

# Where-What Network 3: Developmental Top-Down Attention for Multiple Foregrounds and Complex Backgrounds

Matthew Luciw, *Student Member, IEEE* and Juyang Weng, *Fellow, IEEE*

**Abstract**—The Where-What Network 3 (WWN-3) is an artificial developmental network modeled after visual cortical pathways, for the purpose of attention and recognition in the presence of complex natural backgrounds. It is general-purpose and not pre-determined to detect a certain type of stimulus. It is a learning network, which develops its weights from images using a supervised paradigm and a local Hebbian learning algorithm. Attention has been thought of as bottom-up or top-down. This paper focuses on the biologically-inspired mechanisms of top-down attention in WWN-3, through top-down excitation that interacts with bottom-up activity at every layer within the network. Top-down excitation in WWN-3 can control the location of attention by imposing a certain location or disengaging from the current location. It can also control what type of object to search for. Paired layers and sparse coding deal with potential hallucination problems. Top-down attention in WWN occurs as soon as an action emerges at a motor layer, which could be imposed by a teacher or internally selected. Given two competing foregrounds in the same scene, WWN showed effective performance in all the attention modes tested.

## I. INTRODUCTION

Selective attention refers to some mechanisms by which an agent recodes its sensory information into a simpler, more useful form. Simplified relevant information is necessary for cognitive processes, such as decision making. Attention is essential for artificial agents that learn intelligent behavior in complex unconstrained environments, especially those that utilize vision. Understanding how attention works is essential to design such agents. However, understanding of the detailed mechanisms controlling how attention operates has been elusive.

Selection means that, at some stage of processing, a subset of the information is suppressed, or prevented from being further processed, while another subset of information is let through. What causes selection? In psychology, two types of attention are identified: stimulus-driven, bottom-up, exogenous processing and goal-driven, top-down, endogenous processing. Bottom-up selection is not controlled: foreground objects or locations tend to “pop out” at the viewer. However, if one is looking for something, the selection process becomes subsequently biased. Given the same scene with the same eye fixation, but two different top-down biases, the representation of the information that reaches the later stage can be very different. For example, imagine the differences between what a vehicle’s driver tends to attend to compared to a passenger, even if they look in the same direction.

Matthew Luciw and Juyang Weng are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 (email: {luciwmat, weng}@cse.msu.edu).

Where-What Networks (WWN) are biologically-inspired grounded networks that learn attention and recognition from supervision. By grounded, we mean such a network is internal to an autonomous agent, which senses and acts on an external environment. WWN network models consist of two pathways for identity (what) and location (where), so there are separate “motor” areas for identity and location. The motor areas connect to another later module that controls the actions of this agent. In our supervised paradigm, the agent is taught to attend by being coerced to act appropriately over many cases. For example, a teacher leads it to “say” “car” while the agent is looking at a scene containing a car, and the teacher points out the location of the car. Before learning, the agent does not understand the meaning of car, but it was coerced to act appropriately. Such action causes activation at the Type-Motor and Location-Motor in WWN-3. Top-down excitatory activity from these motor areas, which are concerned with semantic information, synchronize with bottom-up excitatory activity from the earlier areas, concerned with “physical” image information. Bidirectional co-firing is the cause of learning meaning within the network. Neurons on a particular layer learn their representations via input connections from other neurons from three locations: from earlier areas (ascending or bottom-up), from the same area (lateral), and from later areas (descending or top-down). Learning occurs in a biologically-inspired cell-centered (local) way, using an optimal local learning algorithm called Lobe Component Analysis [1].

WWN-3 utilized the following four different attention mechanisms: (1) **Bottom-up free-viewing**, (2) **Attention shift**, (3) **Top-down object-based search**, and (4) **Top-down location-based binding**. Top-down excitation is the impetus of the latter three mechanisms. In WWN-3, top-down excitation serves a modulatory role, while the bottom-up connections are directed information carriers, as are the suspected roles of these connections in the brain [2]. We’ll show how an architecture feature called paired layers is important so that the top-down excitation, which we think of internal expectation, can have appropriate influence without leading to hallucination or corrupting bottom-up (physical) information with top-down semantic bias. In addition to modulation, in WWN-3, top-down connections (along with the bottom-up and lateral connections) allow a network’s internal activity to synchronize. When there are multiple objects in the scene, there will be multiple internally valid solutions for type and location, but some of these solutions will actually be incorrect (mixed up). This is part of the well-known binding problem. But in WWN-3, after a bottom-up pass to select for

candidate locations and candidate types, top-down location bias effectively selects a particular location to analyze and it then synchronizes with the appropriate type, similar to Treisman’s idea of spotlight [3]. Top-down bias can also be introduced at the Type-Motor, causing “object search” — it picks the best candidate location given the particular type bias. After the network has settled on a particular type and location, it can disengage from the current location to try another location, or it can disengage from both location and type.

The first version WWN-1 [4] realized two competencies for single foregrounds over natural backgrounds: type recognition given a location and location finding given a type, but only 5 locations were tested. The second version WWN-2 [5] additionally realized attention and recognition for single objects in natural backgrounds without supplying either position or type (free-viewing), and also used a more complex architecture. Further, all pixel locations were tested. The work reported here corresponds to the third version of WWN — WWN-3. It expands the prior versions WWN-1 and WWN-2 to deal with multiple objects in natural backgrounds, using arbitrary foreground object contours (WWN-1 used square contours), and the test foregrounds are totally disjoint from training foregrounds. We concentrate here on the case where there are multiple possible foregrounds that the network was trained to recognize in each image.

The remainder of this paper is organized as follows. Background is presented in section II. Theory and mathematical descriptions of WWN are introduced in section III. Experiments and results are presented in section IV. Section V concludes the paper.

## II. BACKGROUND

There is a very rich literature available in visual attention and segmentation, but it is still an open challenge to build a biologically plausible developmental model that integrates both bottom-up and top-down modes of attention and recognition, without being limited to a specific task. In the saliency-based approach as implemented by Itti and Koch [6], winner-take-all operation of neurons on a master location map led to both binding and attention location. Saliency methods have been coupled with recognition. An example is NAVIS (Neural Active Vision) by Backer *et al.* [7]. In [8], top-down bias modified gains and selected the scale of processing to produce a biased saliency map, which is specific for the sought after class. CODAM [9] models attention as a control system, using a goal module to generate goal-signals that bias lower-level competitive processing. A few researchers have proposed connectionist models for transferring the image to a master object map. Examples include Olshausen, Anderson and Van Essen [10] and selective tuning(ST) by Tsostos *et al.* [11]. ST is a multilayer pyramidal network that uses a complete feedforward pass to find the best candidate location and type and a complete feedback pass to focus attention by inhibiting non-selected features and locations. Gating units perform selection from the top-down. Differences between ST and

WWN are that ST uses top-down selection (gating) and top-down inhibition through winner-take-all, while WWN uses top-down excitation through weighted connections; additionally WWN uses multiple motor areas, for controlling and sensing an agent’s actions, while ST uses a single area of interpretive output nodes. Deco and Rolls, 2004 [12], created a biologically inspired network for attention and recognition where top-down connections controlled part of attention, but were not enabled in the testing phase. Top-down excitatory connections in a connectionist model creates a highly recurrent (“loopy”) network, where control is an open problem. HTM as discussed by George and Hawkins [13] modeled a Bayesian way of interpreting cortical circuits. In the Bayesian framework, feedback did not directly interact with feedforward activity, whereas in WWN, top-down and bottom-up interact at every layer in the network. Fazl and Grossberg’s ARTSCAN [14] used What and Where pathways and top-down excitation, but focused on a different problem — object detection during unsupervised learning via “attentional shroud” formation to tell eye scans on an object’s surface from those that are not. It was shown to work well for non-overlapping objects over a black background.

### A. WWN Architecture

It is known that our visual system has two major pathways: ventral (“what”) for object identification and dorsal (“where”) that deals more with visuomotor aspects (i.e., where to reach for an object), which presumably codes an object’s location. These pathways separate from early visual areas and converge at prefrontal cortex, which is known to be active in top-down attention. Prefrontal cortex connects to motor areas. WWN was built inspired by the idea of these two separating and converging pathways. Meaningful foregrounds in the scene will compete for selection in the ventral stream, and locations in the scene will compete for processing in the dorsal stream.

There are five areas of computation in WWN-3. The input image is considered as retinal activation. Instead of a multi-area feature hierarchy (ventral pathway), we use a shape-sensitive area we called V4, but we don’t claim the representation is identical to V4. From this area, one path goes through the IT (inferotemporal) and TM (Type-Motor) — possibly analogous to the inferior frontal gyrus [9]. TM is concerned with object type. The other path goes through the PP (posterior parietal) area and LM (Location Motor) — possibly analogous to the frontal eye fields (FEF). LM is concerned with object location. Each of these five areas contains a 3D grid of neurons, where the first two dimensions are relative to image height and width and the third is “depth”, for having multiple features centered at the same location. These neurons compute their firing rates at each time  $t$ . WWN is a discrete-time, rate-coding model, and each firing rate is constrained from zero to one. The pattern of firing rates for a single depth at any time  $t$  can be thought of as an image. Computing inputs to a neuron in an area is equivalent to sampling the image of firing rates from the

input area images. There are two types of input sampling methods for an area — local or global:

- **Local input field:** V4 neurons have local input fields from the bottom-up. This means they sample the retinal image locally, depending on their position in the 2D major neural axes (ignoring depth). A neuron at location  $(i, j)$  with receptive field size  $w$ , will take input vector from a square of sides  $w$  long, centered at location  $(i + \lceil w/2 \rceil, j + \lceil w/2 \rceil)$ .
- **Global input field:** Neurons with global input fields sample the entire input area as a single vector.

An architecture figure for WVN-3 is shown in Fig. 1. We initialized WVN-3 to use retinal images of total size  $38 \times 38$ , having foregrounds sized roughly  $19 \times 19$  placed on them, with foreground contours based on the object’s contours. V4 had  $20 \times 20 \times 3$  neurons, with bottom-up local input fields (of  $19 \times 19$ ) at different locations on the retina (based on the neurons’ 2D locations), and top-down global receptive fields. PP and IT also had  $20 \times 20$  neurons and had bottom-up and top-down input fields that were global. LM had  $20 \times 20$  neurons with global bottom-up input fields, and TM had  $5 \times 1$  neurons (since there were 5 classes) with global bottom-up receptive fields.

### B. Attention Selection Mechanisms at a High-Level

Selective attention is not a single process; instead, it has several components. These mechanisms can be broken down into orienting, filtering and searching. These are not completely independent, but the distinctions are convenient for the following discussion. The Where-What network makes predictions about how each of the mechanisms could work, and, in the following, we will discuss how these mechanisms work in WVN-3.

1) *Orienting:* Orienting is the placement of attention’s location. In covert orienting, one places one’s focus at a particular location in the visual field without moving the eyes; this is different from overt orienting, which would require the eyes to move. Covert orientation is realized in WVN based on sparse firing (e.g., winner-take-all or WTA) in the LM area. LM neurons’ firing is correlated with different attention locations on the retina, which emerged through supervised learning. Changes in attended location can occur in two ways: (1) an attended area emerges through feedforward activity, (2) an attended location is imposed (“location-based” top-down attention, which could be done by a teacher or internally by the network, or (3) a currently attended location is suppressed, by boosting LM neurons in other areas. The effect of this is to de-engage attention, and shift to some other foreground.

2) *Filtering:* Attention was classically discussed in terms of filtering [15]–[17]. In order to focus on a certain item in the environment, the “strength” of the other information seems to be diminished. In WVN-3, there are multiple passes of filtering. 1. Early filtering. This is done without any top-down control. As WVN-3 is developmental, the result of early filtering depends totally on the filters that were

developed in the supervised learning phase. Then, responses of these developed V4 neurons in WVN are an indicator of what interesting foregrounds there are in the scene. If there is no top-down control, internal dynamics will cause to converge to a steady state with a single neuron active in each of TM and LM, representing the type (class) and location of the foreground. A single feedforward pass is thought to be enough in many recognition tasks. But if there are multiple objects, attention seems to focus on each individually [3].

Another filtering process, based on top-down excitation, occurs on the result of the first-pass filtering. 2. Biased filtering. This “second-pass” filtering is in the service of some goal, such as searching for a particular type of foreground. The first-pass filtering has coded the visual scene into firing patterns representing potentially meaningful foregrounds. Binding after the first pass would be done in a purely feed-forward fashion, and an incorrect result could then emerge at the motor layers. The second-pass filtering is due to top-down expectation. It re-codes the result of first pass filtering into biased firing patterns, which then influence the motor areas. The second-pass allows the network to synchronize its state among multiple areas. For example, feedback activity from the Location-Motor causes attention at a particular location, which causes the appropriate type, at that location, to emerge at the Type-Motor.

3) *Searching:* Searching for a foreground type is realized in WVN-3 based on competitive firing (e.g., WTA) of the Type-Motor (TM) neurons. Similar to the link between retinal locations and LM neurons, correlations between foreground type and TM neurons are established in the training phase. Along the ventral pathway, the location information is gradually discarded. Top-down type-based attention will cause type-specific activation to feed back from TM to V4, biasing first-stage filtered foregrounds that match the type being searched for. Afterwards, the new V4 activation feeds forward along the Where pathway, and will cause a single location to become attended based on firing in the location motor area.

## III. CONCEPTS AND THEORY

### A. Bottom-Up and Top-Down Integration

Here, we discuss how bottom-up and top-down information are integrated in the Hebbian network (see Fig. 1).

1) *Paired Input:* For some neuron on some layer, denote its bottom-up excitatory vector by  $\mathbf{x}$ , and its top-down excitatory vector as  $\mathbf{z}$ . We used paired input:

$$\mathbf{p} \leftarrow \left( \rho \frac{\mathbf{x}}{\|\mathbf{x}\|}, (1 - \rho) \frac{\mathbf{z}}{\|\mathbf{z}\|} \right) \quad (1)$$

where  $0 \leq \rho \leq 1$ . The parameter  $\rho$  allows the network to control the influence of bottom-up vs. top-down activation, since the vector normalization fundamentally places bottom-up and top-down on equal ground. Setting  $\rho = 0.5$  gives the bottom-up and top-down equal influence.

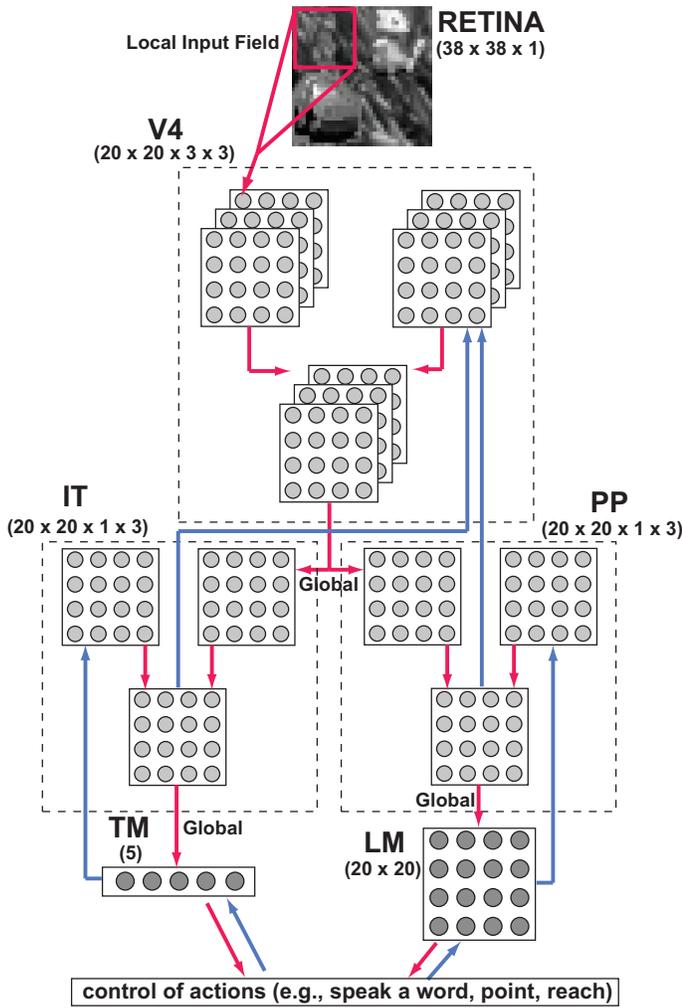


Fig. 1. WWN-3 system architecture. The V4 area has three layers (depths) of feature detectors at all image locations. Paired layers: each area has a bottom-up component, a top-down component, and a paired (integration) component. Within each component, lateral inhibitory competition occurs. Note that in this figure top-down connections point up and bottom-up connections point down.

2) *Paired Layers*: In Fig. 1, we show three internal components within each major area of V4, IT, and PP. This internal organization is called paired layers. Paired layers handle and store bottom-up information separately from the top-down boosted bottom-up information. The paired layer organization is inspired by the six-layer laminar organization which is found in all cortical areas [18], but it has been simplified to its current form. Further discussion of this architecture feature is found in [19].

Paired layers allow a network to retain a copy of its unbiased responses internally. Without paired layers, top-down modulatory effects can corrupt the bottom-up information. Such “corruption” might be useful, when the internal expectation (we consider top-down excitation as an internal expectation) relates to something in the visual scene. However, there are times when such expectations are violated. In such instances, the internal expectation is in opposition to

reality. Then, the incorrect feedback can potentially lead to false alarms, or hallucinations.

To explain further, we present the following formalism to explain how paired layers are implemented. Given a neuronal layer  $l$ , let  $\mathbf{V}$  be the matrix containing bottom-up column weight vectors to neurons in this layer, and  $\mathbf{M}$  be the matrix containing top-down column weight vectors to the layer, where each of the column vectors in these matrices are normalized.  $\mathbf{X}$  is a bottom-up input matrix: column  $i$  contains the input activations for neuron  $i$ 's bottom-up input lines.  $\mathbf{Z}$  is the top-down input matrix. For the following,  $\mathbf{X}$  and  $\mathbf{Z}$  are also column normalized. We use  $\mathbf{diag}(\mathbf{A})$  to mean the vector consisting of the diagonal elements of matrix  $\mathbf{A}$ .

**Non-paired layers**: First compute layer  $l$ 's pre competitive response  $\hat{\mathbf{y}}$ :

$$\hat{\mathbf{y}} = \rho \mathbf{diag}(\mathbf{V}^T \mathbf{X}) + (1 - \rho) \mathbf{diag}(\mathbf{M}^T \mathbf{Z}) \quad (2)$$

where  $\rho$  and  $1 - \rho$  are positive weights that control relative bottom-up and top-down influence. The post-competitive firing rate vector  $\mathbf{y}$  of layer  $l$  is computed after lateral inhibition function  $f$ , controlled by the layer's sparsity parameter  $k$ . The top  $k$  firing neurons will fire and others have firing rate set to zero, giving a sparse response:

$$\mathbf{y} = g(f(\hat{\mathbf{y}}, k)) \quad (3)$$

The lateral inhibition and sparse coding method  $f$  is achieved by sorting the components of the pre-response vector  $\hat{\mathbf{y}}$ . Let  $s_1$  be the highest value,  $s_k$  be the  $k$ -th highest. Then set a neuron's response as:

$$y_i \leftarrow \begin{cases} \hat{y}_i, & \text{if } \hat{y}_i \geq s_k \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where  $g$  is a threshold function that prevents low responses after competition.

**Paired layers**: With paired layers, the firing rate vector  $\mathbf{y}$  of layer  $l$  neurons are computed after lateral inhibition of both the bottom-up part  $\hat{\mathbf{y}}_b$  and top-down part  $\hat{\mathbf{y}}_t$  separately, and additional lateral inhibition for the integrated response:

$$\begin{aligned} \mathbf{y}_b &= g(f(\mathbf{diag}(\mathbf{V}^T \mathbf{X}), k_b)) \\ \mathbf{y}_t &= g(f(\mathbf{diag}(\mathbf{M}^T \mathbf{Z}), k_t)) \\ \mathbf{y} &= g(f(\rho \mathbf{y}_b + (1 - \rho) \mathbf{y}_t, k_p)) \end{aligned} \quad (5)$$

A thorough analysis of paired layers could not fit due to space limitations. A key idea is that the lateral inhibition causes sparse firing within the bottom-up layer and sparser firing in the paired layer (e.g.,  $k_b = 20$  and  $k_p = 10$  where there are 1200 neurons). The top-down layer is generally not as sparse (e.g.,  $k_t = 200$ ) so that it might reach as many potentially biased neurons as possible. In a non-paired layer, such diffuse top-down biasing can match up with many relatively weak responding filters (if there was no top-down influence) and boost them above the stronger filters that have

more bottom-up support. But the intermediate competition step in the paired layer method ensures the diffuse top-down biasing will not significantly boost the relatively weak filters since they were already eliminated in bottom-up competition. Filters with support from both bottom-up and top-down thus receive the most benefit.

Both bottom-up and top-down are highly local in both space and time — they are highly spatiotemporally sensitive. The input image (bottom-up) could change quickly, or the “intent” of the network (top-down) could change quickly. In either case, a paired layer adapts quickly.

### B. Learning Attention

Each bottom-up weight for a V4 neuron was initialized to a randomly selected  $19 \times 19$  training foreground, as seen in Fig 2(b). This greatly quickens training by placing the weights close to the locations in weight space we expect them to converge to. Initial bottom-up weights for IT, PP, TM, LM were randomly set. Top-down weights to V4, IT and PP were set to ones (to avoid initial top-down bias)<sup>1</sup>.

WWN learns through supervised learning externally and local learning internally. The Type-Motor and Location-Motor areas were firing-imposed per-sample in order to train the network. There are  $c$  neurons in TM, one for each class, and we used  $20 \times 20$  neurons in LM, one for each attention location. Each areas’ neurons self-organize in a way similar to self-organizing maps (SOM) [20], but with combined bottom-up and top-down input  $\mathbf{p}$ , and using the LCA algorithm for optimal weight updating. WWN’s top-down connections have useful roles in network development (training). They lead to discriminant features and a motor-biased organization of lower layers. Explaining these developmental effects are out of the scope of this paper, but have been written about elsewhere [21]. Further focus on learning and development in WWN is presented in [5].

For a single area to learn, it requires bottom-up input  $\mathbf{X}$ , top-down input  $\mathbf{Z}$ , bottom-up and top-down weights  $\mathbf{V}$  and  $\mathbf{M}$ , and the parameters  $\rho$  (controlling influence of bottom-up versus top-down),  $k_b$  (the number of neurons to fire and update after competition in the bottom-up layer),  $k_t$  (the same for the top-down layer), and  $k_p$  (for the paired layer). This area will output neuronal firing rates  $\mathbf{y}$ . It updates neuronal weights  $\mathbf{V}$  and  $\mathbf{M}$ .

The non-inhibited neurons update their weights using the Hebbian-learning LCA updating [1]:

$$\mathbf{v}_i \leftarrow \omega(\eta_i) \mathbf{v}_i + (1 - \omega(\eta_i)) \mathbf{x}_i y_i \quad (6)$$

where the plasticity parameters  $\omega(\eta_i)$  and  $(1 - \omega(\eta_i))$  are determined automatically and optimally based on the neuron’s updating age  $\eta_i$ . A neuron increments its age when it wins in competition. Learning rate of each neuron is a function of its firing age<sup>2</sup>. This learning is Hebbian as the strength

<sup>1</sup>Setting to initial positive values (nonzero) mimics the initial overgrowth of connections in early brain areas, later pruned by development.

<sup>2</sup>The above equation is for bottom-up weights. For top-down weights, substitute in  $\mathbf{z}_i$  for  $\mathbf{x}_i$  and  $\mathbf{m}_i$  for  $\mathbf{v}_i$ .

of updating depends on both presynaptic potentials (e.g.,  $\mathbf{x}_i$ ) and postsynaptic potentials (e.g.,  $y_i$ ).

Denote the per-area learning algorithm we have discussed as *LCA*. To train the whole WWN, the following algorithm ran over three iterations per sample. Let  $\theta = (k_b, k_t, k_p, \rho)$ .

1.  $(\mathbf{y}^{V4}, \mathbf{V}^{V4}, \mathbf{M}^{V4}) \leftarrow LCA(\mathbf{X}^{V4}, \mathbf{Z}^{V4}, \mathbf{V}^{V4}, \mathbf{M}^{V4}, \theta^{V4})$
2.  $(\mathbf{y}^{IT}, \mathbf{V}^{IT}, \mathbf{M}^{IT}) \leftarrow LCA(\mathbf{X}^{IT}, \mathbf{Z}^{IT}, \mathbf{V}^{IT}, \mathbf{M}^{IT}, \theta^{IT})$
3.  $(\mathbf{y}^P, \mathbf{V}^P, \mathbf{M}^P) \leftarrow LCA(\mathbf{X}^P, \mathbf{Z}^P, \mathbf{V}^P, \mathbf{M}^P, \theta^P)$
4.  $(\mathbf{y}^{TM}, \mathbf{V}^{TM}, \mathbf{0}) \leftarrow LCA(\mathbf{X}^{TM}, \mathbf{0}, \mathbf{V}^{TM}, \mathbf{0}, \theta^{TM})$
5.  $(\mathbf{y}^{LM}, \mathbf{V}^{LM}, \mathbf{0}) \leftarrow LCA(\mathbf{X}^{LM}, \mathbf{0}, \mathbf{V}^{LM}, \mathbf{0}, \theta^{LM})$

A few more items to note on training: (1) Each area’s output firing rates  $\mathbf{y}$  is sampled to become the next area’s bottom-up input  $\mathbf{X}$ , and the previous area’s top-down input  $\mathbf{Z}$ . (2) For V4, each top-down source from the What or Where path is weighted equally in setting  $\mathbf{Z}^{V4}$ . (3) We used a supervised training mechanism (“pulvinar”-based training [5]) to bias V4 to learn foreground patterns: we set its  $\mathbf{Z}$  based on the firing of the LM area — only neurons with receptive fields on the foreground would receive a top-down boost. (4) For PP (denoted as “P” above) and IT areas, we used  $3 \times 3$  neighborhood updating in the vein of self-organizing maps in order to spread representation throughout the layer. This was done for the first two epochs. (5) The above algorithm is a forward-biased algorithm, in the sense that it takes less time for information to travel from sensors to motors (one iteration) than from motors to V4 (two iterations). Therefore, weight-updating in V4 only occurred on iterations two and three for each image. (6) Parameters:  $\rho^{V4}$  was set to 0.75 (bottom-up activity contributed 75% of the input), and  $\rho^{IT} = \rho^P = 0.25$ . In training, all values of  $k$  were set to one. This ensured a sparse representation to develop.

Neurons further along the What pathway become more invariant to object position, while becoming specific to object type. The opposite is true for the Where pathway. Neurons earlier in each pathway represent both location and type in a mixed way. For more information on learning, see [5]. Tests in feedforward mode with a single foreground in the scene showed the recognition and orientation performance improved after epochs of learning (an epoch is an entire round of training all possible samples).

### C. Attention Selection Mechanisms

Through training, WWN-3 becomes sparsely and selectively wired for attention. Afterwards, manipulation of parameters allows information to flow in different ways. The changing of information flow direction is a key to its ability to perform different attention tasks. Specifically, it involves manipulating the  $\rho$  parameters (bottom-up vs. top-down within an area) and  $\gamma$  (percentage of top-down to V4 from IT vs. PP).

In WWN-3, we examined four different attention modes. Free-viewing mode is completely feedforward. It quickly generates a set of candidate hypotheses about the image based on its learned filters. But there can be no internal verification that the type and location that emerge at the motors match (binding problem). Free-viewing mode is necessary to

reduce complexity of an under-constrained search problem, but it cannot solve the problem itself. Another mode, top-down location-based binding acts as a spotlight. It constrains firing at LM to a winner location neuron and selects for TM appropriately through top-down bias from LM to V4 and bottom-up from V4 to TM. The top-down object-based attention mode allows the network to act in search mode. It constrains firing at TM to a winner neuron and selections for LM appropriately through top-down bias from TM to V4 and bottom-up from V4 to LM. The fourth mode involves a disengage from the currently attended location or both currently attended location and type.

Multiple objects introduce the binding problem for WWN-3. Another problem is the hallucination problem, which could occur when the image changes (containing different foregrounds). The following rules for attention allow WWN-3 to deal with multiple objects and image changes. Whenever a motor area's top neuron changes suddenly, switch to the corresponding top-down mode if it is a strong response, or switch to free-viewing mode if it is a weak response. Whenever both motor areas' top neurons change suddenly to strong responses, go into the top-down location-based mode. If the network focuses on a single location and type for too long, disengage from the current location. The following parameter settings specify how this was done in WWN-3. In all modes we set  $k_b^{V4} = 8$  and  $k_p^{V4} = 4$ .

1). **Bottom-up free-viewing:** In forward mode, there is no top-down since the network does not yet have any useful internal information to constrain its search. WWN-3 sets ( $\rho^{V4} = \rho^P = \rho^{IT} = 1$ ).

2). **Top-down searching (object-based):** In this mode, information must travel from the TM down to V4 and back up to the LM. If  $\gamma = 1$ , all top-down influence to V4 is from the what path. Set  $\gamma = 1$ ,  $\rho^{IT} = 1$ ,  $\rho^P = 0$ , and  $\rho^{V4} = 0.5$ .  $k_t^{V4}$  is large (50% of neurons) to allow wide-spread top-down bias. For IT and PP,  $k_b$  is set small (up to 10% of neurons), for sparse coding, while  $k_t$  and  $k_p$  must be large enough to contain all neurons that may carry a bias. For example  $k_t^{IT} = n/c$  where  $c = 5$  (classes) and  $n = 400$  neurons if we assume an equal number of neurons per class.

3). **Top-down location-based:** In this mode, information must travel from LM to V4 and back up to TM. Thus, the network sets  $\gamma = 0$ ,  $\rho^{IT} = 0$ ,  $\rho^P = 1$ , and  $\rho^{V4} = 0.5$ . The  $k$  parameters are set the same as for object-based attention.

4). **Location and type attention shift:** In this mode, the network must disengage from its current attended foreground to try to attend to another foreground. To do so, the current motor neurons that are firing are inhibited (for the location motor, an inhibition mask of  $15 \times 15$  width was used) while all other neurons are slightly excited until the information can reach V4 (two iterations). The information flows top-down from motors to V4 ( $\rho^{IT} = \rho^{PP} = 1$ ). The top-down activation parameters  $k_t$  must be set much larger since all neurons *except* the current class should be boosted. After two iterations, it re-enters free-viewing mode.

Is it plausible to have multiple different attention modes?

Computationally, attention and recognition is a “chicken-egg” problem, since, for attention, it seems recognition must be done, and for recognition, it seems attention (segmentation) must be done. The brain might deal with this problem by using complementary pathways and use of different internal modes to enforce internal validity and synchronization. For example, Treisman famously showed [3] that there is an initial parallel search followed by a serial search (spotlight), which binds features into object representations at each location. It seems that after feedforward activity, a top-down location bias emerges to focus on a certain spot.

#### IV. EXPERIMENTS AND RESULTS

Each input sample to WWN-3 contains one or more foregrounds superimposed over a natural background. The background patches were  $38 \times 38$  in size, and selected from 13 natural images<sup>3</sup>. The foregrounds were selected from the MSU 25-Objects dataset [21] of objects rotating in depth. The foregrounds were normalized to  $19 \times 19$  size square images, but were placed in the background so that the gray square contour was eliminated (masked). Three training views and two testing views were selected per each of the five classes. The classes and within-class variations of the foregrounds can be seen in Fig. 2(b). Five input samples, with a single foreground placed over different backgrounds, can be seen in Fig. 2(a).

##### A. WWN-3 Learns

The training set consisted of composite foreground and background images, with one foreground per image. We trained every possible training foreground at every possible location (pixel-specific), for each epoch, and we trained over many epochs. So, there are 5 classes  $\times$  3 training instances  $\times$   $20 \times 20$  locations = 6000 different training foreground configurations. After every epoch, we tested every possible testing foreground at every possible location. There are  $5 \times 2 \times 20 \times 20 = 4000$  different testing foreground configurations.

To simulate a shortage of neuronal resource relative to the input variability, we used a small network, five classes of objects, with images of a single size, and many different natural backgrounds. Both the training and testing sets used the same 5 object classes, but different background images. As there are only 3 V4 neurons at each location but 15 training object views, the WWN is  $4/5 = 80\%$  short of resource to memorize all the foreground objects. Each V4 neuron must deal with various misalignment between an object and its receptive field, simulating a more realistic resource situation. Location was tested in all  $20 \times 20 = 400$  locations.

To see how a network does as it learns, we tested a single foreground in free-viewing mode throughout the learning time. As reported in Fig. 2(c), a network gave respectable performance after only the first round (epoch) of practice.

<sup>3</sup>Available from <http://www.cis.hut.fi/projects/ica/imageica/>

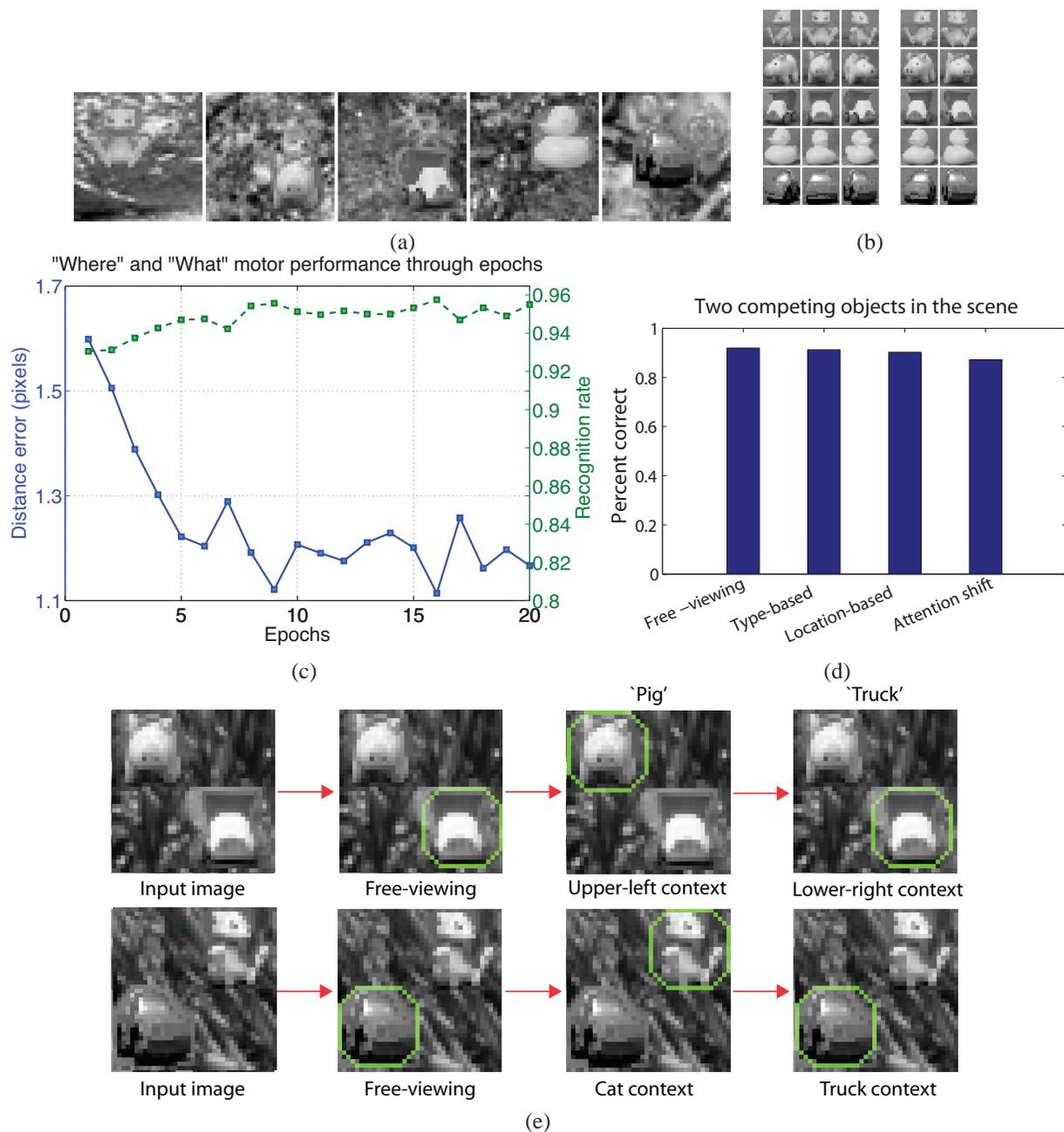


Fig. 2. WWN-3 learning and performance for the joint attention-recognition problem for scenes with two objects. (a) Sample image inputs. (b) The foregrounds used in the experiment. There are three training (left) and two testing (right) foregrounds from each of the five classes of toys: “cat”, “pig”, “dump-truck”, “duck”, and “car”. (c) Free-viewing mode performance when the input contains one object, during the training phase. (d) Performance after training, when the input contains two learned objects. (e) A few examples of operation over different modes by a trained WWN-3. “Context” means top-down goal is imposed. A green octagon indicates the location and type action outputs. The octagon is the default receptive field.

After 5 epochs of practice, this network reached an average location error around 1.2 pixels and a correct disjoint classification rate over 95%.

### B. Two Object Scenes

After training had progressed enough so the bottom-up performance with a single foreground was sufficient, we wished to investigate WWN-3's ability with two objects, and in top-down attention. We tested the above trained WWN-3 with two competing objects in each image, placed at four possible quadrants to avoid overlapping. We placed two different foregrounds in two of the four corners. There were 5 classes  $\times$  4 corners  $\times$  3 other corners (for the second foreground) = 60 combinations. WWN-3 first operated starting in free-viewing mode (no imposed motors), until it converged. If the type and location (within 5 pixels) matched one of the foregrounds, it was considered a success. Next, the type of the foreground that wasn't located was imposed at TM as an external goal, and WWN-3 operated in top-down searching mode to locate the other foreground. Next, WWN-3 would shift its attention back to the first object. Finally, the location of the foreground that wasn't identified in the first phase was imposed at LM as an external goal, and WWN-3 operated in top-down location-based mode to find the other foreground.

As shown in Fig. 2, the success rates for this network were 95% for the free-viewing test. The success rates were 96% when given type context to predict location and 90% when given location context to predict type. It successfully attended to the other object via an attention shift 83% of the time.

## V. CONCLUSIONS AND FUTURE WORK

The work here demonstrated that the WWN-3 can deal with multiple objects by allowing a motor area to provide top-down context bias, as a goal or internal preference. Experiments using the disjoint foreground object subimages with general object contours reached an encouraging performance level by a limited size WWN-3.

The future of this work involves extending WWNs via a larger ventral pathway with more areas. The current networks have "early" receptive fields that are quite large. Other object recognition systems have shown the effectiveness of a local-to-global processing (as in Convolutional Networks [22] or HMAX [23]) in feedforward operation mode. Future work will also involve border selectivity by implementing neurons that develop dynamic shaped receptive fields through a synaptic maintenance (neuromodulation) method. It would also be interesting to look into a motion pathway for sequence or trajectory learning in vision using WWNs. The associative information-filling-in effects of recurrent excitation will be examined. Finally, networks should be embedded into active agents, which will interact with and learn about the world.

## ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their detailed and informative feedback, which was very much appreciated. We could not include some suggestions due to space-limitation, unfortunately.

## REFERENCES

- [1] J. Weng and M. Luciw. Dually-optimal neuronal layers: Lobe component analysis. *IEEE Transactions on Autonomous Mental Development*, 1(1), 2009.
- [2] J. Bullier. Hierarchies of cortical areas. In J.H. Kaas and C.E. Collins, editors, *The Primate Visual System*, pages 181–204. CRC Press, New York, 2004.
- [3] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.
- [4] Z. Ji, J. Weng, and D. Prokhorov. Where-what network 1: "where" and "what" assist each other through top-down connections. In *Proc 7th Int'l Conf on Development and Learning*, Monterey, CA, August 9-12 2008.
- [5] Z. Ji and J. Weng. Where what network-2: A biologically inspired neural network for concurrent visual attention and recognition. In *IEEE World Congress on Computational Intelligence*, Spain, 2010.
- [6] L. Itti and C. Koch. Computational modelling of visual attention. *Nat. Rev. Neurosci.*, 2:194–203, 2001.
- [7] G. Backer, B. Mertsching, and M. Bollmann. Data and model-driven gaze control for an active-vision system. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(12):1415–1429, 2001.
- [8] M.C. Mozer, M.H. Wilder, and D. Baldwin. A Unified Theory of Exogenous and Endogenous Attentional Control. *Department of Computer Science and Institute of Cognitive Science University of Colorado, Boulder, CO 80309*, 430, 2007.
- [9] JG Taylor. CODAM: A neural network model of consciousness. *Neural Networks*, 20(9):983–992, 2007.
- [10] BA Olshausen, CH Anderson, and DC Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700, 1993.
- [11] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545, 1995.
- [12] G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 40:2845 – 2859, 2004.
- [13] D. George and J. Hawkins. Towards a mathematical theory of cortical micro-circuits. *PLoS computational biology*, 5(10):e1000532, 2009.
- [14] A. Fazl, S. Grossberg, and E. Mingolla. View-invariant object category learning, recognition, and search: How spatial and object attention are coordinated using surface-based attentional shrouds. *Cognitive Psychology*, 58(1):1–48, 2009.
- [15] D.E. Broadbent and D.E. Broadbent. Perception and communication. 1958.
- [16] A. Treisman. Monitoring and storage of irrelevant messages in selective attention1. *Journal of Verbal Learning and Verbal Behavior*, 3(6):449–459, 1964.
- [17] D.G. Mackay. Aspects of the theory of comprehension, memory and attention. *The Quarterly Journal of Experimental Psychology*, 25(1):22–40, 1973.
- [18] E. M. Callaway. Local circuits in primary visual cortex of the macaque monkey. *Annu. Rev. Neurosci.*, 21:47–74, 1998.
- [19] M. Solgi and J. Weng. Developmental Stereo: Emergence of Disparity Preference in Models of the Visual Cortex. *IEEE Trans. on Autonomous Mental Development*, 1(4):238–252, 2009.
- [20] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, 3rd edition, 2001.
- [21] M. D. Luciw and J. Weng. Topographic class grouping with applications to 3D object recognition. In *Proc. International Joint Conference on Neural Networks*, Hong Kong, June 1-6 2008.
- [22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [23] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29:411–426, 2007.