A laminar cortical model
for 3D boundary and surface representations of complex natural scenes

Yongqiang Cao* and Stephen Grossberg**

*Intel Labs, 3600 Juliette Ln, Santa Clara, CA 95054

**Center for Adaptive Systems
Graduate Program in Cognitive and Neural Systems
Department of Mathematics & Statistics, Psychological & Brain Sciences,
and Biomedical Engineering, Boston University
677 Beacon Street
Boston, MA 02215

Submitted: May 10, 2018
Revised: August 29, 2018

All correspondence should be addressed to
Professor Stephen Grossberg
Center for Adaptive Systems, Room 213
Boston University
677 Beacon Street
Boston, MA 02215
Phone: 617-353-7858
Fax: 617-353-7755
Email: steve@bu.edu

Abstract

How does the visual cortex process complex natural scenes? The 3D LAMINART model has been developed to clarify how laminar cortical mechanisms interact to create 3D boundary and surface representations that are embodied in conscious 3D percepts, and to thereby explain and predict data from psychophysical, neurophysiological, and neuroanatomical experiments. Here the model is applied to show how the same mechanisms, suitably refined, can generate 3D boundary and surface representations in response to natural scenes. Model accuracy on tested scenes is comparable to state-of-the-art results. The 3D LAMINART model does not, however, directly create a disparity map. Rather, it generates a 3D representation of a scene's boundary groupings and filled-in surfaces that can be used to see and recognize objects in it, as well as to derive such a map for benchmark purposes.

Keywords: visual cortex, stereopsis, natural scenes, binocular vision, surface perception, lightness perception, monocular-binocular interactions, 3D LAMINART model

Acknowledgements

The authors would like to thank Dr. Y. Ohta and Dr. Y. Nakamura for supplying the ground truth data from the University of Tsukuba.

Supported in part by CELEST, an NSF Science of Learning Center (SBE-0354378), and by the SyNAPSE program of the Defense Advanced Research Projects Agency (HR0011-09-3-0001 and HR0011-09-C-0011).

1. Introduction: Different modeling approaches to representing natural scenes

Understanding how humans and other animals see the world in depth is an essential first step in understanding many visual behaviors. Many stereo algorithms for natural images have been developed in the computational community (e.g., Baker and Binford, 1981; Kanade and Okutomi, 1994; Levine, O'handley and Yagi, 1973; Lloyd, Haddow and Boyce, 1987; Marr and Poggio, 1976, 1979; Mori, Kidode and Asada, 1973; Xie, Girshick, and Farhadi, 2016; Zitnick and Kanade, 2000; Zontar and LeCun, 2015). A comparison of the explanatory range of the 3D LAMINART model proposed herein with some leading alternative algorithms is provided in Table 1.

Table 1. Properties of several computational and biological stereovision models.

Model	Natural scenes	3D Surface representation	Panum's limiting case	da Vinci stereopsis (Nakayama and Simojo, 1990; Gillam et al, 1999)	Dichoptic masking (McKee et al, 1994, 1995)	Contrast variations (Smallman and McKee, 1995)	Venetian blind effect (Howard and Rogers, 1995)	Polarity-reversed da Vinci stereopsis (Nakayama and Simojo, 1990)	Sparse images (Fang and Grossberg, 2009)	Bistable percepts
Bayesian diffusion (Scharstein and Szeliski, 1998)	Yes	No	No	No	No	No	No	No	No	No
Cooperative (Zitnick and Kanade, 2000)	Yes	No	No	No	No	No	No	No	No	No
Belief propagation (Sun et al, 2003)	Yes	No	No	No	No	No	No	No	No	No
Semiglobal (Hirschmuller, 2008)	Yes	No	No	No	No	No	No	No	No	No
Disparity energy (Chen and Qian, 2004; Assee and Qian, 2007)	Yes, but no accuracy reported	No	Yes	Nakayama and Simojo, 1990 only	No	No	No	No	No	No
3D LAMINART	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Area-based methods. These algorithms can be divided into two categories: area-based methods and feature-based methods. For area-based methods (e.g., Kanad and Okutomi, 1994; Zitnick and Kanade, 2000), a small window centered at a given pixel is chosen as the basic unit that is matched across two images. A difficulty for this type of method is how to choose the size of supporting windows. A smaller window is usually desirable to avoid unwanted smoothing. However, in areas of homogeneity or low texture, a larger

window is needed so that the window contains enough intensity variation to achieve reliable matching.

In order to obtain a smooth and detailed disparity map, these methods generally use two basic assumptions: uniqueness and continuity. That is, a pixel in one image can correspond to no more than one pixel on the other image (uniqueness), and disparity is continuous for two neighboring pixels (continuity). However, both uniqueness and continuity assumptions cannot be strictly satisfied in most images. Depth is often discontinuous across edges, and the uniqueness constraint is not satisfied in response to many scenes and natural images. These properties are clearly depicted in psychophysical displays that depict Panum's limiting case and Da Vinci stereopsis (Cao and Grossberg, 2005, 2012; Grossberg, 1994; Grossberg and Howe, 2003; Grossberg and McLoughlin, 1997; Mckee et al., 1995, 1995; McLoughlin and Grossberg, 1998; Nakayama and Shimojo, 1990).

Panum's limiting case illustrates how our brains can binocularly match, and fuse, a single feature that is seen in one eye with more than one feature that is seen in the other eye. Da Vinci stereopsis illustrates how part of a depthful scene may be registered by only one retina due to occlusion by a nearer part of a scene. Despite the fact that the monocularly viewed part of the scene carries no depth information, it may be seen at a definite depth due to interactions with parts of the scene that are seen by both eyes. Such percepts are generated by brain mechanisms that violate uniqueness and continuity assumptions in multiple ways.

In general, area-based methods work well for some natural images, but they do not work well for images including large homogeneous regions, such as are found in numerous smooth human-made objects, and around edges where disparity is discontinuous.

With the successful recent application of convolutional neural networks (CNN) to vision problems, CNN-based stereo matching methods have appeared (e.g., Xie, Girshick, and Farhadi, 2016; Zbontar and LeCun, 2015;). These models first train a deep convolutional network on ground-truth stereo pairs of small image patches, and then do inference on other image pairs. Their advantage is the high performance on large datasets such as the KITTI stereo dataset (Geiger et al, 2013). In this paper, we do not aim at

competing in performance with these deep learning based models that need to train their networks on large ground-truth datasets first, but rather show that a biologically plausible laminar cortical model that explains many psychophysiological phenomena (see Table 1) can also be extended to do stereo-matching on natural images without any prior learning.

Feature-based methods. For feature-based methods (e.g., Sherman and Peleg, 1990), image features are the basic units that are matched. Features can include occlusion edges, vertices of linear structures, prominent surface markings, and intensity anomalies. In particular, both edge points and edge contours have been used as matching features. In edge-based methods, a main type of feature-based methods, one common constraint is edge consistency. That is, all matches along a continuous edge must be consistent. The uniqueness assumption is also used. In general, feature-based methods produce only a sparse disparity map.

3D LAMINART: Laminar cortical model of 3D boundary and surface formation. Although many stereo algorithms as aforementioned have been developed, they can neither explain how the visual cortex processes complex natural scenes, nor the percepts induced by many psychophysical displays. The 3D LAMINART model (Figures 1 and 2) has been developed to explain how the visual cortex sees. In particular, the model proposes how visual cortical areas are defined in terms of layered circuits, typically with six characteristic layers (Brodmann, 2009; Felleman and van Essen, 1991; Martin, 1989; Pandya and Yeterian, 1985), and how these laminar circuits interact using bottom-up, horizontal, and top-down connections to generate 3D boundary and surface representations whose properties simulate those of conscious 3D surface percepts. Here the model is used to show how these mechanisms, properly refined, can explain how the brain generates 3D boundary and surface representations in response to complex natural scenes. The model describes how early monocular and binocular cortical cells that carry out bottom-up adaptive filtering (e.g., latereal geniculate nucleus (LGN) and V1 cortical cells in Figure 2) interact with later stages of 3D boundary completion and surface filling-in (e.g., V2 and V4 cortical cells in Figure 2). In particular, the model proposes how interactions between layers 4, 3B, and 2/3 in V1 and V2 contribute to stereopsis (Figure 1), and how binocular and monocular information combine to complete 3D boundary and surface representations at subsequent cortical processing stages.

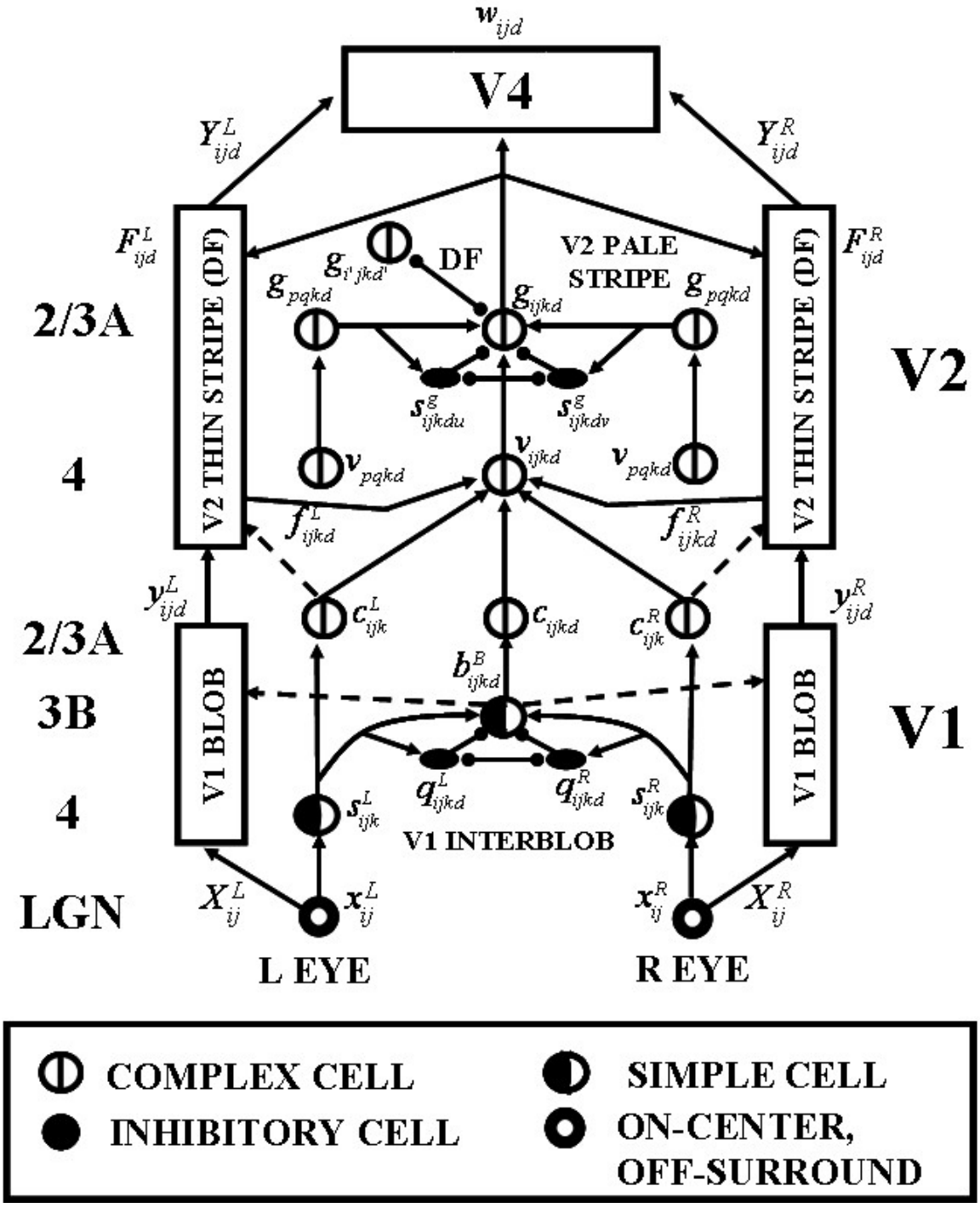

Figure 1. The 3D LAMINART model circuit diagram. The model consists of a V1 Interblob - V2 Pale Stripe stream (Boundary Stream, in green) and a V1 Blob - V2 Thin Stripe stream (Surface Stream, in red). The two processing streams interact to overcome their complementary deficiencies and create consistent 3D boundary and surface percepts. A disparity filter (DF) exists in both V2 pale stripes (boundary network) and thin strips (surface network), with the gray boxes denoting multiple surfaces which inhibit each other across depth in V2. The new connections are denoted in dashed lines.

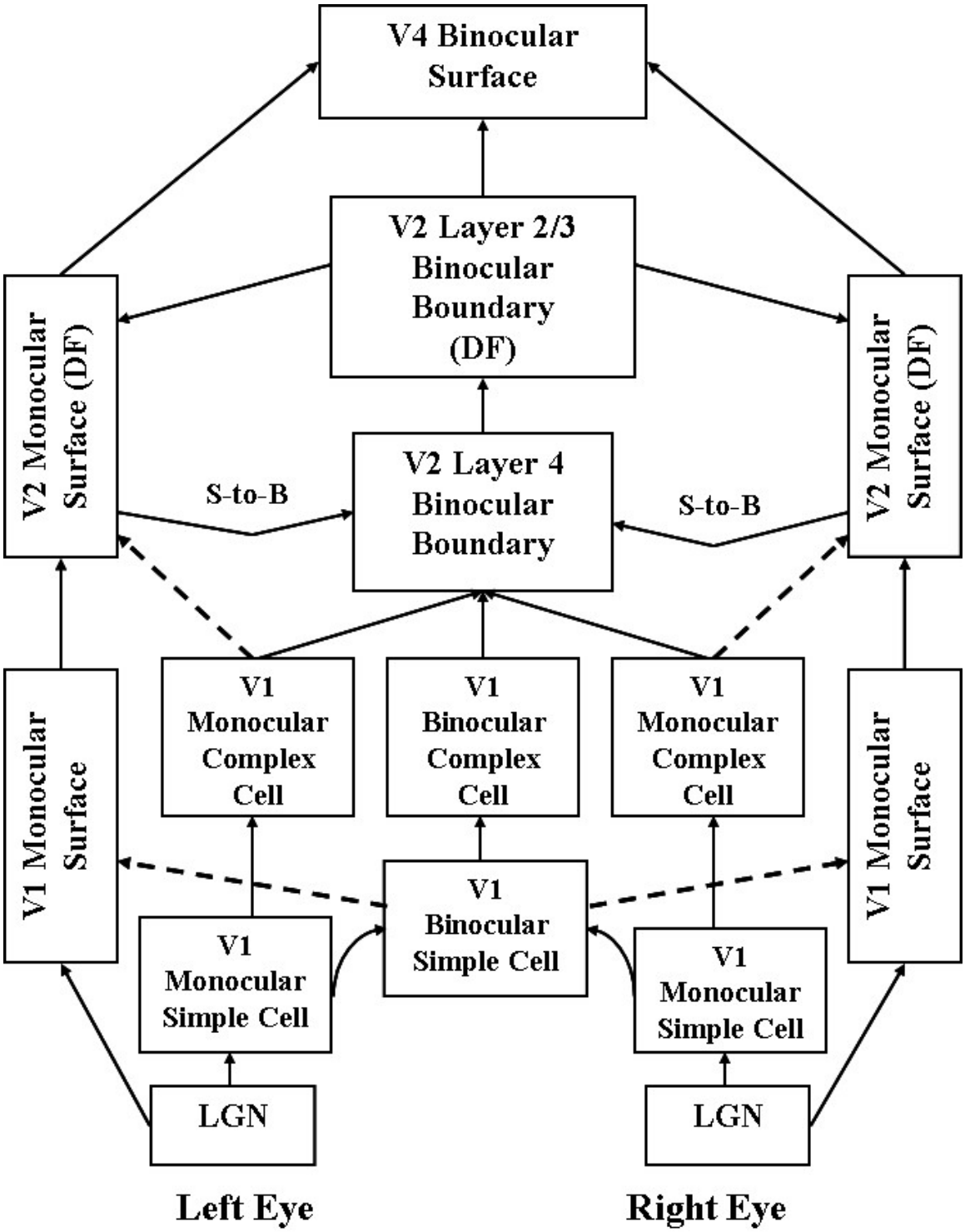

Figure 2. A simplified block diagram of the 3D LAMINART model. The new connections are denoted in dashed lines.

Since its introduction, the 3D LAMINART model has been incrementally refined through experimental and theoretical analyses that have led to the discovery of additional design constraints that explain and predict additional perceptual, anatomical, and neurophysiological data. Each such embodiment illustrates a "method of minimal anatomies" wherein every model process realizes functional properties without which significant bodies of data cannot be explained. By proceeding in this way, as increasingly complex brain processes are modeled, each model process continues to play clear functional roles that are tightly linked to data explanations.

Complementary boundaries and surfaces and complementary consistency. The 3D LAMINART model builds upon the discovery that the visual cortical streams that process perceptual boundaries and surfaces obey computationally *complementary* laws (Figure 3; e.g., Grossberg, 1994). In particular, boundaries are completed *inwardly* between pairs of similarly *oriented* and collinear cell populations (the so-called *bipole* grouping property; see Section 2.4 and equation (26)). This inward and oriented boundary process enables boundaries to complete across partially occluded object features. Boundaries also pool inputs from opposite contrast polarities, so are *insensitive to contrast polarity*. This pooling process enables boundaries to form around objects that are seen in front of backgrounds whose contrast polarities with respect to the object reverse around the object's perimeter (e.g., Figures 3a and 3c). Because boundaries pool across opposite contrast polarities (e.g., light-to-dark and dark-to-light contrasts), they give up the ability to represent visible contrast comparisons (e.g., lighter vs. darker), and thus cannot represent visual qualia. This conclusion can be vividly summarized as the claim that "all boundaries are invisible", or amodal, within the boundary cortical stream, which proceeds from the LGN through the interblobs of cortical area V1, then through the pale stripes of cortical area V2, and on to cortical area V4 (Figure 1)

If all boundaries are invisible, then how do we see the world? The 3D LAMINART model proposes that "all visible percepts are surface percepts" formed within the surface processing stream that passes from the LGN through the blobs of V1, then through the thin stripes of V2, and on to V4 (Figure 1). Within this stream, surface brightness and color fill-in *outwardly* in an *unoriented* manner until they reach object boundaries or dissipate due to their spread across space (Figure 3; Grossberg and Todorovic, 1988).

This filling-in process is also *sensitive* to contrast polarities because it subserves all conscious visual percepts.

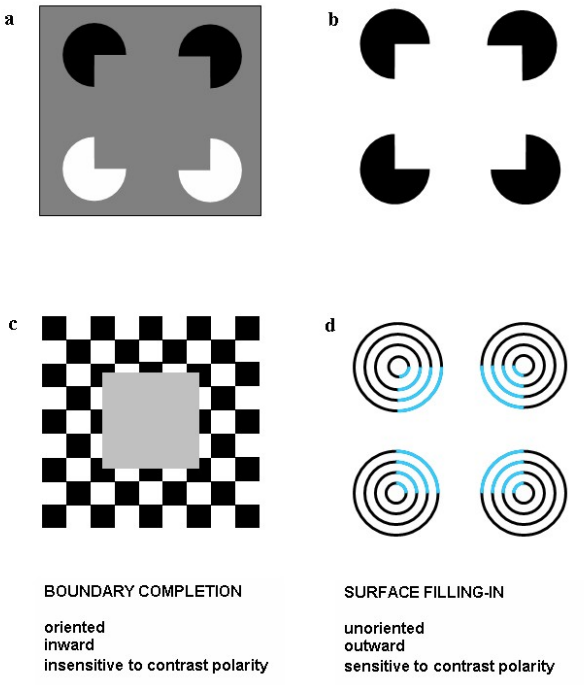

Figure 3. Examples of complementary boundary and surface processes. The square illusory contours in (a) and (b) illustrate that boundaries form in an *oriented* way *inwardly* between pairs or greater numbers of boundary inducers. Figures (a) and (c) also illustrates that boundaries can be completed between opposite contrast polarities, and thus that they combine opposite contrast polarities at each position, thereby becoming *insensitive to contrast polarity*. Figures a, b, and d illustrate how surface brightnesses and colors can fill-in *outwardly* in an *unoriented* way, spreading in all directions, until they hit a boundary or are attenuated by their spatial spread. These surface brightnesses and colors can be seen, and thus are *sensitive to contrast polarity*.

These computational properties of boundaries and surfaces (inward-outward, oriented-unoriented, insensitive-sensitive) are manifestly complementary (Figure 3). Cross-stream interactions between boundaries and surfaces at a series of processing stages (Figures 1 and 2) overcome their complementary deficiencies and generate

consistent percepts of objects in the world, a property called *complementary consistency* (Grossberg, 2008).

Adapting 3D LAMINART to process incomplete 3D boundaries in natural scenes. Two major challenges for processing natural scenes are that pictorial and scenic boundaries are often incomplete, and cluttered scenes incorporate many possibilities of false binocular matches (see Figures 4, 12 and 13). In order to deal with these challenges, the main new developments in the current enhanced model are (see Figures 1 and 2):

(1) Interactions occur from V1 binocular boundaries to V1 monocular surfaces. They help with initial depth assignments (see Section 2.3).

(2) Feedback interactions occur between V2 binocular boundaries and V2 monocular surfaces. In particular, *surface contour* feedback signals from V2 surfaces to V2 boundaries (signals S-to-B in Figure 2) had earlier been used to explain data about 3D figure-ground separation, among others (e.g., Grossberg, 1994, 2016; Grossberg and Yazdanbakhsh, 2005). Herein, such signals also help to deal with broken boundaries and to eliminate false binocular matches (see Sections 2.4 and 2.5).

By (1) and (2), both V1 and V2 now exhibit more homologous interactions between their boundary and surface representations.

(3) A disparity filter (DF in Figure 1) is defined in both the boundary and surface processing streams in V2, rather than just in the boundary stream of previous model instantiations. Together, they help to solve the Correspondence Problem and generate correct 3D surface representations by inhibition along lines-of-sight (see Sections 2.4 and 2.5).

These refinements together create a more symmetric set of interactions within and between the various boundary and surface processing stages of the model. They are used to clarify how the brain may complete the broken boundaries that are often generated by natural images. They do not, however, affect the model explanations of psychophysical percepts that were given in Cao and Grossberg (2005, 2012), which did not require the completion of broken boundaries. The model hereby provides a unified approach to providing both a quantitative explanation of perceptual and neurobiological data about 3D boundary and surface perception, as well as a system for 3D processing of natural scenes in computer vision applications.

Besides overcoming the uniqueness and continuity constraints, and hereby explaining psychophysical data such as those that arise when viewing displays of Panum's limiting case and Da Vinci stereopsis, the 3D LAMINART model also shows how cortical interactions between boundary and surface representations overcome the problem of choosing the size of windows that is faced by area-based stereo methods, thereby avoiding unwanted smoothing across edges, and achieves edge consistency, as sought by edge-based stereo methods. Furthermore, the model generates a 3D surface percept of natural images (e.g., Figures 8 and 11). Alternative stereo algorithms instead have aimed at primarily generating a dense disparity map. Also relevant are explanations and simulations of 3D surface percepts in response to both dense and sparse stereograms, including definite depth assignments to the large ambiguous surface regions in the latter whose uniform white color provides no cues to depth, and figure-ground percepts in response to dense stereograms that represent partially occluded objects, despite the fact that the boundaries that are needed to complete the partially occluded object occur across an occluding gap that is much larger than the defining features of the stereogram (Fang and Grossberg, 2009).

2. Model description

It is known that the visual cortex consists of several parallel processing streams (DeYoe and Van Essen, 1988). As noted in Section 1, 3D LAMINART models two of these streams: a boundary stream and a surface stream. Figure 1 describes an anatomically labelled laminar cortical circuit diagram of the model. The boundary stream passes from the lateral geniculate nucleus (LGN) through the V1 interblobs and then to the V2 pale stripes and V4 to select and complete 3D boundary groupings. The surface stream passes from the LGN through the V1 blobs and then to the V2 thin stripes and V4 to fill-in 3D surface representations of depth, lightness, and color. The two streams interact to overcome their complementary computational deficiencies and thereby create consistent 3D boundary and surface percepts (Grossberg, 1994). Figure 2 provides a block diagram of the model that labels the boundary and surface processing stages that correspond to the anatomically labeled stages in Figure 1. In particular, the boundary stream has five component networks: V1 monocular boundaries, V1 binocular boundaries, and V2

binocular boundaries. The surface stream has three component networks: V1 monocular surfaces, V2 monocular surfaces, and V4 binocular surfaces. A mathematical description of model equations and parameters is provided in Section 4.

A heuristic functional description of these processing stages is provided in this section. This description also includes the equation numbers of the corresponding model equations in Section 4 to facilitate comparison of these functional and mathematical descriptions. The mathematical variables are also provided in Figure 1 to facilitate visualization of the flow of information throughout the model architecture.

2.1. V1 monocular boundaries

The left and right retinal images are first processed by LGN cells that use on-center off-surround networks whose cells obey membrane equation, or shunting, dynamics to compensate for variable illumination levels (i.e., "discount the illuminant") and contrast normalize the response to scenic contrast (Grossberg, 1973, 1980). These equations are solved at equilibrium (equations (2)-(6)), as are various other model processes that represent fast dynamics. LGN cells then input into oriented polarity-selective filters that model monocular simple cells in V1 layer 4 (Hubel and Wiesel, 1968). Simple cells with odd and even symmetry are simulated at six different orientational selectivities (equations (7)-(10)). Due to their polarity-selectivity, simple cells are sensitive to either dark-light or light-dark contrast polarity, but not both. They mutually inhibit one another across orientation and position, hereby contrast normalizing their responses (equation (11); Grossberg and Mingolla, 1985a, 1985b; Heeger, 1992). This divisive normalization process helps to enhance weak boundaries. Simple cells at the same position that are sensitive to the same orientation but opposite contrast polarities generate outputs to monocular complex cells in V1 layer 2/3 (equation (15)). Monocular complex cells in layer 2/3 sum these inputs to implement contrast-invariant boundary detection.

2.2. V1 binocular boundaries

V1 binocular boundaries start to get computed by successive processing stages in layers 4, 3B, and 2/3 in the interblobs of V1. Left and right eye monocular simple cells in layer 4 with the same orientational selectivity and contrast polarity (equation (12)) are binocularly fused in layer 3B to give rise to disparity-selective binocular simple cells (Poggio and Fischer, 1977; Poggio et al., 1988). These binocular simple cells are

selective for both binocular disparity and contrast polarity. The latter property is said to satisfy the *same-sign hypothesis* (Howard and Rogers, 1995). The layer 4 cells also activate inhibitory interneurons in layer 3B (equations (13)-(14)) whose inhibition of each other and of the binocular simple cells ensures that binocular simple cells respond only when their left and right eye inputs are approximately equal in magnitude and of the same contrast polarity. This is called the *obligate* property (Poggio, 1991). The same-sign and obligate properties help to avoid false binocular matches—that is, binocular matches that do not correspond to the same object boundary—and thereby to help solve the *correspondence problem* (Howard and Rogers, 1995; Julesz, 1971). They are not, however, sufficient to completely solve the correspondence problem. For this, more interactions are needed in cortical area V2, as summarized below.

Layer 3B binocular simple cells that are sensitive to the same position and disparity but opposite contrast polarities pool their signals at layer 2/3 binocular complex cells (equation (18)). These binocular complex cells also pool inputs from nearby depths and positions to as part of the process whereby a finite number of cell populations can support percepts of depth that change continuously across a scene (equations (16)-(17)). As in the monocular cases in Section 2.1, layer 2/3 binocular complex cells implement contrast-invariant boundary detection, in addition to being disparity selective.

2.3. V1 Monocular Surfaces

V1 monocular surfaces start to get processed in the V1 blobs. The V1 left (right) surface receives lightness signals from left (right) LGN cells. These surface cells also receive modulatory signals from V1 binocular simple cells, which enhance the activities of surface cells at the corresponding positions (equations (19)-(23)). These interactions help to guide initial depth assignments. A surface filling-in process may exist in V1 (Huang and Paradiso, 2008; Fang and Grossberg, 2009), but it is omitted here for simplicity. However, adding such a process will not undermine the current results.

2.4. V2 boundaries

V1 boundaries accomplish the first stages of binocular fusion, but do not carry out spatially long-range boundary completion, which begins in the V2 pale stripe region and is completed in V2 layer 2/3. This layer 2/3 boundary completion process receives its inputs from V2 layer 4, which combines monocular and binocular inputs from V1 layer

2/3 (equation (24)). In particular, pale stripe V2 layer 4 cells receive inputs from binocular complex cells at the corresponding position, as well as from left and right monocular complex cells, all from V1 layer 2/3. Since the monocular cells are not associated with a particular depth plane, their outputs are added to all depth planes in V2 layer 4 along their respective lines-of-sight. The combination of monocular and binocular information helps to complete depthful percepts at all positions in response to many scenes wherein part of the scene can be seen by only one eye due to occlusion by nearer objects. Such percepts are often described under the rubric of da Vinci stereopsis (Nakayama and Shimojo, 1990). Several different examples of da Vinci stereopsis have been simulated (Cao and Grossberg, 2005, 2012; Grossberg and Howe, 2003).

The V2 layer 4 cells also receive feedback signals from left and right V2 monocular surfaces that are formed in the V2 thin stripe region (equations (24) and (25)). These surface-to-boundary feedback signals help to select consistent percepts despite the computationally complementary laws of boundary completion and surface filling-in (Grossberg, 1994). These feedback signals modulate V2 layer 4 cells in the following way: the activity of an active layer 4 cell is enhanced if it receives either a left or right surface-to-boundary excitatory feedback signal, or both. Its activity is suppressed otherwise. These surface-to-boundary feedback signals play an indispensable role in explaining percepts of some stereo displays, and in figure-ground separation (Cao and Grossberg, 2005; Fang and Grossberg, 2009; Grossberg and Yazdakbakhsh, 2005; Kelly and Grossberg, 2000). They are called *surface contour* signals because they are generated by contrast-sensitive on-center off-surround networks across space and within disparity that respond only at the bounding contours of successfully filled-in surfaces within the V2 thin stripes; that is, surfaces that fill-in within closed boundaries that can contain the filling-in process.

As noted in Section 2.2, V1 layer 3B binocular cells attempt to match every edge in one retinal image with every other like-oriented edge in the other retinal image within its disparity range that has the same contrast polarity and approximately the same magnitude of contrast. These same-sign and obligate properties help to reduce the number of false matches that are computed in V1. However, not all false matches are eliminated in V1. Figure 4 shows nine possible matches if each eye sees three bars. Only

the three matches in the fixation plane are correct (three solid black ellipses), and the others are false. Such false matches are suppressed in V2 via a *disparity filter* (Cao and Grossberg, 2005; Grossberg and McLoughlin, 1997). The disparity filter works as follows: The solid lines in Figure 4 depict the monocular lines of sight of the contrastive inputs from the left and right eyes. The disparity filter encourages unique matching by generating line-of-sight inhibition from each neuron to all other neurons that share either of its monocular inputs.

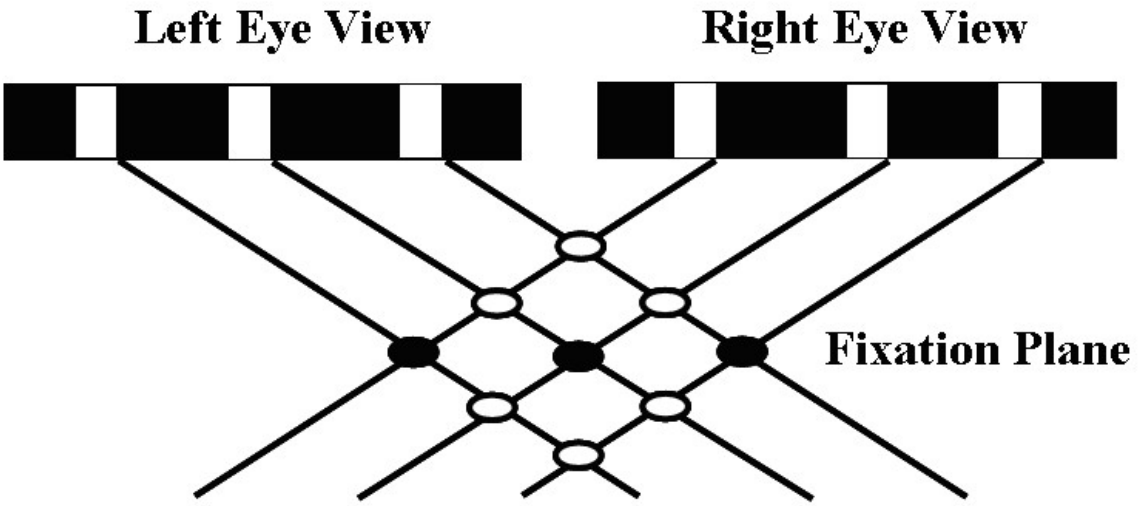

Figure 4. The V2 disparity filter. In response to this image, the V1 boundary network creates nine matches. Only the three matches (filled dots) in the fixation plane are true, others (open dots) are false. These false matches are suppressed by the disparity filter in V2, wherein each neuron is inhibited by every other neuron that shares a monocular line-of-sight represented by the solid lines.

The model proposes that the disparity filter occurs in V2 layer 2/3 (DF in Figure 2), where it is part of the inhibitory interactions that control boundary completion, also called perceptual grouping, by long-range horizontal connections in V2 layer 2/3. The model hereby parsimoniously combines suppression of false matches, and thus a solution of the correspondence problem, with the process of long-range perceptual grouping (Cao and Grossberg, 2005, 2012).

Perceptual grouping is achieved by binocular complex cells in V2 layer 2/3 whose collinear, coaxial receptive fields excite each other via long-range horizontal axons.

These excitatory interactions are balanced by short-range disynaptic inhibition via inhibitory interneurons (Figure 1). This balance of excitation and inhibition helps to control grouping by implementing the *bipole property* (Grossberg, 1999; Grossberg, Mingolla and Ross, 1997; Grossberg and Raizada, 2000; Grossberg and Williamson, 2001). The bipole property ensures that grouping can occur inwardly in response to pairs, or greater numbers, of approximately collinear cells whose orientational tuning is sufficiently similar, but not outwardly in response to individual cell activations.

This combination of excitation and inhibition in V2 is homologous to the one in V1 that realizes the obligate property (Figure 1). It remains to be determined whether both processes have a similar phylogenetic ancestor. This boundary grouping process, together with contrast-invariant boundary detection and the suppression of false binocular matches, allows consistent and connected object boundaries to be formed even in response to noisy textured backgrounds (equations (26)-(37)).

2.5. V2 monocular surfaces

The network that forms V2 monocular surfaces is located in the V2 thin stripes, which receive binocular boundary signals from the V2 layer 2/3 pale stripes and monocular lightness and color signals from the V1 blobs (Figure 1). The boundaries define the regions within which filling-in of lightness and color can occur (Figure 5). Multiple boundary representations exist that are formed in response to image properties are different depth ranges. Each of them attempts to capture monocular surface signals that abut and are collinear with them. This process of surface capture enables the surface signals from V1 to be selectively filled-in within depth-selective Filling-In DOmains, or FIDOs (Grossberg, 1994). Previous simularions have illustrated how 2D surfaces (Grossberg and Hong, 2006; Grossberg and Mingolla, 1985b; Grossberg and Todorovic, 1988) and 3D surfaces (Fang and Grossberg, 2009; Grossberg, 1994, 1997; Grossberg and McLoughlin, 1997; Grossberg and Swaminathan, 2004; Grossberg and Yazdanbakhsh, 2005; Hong and Grossberg, 2004; Kelly and Grossberg, 2000) may be generated by such a boundary-gated filling-in process, and have thereby explained and predicted many psychophysical and neurobiological data about how the brain sees 2D and 3D surfaces. Psychophysical data (e.g., Paradiso and Nakayama, 1991; Pessoa and Neumann, 1998; Pessoa, Thompson and Noe, 1998) and neurophysiological data (e.g.,

Lamme, Rodriguez-Rodriguez and Spekreijse, 1999; Rossi, Rittenhouse and Paradiso, 1996) have supported the existence and predicted properties of such a filling-in process.

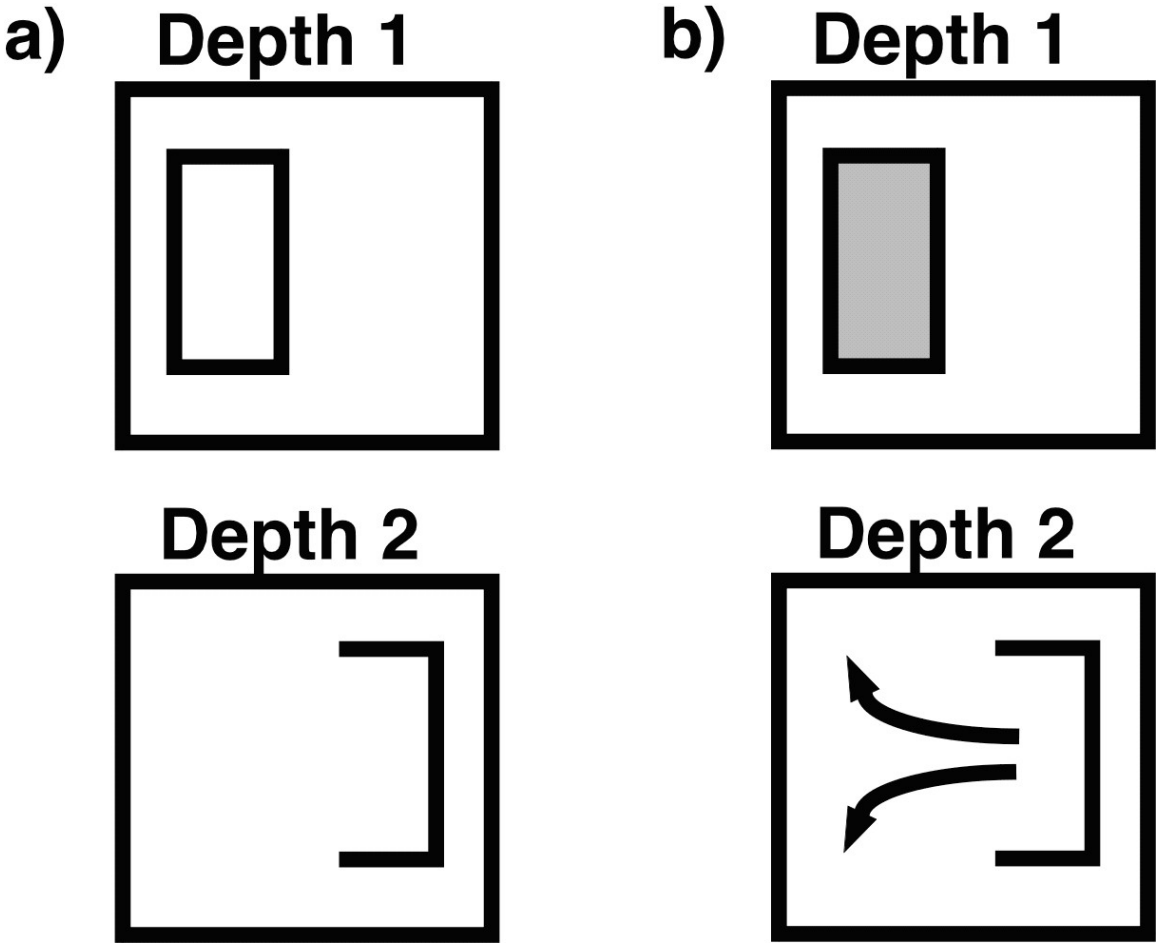

Figure 5. (a) Open and connected boundaries; (b) Filling-in of surface lightness. The connected boundaries in Depth 1 can contain filled-in lightness signals, but open boundaries in Depth 2 cannot.

The filling in process is here, as is often the case, modeled by a boundary-gated diffusion equation (Grossberg and Todorovic, 1988; however, also see Grossberg and Hong (2006) for a much faster filling-in process). The 3D LAMINART model clarifies how only perceptual regions that are surrounded by a closed and connected boundary can become part of a conscious 3D surface percept (Figure 5) due to the fact that surface contour feedback signals are not generated around regions where filling-in spills out of a large open boundary, as illustrated in Figure 5 by the incomplete Depth 2 boundary

(Figure 5a) and uncontained filling-in (Figure 5b). This fact highlights the importance of correctly completing perceptual boundaries.

One reason for broken boundaries in response to natural images is that the object contours at corresponding positions along an object seen in depth may have opposite contrast polarities. This is illustrated in Figure 10 below as part of the analysis of how the model processes such images. By the same-sign hypothesis, such contour positions cannot be binocularly fused. Bipole binocular boundary completion is thus not sufficient to complete binocular boundaries across all such positions. It has already been noted, however, that the model adds V1 monocular boundaries to the V2 layer 2/3 binocular boundaries that gate the V2 monocular surface filling-in process (Figures 2 and 11). These monocular boundaries can provide boundary inputs even at positions where the same-sign hypothesis is violated.

However, this addition adds monocular boundaries to *all* depths within the surface stream. These boundaries can potentially capture monocular surface information from V1 at multiple depths, thereby creating the same kind of correspondence problem in the surface stream that the disparity filter helps to solve in the boundary stream. A disparity filter in V2 is thus added to the surface stream as well to cope with this surface-based correspondence problem.

In summary, as noted in Section 2.4, within the boundary stream, a V2 disparity filter is needed to eliminate spurious boundaries at the incorrect depths. In the current model, a V2 *surface disparity filter* is introduced to eliminate filled-in surface signals at incorrect depths. In other words, each V2 surface cell inhibits all other surface cells that shares one of its monocular lines-of-sight (equations (39), (44) and (45)). This V2 surface disparity filter, together with the filling-in process, creates the 3D monocular surface representations that are captured at different depths by binocular boundaries in the V2 thin stripes.

Thus, the current model embodies a more symmetric organization which includes interactions between boundaries and surfaces in both V1 and V2, and disparity filters in both the boundary and surface streams.

Figure 6. University of Tsukuba scene (courtesy of the University of Tsukuba). Left: Left input image; Right: Right input image.

Successfully filled-in monocular surfaces in V2 then send their surface contour signals—that is, contour-sensitive surface-to-boundary feedback signals—into V2 layer 4 (Figure 2, equations (48)-(50)). These surface-to-boundary signals modulate the activities of V2 boundary cells so that the boundaries that surround the successfully filled-in surfaces are enhanced and other boundaries are suppressed. See equations (38)-(50) for mathematical details.

2.6. V4 surfaces

Area V4 receives boundary signals from the V2 layer 2/3 pale stripes and lightness signals from the LGN, which are modulated by signals from the V2 thin stripes (Appendix equations (52)-(54)). In particular, successfully filled-in features in the V2 thin stripes are subtracted from farther depths (surface pruning) in V4 to ensure that opaque objects do not look transparent (Fang and Grossberg, 2009; Grossberg, 1997, Figure 21). The V4 binocular surface representation then fills-in the final visible depth-selective surface representation (Appendix equations (51)-(56)).

3. Results

The model's ability to process natural images is illustrated using three benchmark images that illustrate different combinations of computational problems.

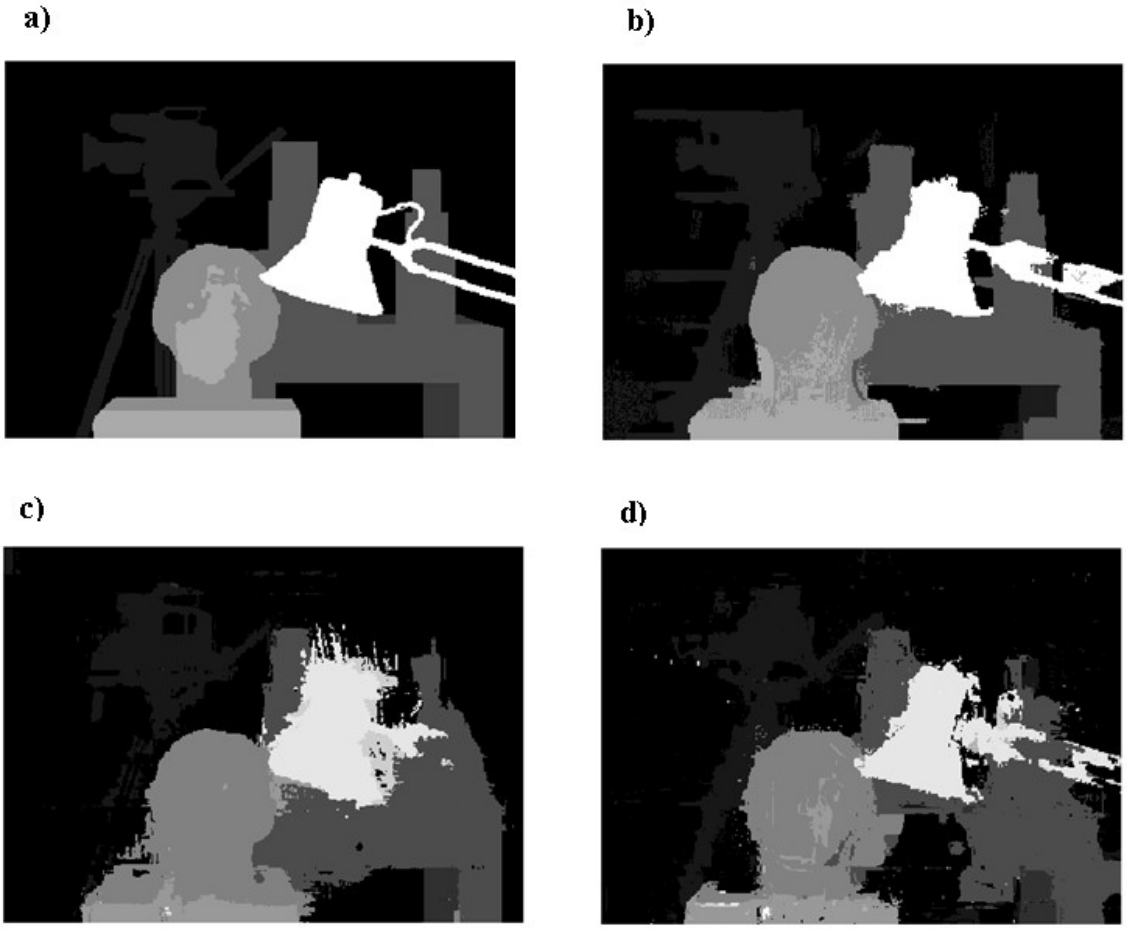

Figure 7. (a) Ground truth disparity map for University of Tsukuba scene (courtesy of the University of Tsukuba); (b) Disparity map found using our 3D LAMINART model; (c) Disparity map found excluding the new connection from V1 monocular boundary to V2 monocular surface (c.f. Figure 2); (d) Disparity map found excluding the new connection from V1 binocular boundary to V1 monocular surface (c.f. Figure 2).

3.1. University of Tsukuba Scene with Ground Truth

The University of Tsukuba's Multiview Image Database is a famous benchmark that provides real stereo image pairs of a complex scene along with ground truth data. Figure 6 shows the stereo image pair, with the ground truth data shown in Figure 7a. Figure 8 summarizes a computer simulation of the model's 3D surface lightness representation.

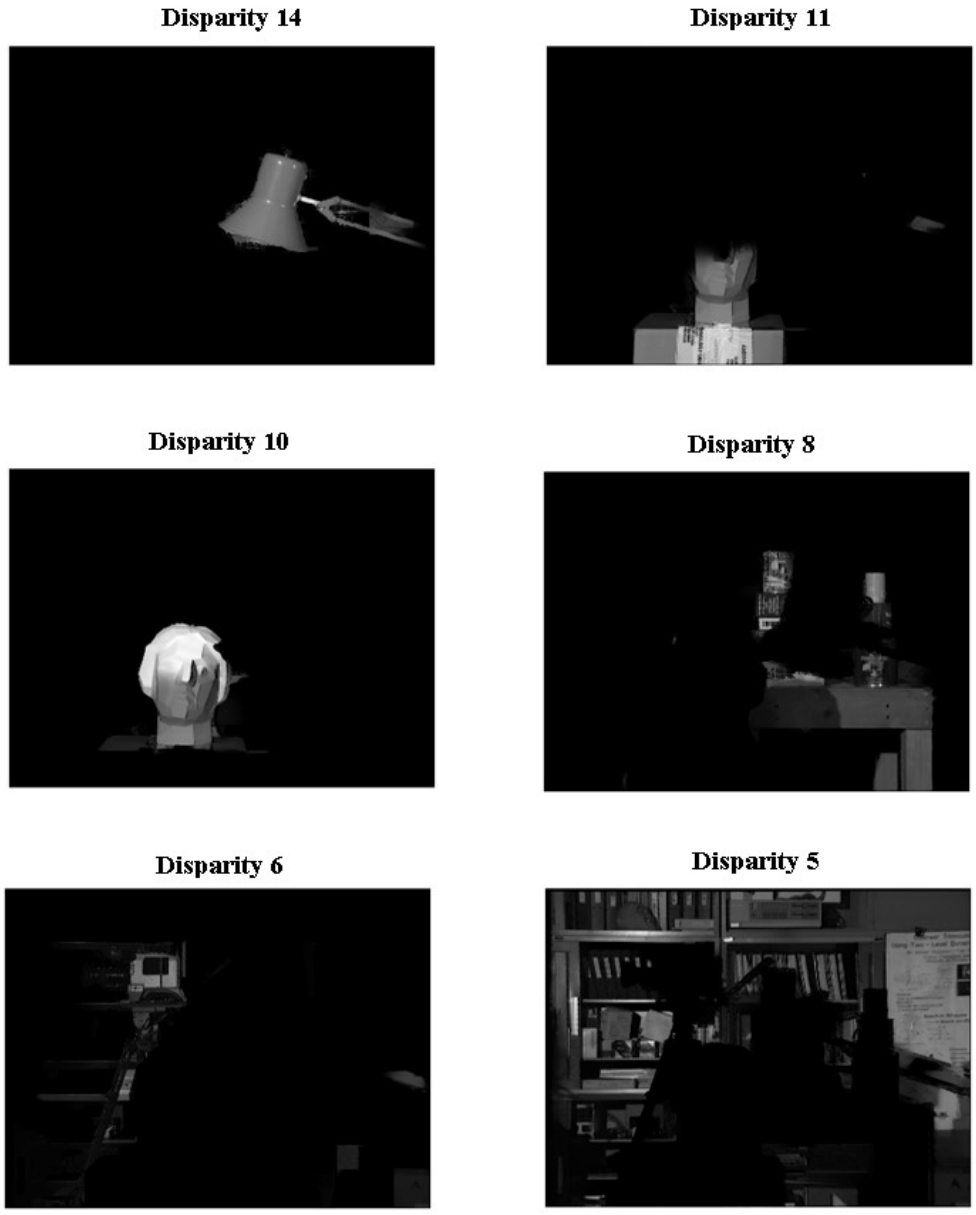

Figure 8. 3D surface representation for University of Tsukuba scene found using the 3D LAMINART model.

Each major part of the scene (lamp, statue of head, bottle, table, camera, bookcase and other background featural details) are correctly separated in depth and filled-in. In order to compare with the ground truth data in Figure 7a, a disparity map of this surface representation is provided in Figure 7b. For each point in the reference image (the left image of Figure 6), its disparity value is computed as the depth (disparity) that has the

maximal lightness signal strength along the left line-of-sight. The accuracy is 96.2%, in which an error is counted when the disparity difference between the resulting disparity map and the ground truth data is greater than one pixel.

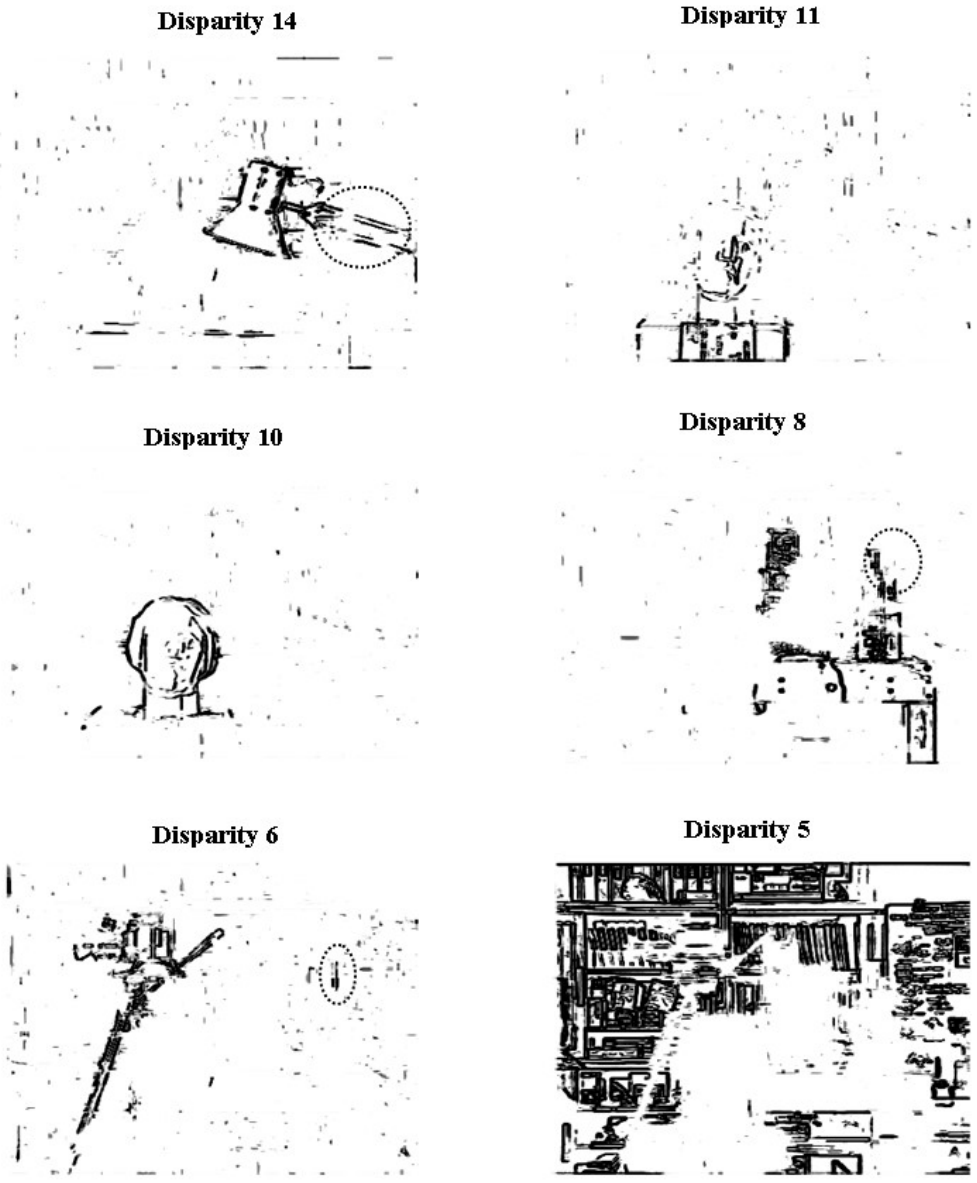

Figure 9. V2 layer 2/3 binocular boundaries for University of Tsukuba scene before the new connections are added. The dotted circles in Disparity 14 and Disparity 8 emphasize some incomplete boundary regions, and the dotted circle in Disparity 6 shows a false match. The incomplete binocular boundaries in circled regions are mainly due to unmatched contrast polarities of left and right monocular boundaries which are caused by cluttered background (cf. Figure 12). In particular, the false match in the circled region in Disparity 6 is created by a same-polarity match between the right edge of the top bottle on the desk in the

left image and the left edge of the poster in the background in the right image. They have the same contrast polarity (cf. Figure 12).

The accuracy is comparable to state-of-the-art stereo algorithms. For example, Zitnick and Kanade (2000) report an accuracy of 98.6%, but only after excluding occluded pixels that are seen by only one eye. In contrast, the current algorithm counts all pixels. The occluded pixels amount to about 2.2% in the current scene. Figure 7c shows the estimated disparity map without the new connection from the V1 monocular boundary to the V2 monocular surface (Figure 2). Here, the accuracy is 91%. Figure 7d shows the estimated disparity map without the new connection from the V1 binocular boundary to the V1 monocular surface (Figure 2). The resulting accuracy is then 90%.

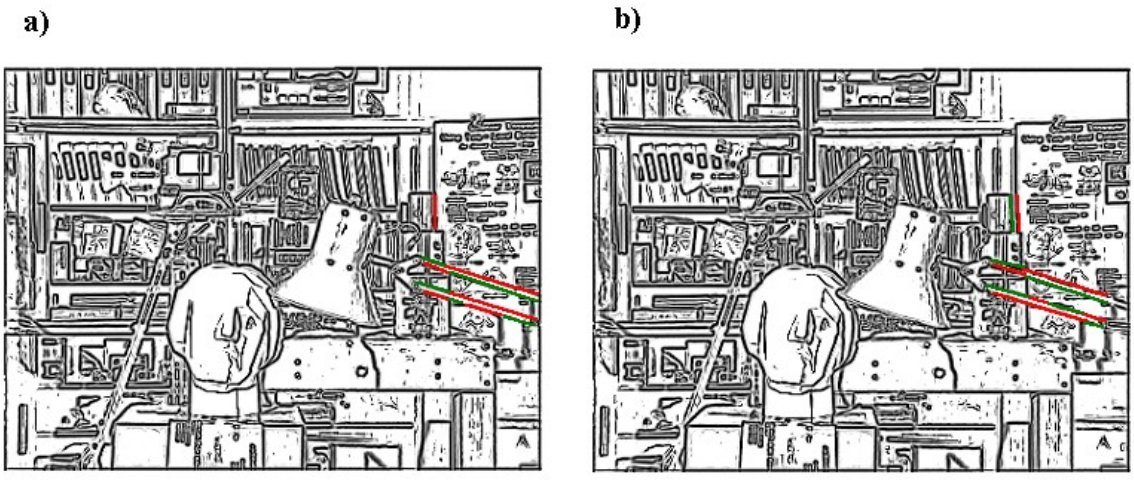

Figure 10. V1 monocular boundaries for University of Tsukuba scene (a) Left eye image; (b) Right eye image. Some investigated boundaries are colored to show their contrast polarities. Red denotes a dark-light polarity and green a light-dark polarity. Only boundaries with the same contrast polarity can be matched in V1 binocular cells according to the same-sign rule.

The University of Tsukuba image illustrates key additional challenges for processing natural scenes; namely: (1) 3D boundaries are often incomplete, either due to noise in the acquisition of the images, or due to unavailability of like-polarity matches at some boundary positions; and (2) cluttered scenes incorporate many possibilities for false binocular matches. For example, V2 layer 2/3 binocular boundaries for the arms of the

lamp (see the circled region of Disparity 14 in Figure 9) and the right edge of the top bottle on the desk (see the circled region of Disparity 8 in Figure 9) are incomplete. This can cause lightness signals to flow out of their respective image regions during the filling-in process. The incomplete binocular boundaries in these circled regions are mainly due to unmatched contrast polarities of left and right monocular boundaries that are caused by the cluttered background (cf. Figure 10).

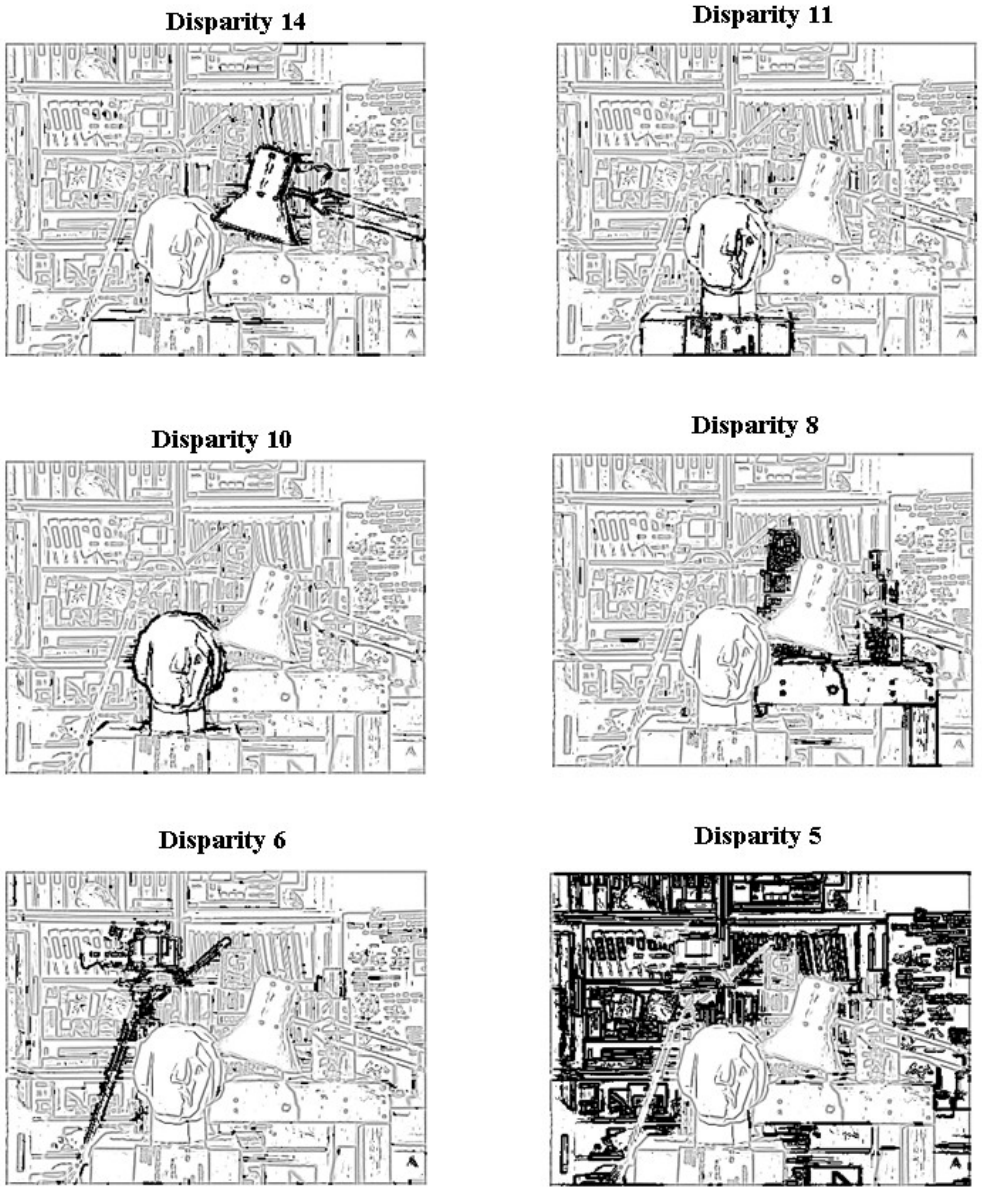

Figure 11. The completed boundaries to the V2 left monocular surface filling-in domain after the new connections are added.

In particular, a false match in the circled region in Disparity 6 is created by the same-polarity match between the right edge of the top bottle on the desk in the left image and the left edge of the poster in the background in the right image. They have the same contrast polarity (cf. Figure 10).

Additional interactions (Figure 2, dashed arrows) were proposed herein within the 3D LAMINART model to overcome these challenges. These interactions led to a more symmetric anatomical organization for the model as a whole, while also elaborating how boundary and surface representations interact to overcome each other's complementary deficiencies. Figure 11 shows the boundaries that are completed by these additional interactions, leading to simulation results, as summarized above, that are comparable to state-of-the-art stereo algorithms, with the additional advantage of the current model that it also generates a 3D surface representation of the consciously seen percept.

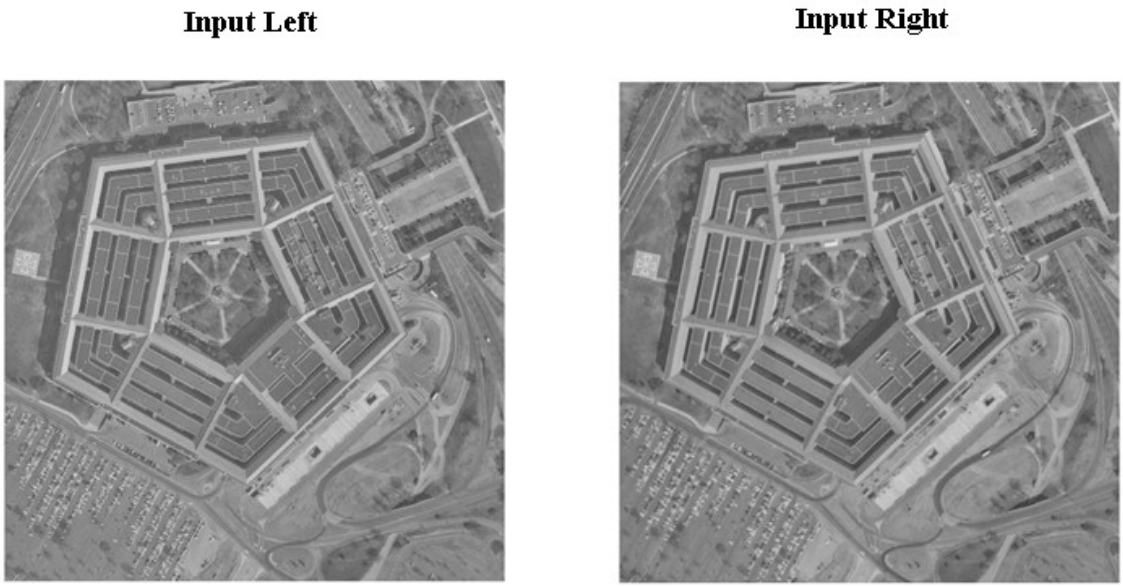

Figure 12. Pentagon scene. Left: Left input image; Right: Right input image.

3.2. Pentagon Images

Figure 12 shows a pair of Pentagon images. No ground truth disparity map for the Pentagon images is available in its public database. Figure 13 shows the model simulation. It can be seen that the Pentagon is actually tilted in this photo, with the highest region being in the lower corner and the lowest being in the upper-left corner of

the Pentagon images. The resulting disparity map from the simulation is shown in Figure 14.

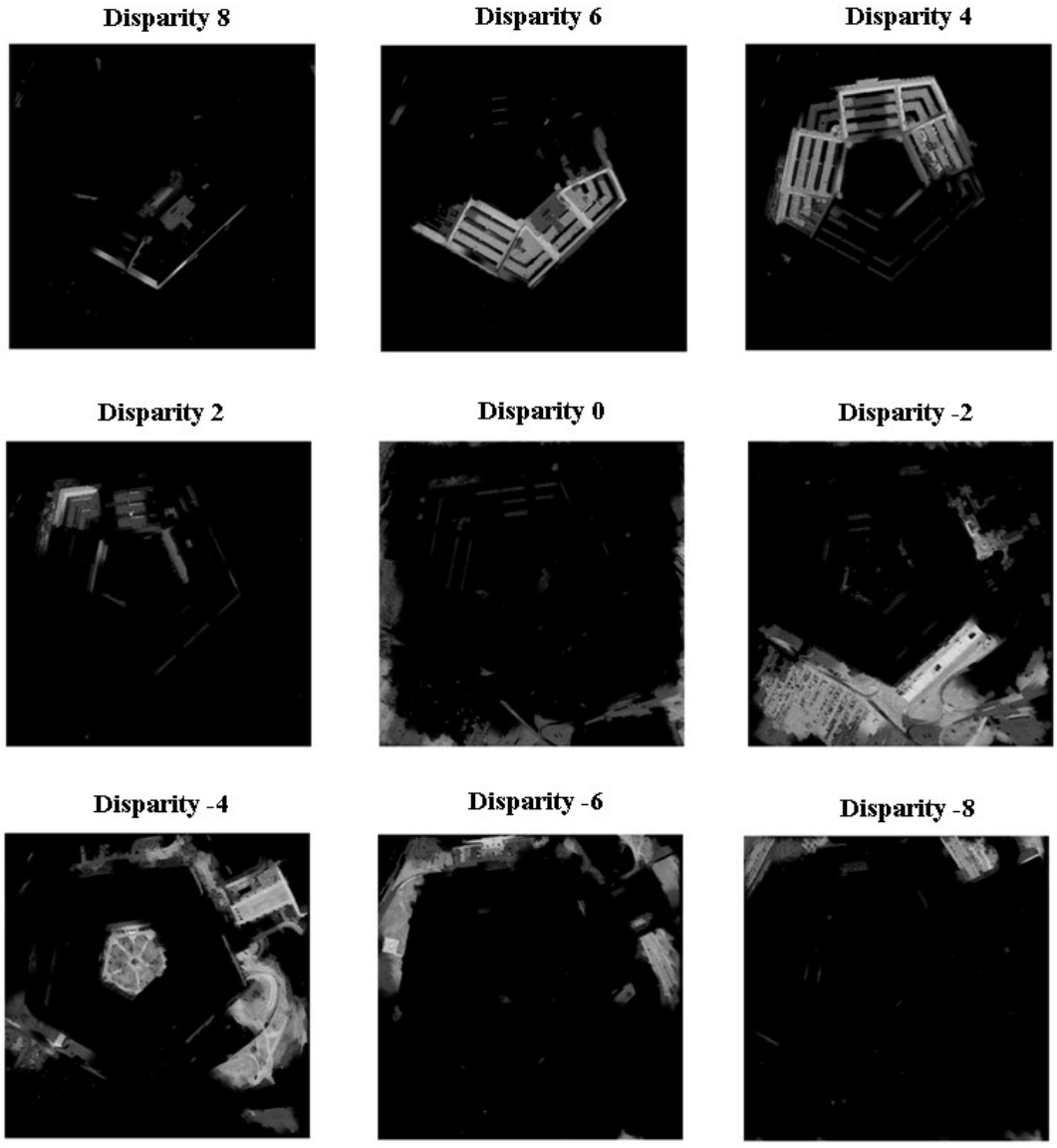

Figure 13. 3D surface representation for Pentagon scene found using the 3D LAMINART model.

It should be noted that the 3D LAMINART boundaries that are modeled in this article are not designed to represent objects that are tilted in depth. Grossberg and Swaminathan (2004) have modeled how tilted boundaries and surfaces can be completed and filled-in, respectively, in depth. Their results include a simulation of the classical

bistable 3D Necker cube percept, including how two 3D boundary and surface representations can form in response to the 2D Necker cube image and switch spontaneously through time from one to the other. These augmented laws for tilted boundary and surface representations can be added to the current model without disrupting how it works in response to 3D scenes that are not tilted in depth.

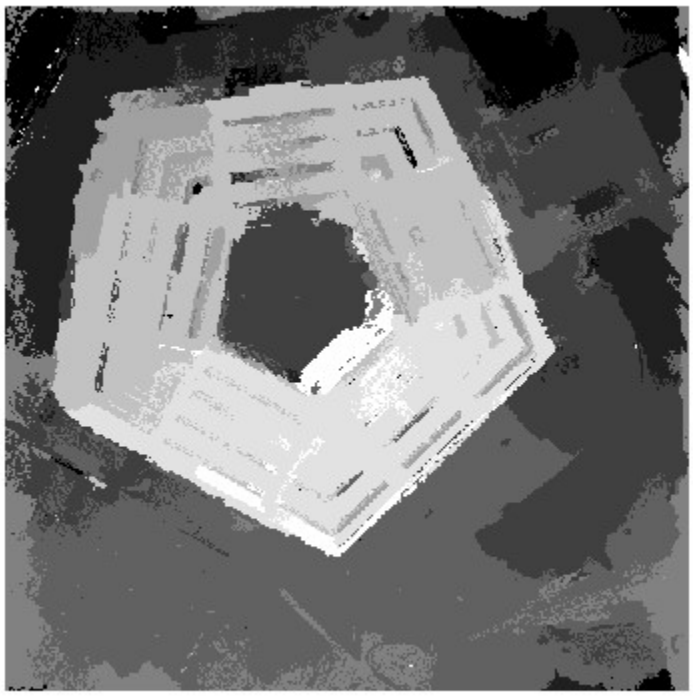

Figure 14. Disparity map found using our 3D LAMINART model for Pentagon scene.

3.3. Barn Images

Figure 15 shows a pair of Barn images (Scharstein and Szeliski, 2002; http://vision.middlebury.edu/stereo/data/scenes2001/data/barn2/). The ground truth map is shown in Figure 16. Figure 17 shows the different disparity-sensitive 3D V2 binocular boundaries that are computed by our model, and Figure 18 shows the 3D surface representations that are captured in depth by these boundaries. The estimated disparity map is shown in Figure 19. Its accuracy is 92%. These images include another example where tilted boundaries occur. Including the capacity for representing such boundaries

should increase the accuracy of the resulting 3D boundary and surface representations, and the disparity map that is derived from them.

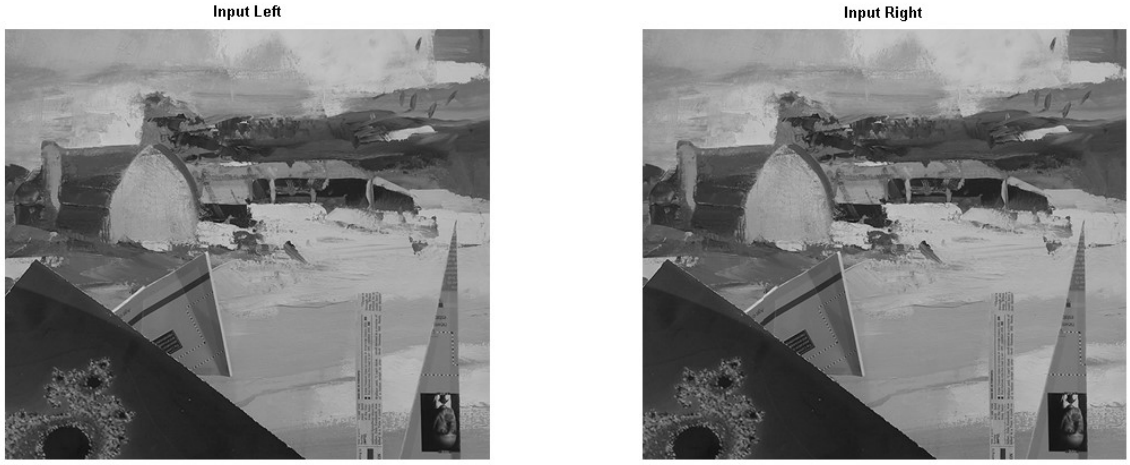

Figure 15. Barn scene. Left: Left input image; Right: Right input image.

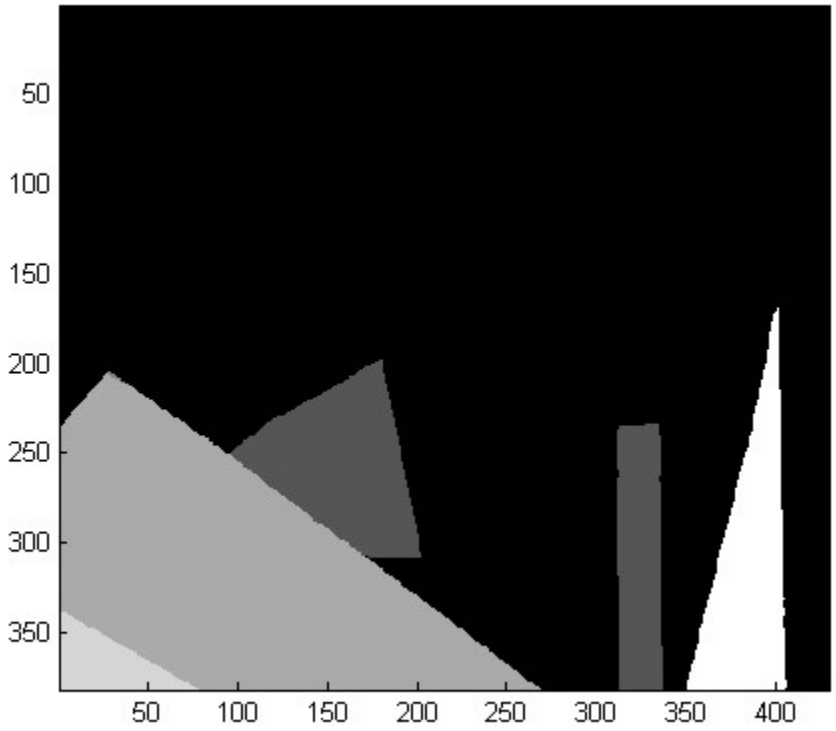

Figure 16. Ground truth disparity map for Barn scene.

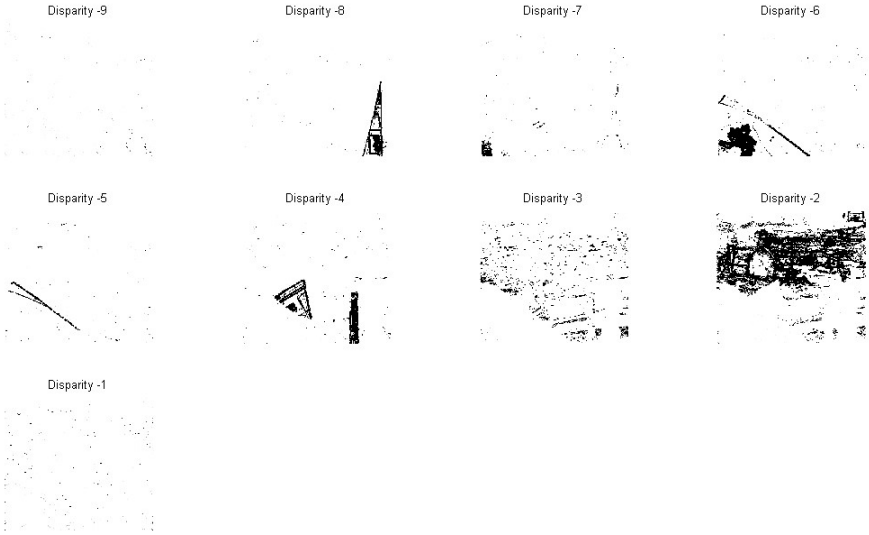

Figure 17. V2 layer 2/3 binocular boundaries for Barn scene.

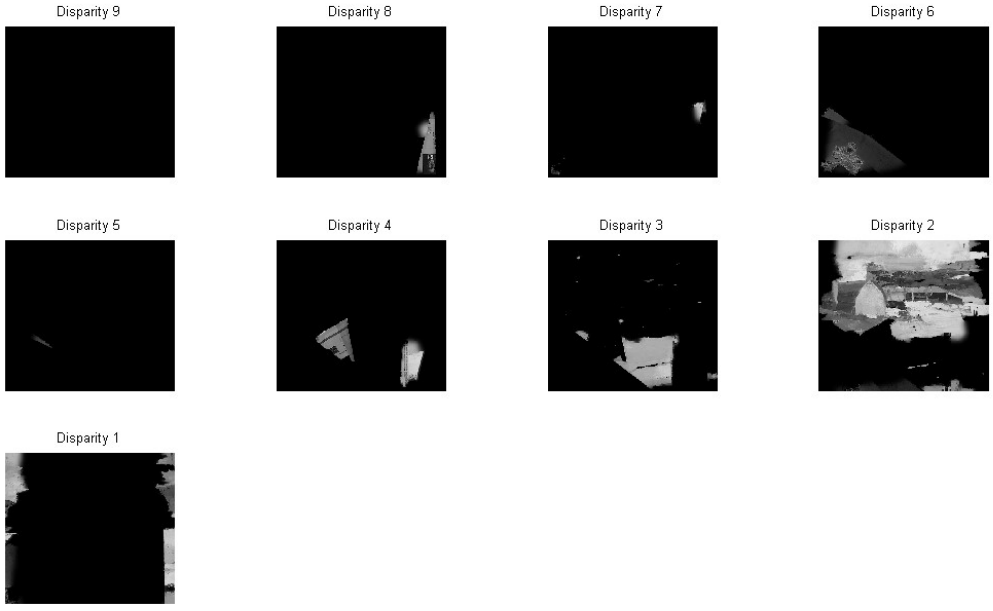

Figure 18. 3D surface representation for University for Barn scene.

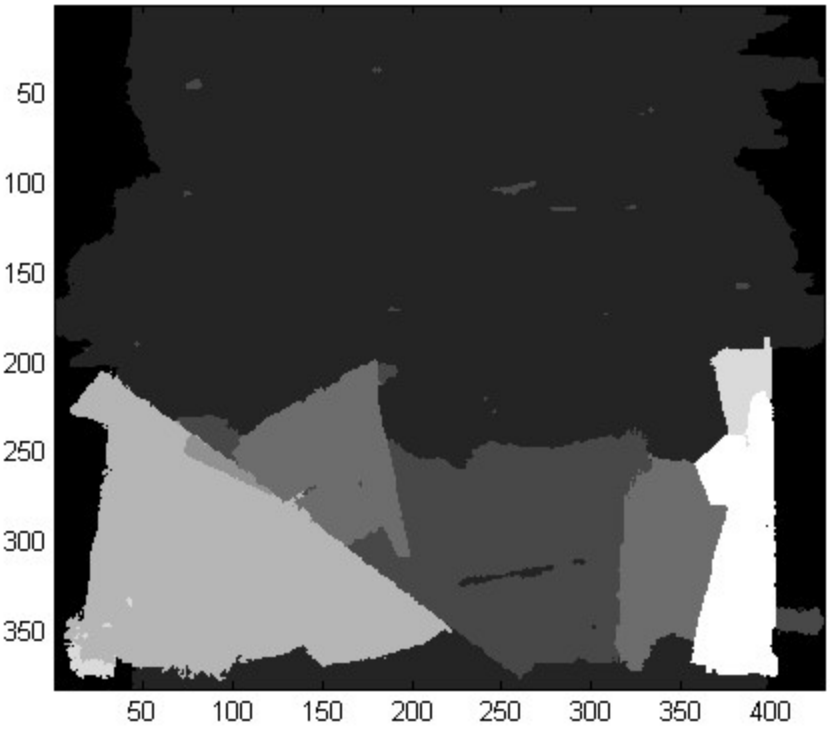

Figure 19. Estimated disparity map for Barn scene using our 3D LAMINART model.

4. Model Equations

This section describes the model equations. The equations (11), (19)-(23), and (38)-(47) represent model refinements to deal with additional computational challenges that are posed by natural images. In particular, equations (19)-(23) represent the new connections from V1 binocular boundaries to V1 monocular surfaces, which help with initial depth assignments. Equations (38)-(47) represent the new connections from V1 monocular boundaries to V2 monocular surfaces and to the new V2 surface disparity filter, which help to complete incomplete binocular boundaries and to eliminate spurious monocular boundaries from 3D monocular surface representations. Figure 7c and 7d show how the simulated results are weakened without these new additions.

Each neuron is typically modeled as a single voltage compartment in which the membrane potential v is given by

$$\frac{dv}{dt} = -Av + (B - v)g_{excit} - (C + v)g_{inhib}, \qquad (1)$$

where A is a constant decay rate, B is the maximum membrane potential, C is the minimum membrane potential, g_{excit} is the total excitatory input, and g_{inhib} is the total inhibitory input.

The new refinements to Cao and Grossberg (2005) are almost all within surface system, which aim at dealing with the broken boundary problem occurred in natural images. As a result, it does not affect the model explanation to psychophysical displays explained in Cao and Grossberg (2005), which has no a broken boundary problem.

LGN. The LGN cells obey membrane equations that receive input from the retina and are assumed to have circularly symmetric on-center, off-surround receptive fields. When these fields are approximately balanced, the network discounts the illuminant and contrast-normalizes its cell responses (Grossberg and Todorović, 1988). The LGN cell membrane potentials, $x_{ij}^{L/R}$, obey the following differential equation.

For a LGN on cell,

$$x_{ij}^{L/R,+} = 5\left[\frac{E + C_{ij}^{L/R} - S_{ij}^{L/R}}{1 + C_{ij}^{L/R} + S_{ij}^{L/R}} - 0.12\right]^{+}, \tag{2}$$

and for a LGN off cell,

$$x_{ij}^{L/R,-} = 5\left[\frac{\overline{E} + S_{ij}^{L/R} - C_{ij}^{L/R})}{1 + C_{ij}^{L/R} + S_{ij}^{L/R}} - 0.2\right]^{+}, \tag{3}$$

where L/R designates that the cell belongs to the left or right monocular pathway, indices i and j denote the position of the input on the retina, on baseline level activity E=0, off baseline level activity $\overline{E} = 1$, total center input

$$C_{ij}^{L/R} = \sum_{p,q} I_{ij}^{L/R} G_{pqij}^{c}, \tag{4}$$

and total surround input

$$S_{ij}^{L/R} = \sum_{p,q} I_{ij}^{L/R} G_{pqij}^{s}, \tag{5}$$

with $I_{ij}^{L/R}$ is the luminance of the left or right retinal image and G_{pqij} is a Gaussian kernel

$$G^{v}_{pqij} = \frac{1}{2\pi{\sigma_v}^2}\exp\left(-\frac{(p-i)^2+(q-j)^2}{2{\sigma_v}^2}\right), \tag{6}$$

where $\sigma_c = 0.3$ and $\sigma_s = 2$, with the kernel size of center 2 and of surround 6.

V1 Layer 4 simple cells. All cells in V1 layer 4 are modeled as monocular simple cells that are sensitive to either dark-light or light-dark contrast polarity, but not both, depending on their receptive field structure. At steady-state, the membrane potentials, $\widetilde{s}^{L/R,odd/even,+/-}_{ijk}$, of odd and even simple cells that respond to dark-light (+) and light-dark (-) contrast polarity are given by:

$$\widetilde{s}^{L/R,odd/even,+}_{ijk} = \left[\sum_{p,q} K^{odd/even}_{pqk}\left(x^{L/R,+}_{i+p,j+q} - x^{L/R,-}_{i+p,j+q}\right)\right]^{+}, \tag{7}$$

$$\widetilde{s}^{L/R,odd/even,-}_{ijk} = \left[\sum_{p,q} K^{odd/even}_{pqk}\left(x^{L/R,-}_{i+p,j+q} - x^{L/R,+}_{i+p,j+q}\right)\right]^{+}, \tag{8}$$

where index k denotes orientation. Six orientations were used in these simulations, the threshold linear function $[x]^{+} = \max(x,0)$, and K_{pqk} is a Gabor function representing the simple cell receptive field kernel. For a horizontal orientation,

$$\mathrm{K}^{odd}_{pqk} = \frac{1}{2\pi\sigma_p\sigma_q}\sin\frac{2\pi(p-0.5)}{T}\exp\left[-\frac{1}{2}\left(\frac{(p-0.5)^2}{\sigma_p^2}+\frac{(q-0.5)^2}{\sigma_q^2}\right)\right], \tag{9}$$

$$\mathrm{K}^{even}_{pqk} = \frac{1}{2\pi\sigma_p\sigma_q}\cos\frac{2\pi p}{T}\exp\left[-\frac{1}{2}\left(\frac{p^2}{\sigma_p^2}+\frac{q^2}{\sigma_q^2}\right)\right], \tag{10}$$

where $\sigma_p = 1.27$, $\sigma_q = 2$, $T = \pi$ for an odd cell in (9). The parameters for an even cell in (10) will be defined later when they are used in (22). Kernels for other orientations are obtained by appropriate rotation.

Divisive Normalization. A Divisive normalization is applied to enhance weak boundaries:

$$s^{L/R,+/-}_{ijk} = \frac{20(\bar{s}^{L/R,+/-}_{ijk})^2}{1+\sum_{p,q,r}[(\bar{s}^{L/R,+}_{pqr})^2+(\bar{s}^{L/R,-}_{pqr})^2]G_{pqij}}, \tag{11}$$

where $\bar{s}^{L/R,+/-}_{ijk} = [\,\widetilde{s}^{L/R,odd,+/-}_{ijk} - 0.2]^{+}$ and G_{pqij} is 6x6 kernel with all value 1.

V1 Layer 3B binocular simple cells. The layer 3B binocular simple cells receive excitatory input from layer 4 and inhibitory input from the layer 3B inhibitory interneurons that correspond to the same position and disparity. The membrane potentials, $b_{ijkd}^{B,+/-}$, of layer 3B simple cells obey the equations:

$$\begin{aligned}\frac{d}{dt}b_{ijkd}^{B,+/-} = &-\gamma_1 b_{ijkd}^{B,+/-} + (1-b_{ijkd}^{B,+/-})\left(\left[s_{ijk}^{L,+/-}-\theta\right]^+ + \left[s_{(i-s)jk}^{R,+/-}-\theta\right]^+\right) \\ &-\alpha\left(\left[q_{ijkd}^{L,+/-}\right]^+ + \left[q_{ijkd}^{L,-/+}\right]^+ + \left[q_{ijkd}^{R,+/-}\right]^+ + \left[q_{ijkd}^{R,-/+}\right]^+\right),\end{aligned} \tag{12}$$

$$\frac{d}{dt}q_{ijkd}^{L,+/-} = -\gamma_2 q_{ijkd}^{L,+/-} + \left[s_{ijk}^{L,+/-}-\theta\right]^+ - \beta\left(\left[q_{ijkd}^{R,+/-}\right]^+ + \left[q_{ijkd}^{R,-/+}\right]^+ + \left[q_{ijkd}^{L,-/+}\right]^+\right), \tag{13}$$

$$\frac{d}{dt}q_{ijkd}^{R,+/-} = -\gamma_2 q_{ijkd}^{R,+/-} + \left[s_{(i-s)jk}^{R,+/-}-\theta\right]^+ - \beta\left(\left[q_{ijkd}^{L,+/-}\right]^+ + \left[q_{ijkd}^{L,-/+}\right]^+ + \left[q_{ijkd}^{R,-/+}\right]^+\right), \tag{14}$$

where γ_1, α, γ_2, β and θ are constants (0.01, 1.01, 1, 0.9, 0) representing the rate of decay of the membrane potential (γ_1, γ_2), the strength of the inhibition (α, β) and the signal threshold (θ), $q_{ijkd}^{L/R,+/-}$ are the membrane potentials of inhibitory interneurons in layer 3B, d is the disparity to which the model neuron is tuned and index *s* is the positional shift between left and right eye inputs that depends on the disparity (shifting one pixel for each increase in disparity). Since the ground truth data is based on the left image, in order to make proper comparison with it the left image is not shifted as is usually done in simulations of biological data (e.g., Cao and Grossberg, 2005). Term ($b_{ijkd}^{B,+/-}$ and $s_{ijk}^{L/R,+/-}$) can be either ($b_{ijkd}^{B,odd,+/-}$ and $s_{ijk}^{L/R,odd,+/-}$) or ($b_{ijkd}^{B,even,+/-}$ and $s_{ijk}^{L/R,even,+/-}$) to denote odd and even cells respectively.

V1 Layer 2/3 monocular and binocular complex cells. V1 layer 2/3 consists of both monocular and binocular complex cells, which pool the cell membrane potentials of monocular/binocular layer 3B simple cells of like orientation and both contrast polarities at each position.

For a monocular cell, its activity obeys

$$c_{ijk}^{L/R} = \sum_{p,q}\left|s_{pqk}^{L/R,odd,+} - s_{pqk}^{L/R,odd,-}\right|. \tag{15}$$

For a binocular cell, its activity obeys

$$c^{B}_{ijkd} = \sum_{p,q} W_{pqijk} (\bar{b}^{B}_{pqkd} + 0.2\bar{b}^{B}_{pqk(d-1)} + 0.2\bar{b}^{B}_{pqk(d+1)}), \tag{16}$$

where W_{pqijk} is the spatial pooling Gaussian kernel

$$W_{pqijk} = \frac{1}{2\pi\sigma_p\sigma_q} \exp\left[-\frac{1}{2}\left(\frac{(p-i)^2}{\sigma_p^2} + \frac{(q-j)^2}{\sigma_q^2}\right)\right], \tag{17}$$

with $\sigma_p = 1$, $\sigma_q = 1$, size of kernel 3, and

$$\bar{b}^{B}_{pqkd} = \left| b^{B,odd,+}_{pqkd} - b^{B,odd,-}_{pqkd} \right|. \tag{18}$$

V1 Surfaces. The activity of V1 left/right surface cells $y^{L/R}_{ijd}$ are modulated by binocular cells:

$$y^{L}_{ijd} = X^{L}_{i,j}\left(0.2 + b^{B}_{ijd}\right), \tag{19}$$

$$y^{R}_{ijd} = X^{R}_{i-s,j}\left(0.2 + b^{B}_{ijd}\right), \tag{20}$$

where $X^{L/R}_{ij}$ is large scale LGN cell output, which can be approximated (Grossberg and Hong, 2006) as

$$X^{L/R}_{ij} = I^{L/R}_{ij}, \tag{21}$$

with $I^{L/R}_{ij}$ is the luminance of the left or right retinal image, and

$$b^{B}_{ijd} = \sum_{m}\sum_{k} (b^{B,even,+}_{ijkd}(m) + b^{B,even,-}_{ijkd}(m)), \tag{22}$$

which sums over all orientations (k) and scales (m). Equation (22) was simulated with three scales m and the parameters for even cells in (10) defined as $\sigma_p(m) = m$, $\sigma_q(m) = 3m$, $T(m) = 3m$ with $m = 1,2,3$. Figure 20 shows the estimated disparity map for University of Tsukuba scene, where the accuracy is 95.5%. In vivo, the number of cell scales can be greater than three and with various possible kernel sizes. Hence, this test result is not necessarily optimal.

The improved benchmark of 96.2% in Figure 7b can be achieved with a computationally simpler single-scale binocular boundary, defined as

$$b^{B}_{ijd} = \exp(-\left(10(I^{L}_{i,j} - I^{R}_{i-s,j})\big/(10^{-5} + I^{L}_{i,j} + I^{R}_{i-s,j})\right)^2), \tag{23}$$

where $I_{ij}^{L/R}$ is the luminance of the left or right retinal image, respectively, and index *s* is the positional shift between left and right eye inputs that depends on the disparity d as defined in Equations (12-14).

Figure 20. Estimated disparity map for University of Tsukuba scene using equation (22) instead of (23).

V2 Layer 4 cells. The left and right monocular inputs are combined in layer 4 of V2. Since the monocular inputs do not yet have a depth associated with them, they are added to all depth planes along their respective lines-of-sight. The V2 layer 4 cells also receive feedback signals from the left and right V2 monocular surfaces (to be defined later) operating from V2 thin stripes to pale stripes. At steady-state, v_{ijkd} is defined by:

$$v_{ijkd} = \left(\left[c_{ijkd}^{B} - 0.1 \right]^{+} + \left(\left[c_{ijk}^{L} \right]^{+} + \left[c_{(i-s)jk}^{R} \right]^{+} \right) \right) (1 + \alpha_f f_{ijkd}) \left(\delta + (1-\delta) h(f_{ijkd}) \right), \tag{24}$$

where α_f is a constant (1.0) that scales the strength of surface-to-boundary feedback signals, and δ is a constant (0.2) that scales the activities of layer 4 cells. *h* is the signal function with *h(x)=1* if *x>0*, *0* otherwise. f_{ijkd} is the total V2 surface-to-feedback signal,

$$f_{ijkd} = [f^L_{ijkd} - 0.03]^+ + [f^R_{ijkd} - 0.03]^+. \tag{25}$$

V2 layer 2/3 complex cells. The V2 layer 2/3 cells receive input from V2 layer 4. The bipole cells in V2 layer 2/3 implement perceptual grouping by long-range horizontal connections, as well as the disparity filter. The membrane potential, g_{ijkd}, of the bipole cell in V2 layer 2/3 at position (i,j) that codes orientation k and disparity d obeys the equation:

$$\begin{aligned}\frac{d}{dt} g_{ijkd} = &-g_{ijkd} + (1 - g_{ijkd})\left(I^g_{ijkd} + 10\left[\sum_v H^{E_g}_{ijkdv} - H^{I_g}_{ijkd} \right]^+ \right), \\ &-(0.2 + g_{ijkd})(G^O_{ijkd} + G^S_{ijkd} + G^P_{ijkd})\end{aligned} \tag{26}$$

where I^g_{ijkd} is the input signal from V2 layer 4 that is given by:

$$I^g_{ijkd} = [v_{ijkd}]^+. \tag{27}$$

The V2 layer 2/3 collinear bipole cells receive long-range input from other (almost) collinear and coaxial bipole cells at nearby positions with the same disparity preference. Term $H^{E_g}_{ijkdv}$ is the input from branch *v* of the bipole cell at position *(i,j)*, orientation *k* and disparity *d*:

$$H^{Eg}_{ijkdv} = \sum_{pq} W^g_{pqijkv} [g_{pqkd} - 0.05]^+, \tag{28}$$

where the long-range connection weights (W^g_{pqijkv}) for the horizontal orientation (*k=1*) are defined as follows (*v=1* for left branch and *v=2* for right branch):

$$W^g_{pqijk1} = \left[sign(i-p) \exp\left(-\left(\frac{(i-p)^2}{\sigma_p^2} + \frac{(j-q)^2}{\sigma_q^2} \right) \right) \right]^+, \tag{29}$$

and

$$W^{g}_{pqijk\,2} = \left[sign(p-i)\exp\left(-\left(\frac{(i-p)^2}{\sigma_p^2} + \frac{(j-q)^2}{\sigma_q^2} \right) \right) \right]^{+}, \tag{30}$$

where $sign(x) = 1$ if $x > 0$, -1 if $x < 0$, and 0 otherwise. The parameters $\sigma_p = 20,\ \sigma_q = 0.2$, and the spatial connection range (diameter) is 11. The connection weights for other orientations are obtained by appropriate rotations.

Term H^{Ig}_{ijkd} is the inhibitory input from the inhibitory interneurons, defined by:

$$H^{Ig}_{ijkd} = \sum_{v} [s^{g}_{ijkdv}]^{+}, \tag{31}$$

with the activity, s^{g}_{ijkdv}, of the inhibitory interneuron for branch *v* being defined by:

$$s^{g}_{ijkdv} = (-B_v + \sqrt{B_v{}^2 + 4\eta H^{Eg}_{ijkdv}})/2\eta, \tag{32}$$

where

$$B_v = 1 + \eta(H^{Eg}_{ijkdu} - H^{Eg}_{ijkdv}), \tag{33}$$

with *u, v* are the two branches of orientation *k*, and parameter $\eta = 100$.

Term G^{O}_{ijkd} is the inhibition across orientation within depth and position:

$$G^{O}_{ijkd} = 0.2\left(\sum_{r \neq k} \sin^2 \frac{(k-r)\pi}{K} [g_{ijrd} - 0.03]^{+} \right), \tag{34}$$

where K is the number of total orientations.

Term G^{S}_{ijkd} is the inhibition across space within depth:

$$G^{S}_{ijkd} = 20 \sum_{p \neq i, q \neq j, r} W_{pqijk} [g_{pqrd} - 0.03]^{+}, \tag{35}$$

where the weight W_{pqijk} for horizontal orientation is defined by

$$W_{pqijk} = \frac{1}{2\pi\sigma_p\sigma_q} \exp\left(-\left(\frac{(i-p)^2}{\sigma_p^2} + \frac{(j-q)^2}{\sigma_q^2} \right) \right), \tag{36}$$

with $\sigma_p = 1.5,\ \sigma_q = 1.5$ with the kernel size 9. The weights W_{pqijk} for other orientations are defined by appropriate rotations.

Each V2 layer 2/3 bipole cell also receives inhibitory input from other bipole cells that share either of its monocular inputs (line-of-sight competition). Term G^P_{ijkd} in (26) is the inhibition across disparities along the lines-of-sight:

$$G^P_{ijkd} = 200\sum_{d'\neq d}([g_{ijkd'} - \beta_g]^+ + [g_{(i+s'-s)jkd'} - \beta_g]^+), \tag{37}$$

where $[g_{ijkd'} - \beta_g]^+$ and $[g_{(i+s'-s)jkd'} - \beta_g]^+$ are V2 layer 2/3 bipole cell inhibitory inputs along the left and right lines-of-sight with positional shifts *s* and *s'* and inhibitory signal threshold β_g (0.03).

V2 thin stripe monocular surfaces. V2 surface cells implement the surface filling-in process. The model adds both V2 layer 2/3 binocular boundary and V1 monocular boundary together to form the connected boundaries as resistive barriers to the filling-in process in response to complex natural scenes. At the same time, each V2 surface cell inhibits all other cells that shares one of its monocular lines-of-sight. The V2 surface disparity filter and together with the filling-in process create 3D monocular surface representations in V2 thin stripes. The equations describing the above processes are as follows

$$\frac{d}{dt}F^{L/R}_{ijd} = -\alpha_1 F^{L/R}_{ijd} + \sum_{(p,q)\in N_{ij}}(F^{L/R}_{pqd} - F^{L/R}_{ijd})\Phi^{L/R}_{pqijd} + I^{L/R}_{ijd}\ , \tag{38}$$

$$\tau\frac{d}{dt}\hat{F}^{L/R}_{ijd} = -\alpha_2 \hat{F}^{L/R}_{ijd} + F^{L/R}_{ijd} - \hat{F}^{L/R}_{ijd} J^{L/R}_{ijd}\ , \tag{39}$$

where decay rates $\alpha_1 = 1,\ \alpha_2 = 10^{-5},\ \tau > 1$ is a temporal scale constant to be defined later, and

$$\Phi^{L/R}_{pqijd} = \frac{1}{1 + 100\left(g^{L/R}_{ijd} + g^{L/R}_{pqd}\right)}, \tag{40}$$

where $g^{L/R}_{ijd}$ is the resistive barrier defined by

$$g^L_{ijd} = \sum_k c^L_{i,j,k}\left[0.1 + \left\{[g_{ijkd} - \theta]^+ + 0.1\sum_{d'<d}\left(\left[g_{ijkd'} - \theta_g\right]^+\right)\right\}\right]^+, \tag{41}$$

$$g^R_{ijd} = \sum_k c^R_{i-s,j,k}\left[0.1 + \left\{[g_{ijkd} - \theta]^+ + 0.1\sum_{d'<d}\left(\left[g_{(i-s+s')jkd'} - \theta_g\right]^+\right)\right\}\right]^+, \tag{42}$$

where $c_{i,j,k}^{L/R}$ is the left/right monocular boundary, and g_{ijkd} is the V2 binocular boundary.

Initially the input $I_{ijd}^{L/R} = y_{ijd}^{L/R}$, where $y_{ijd}^{L/R}$ is the V1 surface signal.

Then after the first iteration,

$$I_{ijd}^{L/R} = f(\hat{F}_{ijd}^{L/R}) y_{ijd}^{L/R}, \tag{43}$$

where f is a signal function defined by

$f(x) = \frac{\Gamma x^{\beta}}{\Gamma + x^{\beta}}$, with $\beta = 1.5$, and when $\Gamma >> x^{\beta}$ we have $f(x) \approx x^{\beta}$, which is used in the simulation.

Terms $J_{ijd}^{L/R}$ in Eq. (39) is defined by

$$J_{ijd}^{L} = \sum_{d'} F_{i,j,d'}^{L}, \tag{44}$$

$$J_{ijd}^{R} = \sum_{d'} F_{i-s+s',j,d'}^{R}, \tag{45}$$

where $F_{i,j,d'}^{L}$ and $F_{i-s+s',j,d'}^{R}$ are V2 surface cell inhibitory inputs along the left and right lines-of-sight with positional shifts *s* and *s'*.

Solving the above equations at equilibria we get

$$F_{ijd}^{L/R} = \frac{I_{ijd}^{L/R} + \sum_{(p,q)\in N_{ij}} F_{pqd}^{L/R} \Phi_{pqijd}^{L/R}}{1 + \sum_{(p,q)\in N_{ij}} \Phi_{pqijd}^{L/R}}, \tag{46}$$

and

$$\hat{F}_{ijd}^{L/R} = \frac{F_{i,j,d}^{L/R}}{10^{-5} + J_{ijd}^{L/R}}. \tag{47}$$

The equilibrium Eqs. (46) and (47) were used in the simulations. The τ in Eq. (39) is chosen such that doing 100 iterations for (46) for each iteration of (47).

Surface-to-boundary feedback signals. The V2 monocular surfaces in the thin stripes generate surface-to-boundary feedback signals to the V2 pale stripes to modulate the activities of corresponding V2 layer 4 cells. Output signals from the filled-in activities in the V2 thin stripes are derived from oriented filters

$$f_{ijkd}^{L/R,+} = \sum_{p,q} \mathrm{K}_{pqk} \left[F_{i+p,j+q,d}^{L/R} \right]^+, \tag{48}$$

$$f_{ijkd}^{L/R,-} = -\sum_{p,q} \mathrm{K}_{pqk} \left[F_{i+p,j+q,d}^{L/R} \right]^+, \tag{49}$$

where the Gabor kernel K_{pqk} is defined in equation (9).

The surface-to-boundary signals $f_{ijkd}^{L/R}$ are finally defined by

$$f_{ijkd}^{L/R} = [f_{ijkd}^{L/R,+}]^+ + [f_{ijkd}^{L/R,-}]^+. \tag{50}$$

V4 surfaces. V4 receives boundary signals from V2 layer 2/3 and lightness signals from the LGN coupled with the signals from the V2 thin stripes.

$$w_{ijd} = \frac{z_{ijd} + \sum_{(p,q)\in N_{ij}} w_{pqd} \Phi_{pqijd}}{1 + \sum_{(p,q)\in N_{ij}} \Phi_{pqijd}}, \tag{51}$$

where

$$z_{ijd} = Y_{ijd}^L + Y_{ijd}^R, \tag{52}$$

with

$$Y_{ijd}^L = X_{ij}^L \left[F_{ijd}^L - \sum_{d'<d} F_{ijd'}^L \right]^+, \tag{53}$$

$$Y_{ijd}^R = X_{(i-s)j}^R \left[F_{ijd}^R - \sum_{d'<d} F_{(i-s+s')jd'}^R \right]^+, \tag{54}$$

where $X_{ij}^{L/R}$ and $F_{ijd}^{L/R}$ are defined in Eq. (21) and Eq. (46), respectively. In Equations (53) and (54), successfully filled-in features in V2 thin stripes are subtracted from farther depths (surface pruning) in V4 to ensure that opaque objects do not look transparent. This process, first proposed in Grossberg (1994), was simulated in several subsequent articles to generate various 3D percepts; e.g., Grossberg and McLoughlin (1997), Grossberg and Yazdanbakhsh (20050, and Fang and Grossberg (2009).

In Eq. (51), term

$$\Phi_{pqijd} = \frac{\delta}{1 + \rho\left(\hat{g}_{ijd} + \hat{g}_{pqd}\right)}, \tag{55}$$

where the spread scale parameter $\delta = 1$, the blocking scale parameter $\rho = 1000$, and the resistive boundary barrier

$$\hat{g}_{ijd} = g^{L}_{ijd} + g^{R}_{ijd}, \tag{56}$$

with $g^{L/R}_{ijd}$ are defined in Eqs. (41) and (42)

5. Discussion

The 3D LAMINART model was developed to explain and predict perceptual and neurobiological data in terms of how laminar cortical mechanisms interact to create 3D boundary and surface representations. The article proposes several refinements whereby to better meet the challenges for processing natural scenes; namely, (1) 3D boundaries are often incomplete, either due to noise in the acquisition of the images, or due to unavailability of like-polarity matches at some boundary positions; and (2) cluttered scenes incorporate many possibilities for false binocular matches. The enhanced model has a more symmetric global anatomical organization, with interactions between blobs and interblobs in V1, as well as between thin stripes and pale stripes in V2, and disparity filters in both the thin stripes and pale stripes of V2.

The existence of disparity filters, in particular, stands as a prediction. It is known, however, that there are long-range bipole-like interactions in V2 (von der Heydt, Peterhans, and Baumgartner, 1984; Peterhans and von der Heydt, 1989), as well as a complex organization of shorter-range recurrent inhibitory interactions (Lund, Yoshioka, and Levitt, 1993; Tamas, Somogyi, and Buhl, 1998) that are consistent with the needs of both bipole grouping and the disparity filter requirements for inhibition along lines-of-sight of spurious boundaries. These various interactions, taken together, propose how the boundary and surface cortical streams may interact to overcome each other's complementary computational deficiencies (Grossberg, 1994, 2017), and to thereby generate a conscious visual percept that realizes the property of complementary consistency.

Some other biological models of various aspects of 3D vision have also been proposed (Parker, 2007). The well-known *energy model* includes binocular complex cells

in V1 and predicts the shape of the binocular receptive field of complex cells in the cat (Fleet, Wagner, and Heeger, 1996; Ohzawa, 1998; Ohzawa, DeAngelis, and Freeman, 1990). Although variants of the energy model, including both phase and positional shifts to compute disparities, have been successfully used to provide the front end for some stereo computations (Assee and Qian, 2007; Chen and Qian, 2004; McLoughlin and Grossberg, 1998; Qian and Zhu, 1997), such models are insufficient to explain how 3D boundary groupings and surface representations form and lead to conscious percepts, including surface percepts of random dot stereograms, da Vinci stereopsis, Panum's limiting cases, transparency, and bistable percepts, percepts that 3D LAMINART can explain and simulate (Cao and Grossberg, 2005; Fang and Grossberg, 2009; Grossberg and Howe, 2003; Grossberg and Swaminathan, 2004; Grossberg and Yazdanbakhsh, 2005; Grossberg, Yazdanbakhsh, Cao, and Swaminathan, 2008). The V1 binocular cells in our model are similar to those of the energy model, but our model goes far beyond that to propose how 3D boundary and surface representations are generated by laminar cortical circuits in V1, V2, and V4. Chen and Qian (2004) have proposed a coarse-to-fine disparity energy model that is capable of estimating disparity maps for natural images, but no estimate of accuracy is reported for their model. See Table 1 for a comparison of various computational and biological stereovision model properties.

Model extensions: Towards conscious seeing and recognition of 3D scenes. Our results could be improved using several model refinements that have been used in other modeling studies of biological vision. One such refinement would be to incorporate boundary and surface computations that can represent tilted and slanted surfaces in depth. Grossberg and Swaminathan (2004) have shown how, in particular, bipole cells can be generalized to model *disparity gradient cells* that can represent a boundary which spans more than one depth, and *angle cells* that are selectively activated by particular angles between straight edges. Interactions of disparity gradient cells and angle cells can disambiguate tilt in response to otherwise ambiguous 2D pictures and 3D scenes. This is a natural generalization of the bipole cell concept if only because bipole cells that represent straight contours within one depth, disparity gradient cells that represent contours that cross several depths, and angle cells can all develop using the same learning laws. This becomes clear when one considers that a perceptually straight edge is not

straight when it is represented in the visual cortex after the cortical magnification factor, or log polar map, transforms its retinal image (Daniel and Whitteridge, 1961; Drasdo, 1977; Schwartz, 1977). Even a straight bipole cell is an "angle cell" within such a cortical map. In like manner, bipole cell connections across cortical map positions that represent a single depth must be learned in just the same way as the connections across map positions of disparity gradient cells that span several depths.

Incorporation of spatial attentional and eye movement control mechanisms for active scanning of a scene may also improve the model's explanatory power. In particular, the brain uses spatial attention and eye movements to fixate areas of interest in a scene. Due to the cortical magnification factor, extrafoveal regions do not provide high resolution vision. The foveal area uses the cortical magnification factor to provide much higher resolution to fixated areas. One of the computational limitations of the current model is the small number of pixels devoted to resolving locally ambiguous pixels and binocular matches in a complex natural scene. For example, local contrasts for the arms of the lamp in the University of Tsukuba Scene are weak (see Figure 7). Foveation can devote more pixels to areas of interest, and spatial attention is known to enhance perceived image contrast (Carrasco, Penpeci-Talgar, and Eckstein, 2000; Reynolds and Desimone, 2003). The ARTSCAN neural model, and its subsequent refinements, uses a log polar mapping to process imagery, followed by a simplified 3D LAMINART front end, to simulate how our brains learn *invariant* object category representations for object attention, recognition and prediction (Cao, Grossberg, and Markowitz, 2011; Chang, Grossberg, and Cao, 2014; Fazl, Grossberg and Mingolla, 2009; Foley, Grossberg, and Mingolla, 2012).

In particular, ARTSCAN and its extension to the positional ARTSCAN, or pARTSCAN, model (Cao, Grossberg, and Markowitz, 2011) and the ARTSCAN Search model (Chang, Grossberg, and Cao, 2014), proposed how spatial and object attention work together to search a scene with eye movements, thereby bringing the fovea and its magnified representation onto regions of interest, and to learn view-, position-, and size-invariant object category representations during such free viewing. Grossberg (2007, 2009) predicted, and Fazl et al. (2009) first simulated, how surface-fitting spatial attention, or an *attentional shroud*, can modulate view-invariant learning by ensuring that

only views of the same object can associated with an emerging invariant object category representation. Grossberg and Huang (2009) showed how the gist of a scene can be rapidly learned as a large-scale texture category, and how attentional shrouds can improve the recognition that gist classification alone can achieve by focusing on, and classifying, a few additional scenic textures.

An attentional shroud is part of a *surface-shroud resonance* that arises due to feedback interactions between a surface representation (e.g., in cortical area V4) and spatial attention (e.g., in posterior parietal cortex, or PPC), which focuses spatial attention upon the object to be learned. ARTSCAN and its generalizations predict that we consciously see surface-shroud resonances; that is, we see the visual qualia of a surface when they are synchronized and amplified within a surface-shroud resonance. This concept helps to explain a wide range of challenging psychophysical and neurobiological data (for reviews, see Grossberg, 2013, 2017).

A different kind of resonance supports conscious recognition of visual objects and scenes. Such a resonance is called a *feature-category resonance*. Such a resonance may, for example, occur between a distributed feature pattern that represents an object (e.g., in cortical areas V2 and/or V4) and a recognition category that classifies it (e.g., in inferotemporal cortex, or IT). When a feature-category resonance synchronizes with a surface-shroud resonance via its shared visual cortical representations, then the object can simultaneously be consciously seen and recognized. A feature-category resonance can be used to recognize objects and scenes whose consciously seen surface representations may be quite incomplete.

The above models used simplified 2D boundary and surface representations to achieve their goals. 2D boundary and surface representations have been used to enhance, and to incrementally learn to recognize, images of natural scenes that have been processed by multiple kinds of artificial sensors, including LADAR, SAR, multispectral IR, and night vision sensors. Adaptive Resonance Theory, or ART, algorithms processed these images and learned to classify the textures, objects, and scenes. Many of these applications were developed in collaborations between Gail Carpenter and Stephen Grossberg and their colleagues with MIT Lincoln Laboratory in the 1990s. Synthetic Aperture Radar, or SAR, images, remote sensing images, and natural textures were given

particular attention in order to provide effective solutions for normalizing input dynamic range, reducing noise, and overcoming the highly pixelated and discontinuous nature of the images by completing coherent boundaries between statistically correlated pixels and filling in surface contrasts within the resulting multiple-scale boundary webs, before inputting them to an ART algorithm for classification (Asfour, Carpenter, and Grossberg, 1995; Bhatt, Carpenter, and Grossberg, 2007; Carpenter, Gjaja, Gopal, and Woodcock, 1997; Carpenter, Gopal, Macomber, Martens, and Woodcock, 1999; Grossberg, Mingolla, and Williamson, 1995; Grossberg and Huang, 2009; Grossberg and Williamson, 1999; Mingolla, Ross, and Grossberg, 1999). 3D generalizations of these boundary and surface properties may also process this expanded range of challenging images.

Other image processing applications using the same model foundations have focused on refining the models' ability to compensate for variable illumination conditions and to automatically *anchor* the resultant image; that is, use its full dynamical range to create an absolute representation of the color "white" that is perceived in a scene. This Anchored Filling-In Lightness Model, or aFILM, is also generalized to process color images under variable illumination conditions and with a much faster mechanism of filling-in (Grossberg and Hong, 2006; Hong and Grossberg, 2004). These generalizations may also be consistently embedded within the current model.

The 3D ARTSCAN model has extended the competence of 3D boundary and surface computations to an *active vision* framework wherein eye, or camera, movements freely scan a 3D scene while perceiving and learning to recognize it. These extensions may also be consistently embedded within the current model (Grossberg, Srinivasan, and Yazdanbakhsh, 2014). In particular, 3D ARTSCAN simulates how binocular fusion can be maintained even as the eyes scan a 3D scene and learn invariant object categories in it. The 3D ARTSCAN model uses a process of *predictive remapping* to maintain the stability of key brain representations during scanning eye movements. These include both the stability of binocularly fused boundaries and the stability of attentional shrouds during eye movements. The ARTSCAN model had previously used predictive remapping for the latter purpose. In order to achieve the stability of 3D percepts as the eyes freely scan a 3D scene, successive eye movements predictively update 3D boundaries that are

computed in head-centered, or spatial, coordinates. Coordinate transformations between spatial and retinotopic coordinates using these remapped binocular boundaries can preserve previously established binocular fusions of object surfaces that are seen in depth within the scene, even though their retinotopic positions have changed, at the same time that predictive remapping of the coordinates that maintain an active shroud in spatial coordinates support learning of invariant 3D object categories as the eyes scan different object views in depth. 3D ARTSCAN hereby clarifies how surface-shroud resonances can support conscious percepts of 3D scenes as the eyes scan and learn about a scene. Although 3D ARTSCAN was used to learn categories of objects from the Caltech 101 image database, these objects were not represented or learned as part of a cluttered scene. This is another useful next step of model development.

In summary, future research can benefit from using the more highly developed 3D boundary and surface representations of the current model, combined with generalizations to tilted and slanted images, spatial and object attention, eye movements and the cortical magnification factor, to provide higher resolution 3D representations of salient scenic objects and textures with which to categorize and understand natural scenes.

References

Asfour, Y.R., Carpenter, G.A., and Grossberg, S. (1995). Landsat satellite image segmentation using the fuzzy ARTMAP neural network. *Proceedings of the World Congress on Neural Networks (WCNN-95),* **I**-150-156.

Assee, A. and Qian, N. (2007). Solving da Vinci stereopsis with depth-edge-selective V2 cells, *Vision Research*, *47*, 2585-2602.

Baker, H.H. and Binford, T.O. (1981). Depth from Edge and Intensity Based Stereo. *Proc. Seventh Int'l Joint Conf. Artificial Intelligence,* 631-636.

Bhatt, R., Carpenter, G., and Grossberg, S. (2007) Texture segregation by visual cortex: Perceptual grouping, attention, and learning. *Vision Research*, 47, 3173-3211.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues.* Leipzig: Barth.

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

Cao, Y., and Grossberg, S. (2012). Stereopsis and 3D surface perception by spiking neurons in laminar cortical circuits: A method of converting neural rate models into spiking models. *Neural Networks*, *26*, 75-98.

Cao, Y., Grossberg, S., and Markowitz, J. (2011). How does the brain rapidly learn and reorganize view- and positionally-invariant object representations in inferior temporal cortex? *Neural Networks*, 24, 1050-1061.

Carpenter, G.A., Gjaja, M.N., Gopal, S., and Woodcock, C.E. (1997). ART neural networks for remote sensing: Vegetation classification from Landsat TM and terrain data. *IEEE Transactions on Geoscience and Remote Sensing,* 35, 308-325.

Carpenter, G.A., Gopal, S., Macomber, S., Martens, S., and Woodcock, C.E. (1999). A neural network method for mixture estimation for vegetation mapping. *Remote Sensing of Environment*, 70, 138-152.

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

Chang, H.-C., Grossberg, S., and Cao, Y. (2014) Where's Waldo? How perceptual cognitive, and emotional brain processes cooperate during learning to categorize and find desired objects in a cluttered scene. *Frontiers in Integrative Neuroscience*, doi: 10.3389/fnint.2014.0043,
https://www.frontiersin.org/articles/10.3389/fnint.2014.00043/full.

Chen, Y. and Qian, N. (2004). A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Computation, 16*, 1545-1577.

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

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

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

DeYoe, E. A , & Van Essen, D . C. (1988). Concurrent processing streams in monkey visual cortex. *Trends in Neurosciences*, *11*, 214-226.

Fang, L. and Grossberg, S. (2009). From stereogram to surface: How the brain sees the world in depth. *Spatial Vision*, *22*, 45-82.

Fazl, A., Grossberg, S., and Mingolla, E. (2009). View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive Psychology*, *58*, 1-48.

Felleman, D. J. and van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1, 1-47.

Fleet, D. J., Wagner, H., and Heeger, D. J. (1996). Encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Res.*, *36*, 1839–1858.

Foley, N.C., Grossberg, S. and Mingolla, E. (2012). Neural dynamics of object-based multifocal visual spatial attention and priming: Object cueing, useful-field-of-view, and crowding. *Cognitive Psychology*, *65*, 77-117.

Gillam, B., Blackburn, S. and Nakayama, K. (1999). Stereopsis based on monocular gaps: Metrical encoding of depth and slant without matching contours. *Vision Research*, *39,* 493-502.

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

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

Grossberg, S. (1994). 3D vision and figure-ground separation by visual cortex. *Perception and Psychophysics*, *55*, 48-120.

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

Grossberg, S. (1999). How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex, *Spatial Vision, 12*, 163-186.

Grossberg, S. (2007). Towards a unified theory of neocortex: Laminar cortical circuits for vision and cognition. For *Computational Neuroscience: From Neurons to Theory and Back Again*, eds: Paul Cisek, Trevor Drew, John Kalaska; Elsevier, Amsterdam, pp. 79-104.

Grossberg, S. (2008). The art of seeing and painting. *Spatial Vision*, 21, 463-486.

Grossberg, S. (2009). Cortical and subcortical predictive dynamics and learning during perception, cognition, emotion, and action. *Philosophical Transactions of the Royal Society of London*, special issue "Predictions in the brain: Using our past to generate a future", *364*, 1223-1234.

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks, 37,* 1-47.

Grossberg, S. (2016). Cortical dynamics of figure-ground separation in response to 2D pictures and 3D scenes: How V2 combines border ownership, stereoscopic cues, and gestalt grouping rules. *Frontiers in Psychology*. 26 January 2016. http://journal.frontiersin.org/article/10.3389/fpsyg.2015.02054/full

Grossberg, S. (2017). Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. *Neural Networks, 87*, 38-95. https://www.sciencedirect.com/science/article/pii/S0893608016301800

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

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

Grossberg, S. and Huang, T.-R. (2009). ARTSCENE: A neural system for natural scene classification. *Journal of Vision*, *9*, 1-19.

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

Grossberg, S. and McLoughlin, N.P. (1997). Cortical dynamics of 3-D surface perception: Binocular and half-occluded scenic images. *Neural Networks*, *10,* 1583-1605.

Grossberg, S. and Mingolla, E. (1985a). Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. *Perception and Psychophysics, 38,* 141-147.

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

Grossberg, S., Mingolla, E. and Ross, W. D. (1997). Visual brain and visual perception: How does the cortex do perceptual grouping? *Trends in Neuroscience, 20,* 106-111.

Grossberg, S., Mingolla, E., and Williamson, J. (1995). Synthetic aperture radar processing by a multiple scale neural system for boundary and surface representation. *Neural Networks,* 8, 1005-1028.

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

Grossberg, S., Srinivasan, K., and Yazdanbakhsh, A. (2014). Binocular fusion and invariant category learning due to predictive remapping during scanning of a depthful scene with eye movements. *Frontiers in Psychology*: *Perception Science,* doi: 10.3389/fpsyg.2014.01457 http://journal.frontiersin.org/Journal/10.3389/fpsyg.2014.01457/full

Grossberg, S. and 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, 1147-1187.

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

Grossberg, S., and Williamson, J. R. (1999). A self-organizing neural system for learning to recognize textured scenes. *Vision Research,* 39**,** 1385-1406.

Grossberg, S. and Williamson, J.R. (2001). A neural model of how horizontal and interlaminar connections of visual cortex develop into adult circuits that carry out perceptual groupings and learning. *Cerebral Cortex, 11,* 37-58.

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

Grossberg, S., Yazdanbakhsh, A., Cao, Y., and Swaminathan, G. (2008). How does binocular rivalry emerge from cortical mechanisms of 3-D vision? *Vision Research*, *48*, 2232-2250.

Heeger, D.J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9,* 181-197.

Hong, S. and Grossberg, S. (2004). A neuromorphic model for achromatic and chromatic surface representation of natural images. *Neural Networks*, 2004, 17, 787-808.

Howard, I. P. and Rogers, B. J. (1995). *Binocular Vision and Stereopsis.* New York: Oxford University Press.

Hirschmuller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30,* 328-341.

Huang, X. and Paradiso, M.A. (2008). V1 response timing and surface filling-in. *Journal of Neurophysiology, 100*, 539-547.

Julesz, B. (1971). *Foundations of Cyclopean Perception.* Chicago: The University of Chicago Press.

Kanade, T. and Okutomi, M. (1994). A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment. *IEEE Trans. Pattern Analysis and Machine Intelligence, 16*, 920-932.

Kelly, F. J. and Grossberg, S. (2000). Neural dynamics of 3-D surface perception: Figure-ground separation and lightness perception. *Perception and Psychophysics, 62*, 1596-1619.

Lamme, V.A.F., Rodriguez-Rodriguez, V. and Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the Macaque monkey. *Cerebral Cortex*, *9*(4), 406-413.

Levine, M., O'Handley, D. and Yagi, G. (1973). Computer Determination of Depth Maps, *Computer Graphics and Image Processing, 2,* 131-150.

Lloyd, S.A., Haddow, E.R. and Boyce, J.F. (1987). A Parallel Binocular Stereo Algorithm Utilizing Dynamic Programming and Relaxation Labelling, *Computer Vision, Graphics, and Image Processing, 39*, 202-225.

Lund, J. S., Yoshioka, T., and Levitt, J. B. (1993). Comparison of intrinsic connectivity in different areas of Macaque monkey cerebral cortex. *Cerebral Cortex*, 3, 148-162.

Marr, D. and Poggio, T. (1976). Cooperative Computation of Stereo Disparity, *Science, 194*, 209-236.

Marr, D. and Poggio, T. (1979). A Computational Theory of Human Stereo Vision, *Proc. Royal Soc. London B, 204*, 301-328.

Martin, J. H. (1989). *Neuroanatomy: Text and atlas*. Norwalk: Appleton and Lange.

McKee, S. P., Bravo, M. J., Smallman, H. S. and Legge, G. E. (1995). The 'uniqueness constraint' and binocular masking. *Perception*, *24,* 49-65.

McKee, S. P., Bravo, M. J., Taylor, D. G. and Legge, G. E. (1994). Stereo matching precedes dichoptic masking. *Vision Research*, *34,* 1047-1060.

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

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

Mori, K., Kidode, M. and Asada, H. (1973). An Iterative Prediction and Correction Method for Automatic Stereo Comparison, *Computer Graphics and Image Processing, 2*, 393-401.

Nakayama, K., and Shimojo, S. (1990). da Vinci stereopsis: depth and subjective occluding contours from unpaired image points. *Vision Research*, *30,* 1811-1825.

Ohzawa, I. (1998). Mechanisms of stereoscopic vision: the disparity energy model. *Current Opinion in Neurobiology, 8,* 509-515.

Ohzawa, I., DeAngelis, G. C., and Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, *249*, 1037–1041.

Pandya, D. N. and Yeterian, E. H. (1985). Architecture and connections of cortical association areas. In A. Peters and E. G. Jones, eds. *Cerebral Cortex* **10** (Plenum Press, New York).

Paradiso, M. A. and Nakayama, K. (1991). Brightness perception and filling-in. *Vision Research*, *31*, 1221-1236.

Parker, A. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience, 8,* 379-391.

Pessoa, L., and Neumann, H. (1998). Why does the brain fill-in? *Trends in Cognitive Sciences*, *2*, 422-424.

Pessoa, L., Thompson, E. and Noë, A. (1998). Finding out about filling-in: a guide to perceptual completion for visual science and the philosophy of perception. *Behavioral and Brain Sciences*, *21*(6), 723-802.

Peterhans, E., and von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. Contours bridging gaps. *The Journal of Neuroscience, 9,* 1749–1763.

Poggio, G. F. (1991). Physiological basis of stereoscopic vision. In *Vision and Visual Dysfunction. Binocular Vision* (pp. 224-238). Boston, MA: CRC Press.

Poggio, G.F., and Fischer, B. (1977). Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *Journal of Neuroscience, 40(6),* 1392-1405.

Poggio, G.F., Gonzalez, F. and Krause, F. (1988). Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *Journal of Neuroscience, 8(12),* 4531 -4550

Qian, N., and Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research, 37,* 1811-1827.

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

Rossi, A. F., Rittenhouse, C. D., and Paradiso, M. A. (1996). The representation of brightness in primary visual cortex. *Science*, *273*, 1104-1107.

Scharstein, D. and Szeliski, R. (1998). Stereo matching with nonlinear diffusion. *International Journal of Computer Vision, 28,* 155–174.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*, 7-42.

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

Sherman, D. and Peleg, S. (1990). Stereo by Incremental Matching of Contours. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 12,* 1102-1106.

Smallman, H. S. and McKee, S. P. (1995). A contrast ratio constraint on stereo matching. *Proceedings of the Royal Society of London B*, *260,* 265-271.

Sun, J., Zheng, N.N., and Shum, H.Y. (2003). Stereo matching using belief propagation. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 25,* 787-800.

Tamas, G., Somogyi, P., and Buhl, E. H. (1998). Differentially interconnected networks of GABAergic interneurons in the visual cortex of the cat. *Journal of Neuroscience*, 18, 4255-4270.

von der Heydt, R., Peterhans, E., and Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science, 224,* 1260–1262.

Xie, J., Girshick, R., and Farhadi, A. (2016). Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks, arXiv:1604.03650 [cs.CV].

Zitnick, C.L. and Kanada, T. (2000). A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 675-684.

Zbontar, J. and LeCun, Y. (2015). Computing the stereo matching cost with a convolutional neural network, Annual conference on *Computer Vision and Pattern Recognition*, 1592-1599.