ground separation.

This boundary-surface feedback loop is proposed to work as follows: 3D boundary signals are topographically projected from V2 Interstripes to V2 Thin Stripes. If a boundary is closed, it can contain the filling-in of lightness and color within it. If the boundary has a sufficiently big gap, surface lightness and color can dissipate through the gap. The surface contour process is sensitive to the contrasts at the border of a successfully filled-in surface within a closed boundary. This contrast-sensitive process is realized by an on-center off-surround network that detects the contours of successfully filled-in surfaces. This is an on-center off-surround network across position and within depth.

The surface contour outputs from successfully filled-in surfaces use topographic excitatory signals to strengthen the boundaries that generated these surfaces, and inhibitory signals to weaken spurious boundaries at the same positions but farther depths. This is an on-center off-surround network within position and across depth. This surface-to-boundary feedback is predicted to arise from V2 Thin Stripes and terminate in V2 Pale Stripes.

By eliminating these spurious boundaries, surface-to-boundary feedback enables occluding surfaces and partially occluded surfaces to be separated onto different depth planes, and allows partially occluded boundaries and surfaces to be amodally completed behind their occluders. Such completed representations can then be more easily recognized in the inferotemporal cortex and beyond. A more detailed explanation and simulations of how this happens are given in Grossberg (1994) and Kelly & Grossberg (2000). In summary, the FACADE model predicts why and how contour-sensitive surface-to-boundary feedback helps to define an object by ensuring that the correct object boundaries and surfaces are consistently bound together to form a pre-attentive object representation. When boundary and surface properties of an object are not consistently bound, illusory conjunctions can form; e.g. of boundary shape and the color that it encloses (Treisman & Schmidt, 1982).

5. Shrouds Coordinate Scanning Eye Movements and Object Category Learning. ARTSCAN predicts that the *same process* that pre-attentively defines and segregates objects in

depth also plays a key role in regulating attentive learning of an object category. It does this by inducing sequences of scanning saccadic eye movements on that object surface whose spatial attentional shroud is active at any given time. This works as follows:

As shown in Figure 3a, LGN tries to pre-attentively activate all the possible surfaces in a scene. The surfaces, in turn, attempt to topographically activate spatial attention to form a surface-fitting attentional shroud (Figure 3b and 4). The spatial attention network contains long-range inhibitory interactions that tend to select the strongest shroud and inhibit weaker ones (Figure 3b). The winning shroud sends topographic feedback to its generative surface, thereby activating it further (Figure 3c). Thus, ARTSCAN predicts that surface representations receive both contrastive bottom-up inputs from areas like LGN and top-down spatial attention inputs from areas like the parietal cortex. Recent data support the view that attention can, in fact, increase the perceived brightness of a surface (Carrasco, Penpeci-Talgar & Eckstein, 2000; Grossberg & Raizada, 2000; Reynolds & Desimone, 2003) and connections from parietal areas to V4 are known (Cavada & Goldman-Rakic, 1989, 1991; Distler, Boussaoud, Desimone & Ungerleider, 1993; Webster, Bachevalier & Ungerleider, 1994).

ARTSCAN predicts that this feedback plays an important role in object learning. In particular, when the winning surface has its activation enhanced by top-down spatial attention, its contrast relative to its surround increases. As a result, *its surface contour signals increase*. As summarized in Figure 3c, stronger surface contour signals generate stronger eye movement target commands to the saccadic eye movement system. We explain below how this enhanced feedback helps to direct scanning eye movements to the object surface whose shroud is active. Thus, the views that the eyes happen to look at tend to belong to the same object surface while its spatial attentional shroud is on, and these are the views that will be learned.

Cortical area V3A is one possible brain area where such surface contour signals may get converted into eye movement target signals. In particular, studies show that it is concerned with relative disparity (Backus, Fleet, Parker & Heeger, 2001), gaze (Galletti & Battaglini, 1989), saccades (Caplovitz & Tse, 2006; Nakamura & Colby, 2000) and prehensile hand movements (Nakamura*, et al.*, 2001). This proposal is offered tentatively due to the sparcity of available data, combined with evidence that the function of macaque V3A differs from that performed by human V3A (Tootell*, et al.*, 1997).

The ARTSCAN proposal differs significantly from *saliency map* models of visual attention; e.g. Itti & Koch (2001). In such models, the units are single spatial locations. If a location contains a stronger feature, such as brightness, color, or orientation, it wins the process of directing attention shifts and scanning eye movements. ARTSCAN, and the FACADE model before it (Grossberg, 1994), explain how 3D boundaries and surfaces work as visual perceptual units. ARTSCAN clarifies, in addition, how a surface-based saliency map works by using surface-fitting attentional shrouds to direct coordinated attention shifts, eye movements, and object learning. Figure 5 summarizes some of the key ARTSCAN processes that have been discussed so far.

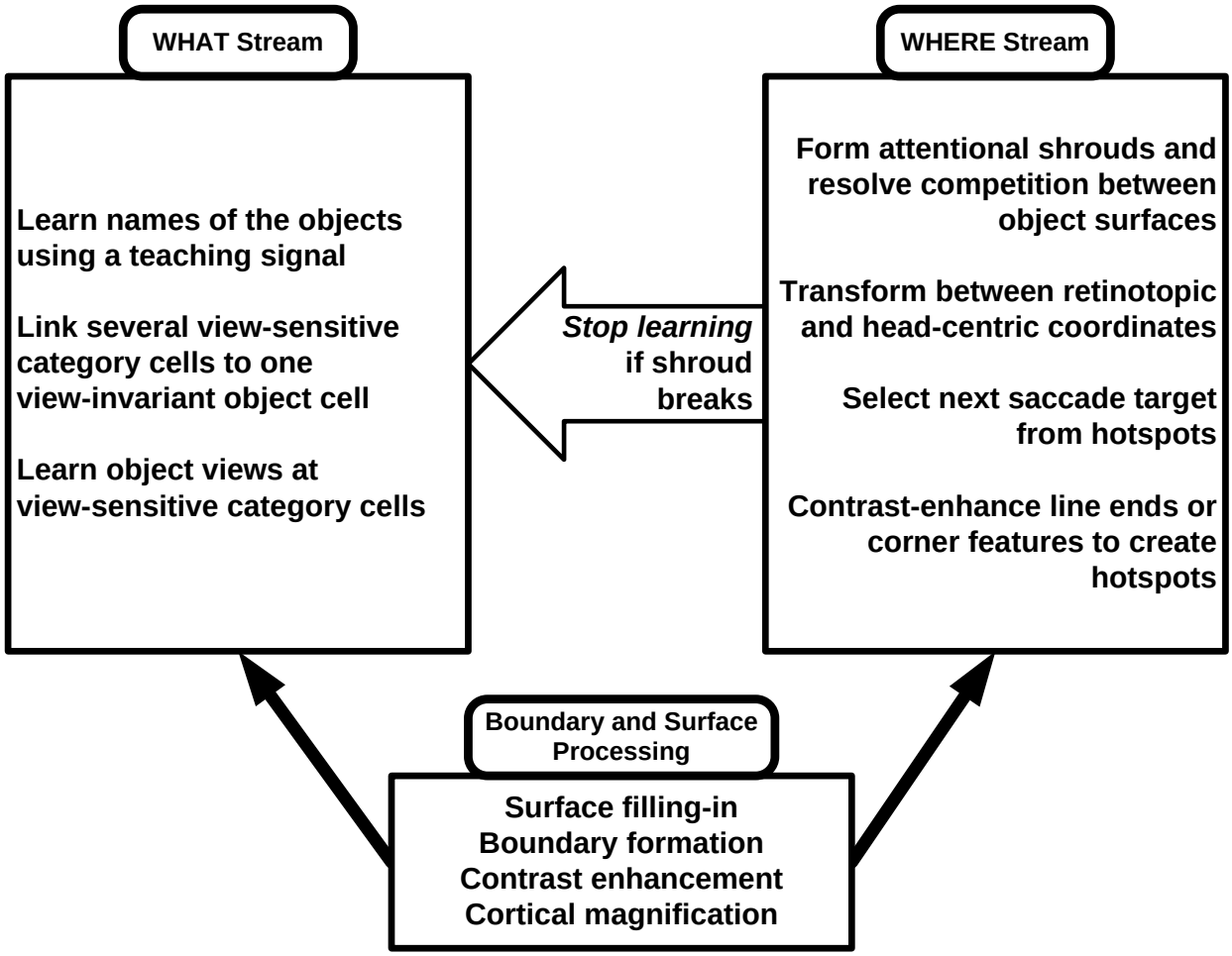

Figure 5: Overview of the ARTSCAN model. There are three main groups of processes in the model: Boundary and Surface, What Stream, and Where Stream processes. Boundary and Surface processes contrast-enhance the input image and perform a log-polar transformation to simulate cortical magnification before computing boundaries and using them to fill-in closed boundaries to form surfaces. The Where Stream changes the coordinates of image surfaces from retinotopic to head-centric, selects one surface and forms an attentional shroud around it. The attentional shroud feedback up-regulates the boundary and surface activities of the corresponding object. The Where Stream also finds hotspots on the attended surfaces and moves the eyes to the next most active one. This attentional shroud can habituate and *break*, at which point a reset signal to the What Stream stops learning and recognition. The What Stream is responsible for learning view-sensitive and view-invariant object categories, as well as the name of the object whose surface is currently attended in the Where Stream.

6. Spatial Object Search: Searching Surface Contour Hotspots with Goal-Oriented Saccades. ARTSCAN provides a way to understand how attention and eye movements can work together to intelligently search a scene. As shown in Figure 3c, attentionally-modulated surface contours input to the eye movement selection process, thus making it unlikely for an unattended boundary to win and be the next saccade target.

ARTSCAN builds upon an earlier model of visual search that was called the Spatial Object Search, or SOS, model (Grossberg, Mingolla & Ross, 1994). SOS showed how the following four interacting processes were sufficient to quantitatively simulate many challenging data about visual search: Boundaries, surfaces, spatial attention, and object attention. However, SOS was not embodied as a neural model. Rather, it is an algorithm whose properties emulate properties of BCS/FCS and ART. The ARTSCAN model begins to fulfill the promise of the SOS model by showing how real-time neural mechanisms can search a scene while learning the objects that are in it. ARTSCAN differs from the Guided Search 1 and 2 models (GS1 and GS2) of visual search (Wolfe, Cave & Franzel, 1989; Wolfe, 1994) in the way it treats objects and space. Objects in GS are localized so that a spotlight of attention can select an entire object. In ARTSCAN, objects are more realistic and occupy an extended area. Indeed, ARTSCAN explains how a spotlight of attention can spread through a surface and precisely select an entire object. Eye movements in GS, as well as in saliency map models (Itti & Koch, 2001), go from one point (or object) to another; whereas in ARTSCAN eyes move both within and between extended objects. Finally, GS models cannot explain object-based attention (Egly, Driver & Rafal, 1994) which in ARTSCAN is a natural by-product of ART category learning, whereby learned top-down expectations focus attention upon a prototype of object critical features.

What evidence supports the idea that spatial attention can direct eye movements? The brain's attentional mechanisms are known to be tightly linked to the control of eye movements. It is nearly impossible to attend to a location when moving our eyes to another location (Deubel & Schneider, 1996; Kowler, Anderson, Dosher & Blaser, 1995). The cortical Where stream has been linked to visual spatial representations, attention, and eye movements (Haxby *et al.*, 1991; Mishkin, Lewis & Ungerleider, 1982). The lateral intraparietal area (LIP) is considered to be a primary area responsible for both eye movements and visual attention (Andersen, Essick & Siegel, 1985; Colby, Duhamel & Goldberg, 1993). LIP neurons are generally active before a saccade and low voltage electrical stimulation in LIP results in an eye movement (Duhamel, Colby & Goldberg, 1992; Thier & Andersen, 1998). LIP is only part of a network that controls attention and eye movements and includes, but is not limited to, frontal eye fields (FEF), superior colliculus (SC), cerebellum, basal ganglia and brainstem nuclei. We do not herein model top-down control of attention on eye movements, notably by FEF to LIP/SC circuitry. For a review of relevant data, see Awh, Armstrong & Moore (2006) and Schall & Boucher (2007). For a recent eye movement control model that does include top-down FEF control of coordinated smooth pursuit and saccadic eye movements, see Grossberg, Srihasam & Bullock (2008) and Srihasam, Bullock & Grossberg (2008).

Is there evidence consistent with the prediction that a spatial-attentionally-enhanced surface representation can, through its surface contour output, guide the selection of eye movement targets on that surface until its attentional shroud collapses? Several studies show that not all parts of a visual scene are equally attractive for eye movements. Eye tracking experiments show that where the eyes look is both dependent on the mandates of the task, and the features of the scene (Findlay, 1995, 1997; Gilchrist, Heywood & Findlay, 2003). Empty and homogenous locations of the scene are seldom foveated, and lines, borders, and especially corners and intersections attract more fixations (Krieger, Rentschler, Hauske, Schill & Zetzsche, 2000). Given that object corners, intersections, and other singular features are the most informative parts of the scene, and that the fovea represents items with the highest resolution, looking at such singular object features ensures that the brain represents the potentially most informative parts of an object with the highest resolution.

The ARTSCAN model clarifies how the surface contour output to the eye movement system computes *hotspots*, or positions of enhanced activation, at singular features of an object, and thus directs eye movements to sequentially foveate with high probability on those object features that promise to be most informative for learning view categories of the object.

Although observers might recognize a simple object category in a single glance, they may scan a complex or unfamiliar object more thoroughly. Since in ARTSCAN surfaces are the units of attention, as long as a shroud up-regulates one object surface and its boundaries, hotspots on that surface can win the competition for the next saccade. The eye movement module is thus able to explore an attended surface at its most informative features.

This discussion raises the question of how spatial attention is organized. The surface representations that compete for spatial attention in shroud formation (Figure 3b) have been called Filling-In Domains, or FIDOs (Grossberg, 1994). FACADE theory predicts that there is a complete set of FIDOs corresponding to each of the depth-selective boundary representations that capture surface lightness and color at prescribed depths. At each such depth, such a complete set of FIDOs has been modeled as a pair of opponent filling-in domains (black vs. white, red vs. green, blue vs. yellow). As noted in Section 3, each FIDO's activity pattern is filtered by an on-center off-surround network that responds to local contrasts in the filled-in pattern; for example, to bounding contours of the object. This is the network that computes the surface contours in Figure 3c. In addition, each pair of opponent FIDOs is predicted to compete at each position; for example, red competes with green at each position to determine a net color. These two types of competition (spatial, opponent), acting together, define a double-opponent field of cells.

Let us imagine that this happens at the highest level of surface filling-in, where object figures are separated from each other and their backgrounds, and only the unoccluded parts of opaque objects are visible. FACADE theory predicts that this processing stage occurs in cortical area V4. Such an organization of surface representations easily explains how a unique conjunction of color and depth can pop out during a search experiment, as Nakayama and Silverman (1986) have reported.

Given this background, it is natural to predict that spatial attention in the parietal cortex inherits at least some of the double-opponent organization of the visible 3D surface representations in V4. If it does, then surface-based spatial attention has a separate region for each color and depth range (Nieman, Hayashi, Andersen & Shimojo, 2005). Suppose that volitionally-activated, spatially diffuse, priming occurs of the spatial attentional region that codes for a particular depth and color. This region can then activate its corresponding depth-and-color FIDO, and thereby preferentially activate all of the surface representations with that depth and color. This sort of mechanism can help to explain many search data about color-specific search; e.g. Egeth, Virzi & Garbart (1984) and Wolfe, Friedman-Hill & Bilsky (1994). By this mechanism, a human observer can learn how to break up a conjunctive search task into a color-priming operation followed by the type of pop-out that is explained in the previous paragraph.

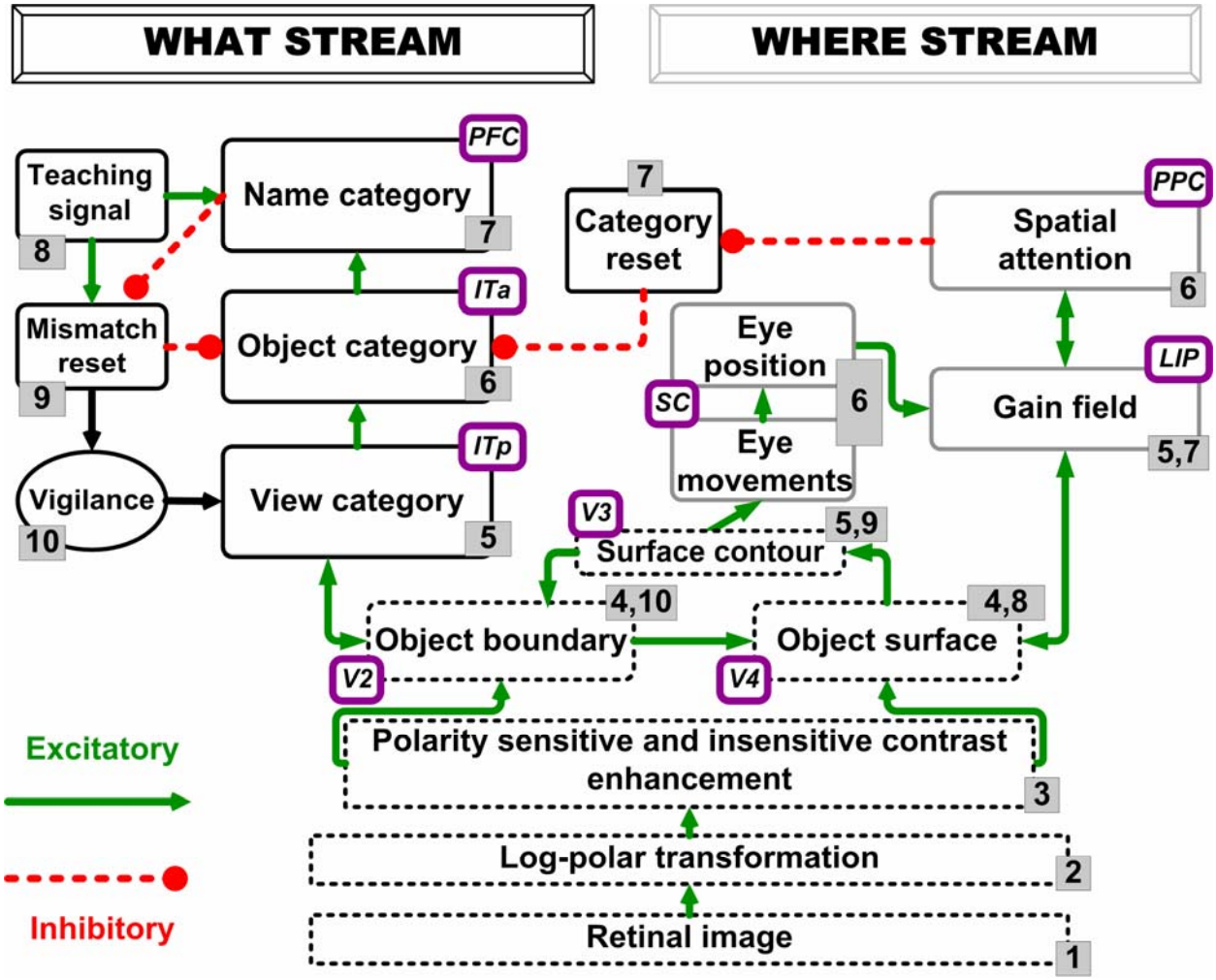

Figure 6: ARTSCAN model diagram. The Boundary and Surface processes have dashed borders and send input to both visual streams. The Where Stream modules have light grey borders and the What Stream modules have black borders. The small white tabs with round edges next to each box represent the anatomical region in which the process is thought to happen. The numbers in the grey boxes next to each module box show the approximate order of first activation in that module after the retina receives an input. If there are two such numbers in a box, the second one represents the time that feedback reaches that module. Solid arrows represent excitatory connections, and dashed connections with a round head represent inhibitory ones. ITa: anterior part of inferotemporal cortex, ITp: posterior part of inferotemporal cortex, LIP: lateral intra-parietal cortex, LGN: lateral geniculate nucleus. PFC: prefrontal cortex, SC: superior colliculus, V1 and V2: primary and secondary visual areas, V3 and V4: visual areas 3 and 4. See text for details.

7. Model Description. ARTSCAN has three main components, whose main operations are summarized in Figure 5: (1) Boundary and Surface Processing; (2) WHAT Stream, and (3) WHERE Stream. Figure 6 provides a macrocircuit of the main model processing stages. Each processing stage is defined by a network of interacting neurons whose membrane potentials vary

dynamically through time in response to inputs and feedback signals. The number beside each processing stage represents the temporal order in which that stage receives inputs in a typical processing cycle. Figure 7 illustrates model circuit interactions more completely.

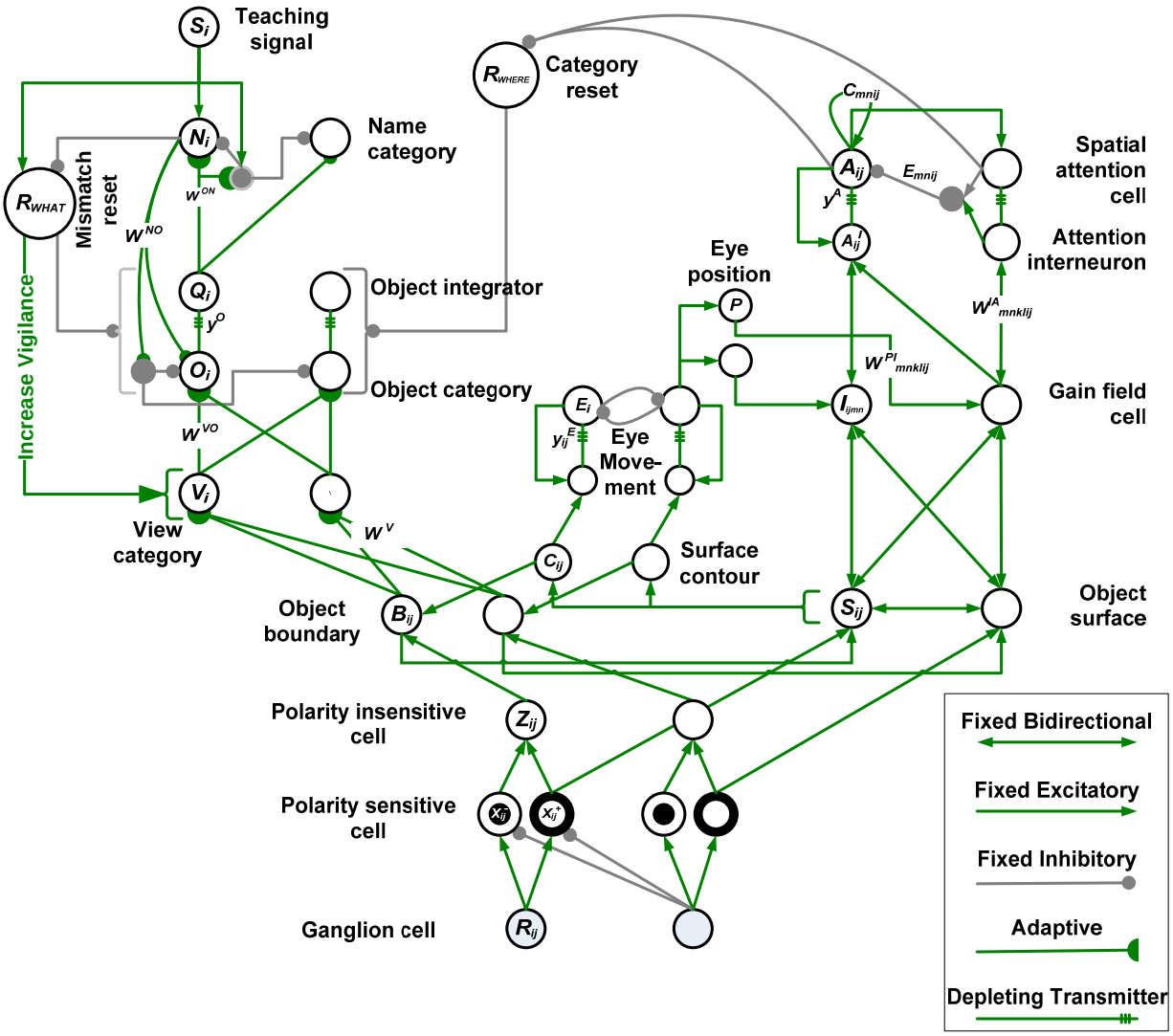

Figure 7: Model connections and variables. The layers correspond to the box diagram in Figure 6, and for convenience, they are placed in approximately the same relative position on that map. Some modules correspond to more than one layer of neurons. For example, the contrast enhancement module in Figure 6 corresponds to polarity-sensitive neurons and polarity-insensitive neurons in Figure 7, the object category module in Figure 6 corresponds to object category neurons and object integrator neurons in Figure 7, and the spatial attention module includes both attention interneurons and spatial attention neurons. The rest of the layers are the same as the modules in Figure 6. Only two neurons are shown for each level. The letter inside each neuron refers to the variable used to represent its activity in Appendix 1. Also shown are the variables used to represent the synaptic strengths between some layers in the Appendix 1.

7.1. Boundary and Surface Processing. The retina samples the image in a space-variant manner via the cortical magnification factor (Daniel & Whitteridge, 1961; Fischer, 1973; Schwartz, 1980; Tootell, Silverman, Switkes & De Valois, 1982; Van Essen, Newsome & Maunsell, 1984), with objects close to fovea represented in high resolution and those in the periphery with low resolution (Figure 8). The high resolution foveal representation facilitates object recognition. The low resolution peripheral representation provides suitable commands for eye movements, and whole field information needed for scene perception. We simulated this property by transforming the space-invariant image by a log-polar transformation (Schwartz, 1980); see Figure 6, steps 1 and 2, and Appendix 1 Equations (4)-(6).

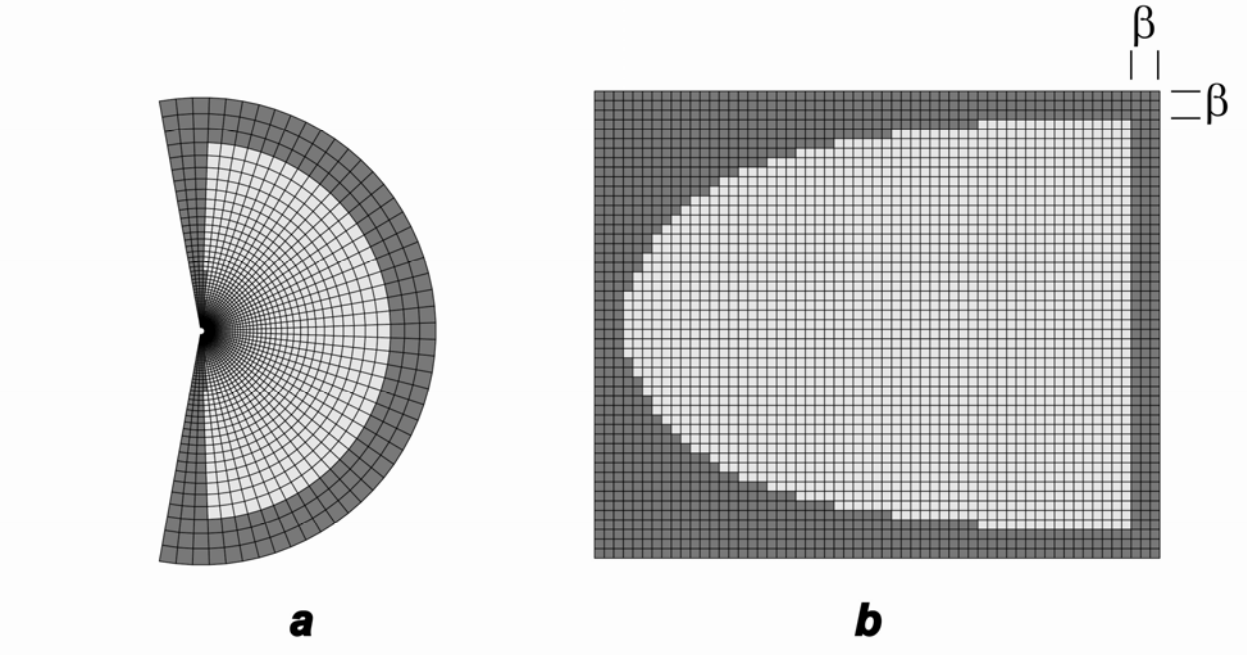

Figure 8: Topographical mapping from half of a retina to its corresponding V1 cortex. (a) The sampling regimen in half of a retina. Each small box on the retina represents the receptive field of one retinal ganglion cell. These receptive fields get larger and the ganglion neuron density gets sparser toward the periphery of the retina. The light grey receptive fields are those that send input to V1. To avoid border sampling artifacts along the vertical meridian in V1, the dark grey cells, which actually belong the opposite half of the retina, are sampled as well to act as border padding. (b) The corresponding V1 representation of (a). The light grey cells are the actual V1 neurons and the dark grey cells acts, which correspond to darker cells in (a), act as padding to offset the sampling bias along V1 borders.

Figure 8a shows the retinal sampling in half of one retina. The light grey cells are the receptive fields of those ARTSCAN retinal cells that send input to the LGN in one hemisphere. Figure 8b shows V1 cortex corresponding to the hemi-retina in Figure 8a. Dark grey cells in Figures 8a and 8b are locations "padded" outside sampled locations in order to avoid image border sampling artifacts; see Appendix 2.

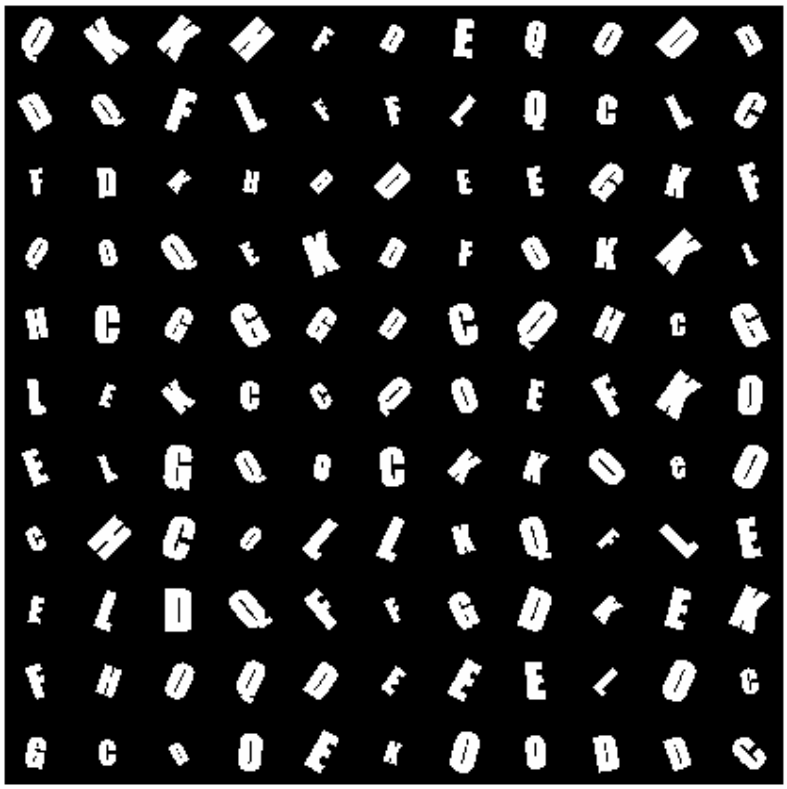

Figure 9: Part of the letter database scene. Each object in the scene is one of ten selected letters of the alphabet and is rotated and/or resized (see text for details). Each letter has a uniform luminance value; the background has zero luminance. Each letter size is about 10˚ x 10˚ in visual angle. The white dashed square shows the area that fits on the model retina. The log-polar representation of this square is shown in Figure 10.

The scene that is the input to the model is filled with 2D forms of letters that do not overlap and are at the same depth (Figure 9). Figure 1 shows such a letter's boundary (Figures 1b-e) and log-polar representation (Figures 1c-f). Objects in this database do not have illusory or missing contours or occlusions. This enables us to simplify several processing stages to speed up simulation times without a loss of critical model properties.

Model retinal cell activities are normalized by the receptive field surface area (Appendix 1 Equation (6)). These normalized signals are the inputs to the model LGN. This is followed by contrast normalization of the input pattern in the model LGN by polarity-sensitive on and off cells. On (off) cells obey cell membrane, or shunting, equations and receive retinal outputs through an on-center off-surround (off-center on-surround) network (Figure 7, Appendix 1 Equations (7) – (10)). The model omits oriented simple cell receptive fields, and properties of ocularity and disparity-sensitivity that are found in the primary visual cortex. Processing the letters in Figure 9 does not require these refinements. These properties are modeled in FACADE articles, and can be added to future model developments. Polarity-insensitive neurons (simplified complex cells; Figure 7, Appendix 1 and Equation (11)) are computed as a sum of rectified signals from polarity-sensitive neurons of opposite polarity at the same position. These cells generate bottom-up inputs to the object boundary stage (Figures 6 and 7, Appendix 1 Equation

(13)), which also receives top-down inputs from the surface contour cells (Figures 6 and 7, Appendix 1 Equation (17)).

The on-center polarity-sensitive cells also provide the bottom-up inputs that drive filling-in of object surface representations. The object boundaries generate signals that gate the diffusive surface filling-in process (Figure 6, Appendix 1 Equations (15) – (16)). Filling-in reconstructs surfaces that are surrounded by closed boundaries and can contain the spread of the surface feature inputs.

The filled-in surfaces generate contour-sensitive output signals via the surface contour process, which also consists of a shunting on-center off-surround network (Figures 6 and 7, Appendix 1 Equation (17)-(19)). Surface contour outputs project back to the object boundaries that induced filling-in of their surface region (Figure 7, Appendix 1 Equation (13)). This excitatory feedback strengthens boundaries that lead to successful filling-in—that is, closed boundaries—and inhibits those boundaries that do not. Figure 10a shows a foveated letter G boundary on the log-polar map before surface contour feedback. Figure 10b shows the foveated G boundary after surface contour feedback is received from a surface region whose activation is enhanced by an attentional shroud. The foveated letter G boundary is more active than the boundaries of all the peripheral letters in the scene, and will become the source of eye movement commands (Figure 3c).

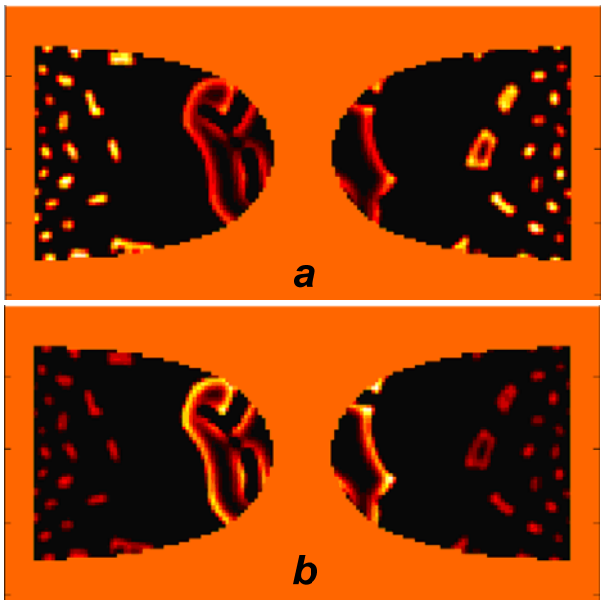

Figure 10: Activity in the object boundary map in response to the dashed part of the letter database scene in Figure 9. Half of a tilted letter **G** falls on each hemi-field. Each half of G is represented in one foveal region, and the nearby letters are compressed to tiny foci of activity in the periphery. The brightness of each pixel on the map shows the level of activity of the object boundary neuron in that location. (a) Activation without any attention feedback from surface contour map. Even those small areas in the periphery that represent a letter have a similar level of activity to the foveated letter; i.e., they have the same brightness on the map. (b) Activity with attention feedback from surface contour. Attention has up-regulated the representation of letter **G** boundaries, which are now far more active than the boundaries of nearby letters.

7.2. WHERE Stream. The ARTSCAN Where cortical stream enables an attentional shroud to be focused on one object surface, out of many, in the What stream. Enhanced surface contour input to the object boundary, again in the What stream, enables a sequence of saccadic eye movements, commanded from the Where stream, to selectively explore that object surface. At the same time, the What Stream learns the view categories resulting from each of those fixations, and links them through associative learning to the same emerging object category, as long as a single shroud remains active. Two factors conspire to ensure that only view categories of the same object are linked to the same object category: The enhanced object boundaries tend to restrict eye fixations to views of the shrouded object surface, and the collapse of the shroud inhibits the active object category. The processing steps in the Where Stream are explained in more detail below.

Retinotopic Surfaces and Head-centric Shrouds. The object surface neurons serve as the input to the Where Stream (Figure 3b; Figure 7, Appendix 1 Equation (15)). Object surfaces are computed in retinotopic coordinates, but surface attention needs to enhance an entire surface regardless of where the eyes are at any moment; that is, in coordinates insensitive to eye movements, notably head-centric coordinates. A neuron in a head-centric map, by definition, responds to a fixed location of the head regardless of where the eyes look. To change coordinates, object surface input is combined with eye position signals in the gain field module to generate a head-centric spatial attention map in the parietal cortex (Figures 6 and 7). Such gain field modulation is known to occur in posterior parietal cortex (Andersen, Essick & Siegel, 1985; Andersen & Mountcastle, 1983; Deneve & Pouget, 2003 ;Gancarz & Grossberg, 1999; Pouget, Dayan & Zemel, 2003). Each gain field neuron's response to a retinal location is thus modulated by eye position. Pouget & Snyder (2000) showed that combining the responses of several such gain field neurons can give rise to a head-centric map. The weights between the gain field neurons and the spatial attention neurons are presumably learned. For simplicity, we used the end product of such a learning process, as suggested by Pouget & Snyder (2000). Appendix 1 Equations (20)-(23) mathematically describe the gain field transformation.

The head-centric spatial attention neurons (Figure 7, Appendix 1 Equations (25)-(31)) receive bottom-up input from gain field neurons. The spatial attention neurons interact via recurrent on-center off-surround interactions whose large off-surround enables selection of a winning attentional shroud. These recurrent on-center interactions enhance the winner shroud, and enable this shroud to remain active as other attentional neurons are persistently inhibited. Figures 11a and 11e show how letters E or L, respectively, can be selectively enhanced by such a shroud.

The spatial attention neurons send top-down feedback to the gain field neurons (Figure 6, Appendix 1 Equation (20)), and from there back down to the object surfaces. The model hereby posits a resonant surface-shroud feedback loop between retinotopic surface representations and head-centric spatial attentional shrouds.

The bottom-up inputs to the spatial attention neurons are gated by habituative chemical transmitters (Appendix 1 Equation (32)), which play an important role in *inhibition of return*. In particular, the level of available transmitter decreases as activity increases in the corresponding gain field and spatial attention neurons (Appendix 1 Equations (25) and (32)). The increased activity of the shroud around the E shape in Figure 11 gradually depletes its habituative transmitter gates, and thereby weakens the net bottom-up inputs that support this shroud (Figure 11b). When the shroud collapses, inhibition of other attentional positions is eliminated, and other

surfaces, in this case, the one corresponding to the inverted letter L, can form a new active shroud (Figure 11e).

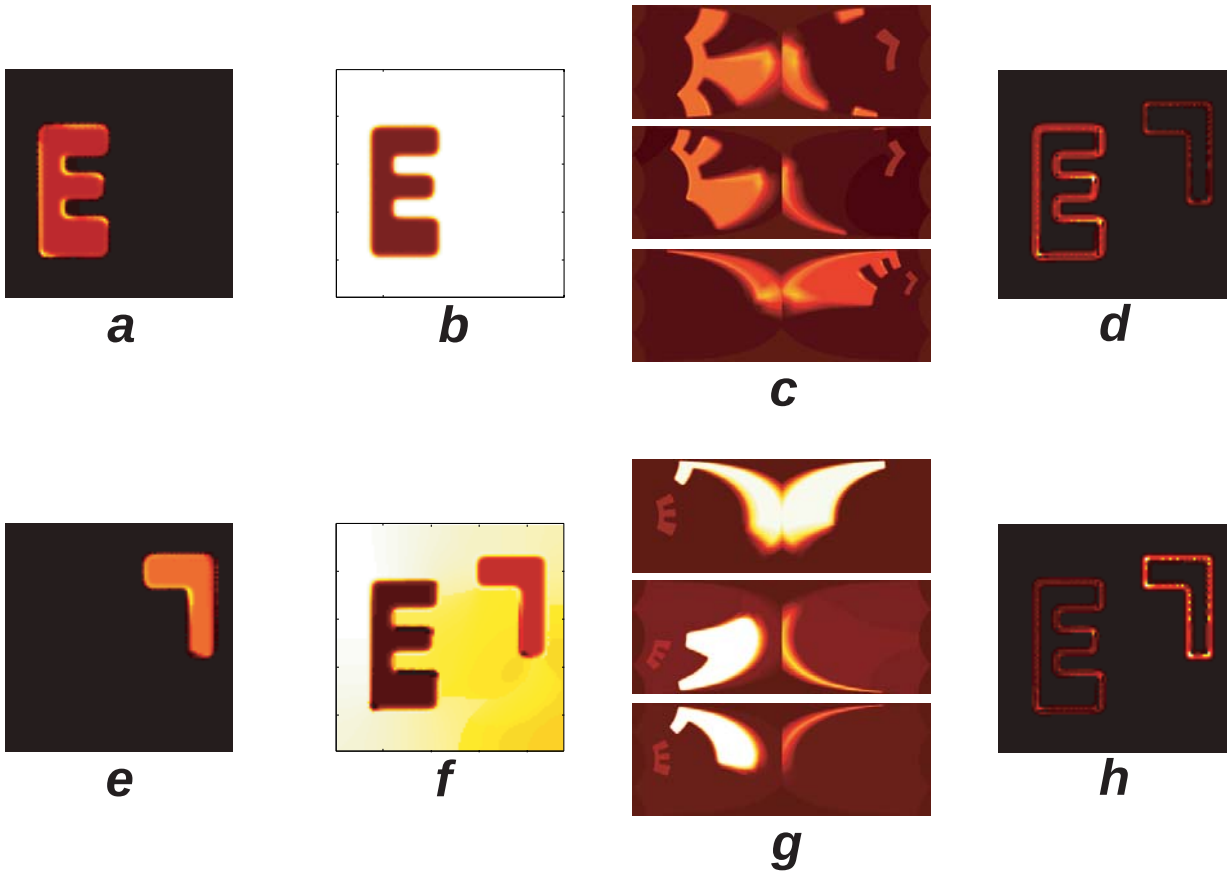

Figure 11: Model simulations of Where stream. The retinal input to this simulation is a small scene containing only two letters: an **E** and an inverted **L**. Neural activity in different ARTSCAN Where stream modules are shown as an attentional shroud first forms around the representation of a letter **E**, runs out of habituative transmitter gate, and moves to the nearby inverted letter **L**. Darker colors represent lower values. (a) Spatial attention map activity when the attentional shroud has formed around letter E. (b) Habituative transmitter levels during the same time as (a). Note that the shroud is running out of transmitter around letter **E**. (c) Three consecutive object surface map activities around the time of (a) and (b) as a result of three saccades and fixations on the letter **E**. Due to faster dynamics in the eye movement module compared to spatial attention module, several exploratory saccades can happen on a surface while it is attended. (d) Surface contour map activity corresponding to the time in (a). The corners of letter **E** are among the most active areas on this map and will serve as consecutive saccade targets while the shroud lasts on this letter. This map is also retinotopic and in log-polar coordinates, but for clarity it is represented in Cartesian coordinates. (e) to (h): the same model stages as in (a) to (d), but in a later point in time. (e) Attentional shroud has collapsed around letter **E** and has moved to around the inverted letter **L**. (f) The transmitter levels have already depleted around the letter **E** and are habituating around the letter **L** as well. (g) Three exploratory saccades on letter L have resulted in three

different surface map activations. (h) Surface contour activity shows higher activity for **L** contours in general, and its corners (hotspots), in particular.

As noted in Section 1, spatial attention may at first cover only a small part of a surface representation, as when a top-down, volitionally-activated, attentional spotlight happens to hit some positions covered by the surface. Top-down feedback from spatial attention to surface (Figure 6) enables the attentionally enhanced locations to spread their activation throughout the surface by filling-in, whence bottom-up activation from surface to attention can cause a form-fitting spatial attentional shroud. If this top-down attention reaches positions that are not enclosed by a closed boundary, it dissipates away and does not enhance that area. Figure 11c shows three such enhanced surface representations in retinotopic, log-polar coordinates that result from three consecutive fixations on the attended letter E.

Object Category Reset. A critical property of the active shroud is to inhibit the category reset model stage (Figures 6 and 7, Appendix 1 Equation (35)). The category reset stage in the Where stream is modeled by a tonically active neuronal population that nonspecifically inhibits all the object neurons in the What cortical stream (Figures 6 and 7, Appendix 1 Equations (40)-(41)). All active cells in the spatial attentional network, notably all the cells in the currently active shroud, inhibit the category reset stage. Thus, while any part of the shroud remains active, category reset remains inhibited, so that the currently active object category can remain active. When the currently active shroud collapses, inhibition of category reset ceases, and the tonic activity of the category reset neurons is disinhibited, thereby enabling category reset neurons to inhibit the currently active object category. Because one shroud needs to collapse before another one can form, there is some time lag between activation of two successive shrouds. The category reset signal is activated between such shifts of attention between surfaces.

A Unified Explanation of How Attention Moves, Engages, and Disengages. Posner (1980) and Posner, Walker, Friedrich, & Rafal (1987;1984) proposed that attention is controlled by three basic operations: move, engage, and disengage. ARTSCAN provides a unified mechanistic account whereby attention can be disengaged, moved, and engaged by different object surfaces. *Engage* attention occurs when an attentional shroud forms around an object's surface representation. *Disengage* attention occurs when an active attentional shroud weakens and collapses. *Move* attention occurs during the time after the breakdown of one shroud and before the full formation of the next active shroud. As noted above, the operations whereby attention can engage, disengage, and move between object surfaces is influenced by the integrity of several parts of the overall ARTSCAN architecture, including the ability of object surfaces to form and the eyes to move across and between surfaces in response to hotspot movement commands.

Eye Movements to the Attended Surface Hotspots. The eye movement map is a retinotopic motor map that gets its input from the surface contour neurons (Figure 7, Appendix 1 Equation (33)). The surface contours are enhanced by the active shroud via the corresponding surface representation (Figure 3c). The attended surface's boundaries are thus typically the most active surface contours (Figures 11c and 11g). The eye movement map is a winner-take-all map, which selects the most active spot ("hotspot") on the attended object's boundaries. In our alphabet database, these hotspots usually correspond to points of high curvature, such as corners and intersections (Figures 11d and 11h), as also often occurs in human and animal eye movements.

The input pathways from surface contours to the eye movement map are gated by habituative transmitters. As the inputs that support foveation of one hotspot habituate, the eye can move

from one hotspot to the next on the attended surface. When the shroud collapses and another shroud forms, hotspots on the newly attended surface will become eye movement targets.

7.3. WHAT Stream. The ARTSCAN What cortical system is responsible for learning view categories, view-invariant object categories, and names of objects in a scene. The What stream stops learning and resets, either when the eyes move off an attended surface, or when it incorrectly guesses the identity of a surface. There are three main layers in the model's What Stream, pertaining to three different areas of the visual system in the brain (Figure 6): (1) The view category neurons corresponds to the posterior parts of inferotemporal visual cortex (TEO or ITp) (Pasupathy, 2006); (2) Object category neurons correspond to anterior parts of inferotemporal cortex (ITa) (Tanaka, 1997, 2000); and (3) name category, or name, neurons, correspond to medial temporal/prefrontal areas of visual system (Rainer & Miller, 2000;Ranganath, 2006).

Learning of View Categories. The inputs to view category neurons are attention-modulated, coarse-coded object boundaries (Figures 3d and 6, Appendix 1 Equations (36)-(39)). They learn to respond to a range of changes in object boundaries due to different sizes, orientations, and nearby gaze points on the same object view. Coarse-coding of the object boundaries increases the tolerance of the view category neurons to such changes. The model uses an Adaptive Resonance Theory, or ART, classifier, namely Fuzzy ARTMAP (Carpenter, Grossberg, Markuzon, Reynolds & Rosen, 1992).

Learning of Object Categories. The object category neurons are activated by multiple view categories through associatively learned connections (Figures 12a-b , Appendix 1 Equation (40)). As noted above, object category neurons do not necessarily get reset when view categories that correspond to the same object get reset. They get reset when the shroud corresponding to a given object gets reset, attention shifts to another object, and the eyes begin to explore the new object. They also get reset when the name category that they represent mismatches the teaching signal (Figure 3h); see below.

Accumulation of View-based Evidence by Object Integrators. Object category neurons are divided into two subpopulations: Object category neurons, which are the neurons just described, and object integrator neurons (Figures 3d-h and 12e, Appendix 1 Equation (41)). The object integrator category neurons encode accumulating evidence for an object by increasing their activity as new viewpoints on the object surface are foveated. The pathways between object category and object integrator neurons contain habituative transmitter gates that convert input increments, no matter how long they are sustained, into stereotyped transient input bursts (Figure 12c-d, Appendix 1 Equation (42)). In effect, the integrator neurons count the number of views by adding up these bursts.

Figure 12 summarizes simulated model dynamics of the What stream after three views and two objects and names have been learned (Figure 12a). The three view categories get activated as a result of three consecutive fixations (Figure 12b, top of the panel). The next panels show the activities of object category neurons (Figure 12b), habituative transmitters (Figure 12c), habituatively gated object category outputs (Figure 12d), and object integrator neurons (Figure 12e). Note the transient nature of the object category output signal, even when these neurons are kept active by bottom-up input from view categories. Also note evidence accumulation in object integrator neuron 1 as a result of consecutive activation of two associated view category neurons. The name category and the mismatch reset neuron activities are shown in Figures 12f and 12g.

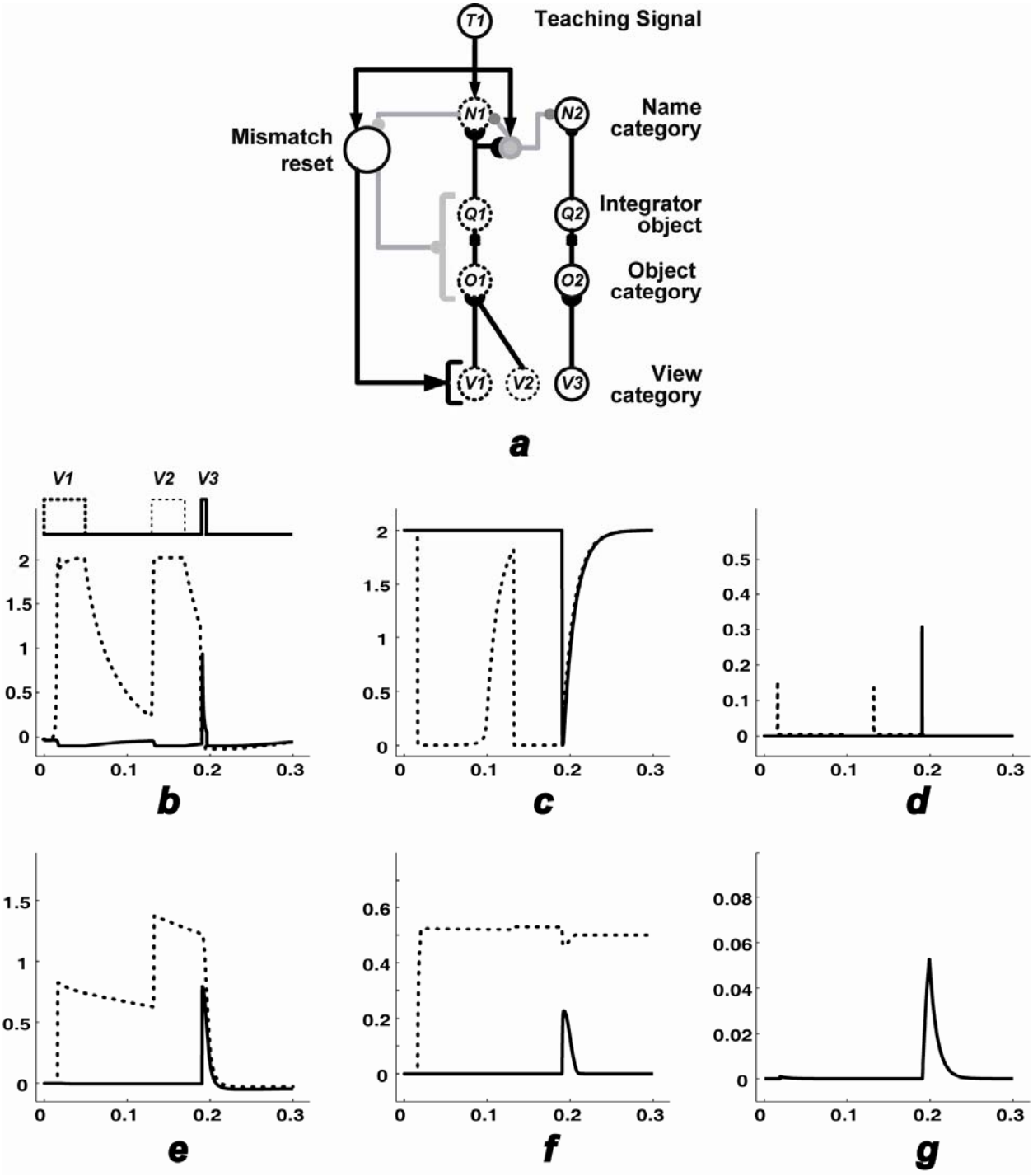

Figure 12: What Stream dynamics. (a) The simplified circuit in this example has just two neurons in each layer except for the view category layer, where it has three neurons. View category 1 and 2 (V1 and V2), object category 1 (O1), object integrator 1 (Q1), and name category 1 (N1) activities are shown in dashed lines in (b) to (f). Previous learning has already associated the V3 with O2 and N2, whose activities are shown in solid lines in (b) to (f). In addition, N1 gets the supervision signal during the entire simulation time. (b) Suppose consecutive fixations activate V1, V2, and V3 in that order (top of the

figure). V1 learns to activate O1 (dashed line) from time 0 to 0.05 seconds. A saccade resets V1 and O1 loses its input and starts to shut down. V2 gets active from time 0.13 to 0.17 and learns to re-activates O1, because O1 is still the most active object neuron. V3 gets activated from time 0.19 seconds and strongly activates O2 (solid line). All object and view neurons are reset by a mismatch reset neuron, see below. (c) The transmitter gate between each object category neuron and its corresponding object integrator neuron habituates when its object category turns on and replenishes after the latter decays below a threshold. (d) The output signal of object category neurons is the product of their activity and their habituative transmitters, which form transient responses. (e) Q1 (dashed line) accumulates the two instances of O1 activities. Q2 activation (solid line) creates a mismatch in the name category layer (f) and results in the mismatch reset signal activation (g). (f) N1 (dashed line) receives a teaching signal T1 as well as the Q1 output. Initiation of activity in Q2 (solid line in (e)) starts to activate N2 (solid line). Due to lateral shunting inhibition, both name neurons down-regulate one another to below a threshold of 0.5 and none of them can inhibit the mismatch reset neuron anymore. This is the time that the mismatch reset neuron gets activated. (g) The mismatch reset neuron is not active before time = 0.20, as N1 activity inhibits the excitatory effect of T1 on that neuron. After t = 0.20, the reset neuron escapes the inhibition of N1, and inhibits the object layers. As all of the object integrator neurons get reset, the input to the name layer is shut off, and N2 is totally turned off (f), but N1 continues its activity because it still gets input from its teaching signal.

Name Neurons and Supervised Learning. The above learning processes work under either unsupervised or supervised learning conditions. The object integrator category neurons input to name category, or name, neurons (Figures 3g-h, 6 and 7, Appendix 1 Equation (43)). The name category neurons also receive a teaching signal on those learning trials that are supervised (Appendix 1 Equations (43)-(44)). Even without supervision, the model explores attended surface hotspots and learns to associate all the resultant view categories with the same object neuron; it just does not label the learned object category with any name.

Two Modes of Object Category Reset: Shroud Collapse and Mimatch Reset. Object category and object integrator neurons can be reset in two ways. We already discussed the reset signal from the Where stream which operates whenever the object's attentional shroud collapses. Another reset signal occurs within the What stream to reset a view neuron when the object's view changes so much that the current view neuron cannot assimilate it into its learning; that is, when the current bottom-up input from the object mismatches the learned top-down prototype of the currently active view category. This sort of reset is called *mismatch reset* and is a basic property of ART models (Carpenter & Grossberg, 1987). Mismatch reset cells receive excitatory input from teaching signals and inhibitory input from all name category neurons (Figures 3g-h, 6, and 7, Appendix 1 Equation (45)). When the total excitatory input exceeds the total inhibitory input by a sufficient amount, the mismatch reset signal is activated and inhibits the currently active object category neurons (Figures 3g-h, 6, and 7). Activation of the mismatch reset stage inhibits an active view neuron by increasing a *vigilance* parameter (Figures 6 and 7, Appendix 1 Equation (38)). Vigilance is increased just enough to cause a search for other view category neurons that can learn how to form a better match with the bottom-up boundary input.

In summary, object category and object integrator neurons can maintain their own activity while multiple views of the object are explored. These neurons can be actively inhibited in two

ways: when attention shifts to another object (via disinhibition of Where stream reset neurons), or when the predicted object category name mismatches the externally supplied object name (via What stream mismatch reset neurons).

Resolution of Predictive Conflicts. Mismatch reset can occur when two or more name category neurons are active simultaneously—that is, represent a predictive conflict—because the predicted object category name is not the same name as the externally supplied name. The name neurons interact via shunting on-center and off-surround interactions (Appendix 1 Equation (43)). Due to the shunting inhibition between name category neurons, the total activity of the name neuron network tends to be conserved (Grossberg, 1980b). As a result, when more than one name neuron is activated, the activity of individual name neurons decreases and none of them can inhibit the mismatch reset neuron (Appendix 1 Equation (45)). The mismatch reset neuron is then excited by the teaching signal, and leads to the reset events described above.

Name Priming Leads to Greater Compression During Supervised Learning. Active name category neurons can bias or prime the object categories through top-down attentional feedback (Figures 3g-h, 6, and 7, Appendix 1 Equation (40)). A primed object category has a greater chance of getting selected and learning a view neuron which is not yet committed to any other object category. Top-down priming hereby helps to compress memory from view neurons to object neurons during supervised learning trials. In the absence of supervision, there is a higher chance that novel views of an object will become associated with different object categories.

Although in the above model description, each processing stage of the model was separately discussed, it should be noted that the entire system is a dynamical system operating in real time. The simulations in Figure 12f and 12g illustrate model dynamics in the name and mismatch reset neurons.

To summarize What stream functions: Automatic formation of an attentional shroud around an object in the Where stream causes the boundaries of that attended object to have more activity than those on other nearby surfaces, thereby directing the eyes to move between hotspots of the attended surface. Each foveated object boundary adequately matches the top-down expectation of a previously learned view category or, if it is a novel input, activates and trains an uncommitted view category. These view neurons get associated with object category neurons, whose repetitive activation accumulates evidence at object integrator neurons. The latter remain active as long as there is no conflict in the name category, or a collapse of the shroud. If the teaching signal provides a name and activates the corresponding name category neuron for the presently viewed object, the activated object integrator neurons get associated with that name. If no name is provided, only the weights between view and object category neurons are learned. A conflict arises when the teaching signal activates one name category neuron and the bottom-up input from the object layer activates a different name category. This mismatch inhibits the object layers, resets the currently active view category, and triggers search for a better matching view category through vigilance control.

8. Simulation Results

8.1. Simulations of Spatial and Object Attention Psychophysical Data. As noted in Section 7.2, Posner (1980) proposed that three processes control attention: (1) *disengage* attention from the current location, (2) *move* attention to the new location, and (3) *engage* attention at the new location. The Posner (1980) terminology addresses spatial attention, where units of attention are single locations, or a spotlight of attention. Since then, attention was shown to have object-based properties (Duncan, 1984). A typical trial in experiments to test these

operations begins with presenting one two objects (bars) and cuing one of its ends so that attention is initially drawn to that end (Egly, Driver & Rafal, 1994). A target then appears in one of four types of locations: on the same cue location (valid cue, Figure 13a), on the other end of the same bar (invalid cue with intra-object attention shift, Figure 13d), on another object (invalid cue with object-to-object attention shift, Figure 13g), or on another location outside the cued object (invalid cue with object-to-location attention shift, Figure 13j).

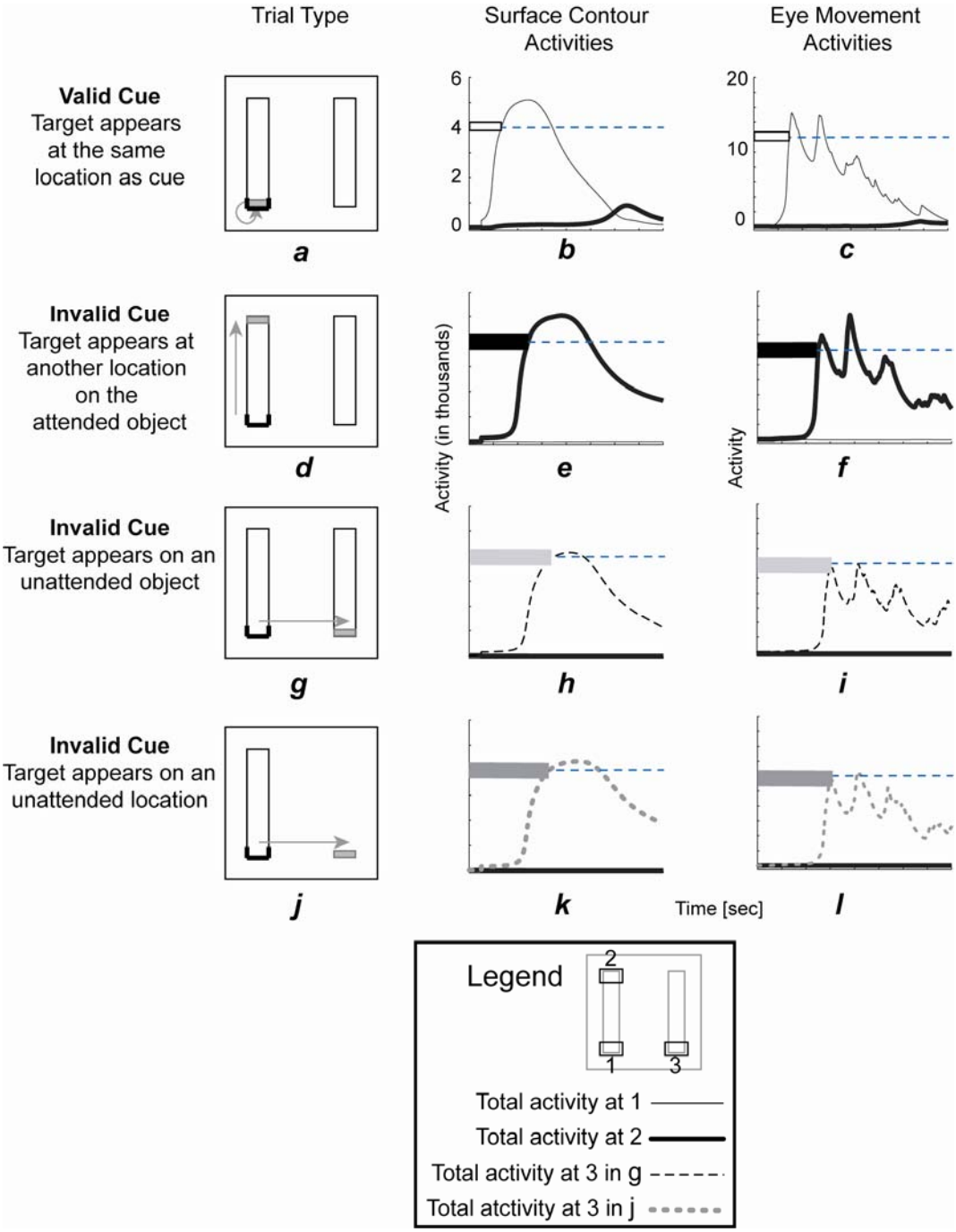

Figure 13: ARTSCAN simulation of object-based attention. The left column shows the trial types described in the text. The middle column shows surface contour (SC), and the

right column shows eye movement (EM) activities. Each row of SC and EM corresponds to the trial type in that row. As shown in the legend, for SC and EM graphs, the total activity around the cue location is shown in solid grey line, that around the other end of the cued object is shown in solid black line, that around the lower side of the unattended object is shown in thin dotted line, and that around the target location in Figure 13j is shown in thick dotted line. Response time is the time when the activity on any map reaches an arbitrary threshold, shown as a dashed horizontal line in all graphs. (a) Valid cue condition. The arrow schematically shows that the attention does not shift in this condition. (b) and (c): Total activity around the cue and target locations in both SC and EM maps in valid trials. While the activity on SC map rises smoothly, that of EM map has several peaks. This is because several points in the averaged area get very active and win as saccade target while attention forms around the object. (d) Invalid-same-object trials, where attention moves from the cued end of an object to its other end. (e) and (f): The CS and EM maps activities in the described locations. (g) and (j): Invalid trials with target either appearing on another object (g) or on another location (j). (h) and (k): CS map activities corresponding to (g) and (j), respectively, and show very similar rise time and peaks. These reaction times are slower than (e), showing the object-based attentional effects. (i) and (l): EM map activities corresponding to (g) and (j), respectively.

Brown and Denney (2007) showed that *inter-object* (Figure 13g) and *object-to-location* (Figure 13j) shifts of attention take longer than *intra-object* shifts (Figure 13d) because of the longer disengagement processes in the former conditions. Moreover, they found that shifting attention from an object to another object, or to another location, takes nearly the same amount of time (369 ± 10 msec versus 376 ± 9 msec, $p > .87$ in Figure 14a). Respecting the Posner terminology, the only event different across Figure 13g and 13j trials was the engagement of attention to an object. Thus Brown and Denney concluded that the engagement of attention is not the time-limiting bottleneck in object-based experiments and that intra-object advantage occurs because we do not need to *disengage* our attention from the object as we move inside it.

ARTSCAN simulates the longer reaction times in the inter-object as well as object-to-location attention shifts compared to intra-object attention shifts. The main reason for this is that it takes time for an attentional shroud to collapse before any other location or object can form a new shroud. Reaction time in each trial was computed in the model as the time it takes for surface contour or eye movement activity at the target location to reach a prescribed threshold. The ARTSCAN model had already been used to learn the letter database that is described in Section 8.2 before the Brown & Denney data became available. The same parameters that were developed to learn the database successfully simulated the Brown & Denney data as well. This fact supports the main hypothesis of the present work that shroud-based attentional control is used to learn view-invariant object categories under both unsupervised and supervised learning conditions.

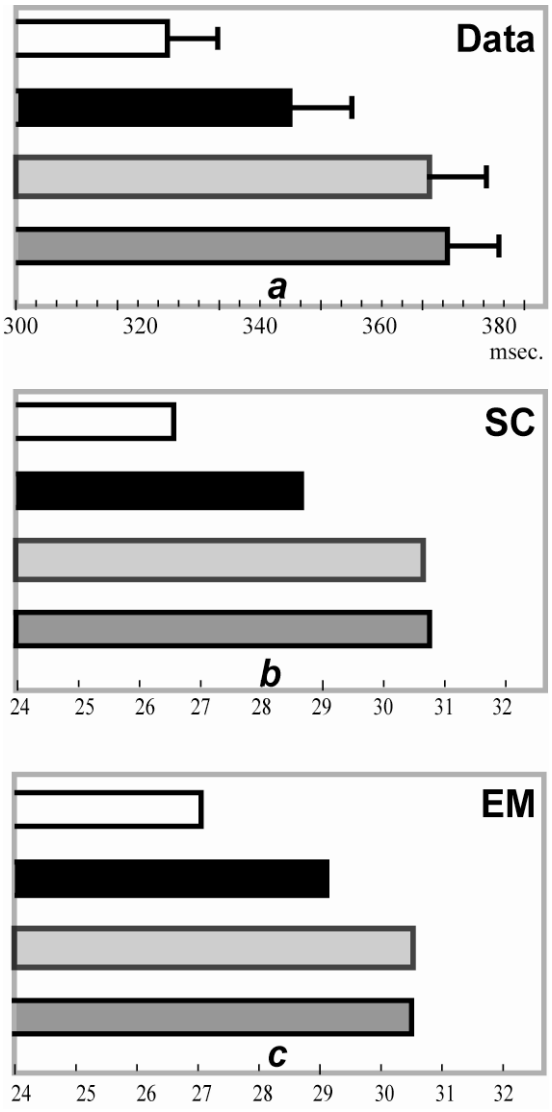

Figure 14: Comparison between the reactions times in experimental data and simulation. In (a)-(c), white bars correspond to valid trials. Black bars, light grey bars, and dark grey bars correspond to invalid within object, invalid between object, and invalid object-to-location trials. (a) The experimental results of Brown & Denney (in press). (b) The simulation reaction times based on surface contour activity, as shown in Figure 13 (b), (e), (h), and (k) respectively. Similar to the experimental data, valid trials (white bar) have the fastest RTs, followed by those in invalid-within-object trials (black bar). The slowest RTs are in the invalid between-object and object-to-location trials (the light and dark grey bars, respectively). The latter are statistically similar in the data (369 ± 10 msec and Object to location shifts were 376 ± 9 msec, $p > .87$), as well in the simulation. (c) The simulation reaction times based on eye movement activity, as shown in Figure 13 (c), (f), (i), and (l), respectively. RTs are similar to (b) as well as to the experimental data. For demonstration purposes, the simulation results are scaled such that valid trial RT in the simulation is equal to the valid trial in the data.

Presenting a cue on the lower end of the bar causes the shroud to form around the entire bar which, through feedback, increases activity at the corresponding locations on object boundaries, surface contours, and eye movement commands (Figure 3c). A target will later appear either on the same location (Figure 13a), on the other end of the bar (Figure 13d), at the same distance outside the bar on another bar (Figure 13g), or at the same distance outside the object (Figure 13j). For our simulations, the size of input display was 95 by 95 pixels, each bar was 57x16 pixels with a border width of 1 pixel. The bar border pixels had a luminance value of 0.5. The cue had a thickness of 2 pixels and a luminance of 1 for each pixel. The target was a small rectangle of 16x3 pixels of luminance 1. The ISI between cue and target was 0.25 seconds. To generate graphs of model responses, and calculate reaction times, the activity on surface contour or eye movement maps inside an imaginary rectangle of 20x7 pixels at the probable target location were summed. This imaginary rectangle is slightly larger than the target (Figure 13 legend). The time for such sum of activity to reach an arbitrary threshold decided the reaction time on any trial. In valid cue trials, the activity of surface contour or eye movement neurons representing the target locations (thin solid lines in Figures 13b and 13c, respectively) peak faster than those representing the other end of the attended bar (thick solid lines in Figures 13b and 13c, respectively), or those representing any location outside the bar (thin and thick dashed lines). Similarly, in all other trial types (Figures 13d, 13g, and 13j), target locations peak fastest in both surface contour (Figures 13e, 13h, and 13k) and eye movement maps (Figures 13f, 13i, and 13l); however, across trial comparisons show that such peak of activity occurs earlier in intra-object shifts of attention, black bars in Figures 14b and 14c, than inter-object shifts, light grey bars in Figures 14b and 14c, or object-to-location shifts of attention, dark grey bars in Figures 14b and 14c. Similar to the data, the two latter conditions have similar reaction times in the simulation.

8.2. Learning View-Invariant Object Categories in a Letter-filled Scene. This section summarizes a series of simulations that illustrate how ARTSCAN learns view-invariant object categories as the eyes autonomously scan a cluttered scene, under different combinations of unsupervised and supervised learning. The simulations illustrate how spatial attentional shrouds interact with cortically magnified images to foveate new features of interest on a surface and categorize them into view categories and view-invariant categories under different levels of supervision. The scene was filled with exemplars of ten letters of the alphabet (LFEHKDCOGQ). The exemplars differed in size, orientation, and spacing (Figure 9). Similar-appearing letters with the "Impact" font were chosen to make the recognition task harder: as a group, letters L, F, E, H, and K are more similar to each other than letters D, C, O, G, and Q as a group. Each letter had a uniform maximum possible luminance value of 1. The background had zero luminance.

Each letter could appear in the scene rotated from -45˚ to +45˚ from the vertical in 5˚ steps and expanded up to twice its size in step sizes of 0.05 times its size. Out of this 4000 letter database (10 letters by 20 rotations by 20 expansions), we randomly chose 440 entries and scattered them in the scene for the training set, as in Figure 9. For the testing set, the scene comprised 100 such randomly selected exemplars. The testing and training sets were disjoint, and the letters did not overlap. Figure 10 shows the activation of object boundaries when looking at a certain point on this scene.

Training and Testing. The model spontaneously scanned such scenes and learned to recognize the letters in them. For some letters during the training phase, a teaching signal was provided with the letter name, and for others, learning proceeded without supervision while ARTSCAN categorized the visited views into view categories and view-invariant object

categories. During the testing phase, neither was the teaching signal provided nor the learning allowed, and the activated name categories were recorded to calculate performance.

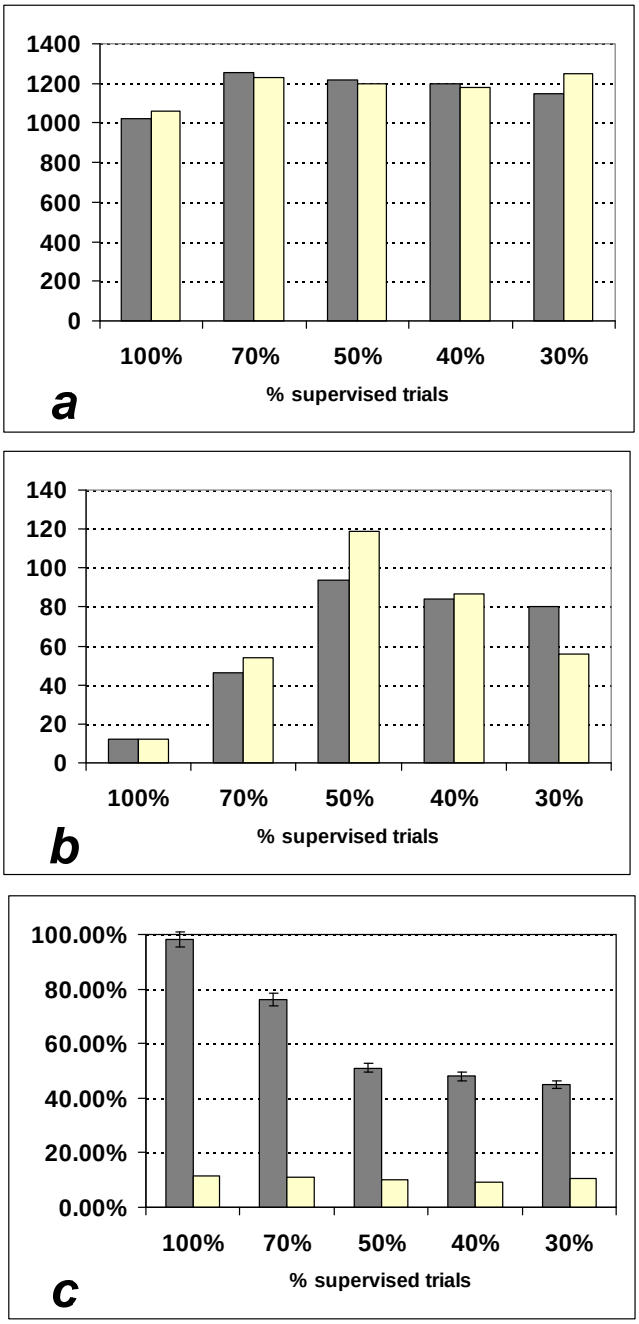

Figure 15: ARTSCAN performance results. (a) The number of generated view category neurons versus the percent of supervised trials in the training mode. The grey bars represent the *normal* condition, whereas the white bars represent the *no-reset* condition in which no category reset signal reaches the What stream from the Where stream. There are still effects of attention on the boundaries (see text). Note that within each supervision mode, there is not much difference between the number of generated view category neurons across the two conditions. (b) The number of generated object category neurons.

The rest is the same as in (a). Note again the similarity of the number of committed object neurons in both *normal* and *no-reset* conditions. (c) The model's performance plotted against the percent of supervised trials in both *normal* and *no-reset* conditions. Under any supervision condition, there is a striking difference in performances for *normal* versus *no-reset* condition. Performance in *no-reset* conditions falls to chance level regardless of percent of supervised trials in the training mode.

Figure 15 summarizes model performance for five different supervision regimens, along with the number of view and object neurons that ARTSCAN uses in each experiment. Two different conditions were tested: in the *normal* condition (grey bars in Figure 15), all model processes were active, whereas in the *no-reset* condition (white bars in Figure 15), the category reset signal was off all the time and did not reset the object categories when the corresponding attentional shroud collapsed. The shroud could still form and down-regulate the boundaries of other surfaces. As the ratio of supervised to unsupervised learning trials in the training phase decreased, the system used more view neurons and recognized letters less accurately during the testing phase. The model usually visited 10-20 hotspots on a letter, thus visiting about 6600 views (440 letters by 15 centers of foveation) in each simulation. Under *normal* conditions, if supervision is 100%, the 6600 views are compressed to approximately 1022 view neurons and 12 object neurons; that is, nearly as many as name category neurons. As fewer trials get supervised, the compression at the level of view neurons decreases; that is, more view neurons are used, but the number of object neurons reaches a peak in the 50% supervision regimen and decreases afterward. The reason for this Inverted U may be that, in the absence of supervision, ARTSCAN is working in an unsupervised mode, merely clustering visited views into similar object tokens. Without a teaching signal in the form of a name label to correct its errors, the model performs unsupervised clustering. The highest number of conflicting guesses occurs when there is 50% supervision. To correct these mistakes, the model uses more object neurons to overcome the guesses of the wrong object neurons.

The compression across view category and object category layers is not very different in the *no-reset* condition compared to the *normal* condition, yet the performance is drastically worse, no matter how many trials are supervised. Even though ARTSCAN can reach 98% performance in *normal* condition under 100% supervision regimen, the performances under *no-reset* conditions are no better than chance level (10%). The reason is that the learned information is transferred from a previously visited object to the next one and confuses the classifier.

9. Discussion and Related Models. ARTSCAN predicts how several known mechanisms in biological visual systems interact to perform active visual search, object learning, and recognition. These mechanisms are: (1) compressive space-variant cortical representation of retinal stimuli, (2) boundary-surface interactions, (3) spatial attention and scanning eye movements in the Where cortical stream, and (4) view-based object learning and recognition in the What cortical stream. This model predicts how attention can differentiate between saccades on the attended surface and those off that surface by using surface-based attentional shrouds, and how saccades can be restricted to a given object surface during view-invariant object learning by using attentionally-enhanced boundary representations to determine saccadic targets.

Features vs. Objects. Itti and Koch (2001) have summarized a computational model of attention. This model contains different feature maps, such as color, orientation, and motion. The center/surround mechanism in each feature map selects the odd feature. A single map then

combines the outputs of the feature maps to build a saliency map, which predicts the probability that a certain position will attract an observer's attention and eye movements. Our stimuli do not possess different features: they are stationary and black and white with uniform contrast. We did not include other features for two reasons: (1) our goal is to understand how spatial and object attention regulate view-invariant object learning as eyes autonomously scan a scene, and how boundary and surface representations help to determine where the eyes will look, not just how combinations of features can be processed at selected fixation points; and (2) the concept of 3D boundary and surface representations, as understood in FACADE theory, includes such features as color, orientation, and depth (Cao & Grossberg, 2005; Grossberg, 1994, 1999a; Grossberg & Mingolla, 1985c), which can be incorporated in a principled way into future generalizations of ARTSCAN. These properties of FACADE theory concern the What cortical stream. Related work on the Motion Boundary Contour System, or Motion BCS, and how it interacts with FACADE mechanisms in the 3D Formotion model, simulates how boundary-derived object motion properties—again, not just features—can also be used to attract attention (e.g., (Baloch & Grossberg, 1997; Berzhanskaya, Grossberg & Mingolla, 2007; Grossberg, Mingolla & Viswanathan, 2001).

Glance vs. Serial Inspection. The Itti & Koch (2001) and Riesenhuber & Poggio (1999) models assume that targets can be identified in one glance, so there is no need to explore the different parts of an object to gather evidence for recognition. Our prediction that the brain treats saccades that stay on the same object differently than saccades that move between objects has gained support in recent experiments (Beauvillain, Vergilino-Perez & Dukic, 2005;Vergilino-Perez & Findlay, 2004). Using gaze-contingent displacements of the stimuli, these investigators showed that, while a saccade off an object takes into account saccade-contingent stimulus displacements, those saccades on the object do not.

Inhibition of Return: Two Mechanisms. Itti and Koch (2001) also emphasize the importance of an inhibition of return (IOR) mechanism for any model of attention, indeed any model of sequential performance, by selecting the maximally activated cell population across a field of cells that simultaneously represents multiple external stimuli or locations, combined with self-inhibitory feedback to prevent perseverative selection of the same cell population over and over again. This combination of mechanisms was introduced in Grossberg (1978a, 1978b; see also Grossberg & Kuperstein, 1986). ARTSCAN proposes that at least two distinct mechanisms contribute to this IOR property, one that weakens, through habituation, an attentional shroud that has been active for awhile (it gets less "interesting"), and another that inhibits the currently fixated location on the eye movement map, so that the eyes can serially move to new target locations and avoid perseveration.

The superior colliculus (SC, see Figure 6) is the main brain region in ARTSCAN to generate eye movements and one locus that mediates space-based IOR, consistent with Fecteau & Munoz, (2003). We further suggest that this IOR is different from the type of object-based IOR that may be due to habituative weakening of an object-shaped shroud in the parietal cortex, which in the model projects via LIP to SC. The weakening of a shroud, and thus its contribution to IOR, interacts with the ability of a shroud to speed reaction times within an object. In addition, there is the possibility of object-based IOR in the inferotemporal and prefrontal cortex, that is due to the mechanisms whereby view categories and view-invariant categories are reset as the eyes scan a scene. The possible IOR effects of reset mechanisms are complicated by the fact that, whereas a view category may be reset, its view-invariant object category may persist, again leading to the possibility of both inhibitory and excitatory contributions to reaction times, depending on

experimental conditions. These mixed object-based excitatory and inhibitory factors are consistent with the conclusion of List & Robertson (2007, p. 1333) that "space based IOR plays a much stronger role in the inhibitory mechanisms that affect search efficiency".

Inter-Saccadic Information Integration. A related line of research that is consistent with ARTSCAN concerns the nature of information that is retained across saccades. The debate concerns whether more high-level cognitive information or low-level detailed information is retained across saccades. The trans-saccadic literature points to abstract high-level information (Irwin, 1991), and there is growing agreement that information integration across saccades is carried out at an abstract level (Deubel, Schneider & Bridgeman, 2002). Our model suggests that one key question is whether successive saccades are on the same surface or between surfaces. In the model as it currently stands, if saccades move between different surfaces, then there is less chance of information being retained and transferred across saccades, because the object and name layers in the model get reset under this condition. If successive saccades land on the same surface, then higher-level conceptual data can integrate across saccades (Carlson-Radvansky, 1999; Deubel, Schneider & Bridgeman, 2002; Henderson & Hollingworth, 2003).

Feedforward vs. Feedback; Certainty vs. Uncertainty. Being purely feedforward or having feedback connections is another distinction between models of biological visual systems. It is well-known that top-down feedback connections are ubiquitous in the brain. However, object recognition latencies and neural responses indicate that some scenic properties can be recognized through a fast feedforward sweep of activation (Thorpe, Fize & Marlot, 1996). LAMINART models of perceptual grouping clarify how unambiguous images may be processed in a fast feedforward manner and ART models demonstrate stable category learning and fast feedforward, globally-best-match recognition of familiar objects (Carpenter & Grossberg, 1987). However, the role of feedback cannot be ruled out in the control of learning, search, or recognition of ambiguous data. ARTSCAN illustrates how both feedforward and feedback processes can work together. For example, when the identity of an input is not ambiguous, the system can recognize the object by visiting just one view and does not require feedback. If the model gets confused between several ambiguous category choices, it uses feedback mechanisms to code the stimulus with more scrutiny. ARTSCAN hereby illustrates a tradeoff between certainty and speed.

Related Concepts of Attentional Shroud. Tyler & Kontsevich (1995) introduced the concept of attentional shroud to clarify how spatial attention can configure itself to the shape of an object in depth. Tyler & Konsevich (1995) also discussed the uniqueness of such a shroud during any given percept. For example, in response to a transparent display, they proposed that an observer perceives only one depth plane at a time within the perceptual moment, rather than a percept of simultaneous transparency. Tyler & Kontsevich (1995) did not, however, propose a mechanism whereby these properties can be obtained. Above we have noted how exogenously activated spatial attention may initially try to activate multiple attentional foci, corresponding to several different objects in a scene, but how a properly tuned long-range competition at an attentional processing stage, such as parietal cortex, can select a unique shroud. Likova & Tyler (2003) have, moreover, provided evidence that shroud formation involves surface filling-in. Again, they do not provide a mechanistic explanation. The ARTSCAN model provides a mechanistic account of how this can occur in response to either exogenously activated attention, via bottom-up inputs from objects in a scene, or endogenously activated attention, via a top-down attentional spotlight. Either form of attentional activation can shape itself to the winning object's form via feedback interactions with surface filling-in processes, as in Figure 4. The surface-filling in may occur in visual cortical areas such as V4, and the spatial attention may occur in parietal cortex.

Prior models have addressed aspects of attention, object recognition, and/or eye movements; for a review, see Itti & Koch (2001). The pioneering work of LaBerge & Brown (1989) modeled attention as a gradient across the visual field with the peak at the expected target location, as opposed to a moving spotlight of attention, to better explains shifts of attention, and also discussed how such a system could help object recognition. Logan (1996) showed how such a gradient model can capture object-based properties of attention. The ARTSCAN model predicts how an object-fitting attentional shroud can control, not just object recognition, but view-invariant object category learning. It also emphasizes the computational challenge that the brain needs to overcome in response to extended visual objects that undergo cortical magnification before being further processed for figure-ground separation and object learning. When a top-down attentional spotlight remains stable through time due to persistent volitional gain control, the parietal shroud can have a gradient shape due to the combination of top-down spotlight with bottom-up enhanced surface filling-in due to surface-shroud resonance. The cortical magnification factor is another factor that leads to a "gradient" across the visual field that is peaked at the expected target location.

Model extensions*: *3D shrouds, multiple shrouds, spatial and object working memory. As noted in Section 8.1, the present article does not simulate the formation of shrouds in depth. It simulates only one depth plane in response to the planar stimuli that are studied. However, ARTSCAN model concepts are consistent with those of the FACADE theory of 3D vision and figure-ground separation (Grossberg, 1994; Grossberg & McLoughlin, 1997; Grossberg & Yazdanbakshsh, 2005; Kelly & Grossberg, 2000). FACADE theory predicts how 3D boundaries can capture visible surfaces in depth. The boundaries act both as filling-in generators which can trigger depth-selective filling-in when boundary inducers are co-linear with, and interpolate, surface features, as well as filling-in barriers which can restrict filling-in within the surface regions that the boundaries surround. The features that fill in may either be bottom-up object features or a top-down attentional spotlight. Thus the process that defines the depth-selective filled-in object surfaces, within cortical areas such as V4, also localizes spatial attention to these objects within regions such as parietal cortex. Such a 3D surface-shroud resonance proposes a mechanistic revision and explanation of the Tyler suggestion that, once the attentional shroud fits itself to surface input signals, it *is* the representation of object structure. In our view, the *3D surface-shroud resonance* is one representation of object structure, not just the shroud taken alone. Other aspects of object structure include boundary-category, surface-category, and fused boundary-surface-category resonances whereby 3D boundary and surface representations interact reciprocally with their corresponding object category representations.

Even though we have quantitatively simulated only the Egly *et al* (1994) data, the model can explain many similar types of data. Its predictive range will be further enhanced when the model is generalized to 3D vision and figure-ground separation of overlapping scenic figures. Consider, for example, the classical Duncan (1984) experimental results, where two objects are overlapping. The FACADE model has explained and quantitatively simulated how the surfaces of overlapping figures can be preattentively separated from one another in depth, both in response to a 2D picture and a 3D scene (e.g., Fang & Grossberg, 2008; Grossberg, 1994, 1997; Grossberg & Yazdanbakhsh, 2005; Kelly & Grossberg, 2000). ARTSCAN can be consistently generalized to include these neural mechanisms. With them in place, the Duncan (1984) stimuli will be separated into a rectangle in the near depth plane and a thin rectangle (the line) in the far depth plane, and the input of the two surfaces will compete to form an attentional shroud for one object at a time, consistent with the results of Duncan (1984) that a parallel, preattentive

processes serves to segment the field into separate objects, followed by a process of focal attention that deals with only one object at a time. Implementing the circuits needed to also incorporate 3D vision and figure-ground separation of overlapping figures is a research goal for the ARTSCAN project. In this first article, given the number of brain processes and regions that need to be simulated to clarify basic relationships between perception, attention, and action, we simplified the model front end to simulate its dynamics using non-overlapping stimuli.

Other challenging types of data and issues about the relationship between figure-ground processing and attention are also clarified by the model. For example, Peterson, Harvey, & Weidenbacher (1991) have carried out experiments to show that object-based memories can influence how a figure-ground percept reverses through time, and have noted a possible role of "top-down activation (or 'priming') to a structural representation (cf. Carpenter & Grossberg, 1987). The effects of intentional priming should not be evident until bottom-up activation is present (Carpenter & Grossberg, 1987);" see their p. 1087. Such a top-down prime can act via learned category representations in the model cortical area IT to influence how an ambiguous figural boundary will become linked, through a process of border ownership, to one or another possible object interpretations of a scene via figure-ground interactions within model cortical area V2; see Figure 6. Interactions between such bottom-up and top-down processes can clarify controversies about whether figure-ground perception occurs before object representations or after. In Figure 6, it comes before, but it can nonetheless be reorganized by top-down processes to influence what is experienced in conscious perception, consistent with results of Baylis & Driver (2001). In addition, spatial attention that is directed to the surface on one side of a reversible figure boundary, or onto one part of such a boundary, can bias border ownership to influence which of two figural interpretations is seen. The effects of such attentionally-modulated figure-ground effects on Necker cube reversals were simulated by Grossberg & Swaminathan (2004) in a manner that is consistent with data of Peterson & Gibson (1991). Spatial attention can also influence the course of bistable transparency; see Grossberg & Yazdanbakhsh (2005) for simulations that are anticipated the data of Tse (2005).

Spatial attention need not form a shroud around only one object. A large literature shows that spatial attention can form over more than one object (Downing, 1988; Eriksen & Yeh, 1985; LaBerge & Brown, 1989; McMains & Somers, 2005; Pylyshyn & Storm, 1988; Yantis, 1992). Grossberg (2008) predicted that this is possible because the inhibitory gain that determines the strength of inhibition across shrouds is under volitional control by the basal ganglia. Weaker inhibition allows more than one shroud to exist at a time. One possible target of such volitional control is the inhibitory interneurons in cortical layer 4. Volitional control of the balance between cortical excitation and inhibition is predicted to be a general brain mechanism that expresses itself in behavior in strikingly different ways.

For example, in visual cortex, top-down expectations provide attentional modulation of bottom-up inputs. They do so via a top-down, *modulatory* on-center, off-surround network that has its effect on layer 4 cells (see Grossberg, 1999b, 2003 for supportive experimental data). The modulatory on-center can sensitize target cells to respond more vigorously and synchronously to attended visual feature combinations. If increasing volitional gain inhibits the inhibitory interneurons, it can convert the modulatory on-center into one that can drive suprathreshold activation of its target cells via a top-down expectation. This volitional mechanism has been predicted to enable top-down expectations to generate suprathreshold conscious percepts of visual imagery and fantasy, rather than merely modulatory attentional feedback (Grossberg, 2000). When this type of phasic volitional control over visual imagery and fantasy is replaced by

tonic hyperactivity of the gain control source, hallucinations can occur that have many of the properties of schizophrenic hallucinations.

The same sort of volitional control mechanism has been predicted to exist in the ventrolateral prefrontal cortex where it controls whether a sequence of items are stored in a cognitive working memory (Grossberg & Pearson, 2008). If these predictions about volitional control are supported, then they will provide an important example whereby homologous mechanisms within a broadly used neocortical circuit design can carry out different functions: allocation of spatial attention in the parietal cortex, visual imagery and fantasy in the visual cortex, and working memory storage in the prefrontal cortex. Volitionally-mediated inhibition of inhibitory gain is not the only way in which the balance between excitatory and inhibitory signals in these various cortical circuits might be stored and reset. More experiments are needed to study the locus and action of the predicted nonspecific gain control mechanism, whose source is anticipated to be in the basal ganglia.

A recent application of how multiple shrouds can contribute to scene understanding has been developed by Grossberg & Huang (2008). These authors show how a single glance at a scene can learn to activate a "gist" category of the scene, and how subsequent shroud-selected texture categories of the scene can augment the initial gist prediction to realize state-of-the-art scene understanding benchmarks.

Results of Hollingworth, Richard & Luck (2008) about inter-saccadic information integration are relevant to the simultaneous existence of multiple shrouds, and the interactions of shrouds with stored spatial representations in visual short term memory (VSTM). Hollingworth et al. (2008) have shown that object features in VSTM can help to correct saccadic errors, even when object location is altered by a surreptitious shift of target location during a saccade. In order to fully explain such data, one would need to show how What stream working memories, that encode object features (see Grossberg & Pearson (2008) for a recent model), can interact with Where stream spatial working memories to maintain their activity across time and guide the search for desired objects and their critical features.

Prinzmetal & Keysar (1989) have probed another type of phenomenon where top-down attentional effects can also influence boundary and surface interactions. They studied neon color spreading within an ambiguously colored middle letter of a word, such as the letter Y in MAYBE. In one study, the letters M and A were overlaid with a grid of red lines, B and E were overlaid with a grid of green lines, and Y was overlaid with an alternating grid of red and green lines. The M and A surfaces look pinkish and the A and Y surfaces look greenish, due to the effects of neon color spreading (see Grossberg & Mingolla (1985a), Grossberg (1994), and Grossberg & Yazdanbakhsh (2005) for explanations of how neon color spreading can occur). When such a letter Y is seen alone, out of the word context, it appears to be ambiguously colored, with both red and green tinged regions. Within the word MAYBE, however, the Y tended to look more green, consistent with its being in the syllable MAY. Such an effect can also be traced to top-down priming, in this case by learned word category representations.

Figure 6 shows only top-down priming of object boundaries, and does not include some of the known model mechanisms that would be needed to explain data of this kind. For example, Figure 6 does not include a temporal order working memory to temporarily store sequences of letters in working memory, and list chunking networks to learn to group letters into word, or syllable, categories. Grossberg & Myers (2000) and Grossberg & Pearson (2008) have developed a cortical model of such mechanisms. In addition, a more sophisticated representation of surface representations is needed that can conjointly represent both surface color and depth, and how

such surface representations compete for spatial attention, as proposed in Grossberg (1994) and used to quantitatively simulate visual search data in Grossberg, Mingolla, & Ross (1994). With such mechanisms added to the current, simplified ARTSCAN model, the Prinzmetal & Keysar (1989) data may be explained as follows using circuits homologous to those already simulated in Figure 6: A top-down word category for MAY can, through top-down priming, selectively strengthen the boundaries corresponding to the letters M, A, and Y. These boundaries can, in turn, strengthen their corresponding surface representations. These color filled-in surface representations can try to activate attentional shrouds that are specific to the colors in their surfaces. The color-specific shrouds compete via broad inhibitory surrounds, including competition across different color-specific shrouds. The winning color-specific shroud sends top-down priming signals to the corresponding surface regions, and thereby adds some additional red color to all of them, much as in the achromatic attentional enhancement that was reported by Reynolds & Desimone (2003) and Reynolds et al. (2000).

Unifying Cortical Magnification, Spatial and Object Attention, View-Invariant Learning, Eye Movements, and Search. ARTSCAN provides a unifying conceptual framework and neural architecture that combines and coordinates several key visual processes. This architecture helps to formulate and solve problems that might otherwise not even be posed. In the present article, the most pressing problem has been how a brain knows how to learn a view-invariant object representation as eyes scan a scene, even if there is no external supervision to help define the object. In approaching this general problem, ARTSCAN articulates functional roles for several processes that are not usually brought together in a single analysis. Although ARTSCAN is undoubtedly incomplete, it provide a conceptual and mechanistic framework within which many outstanding problems about visual perception, attention, learning, and eye movement can be more clearly discussed and solved.

Appendix 1: Model Equations

The model is a network of point neurons whose single compartment membrane voltage *V(t)* obeys:

$$C_m \frac{dV(t)}{dt} = -[V(t) - E_{leak}]\gamma_{leak}(t) - [V(t) - E_{excit}]\gamma_{excit}(t) - [V(t) - E_{inhib}]\gamma_{inhib}(t), \quad (1)$$

(Grossberg, 1973). Constant C_m is the membrane capacitance, the γ_{leak} term is a constant leakage conductance while the time-varying conductances $\gamma_{excit}(t)$ and $\gamma_{inhib}(t)$ represent, respectively, the total excitatory and inhibitory inputs, determined by the model architecture in Figures 6 and 7. The *E* terms represent reversal potentials. At equilibrium, the above equation can be written as:

$$V = (E_{excit}\gamma_{excit} + E_{inhib}\gamma_{inhib} + E_{leak}\gamma_{leak})/(\gamma_{excit} + \gamma_{inhib} + \gamma_{leak}). \quad (2)$$

Thus, increases in the excitatory and inhibitory conductance depolarize and hyperpolarize the membrane potential, respectively, and all conductances contribute to divisive normalization of the membrane potential, as shown by the denominator. This divisive effect includes the special case of pure "shunting" inhibition when the reversal potential of the inhibitory channel is close to the neuron's resting potential (Borg-Graham, Monier & Fregnac, 1998). Equation (1) can be re-written as:

$$\frac{dX}{dt} = -A_X X + (B_X - X)\gamma_{excit} - (C_X + X)\gamma_{inhib}, \quad (3)$$

by setting $X=V$, $A_X = \gamma_{leak}$, $E_{leak} = 0$, $B_X = E_{excit}$, and $C_X = -E_{inhib}$. Signal functions that are sometimes used in γ_{excit} or γ_{inhib} are usually denoted by *f*, *g*, or *h*. The connection weight from neuron with activity X_{ij} to the neuron with activity Y_{pq} is denoted by W_{ijpq}^{XY}.

Figure 7 summarizes the model interactions and the variables at every model stage. For example, the object boundary membrane potential is labeled B_{ij}.

A. Retina and Primary Visual Cortex Processes

A1. Retina. Because our modeling focuses on the higher-level interactions of the cortical What and Where stream, the front end of the model is simplified. Each half of retinal image undergoes a log-polar transformation when its signal arrives at the opposite primary visual cortex. The spatial relationship of a retinal ganglion cell position *(m,n)* to the location of it corresponding V1 cell *(p,q)* can be approximated by a logarithmic compression in the complex domain:

$$W = b\log(Z + a), \quad (4)$$

where *W* and *Z* are complex numbers such that $W = p + iq$, $Z = m + in$, $b=7$ and $a=0.3$ (Schwartz, 1980). The receptive field of a retinal ganglion neuron at retinal location *(m,n)* is defined as the set of all retinal locations *(i,j)* that are closer to that ganglion cell than to any other cell. The index Φ_{ijmn} shows whether location *(i,j)* is inside the receptive field of ganglion cell *(m,n)* and is defined as:

$$\Phi_{ijmn} = \begin{cases} 1 & \text{if } (m-i)^2 + (n-j)^2 \le (v-i)^2 + (w-j)^2, \quad m \ne v, n \ne w \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

As such, the retinal ganglion cells tessellate the retina into Voronoi cells and each Voronoi cell is the receptive field of one ganglion cell. A Matlab R14 code was used to perform this Voronoi, or

Dirichlet, tessellation (Barber, Dobkin & Huhdanpaa, 1996). The receptive field surface area ϖ_{mn} of retinal ganglion neuron *(m,n)* was calculated using standard geometrical equations.

Retinal cells sample the image cast on the retina in a discrete and space-variant way; their receptive fields are smaller in the fovea and larger toward the periphery. A peripheral retinal cell, therefore, receives more light in its larger receptive field than a foveal retinal cell (Figure 8a). Since the sensitivity of a retinal ganglion cells is inversely proportional to its receptive field size, the responses of ganglion cells at different eccentricities are proportional to the luminance of a patch of light filling their entire receptive fields (Wassle, Grunert, Rohrenbeck & Boycott, 1989). That is, the ganglion cells normalize their total input by their receptive field surface area. The retinal ganglion cell activity R_{mn} at position (m,n) thus obeys:

$$R_{mn} = \frac{1}{\varpi_{mn}} \sum_{ij} \Phi_{ijmn} L_{ij} , \tag{6}$$

where ϖ_{mn} is the receptive field surface area, index Φ_{ijmn} shows whether the location (i,j) on the retina falls in the receptive field of the ganglion cell at location (m,n) as shown in (5), and L_{ij} is the input image luminance at retinal location (i,j).

A2. LGN Polarity-Sensitive Cells. The LGN polarity-sensitive cells are of two types: On and off. The on-cells have a small excitatory center and a broader inhibitory surround. Off-cells have a small inhibitory center and a broader excitatory surround. The activity X_{ij}^{+} of the LGN polarity-sensitive on-cell at position (i,j) has an equilibrium value:

$$X_{ij}^{+} = \frac{\sum_{pq} R_{pq} (D_{pqij}^{+} - D_{pqij}^{-})}{1 + \sum_{pq} R_{pq} (D_{pqij}^{+} + D_{pqij}^{-})} , \tag{7}$$

where R_{pq} is the retinal ganglion cell activity in position (p,q), D_{pqij}^{+} is the Gaussian on-center receptive field:

$$D_{pqij}^{+} = \exp\left(- \frac{(i-p)^2 + (j-q)^2}{2 \cdot 0.2^2} \right), \tag{8}$$

and D_{pqij}^{-} is the Gaussian off-surround receptive field:

$$D_{pqij}^{-} = 14.13 \exp\left(- \frac{(i-p)^2 + (j-q)^2}{2 \cdot 1.5^2} \right). \tag{9}$$

The coefficients of the excitatory and inhibitory kernels normalize these kernels. As a result, a uniform input pattern does not activate the LGN cells.

An LGN off-cell at position (i,j) has the opposite excitatory and inhibitory kernels than does the corresponding on-cell. The activity X_{ij}^{-} of the LGN off-cell at (i,j) obeys:

$$X_{ij}^{-} = \frac{\sum_{pq} R_{pq} (D_{pqij}^{-} - D_{pqij}^{+})}{1 + \sum_{pq} R_{pq} (D_{pqij}^{-} + D_{pqij}^{+})} , \tag{10}$$

A3. V1 Polarity-Insensitive Cells. Because model simulations use binary images with continuous borders that are viewed monocularly, orientation and depth sensitivity were not required for the model. Contrast enhancement and edge detection are sufficient in the model to detect stimulus borders and perform filling-in and surface completion.

The activity z_{ij} of polarity-insensitive cells simplify complex cell properties in the primary visual cortex:

$$z_{ij} = \left[X_{ij}^{+}\right]^{+} + \left[X_{ij}^{-}\right]^{+}, \tag{11}$$

where $\left[X_{ij}^{+}\right]^{+}$ and $\left[X_{ij}^{-}\right]^{+}$ are the on-center and off-center LGN outputs at position (i,j), respectively, and threshold half-rectified output signal $\left[a\right]^{+} = \max(a,0)$. Output of the polarity-insensitive cell is:

$$Z_{ij} = \left[z_{ij} - 0.2\right]^{+}. \tag{12}$$

The threshold 0.2 helps to sharpen the Z_{ij} boundaries around an object, given that the model omits many recurrent circuits that complete V2 boundary cells and sharpen boundaries in response to more complex inputs, e.g. Mingolla, Ross & Grossberg (1999). Equations (7)-(12) are similar to boundary equations in Grossberg & Todorović (1988), as are the surface filling-in equations discussed below.

A4. Object Boundary. FACADE theory and the 3D LAMINART model propose how 3D boundaries are completed and enable figure-ground segregation (Grossberg, 1999a; Grossberg & Kelly, 1999; Grossberg & Yazdanbakhsh, 2005). The surface contour feedback from filled-in surfaces to their inducing boundaries in those models enhance the edges corresponding to filled-in closed boundaries at a certain depth. This is a key process in beginning to separate, and thereby define, object surfaces in depth. Feedback from object surface also plays a role in ARTSCAN. As discussed in Section A5, object surface activities are modulated by top-down effects of attention. As a result, the feedback from object surfaces to object boundaries strengthens boundaries that belong to the attended surface. Each object boundary cell receives bottom-up input from polarity-insensitive or complex cells, as well as modulatory surface contour feedback. The object boundary activity B_{ij} at position (i,j) at equilibrium is thus:

$$B_{ij} = \frac{Z_{ij}(1+10\sum_{pq} C_{pq} F_{pqij}^{+}) - \sum_{pq} C_{pq}}{0.001 + Z_{ij}(1+10\sum_{pq} C_{pq} F_{pqij}^{+}) + \sum_{pq} C_{pq}}, \tag{13}$$

where Z_{ij} is the bottom-up complex cell output, C_{pq} is the surface contour cell activity at position (p,q) defined in (17), and F_{pqij}^{+} is the Gaussian kernel from position (p,q) on the surface contour to position (i,j) on the object boundary:

$$F_{pqij}^{+} = \exp\left(-\frac{(p-i)^2 + (q-j)^2}{2 \cdot 3^2}\right). \tag{14}$$

By (13), surface contour feedback amplifies the boundaries that are activated by complex cells.

A5. Object Surface Filling-in. Object surface activity S_{ij} at position (i,j) obeys the diffusion equation:

$$\frac{dS_{ij}}{dt} = -40S_{ij} + \sum_{(pq \in N_{ij})} P_{pqij}(S_{pq} - S_{ij}) + \left[X_{ij}^{+}\right]^{+} + 7\sum_{mnkl} h(I_{mnkl}) W_{mnklij}^{IS}. \tag{15}$$

In (15), N_{ij} is the set of nearest neighbor cells around *(i,j)*, and P_{pqij} is the boundary-gated permeability between the locations (p,q) and (i,j):

$$P_{pqij} = \frac{10^4}{1 + 40(B_{pq} + B_{ij})}, \tag{16}$$

where B_{pq} and B_{ij} are the activities of boundary cells at positions (p,q) and (i,j) respectively, as in (13). The gate P_{pqij} has a small value whenever B_{ij} or B_{pq} is large. Thus, diffusive filling-in is gated by the boundaries. In (15), $\left[X_{ij}^{+}\right]^{+}$ is the bottom-up input signal from the on-center polarity-sensitive cell at location (i,j), and I_{mnkl} is the top-down attentional input from the gain field cell in location (m,n,k,l). As explained in Section B1, top-down spatial attention reaches object surface cells via gain field cells that transform head-centered attention signals into retinotopic surface inputs. $W_{mnklij}^{IS} = W_{ijmnkl}^{SI}$ is the weight between the gain field cell in location (m,n,k,l) and the surface filling-in cell at position (i,j), as defined in (21).

No matter where on a closed surface region an attentional signal is received, it diffuses across the entire surface. Any space-based spotlight of attention directed to any part of a surface can thereby boost the activity of the whole surface and thereby fit itself to the surface form.

A6. Surface Contours. The activity of object surface is contrast-enhanced by on-center and off-center networks to generate surface contour output signals that modulate object boundaries, as in (13), and thereby control eye movements, as in (33). Surface contour signals occur only at boundary contours of the surface. The surface contour output signal C_{ij} at each location *(i,j)* is the sum of rectified On- and off-channel responses to filled-in object surface activities:

$$C_{ij} = \left[\frac{\sum_{pq} S_{pq}(K_{pqij}^{+} - K_{pqij}^{-})}{.01 + \sum_{pq} S_{pq}(K_{pqij}^{+} + K_{pqij}^{-})}\right]^{+} + \left[\frac{\sum_{pq} S_{pq}(K_{pqij}^{-} - K_{pqij}^{+})}{.01 + \sum_{pq} S_{pq}(K_{pqij}^{-} + K_{pqij}^{+})}\right]^{+}, \tag{17}$$

where S_{pq} is the object surface cell activity at position (p,q), and K_{pqij}^{+} and K_{pqij}^{-} are on-center and off-center Gaussian kernels, respectively, that detect surface contours:

$$K_{pqij}^{+} = \frac{1}{1.69}\exp\left(-\frac{(i-p)^2 + (j-q)^2}{2 \cdot 0.3^2}\right), \tag{18}$$

$$K_{pqij}^{-} = \frac{1}{10.04}K^{-}\exp\left(-\frac{(i-p)^2 + (j-q)^2}{2 \cdot 2^2}\right). \tag{19}$$

B. WHERE Stream

B1. Gain Field. Model processes prior to the spatial attention map are all in retino-centric coordinates. Consequently, each eye movement changes the activity on those maps, even if the scene does not change. The spatial attention map, on the other hand, is in head-centric coordinates and invariant under changes in eye position. Gain fields may mediate the coordinate change from a retino-centric object surface representation to a head-centric spatial attention map (Andersen, Essick & Siegel, 1985; Andersen & Mountcastle, 1983; Deneve & Pouget, 2003; Gancarz & Grossberg, 1999; Grossberg & Kuperstein, 1986; Pouget, Dayan & Zemel, 2003). Here we adopt a pre-wired gain field. In a subsequent development of the model, the weights to and from the gain field will be learned, as in Gancarz and Grossberg (1999) and Elder, Grossberg & Mingolla (2005). In Pouget and Snyder (2000), both retinal and eye position maps are one-dimensional and the gain field is two-dimensional. The gain field connects to each of those 1D

maps using one of its dimensions. For example, if in their implementation the retinal cell *i=5* and the eye position cell *j=3* are active, all the gain field cells *(5,j)* and *(i,3)* will get active, but the particular gain field cell *(5,3)* will be the most active. A head-centric cell *k=8* is associated with all the gain field cells (*i,j*) such that *i+j=8*, e.g. *(6,2)* or *(4,4)* and so on. The rational is that, for any stationary stimulus in head-centric coordinates, an eye movement of *n* units in one direction will result in a shift of *–n* units to the other direction in the retinal representation. The *sum* of eye position and retinal position indices will thus remain constant. Gain field cells along the same diagonal or para-diagonal, where the sum of indices is equal, activate the same head-centric cell. When both retinal and eye position maps are two-dimensional, the gain field will be four-dimensional and not easy to visualize. An example of the weights to a gain field cell is shown in Figure 16.

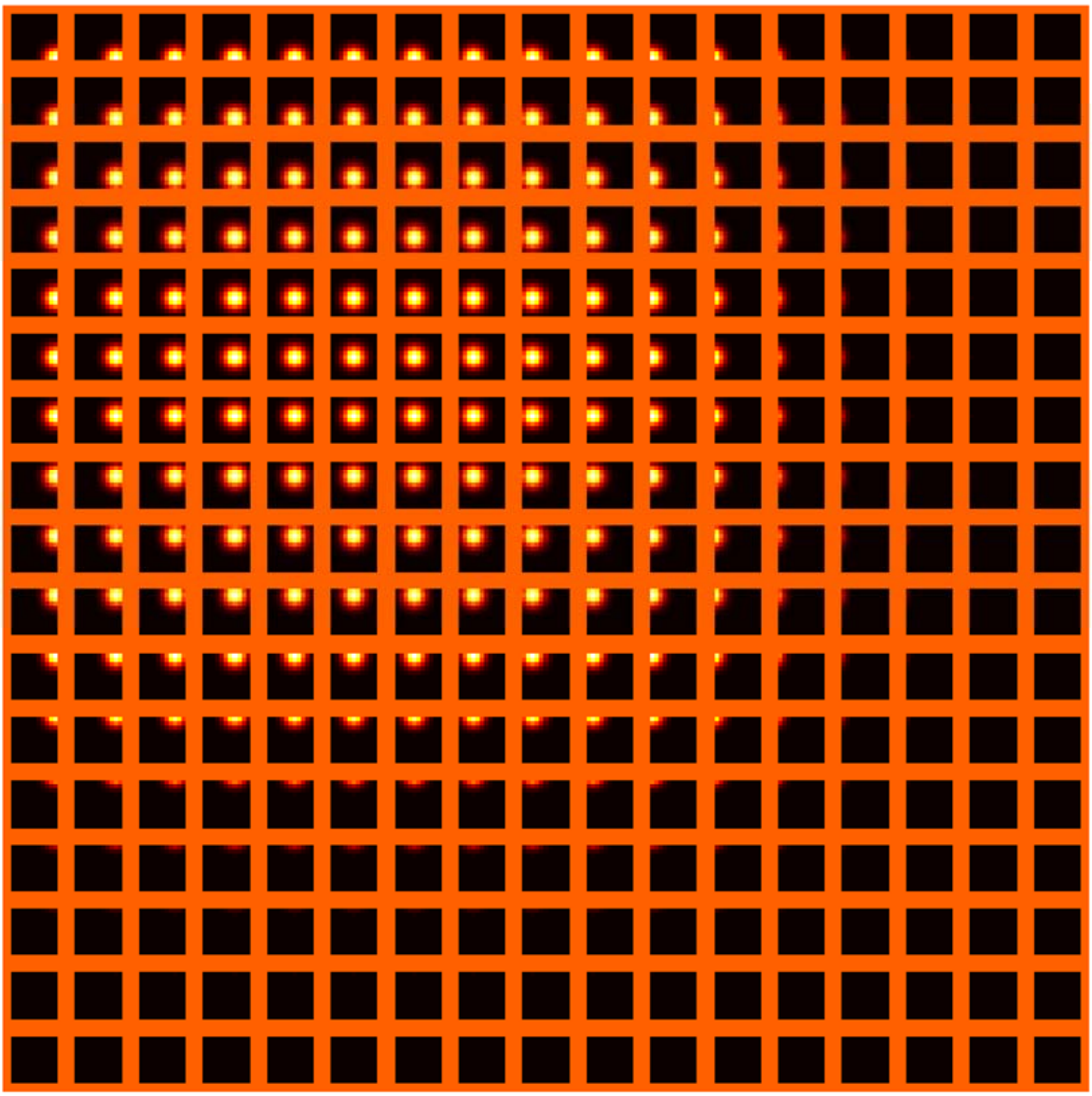

Figure 16: The weights of a gain field cell that connects it to eye position and retino-centric object surface maps. The weights are a four-dimensional entity *(k,l,m,n)*. Each

small tile represents the 2D weights to the retinotopic surface map. The place of that small tile within the big image shows the eye position. For example the demarcated image on the sixth row and sixth column shows the connection weight to the retinotopic surface map when the eyes are in the position *(6,6)* in the orbit. In the eye position *(6,6)*, this particular gain field neuron happens to have the highest sensitivity to stimuli around the fovea, making it a suitable cell to respond best to position *(6,6)* in a head-centric coordinate. Note how the cell is changing its retinal sensitivity in the opposite direction to the eye positions as the latter deviates from the optimal *(6,6)* position. In this manner, it keeps signaling the same head-centric location. A spatial attention map in the head-centric position *(6,6)* has the highest sensitivity to this gain field cell and its weights decrease to its neighbors in a Gaussian manner.

The activity I_{mnkl} of gain field cell at position (*m,n,k,l*) is affected by three inputs through Gaussian weights: one from the 2-dimensional object surface map described in A5, one from the eye position map, and one top-down input from the spatial attention map, as explained in B2,

$$\frac{dI_{mnkl}}{dt} = (1 - I_{mnkl})\left(\sum_{ij} S_{ij} W^{SI}_{ijmnkl} + \sum_{ij} P_{ij} W^{PI}_{ijmnkl} + \sum_{ij} A^{I}_{ij} W^{AI}_{ijmnkl} \right) - (I_{mnkl} + 0.2) \sum_{mnkl} I_{mnkl} \,. \qquad (20)$$

In (20), S_{ij} is the object surface cell activity at position (*i,j*), W^{SI}_{ijmnkl} is the Gaussian weight between object surface cell at position (*i,j*) and the gain field cell at position (*m,n,k,l*):

$$W^{SI}_{ijmnkl} = \exp\left(- \frac{(i-m)^2 + (j-n)^2}{1.7^2} \right), \qquad (21)$$

where P_{ij} is the eye position map activity at location (*i,j*), as defined in (24). W^{PI}_{ijmnkl} is the weight between the eye position cell at position (*i,j*) and the gain field cell at position (*m,n,k,l*):

$$W^{PI}_{ijmnkl} = \exp\left(- \frac{(i-k)^2 + (j-l)^2}{1.7^2} \right). \qquad (22)$$

In (20), A_{ij} is the spatial attention cell activity at location (*i,j*)—see Section B2y—and W^{AI}_{ijmnkl} is the Gaussian weight between the spatial attention map interneuron at (*i,j*) and the gain field cell at (*m,n,k,l*):

$$W^{AI}_{ijmnkl} = \exp\left(- \frac{(i-m-k)^2 + (j-n-l)^2}{2.5^2} \right). \qquad (23)$$

For simplicity, the fixed weights used here are assumed to be the result of training. The eye position cell activity P_{ij} at position (*i,j*) in (20) has a binary value:

$$P_{ij} = \begin{cases} 1 & \text{if eyes are at position } (i,j) \text{ in the orbit} \\ 0 & \text{otherwise} \end{cases} \,. \qquad (24)$$

Wherever the source of the eye position information might be, we assume that the parietal gain field cells have access to it, as is the case in other models (Pouget & Snyder, 2000).

B2. Spatial Attention. The spatial attention map includes spatial attention neurons and interneurons. Object surface inputs (S_{ij}) reach attention inter-neurons (A^{I}_{ij}) via the gain field map (I_{mnkl}); see Figure 7. Attention interneurons project to spatial attention neurons, which compete for attention. The spatial locus of winning activity is called the *attentional shroud*. The shroud, in

turn, feeds back via gain field cells to object surface representations and thereby boosts the activities of the winning surface.

Stated mathematically, the spatial attention interneuron activity A_{ij}^I at head-centric position (*i,j*) receives bottom-up input $\sum_{mnkl} h(I_{mnkl}) W_{mnklij}^{IA}$ from the gain field cells and top-down feedback $f(A_{ij})$ from the corresponding spatial attention cell:

$$\frac{dA_{ij}^I}{dt} = -A_{ij}^I + \sum_{mnkl} h(I_{mnkl}) W_{mnklij}^{IA} + f(A_{ij}), \tag{25}$$

where $W_{mnklij}^{IA} = W_{ijmnkl}^{AI}$ is the weight between the gain field cell at position (*m,n,k,l*) and attention interneuron at position (*i,j*), and A_{ij} is the activity of spatial attention cell at position (*i,j*). Equation (25) was solved in equilibrium:

$$A_{ij}^I = \sum_{klmn} h(I_{klmn}) W_{klmnij}^{IA} + f(A_{ij}). \tag{26}$$

In (25) and (26), the signal function *h* is defined by the threshold-linear function:

$$h(a) = [a - .2]^+, \tag{27}$$

and the signal functions *f* is defined by the sigmoid function:

$$f(a) = \frac{4}{1 + e^{-50a+8}}. \tag{28}$$

The spatial attention cells receive excitatory input from the corresponding attention interneurons through habituative transmitter gates (Grossberg, 1972, 1980b), as well as lateral excitation from other spatial attention cells. Each spatial attention cell also receives long-range inhibition from other attention interneurons and spatial attention cells. The spatial attention cell activity A_{ij} at position (*i,j*) obeys:

$$\frac{1}{10}\frac{dA_{ij}}{dt} = -0.1A_{ij} + (1 - A_{ij})\left(A_{ij}^I y_{ij}^A + \sum_{mn} f(A_{mn}) C_{mnij} \right) - A_{ij}\left(\sum_{mn} \left(A_{mn}^I + f(A_{mn}) \right) E_{mnij} \right), \tag{29}$$

where A_{ij}^I is the attention interneuron output, y_{ij}^A is the excitatory habituative transmitter that gates the attention interneuron signal at position (*i,j*) on its way to the corresponding spatial attention cell; see (32). Sigmoid signal function *f* is defined in (28), C_{mnij} is the Gaussian excitatory weight from the position (*m,n*) to (*i,j*):

$$C_{mnij} = 0.01 \exp\left(-\frac{(m-i)^2 + (n-j)^2}{2 \cdot 0.6^2} \right), \tag{30}$$

and E_{mnij} is the broad inhibitory Gaussian weight for both the bottom-up input of attention interneurons and spatial attention cells at (*m,n*) to the spatial attention cell at (*i,j*) :

$$E_{mnij} = 1.62 \exp\left(-\frac{(m-i)^2 + (n-j)^2}{2 \cdot 400^2} \right). \tag{31}$$

In (29), the habituative transmitter y_{ij}^A that mediates between the attention interneuron and its spatial attention cell at position (*i,j*) obeys:

$$\frac{dy_{ij}^{A}}{dt} = K_A\left(2 - y_{ij}^{A} - 3\cdot 10^{6} A_{ij}^{I} y_{ij}^{A}\right), \tag{32}$$

where $K_A = 7\cdot 10^{-9}$ is a very slow rate of decay, and A_{ij}^{I} is the attention interneuron activity. The more input activity a spatial attention cell receives from its interneuron, the faster its transmitter habituates, and as a result, the most active spatial attention cells that initially form a shroud can collapse and let another group of cells form a new shroud around another surface representation.

B3. Eye Movement. The eye movement map is a winner-take-all map which selects the next target for fixation. It receives input from the surface contour cells, along with self-excitation, both gated by a habituative transmitter. Shortly after an eye movement cell wins the competition and selects the saccade target, its neurotransmitter habituates, its activity crashes, and another cell can win to determine the next saccade target. Eye movement cell activity E_{ij} corresponding to target position (i,j) obeys:

$$\frac{dE_{ij}}{dt} = -20E_{ij} + \left(1 - E_{ij}\right)\left(\left[C_{ij}\right]^{+} + 625E_{ij}^{2}\right)y_{ij}^{E} - .02E_{ij}\sum_{ij}\left(\left[C_{ij}\right]^{+} + E_{ij}^{2}\right), \tag{33}$$

where C_{ij} is the surface contour cell output at location (i,j), as defined in (17), and y_{ij}^{E} is the habituative transmitter that gates the input to the eye movement cell at *(i,j)*:

$$\frac{dy_{ij}^{E}}{dt} = K_E\left(2 - 10^{7} y_{ij}^{E}\left(\left[C_{ij}\right]^{+} + 625E_{ij}^{2}\right)\right), \tag{34}$$

where $K_E = 10^{-8}$. K_E in (34) is much larger than K_A in (32), so it takes an attentional shroud a longer time to collapse compared to a saccade target. Thus, when a surface is attended, it can be explored by several eye movements.

B4. Category Reset. This cell population inhibits, and thus resets, the object cells in the What stream when the attentional shroud breaks; see Figure 7, (40), and (41). Category reset cells have a tonic activity that is inhibited by the total activity across the spatial attention map. If there is little activity in the spatial attention map, say due to shroud collapse, the reset cells get active and non-specifically inhibit both object layers. The activity R_{WHERE} of these cells obey:

$$R_{WHERE} = 1000\left[50 - \sum_{ij} f(A_{ij})\right]^{+}, \tag{35}$$

where A_{ij} is the spatial attention cell activity at location (i,j) and f is the signal function defined in (28).

C. WHAT Stream

The main inputs to the What stream are the object boundary cell outputs, which are connected to the view category neurons through adaptive weights. The view-invariant object category neurons learn to be activated by an appropriate set of view category neurons and are associated with name category neurons to learn the names of objects.

C1. View Categories. The view category neurons learn to respond to a certain size or orientation of an object. They receive input from object boundary neurons through adaptive weights. Fuzzy ART learns the view-sensitive categories (Carpenter, Grossberg, Markuzon, Reynolds & Rosen, 1992). ART, or Adaptive Resonance Theory, explains how category learning occurs when a bottom-up input pattern matches a top-down expectation through a process of competitive attentional matching (Carpenter & Grossberg, 1987, 1993).

Each view category neuron initially receives input from all object boundary neurons, although it ends up learning and using just the foveal regions of that map, because the attentional shroud and eye movements direct the fovea to the object of interest. The activity V_i of the i^{th} view category neuron in response to the boundary input $\vec{B}$ obeys:

$$V_i(\vec{B}) = \frac{\left|\vec{W}_i^{BV} \wedge \vec{B}\right|}{0.001 + \left|\vec{W}_i^{BV}\right|}, \tag{36}$$

where $\vec{B} = \left\{B_{pq}, 1 - B_{pq}\right\}$. The terms $1 - B_{pq}$, are called *complement coding*. They represent Off-cell responses and lead to useful category learning properties. $\vec{W}_i^{BV}$ is the weight vector between the complement-coded object boundary map, $\vec{B}$ (see (13)) and the i^{th} view category. The fuzzy AND operator, $\wedge$, between two vectors p and q is defined as $(p,q)_i \equiv \min(p_i, q_i)$. The L_1 norm operator, $|\bullet|$, is defined as $|p| \equiv \sum_i^M |p_i|$ for any M dimensional vector p. Equation (36) computes the normalized distance between the weights of a view category and a certain boundary map. The more similar they are, the more active that view category neuron becomes.

The most selective, and thus the most highly activated, view category wins the competition among all view neurons and sends its output to the object category layer. Two conditions must be satisfied for a view category V_J to win and output to object category layer: it must be the most active view neuron, that is $V_J = \max(V_i)$, and its activity must satisfy the inequality:

$$V_J(\vec{B}) > \rho, \tag{37}$$

where ρ is a goodness of match criterion called *vigilance* (Carpenter, Grossberg & Reynolds, 1991). Vigilance is defined as:

$$\rho = \left(1 - \Psi(R_{WHAT})\right)\rho_{base} + \Psi(R_{WHAT})\left(\frac{\left|W_J^{BV} \wedge \vec{B}\right|}{0.001 + \left|W_J^{BV}\right|} + \varepsilon\right), \tag{38}$$

where the sign function $\Psi(x)$ is 1 if $x \geq 0$ and 0 otherwise. Equation (38) shows that, if there is no mismatch reset neuron activity, R_{WHAT}, then vigilance is equal to a baseline quantity called $\rho_{base} = 0.85$. A mismatch between an ARTSCAN predicted name and the name label provided by a teaching signal activates a What stream category reset R_{WHAT} (see Section C5), and causes ρ to increase to a slightly higher level, by $\varepsilon = 0.0001$ than the activity of the winning view category neuron. Vigilance hereby carries out *match tracking* during a reset episode. This shuts off the winning view neuron and allows the next most active view neuron to try to satisfy the two above conditions and so on.

If the normalized activity of the winning view category in (36) exceeds ρ, then the view layer is said to be in a *resonant* state and learns the weights W_J^{BV} between object boundary neurons $\vec{B}$ and the winning view category neuron J:

$$W_J^{BV(new)} = \beta(W_J^{BV(old)} \wedge \vec{B}) + (1 - \beta)W_J^{BV(old)}, \tag{39}$$

where β is the learning rate, here set to 1 to accelerate learning. Equation (39) shows that the new weights are the intersection of the old weights and the active boundaries.

C2. View-Invariant Object Categories. Object category neurons are associated with several view neurons that represent different poses of the same object. Thus, the spatial object

transformations that change the responses of individual view category neurons do not change the response of the corresponding object category neuron. The object layer has two neuron types: (1) object category neurons receive bottom-up input from view category neurons and a modulatory top-down attentional matching input from the name category neurons, and (2) object integrator neurons are connected one-to-one to object category neurons by habituatively gated pathways, and integrate the impulses coming from the object category neurons each time one of their view categories becomes active. The habituative transmitter ensures that, no matter how long the eyes fixate a view, only one pulse of activity reaches the corresponding object integrator neuron and the appropriate name neuron. The object integrator neurons are linked to the name category neurons via associative learning (Figure 7). The object category neuron activity O_i obeys:

$$\frac{1}{2000}\cdot\frac{dO_i}{dt} = -0.01O_i + 4.2V_J^2 W_{Ji}^{VO} + \sum_j \left[N_j - 0.5\right]^+ W_{ji}^{NO} - (O_i + 0.1)\left(0.1\sum_j \sum_i \left(\left[N_j - 0.5\right]^+ W_{ji}^{NO}\right) + 2\sum_k V_k^2 + R_{WHAT} + R_{WHERE} \right), \quad (40)$$

where V_J^2 is the output signal from the winning view neuron J, W_{Ji}^{VO} is the weight between the J^{th} view neuron and the i^{th} object category neuron, N_j is the j^{th} name neuron activity, W_{ji}^{NO} is the weight from the j^{th} name neuron to the i^{th} object category neuron, $2\sum_k V_k^2$ is the off-surround input from the view neurons which normalize the effect of the excitatory term $4.2V_J^2 W_{Ji}^{VO}$, and R_{WHAT} and R_{WHERE} are reset signals coming from mismatch reset neurons in the What Stream (Equation 45) and the category reset neurons in the Where stream (Equation 35), respectively.

Activity of the i^{th} object integrator neuron Q_i is influenced only by its corresponding object category neuron through a habituative transmitter gate:

$$\frac{1}{2000}\cdot\frac{dQ_i}{dt} = -.01Q_i + 400\left[O_i - 0.5\right]^+ y_i^O - \left(0.1 + Q_i\right)\left(R_{WHAT} + R_{WHERE}\right), \quad (41)$$

where O_i is the i^{th} object category neuron activity and y_i^O is the habituative transmitter described in (42). As in the object category neuron, both R_{WHAT} and R_{WHERE} can reset the object integrator neuron.

C3. Habituative Transmitter Gate between Object Category and Integrator Neurons. A habituative transmitter gate mediates between the object category neuron and its corresponding object integrator neuron. As the activity of the object category neuron increases, the habituative transmitter slowly decreases, and thus the product of object category activity and transmitter level will first increase and then decrease. This will send a pulse of activity to the object integrator neuron in response to each object category neuron activation, no matter how long the object category neuron remains active. The habituative gate y_i^O between the i^{th} object category neurons and its corresponding object integrator neuron obeys:

$$\frac{dy_i^O}{dt} = 70\left(2 - y_i^O - 5000 y_i^O \left[O_i - 0.5\right]^+\right). \quad (42)$$

The habituative gate enables an object category neuron to send just a pulse to its object integrator neuron, while the latter remains active to get associated through time with the view category and name category neurons if there is no mismatch in the What stream.

C4. Name Categories. Name category neurons receive their inputs from both object integrator neurons and teaching signal within a center-surround network. During training, if a

teaching signal activates a name neuron, that name neuron can selectively learn to be associated with the active object integrator neuron(s) at that time. The center-surround network represents a competition to select a winning name category. The activity of the i^{th} name category N_i obeys:

$$\frac{1}{200}\frac{dN_i}{dt} = -3N_i + (1-N_i)\left(\sum_j 15[Q_j]^+ W_{ji}^{ON} + T_i\right) - 0.8N_i\left(\sum_j\sum_i 15[Q_j]^+ W_{ji}^{ON} + \sum_i T_i\right), \quad (43)$$

where Q_j is the j^{th} object integrator neuron activity. Its signal function ensures that even small inputs from the object integrator layer will strongly activate the name layer so that a mismatch will be detected if a different name is externally supplied; see Section C5. W_{ji}^{ON} is the weight of the learned excitatory connection from the object integrator neuron j to the name neuron i (see (48)). T_i is the teaching signal denoting name category i:

$$T_i = \begin{cases} 1 \text{ if the name of the object is category } i \\ 0 \qquad\qquad\qquad \text{otherwise.} \end{cases} \quad (44)$$

C5. Mismatch Reset. After some supervised learning trials, ARTSCAN learns the names of some objects and activates a name category neuron in response to novel objects that are similar to those it has already seen; that is, it guesses the names of those objects. As long as a correct guess occurs, the model learns that the newly observed object has the same name. If the input activates a name category neuron that is different from the one activated by the teaching signal, a mismatch occurs. When this happens, the system should stop learning, correct its error by resetting both object category and object integrator layers, and either come up with another name for of the viewed object or learn the one provided. The reset signal that responds to this mismatch is called *mismatch reset*. The activity R_{WHAT} of the mismatch reset population obeys:

$$\frac{dR_{WHAT}}{dt} = -100R_{WHAT} + 10^4\left[\sum_i T_i - 2000\sum_j [N_j - 0.5]^+\right]^+, \quad (45)$$

where T_i is the teaching signal to the i^{th} name category in (44), and N_j is the j^{th} name category cell in (43). By (45), if any name category neuron gets more active than half of its maximum possible activity, it can inhibit the reset signal with a high gain. The name category neurons exhibit shunting normalization. Thus, if there are two equally active name neurons, they mutually inhibit each other. When such a mismatch occurs, none of the simultaneously active name neurons activity can exceed the 0.5 output threshold. If at the same time any teaching signal is present, it can activate the mismatch reset neuron. Note that there will be no mismatch reset signal in the absence of a teaching signal and erroneous learning can go on: Activated view and object neurons may sometimes get associated with more than one name neuron and thus become ambiguous views and objects. A teaching signal is not essential for learning, but it helps to correct mistakes made by the system. Also note that, if the teaching signal is active while there are no bottom-up naming activities, the mismatch reset neuron will *start* to get activated, but will shut off quickly, because the teaching signal will also activate its corresponding name category neuron, which in turn will inhibit the effect of the teaching signal on the mismatch reset neuron.

The mismatch reset signal R_{WHAT} (Figure 12g) is strong enough to totally shut down the object category and integrator layers (Figures 12b and 12e) and consequently those name neurons that rely on the bottom-up activity form the object layer (Figure 12f, solid line).

The mismatch reset signal also increases the vigilance in the view category layer, as noted in (38). Increasing vigilance shuts off the active view category, and starts a search for a better view category to represent the input (top of Figure 12b).

C6. WHAT Stream Learning. The weights between view, object, and name category layers obey a double-gated *instar* learning rules (Grossberg, Hwang & Mingolla, 2002). Such learning is gated by both presynaptic and postsynaptic neural activities. If either is inactive, the weight between them does not change. The weights increase or decrease until they match the activity of the presynaptic neurons.

$$\frac{dW_{ij}^{VO}}{dt} = 50\left[V_i\right]^+\left[O_j\right]^+\left(\left[V_i\right]^+ - W_{ij}^{VO}\right), \tag{46}$$

$$\frac{dW_{ij}^{ON}}{dt} = 50\left[Q_i\right]^+\left[N_i - 0.5\right]^+\left(\left[Q_i\right]^+ - W_{ij}^{ON}\right), \tag{47}$$

$$\frac{dW_{ji}^{NO}}{dt} = 24\left[N_j - 0.5\right]^+\left[O_i\right]^+\left(\left[N_j - 0.5\right]^+ - W_{ji}^{NO}\right), \tag{48}$$

The superscripts *V*, *O*, and *N* refer to view, object and name category neurons, respectively. Thus, W_{ij}^{VO} is the excitatory weight from the i^{th} view category neuron to the j^{th} object category neuron, W_{ij}^{ON} is the weight from the i^{th} object integrator neuron to the j^{th} name category neuron, and W_{ji}^{NO} is the weight from the j^{th} name category neuron to the i^{th} object category neuron. The weight between the view layer and the object boundaries map, W^{BV}, was described in (39).

C7. Model Implementation Issues. The number of equations in the model and the extent of the input images make it computationally heavy to simulate the entire model at the same time. To overcome this, the model Where stream was simulated separately from the What stream to show the validity of its equations. Once it was verified that the Where stream produces correct results at the correct times (see below), the What stream was simulated in isolation, receiving the desired Where stream inputs at the appropriate times.

The two streams of the model interact in a predictable way. The Where stream decides (1) which object to attend, (2) which hotspots to look at on that attended object, (3) when to send a reset signal to the object neurons in the What stream if the attentional shroud around that object breaks, and (4) to down-regulate the boundaries of other objects. If we show that these properties hold in the Where stream on a small scene, we can avoid simulating the entire dynamics of the Where stream on the full database and feed the end results of Where stream dynamics at their predicted times to the What stream.

We thereby implemented the full Where stream equations using different sets of two to four letters, rather than the entire scene of 440 letters, as the input image to the model. In this small version we tested all 10 letters in the database (LFEHKDCOGQ), in at least two poses for each letter: one small tilted to left or right, and one large size tilted to left or right. Each such exemplar of a letter was presented in a scene alongside one to three other exemplars. To verify the generality of Where stream parameters, we tested some hand-written characters and simple images as well. This small version showed that (1) the attentional shroud promptly forms around one letter and then breaks and moves to another one, (2) while the shroud forms on one letter, hotspots (corners, intersections, and high curvatures) of that letter are serially visited, (3) the category reset signal only gets active when the shroud breaks, and (4) while the shroud forms around one letter, the feedback from spatial attention can down-regulate all other letters' boundaries.

Once the selected parameters were such that these conditions were met, we fed the end-results of the Where stream equations to the What stream. We moved the fixation from one hotspot to the next most active one every 0.3 seconds, generated the log-polar map of the scene with the fovea in that fixation as the input to the What stream, and attenuated any boundary that belonged to other letters in this input. The time constants of the shroud in our simulations allowed the shroud to last long enough for the eye movement map to visit about 15-20 hotspots, so after we moved the fixations between this many hotspots, we sent a R_{WHERE} category reset signal to both object layers in the What stream and placed the shroud on another object.

A concern in this method of implementation is whether the next saccade target can be predicted only based on the input without actually running all the equations in the Where stream. As observed in the small version of the simulation, *where* the model is looking at on an attended object only slightly affects the selection of the *next* saccade target. The reason is that the hotspots are so active that even placing them peripherally on the log-polar map cannot attenuate them to lose in the competition. In order to respect the order of selecting the hotspots, when running the What stream simulation, we selected the next saccade target as the most active location on the log-polar map after *each* fixation, i.e. we moved the fovea to a new hotspot, computed the object boundary map activity of the *attended* object to compute hotspots, and selected the most active one other than the fovea to be the next fixation point, and repeated the process. This is how the hotspots are selected in the actual small version simulation of the Where stream.

Appendix 2: Obtaining a Regular Map for V1 and Dealing with Border Effects

This section describes how the log-polar map in Equation (3) was implemented. Retinal neuron densities are higher in the fovea than in the periphery of the retina, yet these neurons project to a regular grid on V1. How should the neurons on each half of a retina be arranged in order for their corresponding V1 neurons to form a regular grid? One solution is to (1) define retinal ganglion cell locations on the boundaries of a half-disc, (2) use the *forward transformation* in Equation (3) to obtain the corresponding boundaries of V1, (3) define locations of V1 cells to form a regular grid covering these boundaries, as well as a padding, (4) use the *inverse transform* of the log-polar mapping to calculated the locations of the corresponding retinal ganglion cells, and (5) tessellate the retina between these retinal ganglion cells to obtain each cell's receptive field.

We defined the hemi-retina to have a radius of η (the light grey area in Figure 8a). The corresponding V1 cortex will have a size of $\gamma \times \kappa$ such that $\gamma = 7\log(\eta + 0.3)$ and $\kappa = 2 \cdot \text{imag}(7\log((i\eta + 0.3))$, where imag($a$) is the imaginary part of complex number a. A regular grid for V1 was defined as a set of discrete points $\left\{ p,q : -\beta < p < \gamma + \beta \text{ and } -\frac{\kappa}{2} - \beta < q < \frac{\kappa}{2} + \beta \right\}$, where β is the size of the padding. Each point on V1 can be represented by a complex number $W=p+iq$. The reverse transform of $W = b\log(Z + a)$ in Equation (3) is $Z = e^{\frac{W}{b}} - a$ and gives the corresponding location *(m,n)* for each retinal ganglion cells in the form of the complex number $Z=m+in$. These locations were used to tessellate the retina and obtain each cell's receptive field, using a MATLAB® algorithm (Barber, Dobkin & Huhdanpaa, 1996). This receptive fields are shown in Figure 8a.

In the above method, the actual V1 cells are the light grey cells in Figure 8b and correspond to the actual hemi-retina of radius η (the light grey area in Figure 8a). The padding around this

region (the dark grey cells of Figure 8b) serves as a border to offset the "border artifact" problem in image processing. The retinal cells that correspond to these V1 padding cells are shown as dark grey cells in Figure 8a. Note that on the left of Figure 8a, these cells extend beyond the vertical meridian of retina, and therefore fall on the other hemi-retina. We have therefore sampled the opposite hemi-field along the vertical meridian to avoid border effects in the V1 cortex. This actually happens in biology where the vertical meridian neurons in V1 are connected to their counterparts in the other V1 through corpus callosum (Essen & Zeki, 1978).

REFERENCES

Andersen, R.A., Essick, G.K., & Siegel, R.M. (1985). Encoding of spatial location by posterior parietal neurons. *Science, 230* (4724), 456-458.

Andersen, R.A., & Mountcastle, V.B. (1983). The influence of the angle of gaze upon the excitability of the light- sensitive neurons of the posterior parietal cortex. *Journal of Neuroscience, 3* (3), 532-548.

Ashby, F.G., & Ell, S.W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5* (5), 204-210.

Awh, E., K. M. Armstrong, Moore, T. (2006). "Visual and oculomotor selection: links, causes and implications for spatial attention." *Trends in Cognitive Science* **10**(3): 124-30.

Backus, B.T., Fleet, D.J., Parker, A.J., & Heeger, D.J. (2001). Human cortical activity correlates with stereoscopic depth perception. *Journal of Neurophysiology, 86* (4), 2054-2068.

Baloch, A.A., & Grossberg, S. (1997). A neural model of high-level motion processing: line motion and formotion dynamics. *Vision Research, 37* (21), 3037-3059.

Baloch, A.A., & Waxman, A.M. (1991). Visual learning, adaptive expectations, and behavioral conditioning of the mobile robot MAVIN. *Neural Networks, 4* (3), 271-302.

Barber, C.B., Dobkin, D.P., & Huhdanpaa, H. (1996). The Quickhull algorithm for convex hulls. *Acm Transactions on Mathematical Software, 22* (4), 469-483.

Baylis, G.C., & Driver, J. (2001). Shape-coding in IT cells generalizes over contrast and mirror reversal, but not figure-ground reversal. *Nature Neuroscience, 4* (9), 937-942.

Beardslee, D.C., & Wertheimer, M. (1958). Readings in perception. In: *The university series in psychology* (pp. 194-203, 751 p.). Princeton, N.J., Van Nostrand.

Beauvillain, C., Vergilino-Perez, D., & Dukic, T. (2005). Spatial object representation and its use in planning eye movements. *Experimental Brain Research, 165* (3), 315-327.

Berzhanskaya, J., Grossberg, S., & Mingolla, E. (2007). Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. *Spatial Vision, in press.*

Blaser, E., Pylyshyn, Z.W., & Holcombe, A.O. (2000). Tracking an object through feature space. *Nature, 408* (6809), 196-199.

Booth, M.C., & Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex, 8* (6), 510-523.

Borg-Graham, L.J., Monier, C., & Fregnac, Y. (1998). Visual input evokes transient and strong shunting inhibition in visual cortical neurons. *Nature, 393* (6683), 369-373.

Bradski, G., & Grossberg, S. (1995). Fast-learning VIEWNET architectures for recognizing three-dimensional objects from multiple two-dimensional views. *Neural Networks, 8* (7-8), 1053-1080.

Brown, J. M., & Denny, H.I. (2007). Shifting attention into and out of objects: Evaluating the processes underlying the object advantage. *Perception & Psychophysics*, 69, 608-618.

Bulthoff, H.H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences of the United States of America, 89* (1), 60-64.

Bulthoff, H.H., Edelman, S.Y., & Tarr, M.J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 5* (3), 247-260.

Cao, Y., & Grossberg, S. (2005). A laminar cortical model of stereopsis and 3D surface perception: closure and da Vinci stereopsis. *Spatial Vision, 18* (5), 515-578.

Caplovitz, G.P., & Tse, P.U. (2006). V3A processes contour curvature as a trackable feature for the perception of rotational motion. *Cerebral Cortex, in press.*

Carlson-Radvansky, L.A. (1999). Memory for relational information across eye movements. *Perception and Psychophysics, 61* (5), 919-934.

Carpenter, G.A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern-recognition machine. *Computer Vision Graphics and Image Processing, 37* (1), 54-115.

Carpenter, G.A., & Grossberg, S. (1993). Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences, 16* (4), 131-137.

Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., & Rosen, D.B. (1992). Fuzzy ARTMAP - a neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks, 3* (5), 698-713.

Carpenter, G.A., Grossberg, S., & Reynolds, J.H. (1991). ARTMAP - Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks, 4* (5), 565-588.

Carpenter, G.A., & Ross, W.D. (1993). ART-EMAP: A neural network architecture for learning and prediction by evidence accumulation. *Proceedings of the World Congress on Neural Networks, (WCNN- 93),* III (pp. 649-656).

Carrasco, M., Penpeci-Talgar, C., & Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Research, 40* (10-12), 1203-1215.

Cavada, C., & Goldman-Rakic, P.S. (1989). Posterior parietal cortex in rhesus monkey: II. Evidence for segregated corticocortical networks linking sensory and limbic areas with the frontal lobe. *Journal of Comparative Neurology, 287* (4), 422-445.

Cavada, C., & Goldman-Rakic, P.S. (1991). Topographic segregation of corticostriatal projections from posterior parietal subdivisions in the macaque monkey. *Neuroscience, 42* (3), 683-696.

Cavanagh, P., Labianca, A.T., & Thornton, I.M. (2001). Attention-based visual routines: sprites. *Cognition, 80* (1-2), 47-60.

Cohen, M.A., & Grossberg, S. (1984). Neural dynamics of brightness perception: features, boundaries, diffusion, and resonance. *Perception and Psychophysics, 36* (5), 428-456.

Colby, C.L., Duhamel, J.R., & Goldberg, M.E. (1993). The analysis of visual space by the lateral intraparietal area of the monkey: the role of extraretinal signals. *Progress in Brain Research, 95*, 307-316.

Cooper, E.E., Biederman, I., & Hummel, J.E. (1992). Metric invariance in object recognition - a review and further evidence. *Canadian Journal of Psychology-Revue Canadienne De Psychologie, 46* (2), 191-214.

Daniel, P., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *Journal of Physiology, 159*, 203--221,.

Deneve, S., & Pouget, A. (2003). Basis functions for object-centered representations. *Neuron, 37* (2), 347-359.

Deubel, H., & Schneider, W.X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Research, 36* (12), 1827-1837.

Deubel, H., Schneider, W.X., & Bridgeman, B. (2002). Transsaccadic memory of position and form. *Progress in Brain Research, 140*, 165-180.

Distler, C., Boussaoud, D., Desimone, R., & Ungerleider, L.G. (1993). Cortical connections of inferior temporal area TEO in macaque monkeys. *Journal of Comparative Neurology, 334* (1), 125-150.

Downing, C.J. (1988). Expectancy and visual-spatial attention: effects on perceptual quality. *Journal of Experimental Psychology: Human Perception and Performance, 14* (2), 188-202.

Drasdo, N. (1977). The neural representation of visual space. *Nature, 266* (5602), 554-556.

Driver, J., & Baylis, G.C. (1996). Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognitive Psychology, 31* (3), 248-306.

Duhamel, J.R., Colby, C.L., & Goldberg, M.E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science, 255* (5040), 90-92.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113* (4), 501-517.

Edelman, S., & Poggio, T. (1991). Models of object recognition. *Current Opinion in Neurobiology, 1* (2), 270-273.

Egeth, H.E., Virzi, R.A., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance, 10* (1), 32-39.

Egeth, H.E., & Yantis, S. (1997). Visual attention: control, representation, and time course. *Annual Review of Psychology, 48*, 269-297.

Egly, R., Driver, J., & Rafal, R.D. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General, 123* (2), 161-177.

Elder, D.M., Grossberg, S., & Mingolla, E. (2005). A neural model of visually-guided steering, obstacle avoidance, and route selection. *Society for Neuroscience,* Program No. 390.7 (Washington, DC: Abstract Viewer/Itinerary Planner).

Elder, J., & Zucker, S. (1993). The effect of contour closure on the rapid discrimination of 2-dimensional shapes. *Vision Research, 33* (7), 981-991.

Eriksen, C.W., & Yeh, Y.Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance, 11* (5), 583-597.

Essen, D.C., & Zeki, S.M. (1978). The topographic organization of rhesus monkey prestriate cortex. *Journal of Physiology, 277*, 193-226.

Fang, L. and Grossberg, S. (2007) From stereogram to surface: How the brain sees the world in depth. *Spatial Vision*, in press.

Fazl, A., Grossberg, S., & Mingolla, E. (2004). Invariant object learning and recognition using active eye movements and attentional control. *Society for Neuroscience,* Program No. 605.10 (Washington, DC: Abstract Viewer-Online).

Fazl, A., Grossberg, S., & Mingolla, E. (2005). Invariant object learning and recognition using active eye movements and attentional control. *Journal of Vision, 5* (8), 738-738.

Fazl, A., Grossberg, S., & Mingolla, E. (2006). View-invariant object category learning: How spatial and object attention are coordinated using surface-based attentional shrouds. *Journal of Vision, 6* (6), 315.

Fecteau, J. H., & Munoz, D. P. (2003). Exploring the consequences of the previous trial. *Nature Reviews, Neuroscience,4*(6), 435-43.

Findlay, J.M. (1995). Visual-Search - eye-movements and peripheral-vision. *Optometry and Vision Science, 72* (7), 461-466.

Findlay, J.M. (1997). Saccade target selection during visual search. *Vision Research, 37* (5), 617-631.

Fischer, B. (1973). Overlap of receptive field centers and representation of the visual field in the cat's optic tract. *Vision Research, 13* (11), 2113-2120.

Galletti, C., & Battaglini, P.P. (1989). Gaze-dependent visual neurons in area V3A of monkey prestriate cortex. *Journal of Neuroscience, 9* (4), 1112-1125.

Gancarz, G., & Grossberg, S. (1999). A neural model of saccadic eye movement control explains task-specific adaptation. *Vision Research, 39* (18), 3123-3143.

Gauthier, I., & Tarr, M.J. (1997). Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vision Research, 37* (12), 1673-1682.

Gilchrist, I.D., Heywood, C.A., & Findlay, J.M. (2003). Visual sensitivity in search tasks depends on the response requirement. *Spatial Vision, 16* (3-4), 277-293.

Grossberg, S. (1972). A neural theory of punishment and avoidance, II: Quantitative theory. *Mathematical Biosciences, 15*, 253-285.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics, 52*, 213-257.

Grossberg, S. (1978a). Decisions, patterns, and oscillations in nonlinear competitive systems with applications to Volterra-Lotka systems. *Journal of Theoretical Biology, 73* (1), 101-130.

Grossberg, S. (1978b). Do all neural models really look alike? A comment on Anderson, Silverstein, Ritz, and Jones. *Psychological Review, 85* (6), 592-596.

Grossberg, S. (1980a). Biological competition: Decision rules, pattern formation, and oscillations. *Proceedings of the National Academy of Sciences of the United States of America, 77* (4), 2338-2342.

Grossberg, S. (1980b). How does a brain build a cognitive code? *Psychological Review, 87* (1), 1-51.

Grossberg, S. (1984). Some psychophysiological and pharmacological correlates of a developmental, cognitive and motivational theory. *Annals of the New York Academy of Sciences, 425*, 58-151.

Grossberg, S. (1987a). Cortical dynamics of 3-dimensional form, color, and brightness perception .1. monocular theory. *Perception and Psychophysics, 41* (2), 87-116.

Grossberg, S. (1987b). Cortical dynamics of 3-dimensional form, color, and brightness perception .2. binocular theory. *Perception and Psychophysics, 41* (2), 117-158.

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Perception and Psychophysics, 55* (1), 48-121.

Grossberg, S. (1997). Cortical dynamics of three-dimensional figure-ground perception of two-dimensional figures. *Psychological Review*, *104* , 618-658.

Grossberg, S. (1999a). How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. *Spatial Vision, 12* (2), 163-185.

Grossberg, S. (1999b). The link between brain learning, attention, and consciousness. *Consciousness and Cognition, 8* (1), 1-44.

Grossberg, S. (2000). How hallucinations may arise from brain mechanisms of learning, attention, and volition. Invited article for the *Journal of the International Neuropsychological Society*, 6**,** 579-588.

Grossberg, S. (2003). How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, *2*, 47-76.

Grossberg, S. (2008). Cortical dynamics of attentive object recognition, scene understanding, and decision making. SperlingFest book.

Grossberg, S., & Hong, S. (2006). A neural model of surface perception: lightness, anchoring, and filling-in. *Spatial Vision, 19* (2-4), 263-321.

Grossberg, S., & Howe, P.D. (2003). A laminar cortical model of stereopsis and three-dimensional surface perception. *Vision Research, 43* (7), 801-829.

Grossberg, S., & Huang, T.-R. (2008). ARTSCENE: A neural system for natural scence classification. *Journal of Vision*, in press.

Grossberg, S., Hwang, S., & Mingolla, E. (2002). Thalamocortical dynamics of the McCollough effect: boundary-surface alignment through perceptual learning. *Vision Research, 42* (10), 1259-1286.

Grossberg, S., & Kelly, F. (1999). Neural dynamics of binocular brightness perception. *Vision Research, 39* (22), 3796-3816.

Grossberg, S., & Kuperstein, M. (1986). Neural dynamics of adaptive sensory-motor control: ballistic eye movements. In: (pp. xvi, 336 p.). Amsterdam, New York North-Holland.

Grossberg, S. and Myers, C.W. (2000) The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review, 107*, 735-767.

Grossberg, S., & Mingolla, E. (1985a). Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. *Psychological Review, 92* (2), 173-211.

Grossberg, S., & Mingolla, E. (1985b). Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. *Perception and Psychophysics, 38* (2), 141-171.

Grossberg, S., & Mingolla, E. (1987). Neural dynamics of surface perception - boundary webs, illuminants, and shape-from-shading. *Computer Vision Graphics and Image Processing, 37* (1), 116-165.

Grossberg, S., & Mingolla, E. (1993). Neural dynamics of motion perception: direction fields, apertures, and resonant grouping. *Perception and Psychophysics, 53* (3), 243-278.

Grossberg, S., Mingolla, E., & Ross, W.D. (1994). A neural theory of attentive visual search: interactions of boundary, surface, spatial, and object representations. *Psychological Review, 101* (3), 470-489.

Grossberg, S., Mingolla, E., & Viswanathan, L. (2001). Neural dynamics of motion integration and segmentation within and across apertures. *Vision Research, 41* (19), 2521-2553.

Grossberg, S. and Pearson, L. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: Toward a unified theory of how the cerebral cortex works. *Psychological Review,* in press.

Grossberg, S., & Raizada, R.D. (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research, 40* (10-12), 1413-1432.

Grossberg, S., & Repin, D.V. (2003). A neural model of how the brain represents and compares multi-digit numbers: spatial and categorical processes. *Neural Networks, 16* (8), 1107-1140.

Grossberg, S., Srihasam, K., & Bullock, D. (2008). Neural dynamics of saccadic and smooth pursuit eye movement coordination during visual tracking of unpredictably moving targets. Technical Report CAS/CNS TR-2007-018. Submitted for publication.

Grossberg, S., & Swaminathan, G. (2004). A laminar cortical model for 3D perception of slanted and curved surfaces and of 2D images: development, attention, and bistability. *Vision Research, 44* (11), 1147-1187.

Grossberg, S., & Todorovic, D. (1988). Neural dynamics of 1-D and 2-D brightness perception: a unified model of classical and recent phenomena. *Perception and Psychophysics, 43* (3), 241-277.

Grossberg, S., & Yazdanbakhsh, A. (2005). Laminar cortical dynamics of 3D surface perception: stratification, transparency, and neon color spreading. *Vision Research, 45* (13), 1725-1743.

Haxby, J.V., Grady, C.L., Horwitz, B., Ungerleider, L.G., Mishkin, M., Carson, R.E., Herscovitch, P., Schapiro, M.B., & Rapoport, S.I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences of the United States of America, 88* (5), 1621-1625.

He, Z.J., & Nakayama, K. (1995). Visual attention to surfaces in three-dimensional space. *Proceedings of the National Academy of Sciences of the United States of America, 92* (24), 11155-11159.

Henderson, J.M., & Hollingworth, A. (2003). Global transsaccadic change blindness during scene perception. *Psychological Science, 14* (5), 493-497.

Hochberg, J., & Peterson, M. A. (1987). Piecemeal perception and cognitive components in object perception: Perceptually coupled responses to moving objects. *Journal of Experimental Psychology*: *General*, *116*, 370-380.

Hollingworth, A., Richard, A. M., & Luck, S. J. (2008). Understanding the function of visual short-term memory: transsaccadic memory, object correspondence, and gaze correction. *Journal of Experimental Psychology: General*, *137*, 163-181.

Irwin, D.E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology, 23* (3), 420-456.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2* (3), 194-203.

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 181-211). Hillsdale, NJ: Erlbaum.

Kahneman, D., Treisman, A., & Gibbs, B.J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology, 24* (2), 175-219.

Kelly, F., & Grossberg, S. (2000). Neural dynamics of 3-D surface perception: figure-ground separation and lightness perception. *Perception and Psychophysics, 62* (8), 1596-1618.

Kourtzi, Z., & Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science, 293* (5534), 1506-1509.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35* (13), 1897-1916.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision, 13* (2-3), 201-214.

LaBerge, D. (1995). Attentional processing: The brain's art of mindfulness. Cambridge, Mass., Harvard University Press.

LaBerge, D., & Brown, V. (1989). Theory of attentional operations in shape identification. *Psychological Review, 96* (1), 101-124.

Likova, L.T., & Tyler, C.W. (2003). Peak localization of sparsely sampled luminance patterns is based on interpolated 3D surface representation. *Vision Research*, *43*, 2649-2657.

List, A. and Robertson, L. C. (2007). Inhibition of return and object-based attentional selection. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1322-34.

Logan, G. D. (1996). The CODE theory of visual attention: An integration of space-based and object-based attention. *Psychological Review,* 103 (4), 603-49.

Logothetis, N.K., Pauls, J., Bulthoff, H.H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4* (5), 401-414.

Lueschow, A., Miller, E.K., & Desimone, R. (1994). Inferior temporal mechanisms for invariant object recognition. *Cerebral Cortex, 4* (5), 523-531.

Marr, D., & Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 200* (1140), 269-294.

McLoughlin, N.P., & Grossberg, S. (1998). Cortical computation of stereo disparity. *Vision Research, 38* (1), 91-99.

McMains, S.A., & Somers, D.C. (2005). Processing efficiency of divided spatial attention mechanisms in human visual cortex. *Journal of Neuroscience, 25* (41), 9444-9448.

Mingolla, E., Ross, W., & Grossberg, S. (1999). A neural network for enhancing boundaries and surfaces in synthetic aperture radar images. *Neural Networks, 12* (3), 499-511.

Mishkin, M., Lewis, M.E., & Ungerleider, L.G. (1982). Equivalence of parieto-preoccipital subareas for visuospatial ability in monkeys. *Behavioral and Brain Sciences, 6* (1), 41-55.

Moore, C.M., & Fulton, C. (2005). The spread of attention to hidden portions of occluded surfaces. *Psychonomic Bulletin & Review, 12* (2), 301-306.

Nakamura, H., Kuroda, T., Wakita, M., Kusunoki, M., Kato, A., Mikami, A., Sakata, H., & Itoh, K. (2001). From three-dimensional space vision to prehensile hand movements: the lateral intraparietal area links the area V3A and the anterior intraparietal area in macaques. *Journal of Neuroscience, 21* (20), 8174-8187.

Nakamura, K., & Colby, C.L. (2000). Visual, saccade-related, and cognitive activation of single neurons in monkey extrastriate area V3A. *Journal of Neurophysiology, 84* (2), 677-692.

Nakayama, K., & Silverman, G.H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320* (6059), 264-265.

Nieman, D.R., Hayashi, R., Andersen, R.A., & Shimojo, S. (2005). Gaze direction modulates visual aftereffects in depth and color. *Vision Research, 45* (22), 2885-2894.

O'Craven, K.M., Downing, P.E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature, 401* (6753), 584-587.

Paradiso, M.A., & Nakayama, K. (1991). Brightness Perception and Filling-In. *Vision Research, 31* (7-8), 1221-1236.

Pasupathy, A. (2006). Neural basis of shape representation in the primate brain. *Progress in Brain Research, 154*, 293-313.

Peterson, M. A. (1994). Object recognition processes can and do operate before figure-ground organization. *Current Directions in Psychological Science, 3,* 105-111.

Peterson, M. A., & Gibson, B. S. (1991) Direction spatial attention within an object: Altering the functional equivalence of shape description. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 170-182.

Peterson, M. A., Harvey, E. H., and Weidenbacher, H. L. (1991). Shape recognition inputs to figure-ground organization: Which route counts? *Journal of Experimental Psychology: Human Perception and Performance, 17*, 1075-1089.

Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32* (1), 3-25.

Posner, M.I., Walker, J.A., Friedrich, F.A., & Rafal, R.D. (1987). How do the parietal lobes direct covert attention? *Neuropsychologia, 25* (1A), 135-145.

Posner, M.I., Walker, J.A., Friedrich, F.J., & Rafal, R.D. (1984). Effects of parietal injury on covert orienting of attention. *Journal of Neuroscience, 4* (7), 1863-1874.

Pouget, A., Dayan, P., & Zemel, R.S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience, 26*, 381-410.

Pouget, A., & Snyder, L.H. (2000). Computational approaches to sensorimotor transformations. *Nature Neuroscience, 3 Suppl*, 1192-1198.

Prinzmetal, W. and Keysar, B. (1989). Functional theory of illusory conjunctions and neon colors. *Journal of Experimental Psychology, General*, 118(2), 165-90.

Pylyshyn, Z.W., & Storm, R.W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision, 3* (3), 179-197.

Rainer, G., & Miller, E.K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron, 27* (1), 179-189.

Raizada, R.D., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: computational clues from the visual system. *Cerebral Cortex, 13* (1), 100-113.

Ranganath, C. (2006). Working memory for visual objects: complementary roles of inferior temporal, medial temporal, and prefrontal cortex. *Neuroscience, 139* (1), 277-289.

Reynolds, J.H., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron, 37* (5), 853-863.

Reynolds, J.H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26* (3), 703-714.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2* (11), 1019-1025.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3 Suppl*, 1199-1204.

Roelfsema, P.R., Lamme, V.A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395* (6700), 376-381.

Rogers-Ramachandran, D.C., & Ramachandran, V.S. (1998). Psychophysical evidence for boundary and surface systems in human vision. *Vision Research, 38* (1), 71-77.

Rolls, E.T., Judge, S.J., & Sanghera, M.K. (1977). Activity of neurones in the inferotemporal cortex of the alert monkey. *Brain Research, 130* (2), 229-238.

Rubin, E. (1921). Visuell Wahrgenommene Figuren; Studien in Psychologischer Analyse. In.Copenhagen Gyldendalske Boghandel.

Schall, J. D. & Boucher, L. (2007). Executive control of gaze by the frontal lobes. *Cognitive, Affective, & Behavioral Neuroscience,* **7**(4), 396-412.

Scholte, H.S., Spekreijse, H., & Roelfsema, P.R. (2001). The spatial profile of visual attention in mental curve tracing. *Vision Research, 41* (20), 2569-2580.

Schwartz, E.L. (1977). Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biological Cybernetics, 25* (4), 181-194.

Schwartz, E.L. (1980). Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research, 20* (8), 645-669.

Seibert, M., & Waxman, A.M. (1992). Adaptive 3-D Object Recognition from Multiple Views. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14* (2), 107-124.

Serences, J.T., Schwarzbach, J., Courtney, S.M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex, 14* (12), 1346-1357.

Srihasam, K., Bullock, D., & Grossberg, S. (2008). Target selection by frontal cortex during coordinated saccadic and smooth pursuit eye movements. *Journal of Cognitive Neuroscience*, in press.

Tanaka, K. (1997). Mechanisms of visual object recognition: monkey and human studies. *Current Opinion in Neurobiology, 7* (4), 523-529.

Tanaka, K. (2000). Mechanisms of visual object recognition studied in monkeys. *Spatial Vision, 13* (2-3), 147-163.

Tarr, M.J., & Bulthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology-Human Perception and Performance, 21* (6), 1494-1505.

Tarr, M.J., & Bulthoff, H.H. (1998). Image-based object recognition in man, monkey and machine. *Cognition, 67* (1-2), 1-20.

Tarr, M.J., Williams, P., Hayward, W.G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience, 1* (4), 275-277.

Thier, P., & Andersen, R.A. (1998). Electrical microstimulation distinguishes distinct saccade-related areas in the posterior parietal cortex. *Journal of Neurophysiology, 80* (4), 1713-1735.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381* (6582), 520-522.

Tootell, R.B., Mendola, J.D., Hadjikhani, N.K., Ledden, P.J., Liu, A.K., Reppas, J.B., Sereno, M.I., & Dale, A.M. (1997). Functional analysis of V3A and related areas in human visual cortex. *Journal of Neuroscience, 17* (18), 7060-7078.

Tootell, R.B., Silverman, M.S., Switkes, E., & De Valois, R.L. (1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science, 218* (4575), 902-904.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, *14*, 107-141.

Tse, P. U. (2005). Voluntary attention modulates the brightness of overlapping transparent surfaces. *Vision Research, 45*, 1095-1098.

Tyler, C.W., & Kontsevich, L.L. (1995). Mechanisms of stereoscopic processing: stereoattention and surface perception in depth reconstruction. *Perception, 24* (2), 127-153.

Van Essen, D.C., Newsome, W.T., & Maunsell, J.H. (1984). The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. *Vision Research, 24* (5), 429-448.

Vergilino-Perez, D., & Findlay, J.M. (2004). Object structure and saccade planning. *Brain Research. Cognitive Brain Research, 20* (3), 525-528.

Vuilleumier, P., Henson, R.N., Driver, J., & Dolan, R.J. (2002). Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nature Neuroscience, 5* (5), 491-499.

Wassle, H., Grunert, U., Rohrenbeck, J., & Boycott, B.B. (1989). Cortical magnification factor and the ganglion cell density of the primate retina. *Nature, 341* (6243), 643-646.

Webster, M.J., Bachevalier, J., & Ungerleider, L.G. (1994). Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cerebral Cortex, 4* (5), 470-483.

Wolfe, J.M. (1994) Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulletin & Review*, *1*(2), 202-238.

Wolfe, J.M., Cave, K.R., & Franzel, S.L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15* (3), 419-433.

Wolfe, J.M., Friedman-Hill, S.R., & Bilsky, A.B. (1994). Parallel processing of part-whole information in visual search tasks. *Perception and Psychophysics, 55* (5), 537-550.

Yantis, S., & Serences, J.T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology, 13* (2), 187-193.

Yantis, S. (1992). Multielement visual tracking: attention and perceptual organization. *Cognitive Psychology, 24* (3), 295-340.