Modulatory feedback explain object segmentation by attention

Paolo Papale[1,‡], Jonathan R. Williford[1,‡], Stijn Balk[1] and Pieter R. Roelfsema[1,2,3,4,*]

[1] Department of Vision & Cognition, Netherlands Institute for Neuroscience (KNAW), 1105 BA Amsterdam, Netherlands`

[2] Department of Integrative Neurophysiology, VU University, De Boelelaan 1085, 1081 HV Amsterdam, Netherlands.

[3] Department of Psychiatry, Academic Medical Centre, Postbus 22660, 1100 DD Amsterdam, Netherlands.

[4] Laboratory of Visual Brain Therapy, Sorbonne Université, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique, Institut de la Vision, Paris F-75012, France

[‡] shared first author contribution

*Correspondence should be addressed to P.R.R. (p.roelfsema@nin.knaw.nl)

Studies in neuroscience inspired progress in the design of artificial neural networks (ANNs) and vice versa ANNs have also started to provide new insights into the functioning of brain circuits. So far, the focus has been on how ANNs can help to explain the tuning of neurons at various stages of the visual cortical hierarchy. However, the role of modulatory feedback connections, which play a role in attention and perceptual organization, has not been tested yet. Here, we present a biologically plausible neural network that performs scene segmentation and can shift attention using modulatory feedback connections from higher to lower brain areas. The model replicates several neurophysiological signatures of recurrent processing. Specifically, figural regions elicit more neuronal activity in model units than the background. The modulation of neuronal activity by figure and ground occurs at a delay after the first feedforward response, because it depends on a loop through the higher model areas. Importantly, the enhancement of the figural response is controlled by object-based attention, which stays focused on the figural regions and does not spill over to the adjacent background, as observed in the visual cortex. Our results indicate how progress in deep learning can be used to garner insight into the recurrent cortical processing for scene segmentation and object-based attention.

Introduction

Objects are the building blocks of our perception: we perceive a visual scene as a layout of objects in the physical space (O'Craven et al., 1999). We tend to focus on one object at a time when scanning a visual scene with our eyes. This process, called object attention, is essential to organize our visual information and to interact with the external world (Behrmann et al., 1998; Duncan, 1984; Kramer et al., 1997). Object attention is indeed the mechanism through which our brain identifies the portions of the scene that belong to an object and segment them against the background, and process each object with the higher spatial detail that is needed to e.g., grasp and manipulate it (Roelfsema, 2006; Roelfsema & Houtkamp, 2011).

Although they appear entirely effortless to us, object identification and attention in the primate visual system rely on intensive computations in a hierarchy of cortical regions. The retinal input reaches first the primary visual cortex (V1) with neurons that have small receptive fields (RFs) and are tuned to simple features. Information is propagated through feedforward connections to higher-order regions in the occipital cortex (areas V2 and V4) and then to inferotemporal (IT) and frontal (FC) cortices (van Essen et al., 1992). In higher areas the neurons become selective to more complex features (e.g. object parts) and have larger receptive fields.

However, feedforward connections are only half of the story (Kar et al., 2019). After reaching the higher visual areas, visual information is also carried back to the lower areas through feedback connections (Felleman & van Essen, 1991) (Fig. 1A). Whereas feedforward connections drive the early response of neurons and shape neural selectivity, feedback connections modulate the level of sustained neural activity. Feedback connections enhance the neuronal activity elicited by relevant figural regions over the activity elicited by the background, a process called figure-ground modulation (Poort et al., 2016). Furthermore, feedback connections mediate the influence of top-down attention that amplify neuronal activity elicited by task-relevant objects (Roelfsema, 2006; Spitzer et al., 1988; Vecera & Farah, 1997).

At first sight, many of these neurophysiological observations may seem disconnected, because previous studies focused on only one or a few of these feedback influences (Carandini et al., 2005). We therefore sought to integrate these findings into a coherent framework, capitalizing on the recent advances in the development of artificial neural networks (ANNs) (Serre, 2019).

State of the art ANNs can achieve human level performance on many complex visual tasks, ranging from object classification to segmentation (Serre, 2019). Similar to the brain, an efficient neural architecture to segment objects from the ground in ANNs rely on two different paths: 1) a feedforward direction, comprising a hierarchy of layers (i.e. the homologous of brain regions) whose units (i.e., the homologous of visual neurons) have increasingly larger RFs and 2) a second opposite direction, mirroring the feedforward stream and marginally inspired by top-down feedback connections in the brain (Ullman, 1995), where the hierarchy is reversed and units have decreasing RFs along the path

(Hong et al., 2015). These networks have shown a high degree of similarity with brain activations (Guclu & van Gerven, 2015; Yamins et al., 2014). However, their biologically plausibility is limited: a key difference with the brain, common to all networks, is that feedback are always either totally independent from feedforward information (e.g. (Hong et al., 2015)) or multiplicatory (e.g. (Lei et al., 2021)). Feedback in these models would have a stronger impact on the activity of units than what observed in real neurons, where feedback are weaker, merely modulatory, and subtractive rather than multiplicatory (Poort et al., 2012, 2016; Roelfsema et al., 2002; Self et al., 2012).

Thus, it remains unclear whether modulatory feedback are enough to explain object segmentation by attention or not. In the latter case, our understanding of the role of feedback in the brain would need to be revised. To answer this question, we built a biologically inspired ANN and examined if it can explain the interaction between figure-ground segregation and selective attention that is observed in the visual cortex of monkeys (Poort et al., 2012) based solely on modulatory feedback. Our network comprised several areas modeled after the brain regions involved in object processing – denoted here with an 'm' suffix – from V1m to FCm (Fig. 1B). We included both a feedforward path, driving the activity in successively higher areas for object identification, and a feedback pathway that modulated neuronal activity (Fig. 1C) aiming to highlight the figural image regions when they were attended.

The model produced an accurate segmentation of attended objects by increasing the firing rate of image elements in V1m that represent the relevant figure at a high spatial resolution, whereas the activity elicited by nearby background elements was suppressed, just as is observed in the visual cortex of monkeys (Poort et al., 2012; Self et al., 2019). Moreover, our model independently evolved several tuning signatures that mimicked key neurophysiological observations (e.g. orientation tuning and surround suppression) and offered us the chance to gain novel insights on the properties of visual neurons.

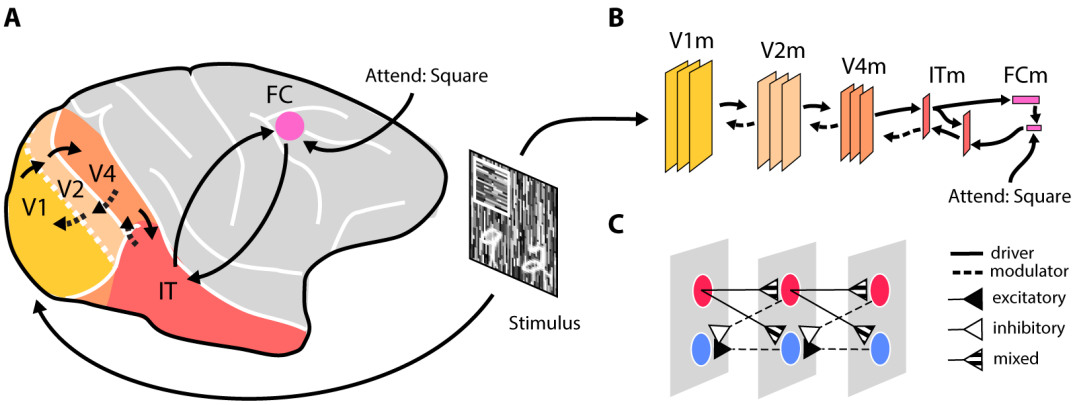

Figure 1: A neural network to model the primate visual cortex. **(A)** Hierarchically organized visual cortical areas that perform increasingly complex transformations of the visual input. V1 registers the information coming from the retina. Downstream regions (V2, V4) further processes information and object-selective neurons in IT identify the elements in a scene and send this information to the frontal cortex (FC) that represents behavioral goals. Frontal cortex sends attentional signal through feedback connections that modulate the activity of earlier brain regions, resulting in the segmentation of task-relevant objects at a high spatial resolution. **(B)** In the model, the stimulus enters in area V1 (V1m) and feedforward connections (solid arrows) activate downstream regions that classify the objects. The network can select one of the object classes and use feedback connections for class-specific object segmentation in V1m. **(C)** The pattern of connections between V1m, V2m and V4m. The network contains units that participate in propagating information in the feedforward direction (red) and other units (blue) that propagate information in the feedback direction. The signal flowing in the feedback direction is based on a comparison (subtraction) of the activity of feedforward and feedforward units.

Materials and Methods

Model Architecture

Figure 1B gives an overview of the model. The architecture was inspired by the decoupled deep neural network (Hong et al., 2015) that performs semantic segmentation, i.e. the task of assigning an object class to every pixel of the input. The architecture introduced by Hong and colleagues (Hong et al., 2015) achieved this by first mapping pixels into object classes in a feedforward network. It then selected one of the classes at a time for a given image, and retrieved the pixels belonging to that class in an additional feedforward network that mapped classes back onto pixels. Hence, this ANN carried out an object-based attention task in a feedforward network. Our model, instead, used feedforward and modulatory feedback connections and can therefore give insight into the different phases of activity of neurons in the visual cortex.

Feedforward pathway

The interactions between the feedforward and feedback pathway ensured that the feedback from the selected shape in FCm modulated the appropriate units at the low-levels. A 12x1x1 layer encoded the object that should be attended (the "attend" label in Fig. 1B). This was multiplied (elementwise) with the classification vector from the feedforward pathway. A series of fully connected layers were then used in the feedback pathway that mirrored the feedforward pathway, followed by a "deconvolution" (transposed convolution) layer that produced output with dimensions of 100x7x7. This output was then concatenated with the corresponding 100x7x7 layer of the feedforward to produce a 200x7x7 tensor, which was passed to a 1x1 convolution reducing the tensor back to 100x7x7.

Each unit in the feedforward pathway had a corresponding (deconvolutional) unit in the feedback pathway in V1m, V2m and V4m. Feedback connections did not drive units in these areas but they could only modulate the feedforward response, in accordance with the effects of attention in the brain (e.g. (Self et al., 2012)). In our model the modulatory feedback interaction was implemented as:

$$b_i = max(0, a_i(z_{i+1} + \beta_i))$$

where a_i is the firing rate (non-negative) of a feedforward neuron in layer i, z_{i+1} is the feedback received from a higher layer (via a strided convolution), β_i is a learned parameter, and b_i is the activity of a unit of the feedback pathway. The "pure" feedback signal, propagated to the next layer, is the difference between the activity of the feedforward and feedback unit:

$$c_i = b_i - a_i$$

Stimulus sets

Based on the MNIST dataset (LECUN & Y., n.d.), we created an augmented stimulus set *SegMNIST* (Fig.2). The new stimulus set makes the MNIST set more challenging and cast it as an object localization task, where the relevant shape can be outside the center of the image. The size of the image was increased to twice the size of the MNIST image (56x56 pixels

instead of 28x28) and MNIST digits were translated, scaled, and positioned on top of a texture. We also added squares and rectangles at different rotation angles. We applied two textures types, Gaussian white noise (spatially IID) and oriented line textures as those used in monkey experiments on texture segregation (Poort et al., 2012; Self et al., 2012).

Training of the feedforward and feedback pathways

The feedforward pathway was trained on classifying the objects in the textured and expanded SegMNIST stimuli using Adam (Kingma & Ba, 2015) with learning rate of 0.001, L2-regularization (weight decay = 0.0005), and as loss a sigmoid combined with binary cross-entropy (pytorch's *BCEWithLogitsLoss*) in 2,500 epochs (batch size = 1024). There were 12 classes: square, rectangle, and the 10 digits, and one to three objects in each image. The task of the network was to determine which object classes appeared in the stimuli. We trained the feedback pathway once the weights of the feedforward pathway were fixed. The feedback pathway was trained on object segmentation using Adam (Kingma & Ba, 2015) with learning rate of 0.001, L2-regularizations (weight decay = 0.0005), and as loss a sigmoid combined with binary cross-entropy (pytorch's *BCEWithLogitsLoss*) in 2,500 epochs (batch size = 256). The network was trained using one GPU.

Simulating the activity propagation in V1m

In order to evaluate the dependence of the attention modulation on the feedforward activity, we allowed the feedforward activity to propagate to different number of layers. In this way, we simulated the feedback modulation as the feedforward activity propagated along the network (Fig.3). We simulated a single feedforward pass and a single feedback pass and we did not model recurrence. We aimed to simulate a subject that is waiting for a stimulus to appear and expecting (i.e. attending) a specific shape. We started a trial in the simulation by first presenting the activity level when presenting a white noise stimulus to the network, with the attention state (the to-be-selected shape) already set. We then presented the stimulus which propagate t layers, where t corresponds to the number of simulated time steps. We repeated the same procedure with 1,000 different stimuli (attending to either figures or digits) and average the obtained responses. Time-courses showed in Figure **Errore. L'origine riferimento non è stata trovata.**B and **Errore. L'origine riferimento non è stata trovata.**D were smoothed with a 3rd order Savitzky–Golay filter.

RF position and tuning of network units

To examine the tuning network units to space, orientation, border-ownership and shape as well as surround suppression, we monitored their activity while varying the stimuli systematically, as would be done in electrophysiology experiments.

We computed the RF of each unit using a light-spot mapping (e.g. (Poort et al., 2016)). In brief, we recorded responses from a set of ON and OFF stimuli, where a single pixel was either white (ON) or black (OFF) against a gray background. We reconstructed the spatial RF of the neuron and determined the RF center, as the mean position of the most influential ON and OFF pixels.

We determined orientation tuning with full-contrast Gabor gratings with 8 different orientations (22.5 degrees apart), 3 wavelengths (2.5, 5 and 10 cycles per stimulus) and 8 phases (45 degrees apart). Following (Williford & von der Heydt, 2016), the orientation tuning index (OI), going from 0 (no tuning) to 1, was defined as following:

$$\mathrm{OI} = \frac{R_{pref} - R_{orth}}{R_{pref} + R_{orth}}$$

Where R_{pref} was the response to the preferred orientation, wavelength and phase, and R_{orth} was the response to the orthogonal orientation with preferred wavelength and phase.

To measure surround suppression, we recorded responses from a set of Gabor patches with increasing sizes (2 pixel steps). We centered the patch on the RF and used the preferred combination of orientation, wavelength and phase. The surround suppression index (SSI) varies between 0 (no suppression) and 1 and was computed as follows:

$$\mathrm{SSI} = \frac{R_{pref} - R_{full}}{R_{pref} + R_{full}}$$

Where R_{pref} was the response to the preferred size and R_{full} was the response to a grating that filled the entire image.

To measure border ownership tuning we placed the edge of a square on the center of the RF of each neuron. The squares had 64 possible orientations (5.625 degrees apart), three different sizes ($Preferred\ size, 2 \times Preferred\ size, 4 \times Preferred\ size$) and two possible sides (i.e., square on the right or on the left). As in (Zhou et al., 2000), the border-ownership index (BOI), going from 0 (no selectivity) to 1, was defined as follows:

$$\mathrm{BOI} = \frac{R_{same} - R_{diff}}{R_{same} + R_{diff}}$$

Where R_{same} is the response to the preferred stimulus, and R_{diff} was the response to the stimulus having the same properties of R_{same} (i.e., square size, orientation and contrast polarity) but was on the opposite side.

To determine shape selectivity, we used 2,000 stimuli with a single object. We averaged the activity across each shape and quantified shape-selectivity as the sparsity of the distribution. Following (Zoccolan et al., 2007), the shape index (SI), going from 0 (no selectivity) to 1, computed as following:

$$\mathrm{SI} = [1 - \frac{\sum({R_i}/{12})^2}{(\sum {R_i}^2/{12})}] \Big/ [1 - (\frac{1}{12})]$$

Where R_i is the response to i^{th} class and 12 is the number of classes.

Finally, we generated a visualization of each unit's tuning with an ANN feature visualization method (e.g. (Olah et al., 2017)). Contrast and luminosity in the images in the main figures have been slightly changed to enhance visualization, the original ones are visible in the Github page of the project (see below).

Code availability

Model and analyses were implemented using custom Python code and NumPy (van der Walt et al., 2011), SciPy (SciPy.org) and PyTorch (Paszke et al., 2019). The code to generate the stimuli is available at the SegMNIST Github repository: http://github.com/williford/seg-mnist-dataset. The code to train the model and reproduce the results is available at the Github repository of the project: [available upon paper acceptance].

A Google Colab of the project to explore the model results and all the individual units is available at: https://colab.research.google.com/drive/13O0-4uYq3l1tZupVn-NLP8C-WvakajCG?usp=sharing.

Results

Modulatory feedback connections achieve accurate segmentations of attended objects

We trained an ANN to recognize and segment the digits 0 through 9, squares and rectangles in display with a background of texture elements. First, we trained the feedforward path of the network to recognize these shapes. Then we trained the feedback pathway to attend and segment one of the shapes from the background. In this training phase, we used a task that was inspired by (Poort et al., 2012), who trained monkeys to either detect texture-based figures or to perform another task in which they could ignore the texture-defined figures.

A few examples of the stimuli and model segmentations are shown in Fig. 2. As evident, the stimuli (left column) were particularly challenging, since the edges of the objects were easy to mix with the line texture in the background. Despite the clutter in the stimuli, the feedforward path of the model could lead to a near-perfect prediction in FCm of the amount and the classes of the objects in the image (mean absolute error (cross validated) = 0.023 – ranging between 0 (all objects correctly classified in each image) and 12 (none of the objects correctly classified in each image)).

Also, the activity propagating backwards from FCm to modulate the response of the units in lower regions, could lead to highly accurate segmentations (left columns in Fig.

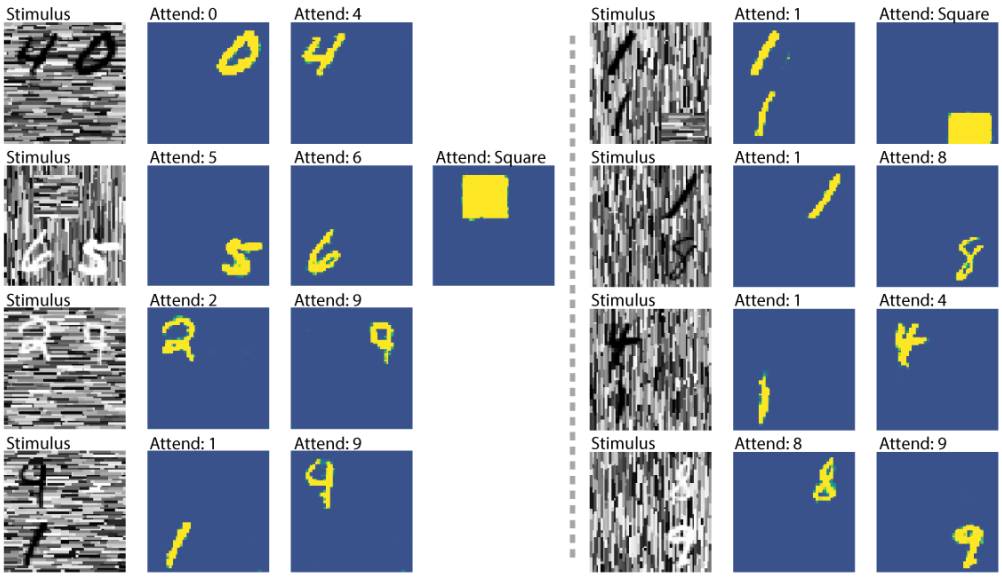

Figure 2: Modulatory feedback achieve highly accurate segmentations of attended objects in cluttered stimuli.
Examples of eight stimuli and the corresponding segmentations achieved by the model. Each row shows one example stimulus and the segmentation generated by the model when instructed to attend to a specific object class (blue: background, yellow: segmented object). The segmentations are accurate, demonstrating that modulatory feedback connections are sufficient to enable a precise interaction between attention and figure-ground segmentation.

2; yellow: regions labeled as the attended object; blue: regions labeled as background) of one attended object class in the image (mean pixel-wise absolute error (cross validated) = 0.0015 – ranging between 0 (all object pixels correctly segmented in each image) and 1 (none of the object pixels correctly segmented in each image). These results demonstrate that attentional modulatory feedback are a theoretically sufficient explanation of the behavior observed in monkey.

We next compared the sequence of activations in the model to the neurophysiological results of (Poort et al., 2012). That study revealed that the response of neurons in V1 depends on whether the RF fell on a texture-defined figure or background and also on its behavioral relevance, i.e. on attention. The task of the monkey was to attend either a square figure (visually identical to the stimuli fed to our network and shown in Fig.2) or to perform another task in which they would ignore the square (i.e. mentally tracking a curve far away from the square). Their main results are summarized in Fig. 3A, showing the time-course of the response of V1 neurons when their RF was on the square figure (red), which was much larger then the neurons' RF, on the ground (blue) or on the edge between figure and ground (black): on top, the V1 response while the monkey was attending the square figure ("attend: square") while on the bottom, the response while attending away, i.e. a curve far from the square figure ("attend: away"). Around 40ms after the presentation of the stimulus, feedforward activations drive the V1 neurons. During this early response phase, their activity depend on the visual information in the RF, and after a short additional delay the activity of neurons with a RF on the edge between figure and ground is enhanced over the activity of neurons with an RF on the background. During a later response phase, around 100ms after the stimulus onset, the activity by image elements that belong to the center of the figure start to be also enhanced over the activity elicited by background elements. Finally, approximately 200ms after onset, the influence of attention to the figure (Fig. 3A, top) kicks in. Behavioral relevance causes a further enhancement of the activity of V1 neurons with a RF on figure region, compared to when monkeys ignored the figure (Fig. 3A, bottom, note the distance between black and red traces).

We next examined the activity of V1m units of the feedback pathway in the trained network (Serre, 2019). We examined the influence of time, by propagating the activity in the feedforward and feedback directions by one connection at a time. V1m units showed similar dynamics as those observed in area V1 of monkeys. The model units exhibited a first, transient peak response that was elicited by the appearance of texture elements in the RF (Fig. 3B). The early response elicited by the edge (black in Fig. 3B), was higher than that elicited by the figure (red) and ground (blue) and the response elicited by the figure was also enhanced compared to the ground. At a later phase, attention to the textured square enhanced the activity evoked by the figure further (Fig. 3B, top), compared to when attention was directed to one of the digits (Fig. 3B, bottom). The influence of attention on neuronal activity in the model feedback pathway was stronger than in monkey V1. However, the order of the effects of figure-ground organization and attention was similar to that in monkey V1.

Also, not just the overall time-courses of response were similar, but also the unfolding of figure-ground modulation (i.e., the difference between activity in figure and ground) in time followed an analogous pattern. In Fig. 3C, D, as in the panels above, the effect is shown separately for the two different attention conditions (i.e. "square" or "away"), only this time the square stimulus in each trial was placed in different positions relative the RFs of either V1 neurons (Fig. 3C) or V1m units (Fig 3D). The difference between the two attention conditions is evident both in monkey V1 and in the model: while response to the borders is enhanced early after the stimulus onset, the enhancement of the figure occurs later, when top-down feedback kicks in.

Tuning of units in the feedforward and feedback paths of the network

Given the similarity between the dynamics of activity in V1m and monkey V1, we investigated the tuning of units, following the same procedures used when studying real

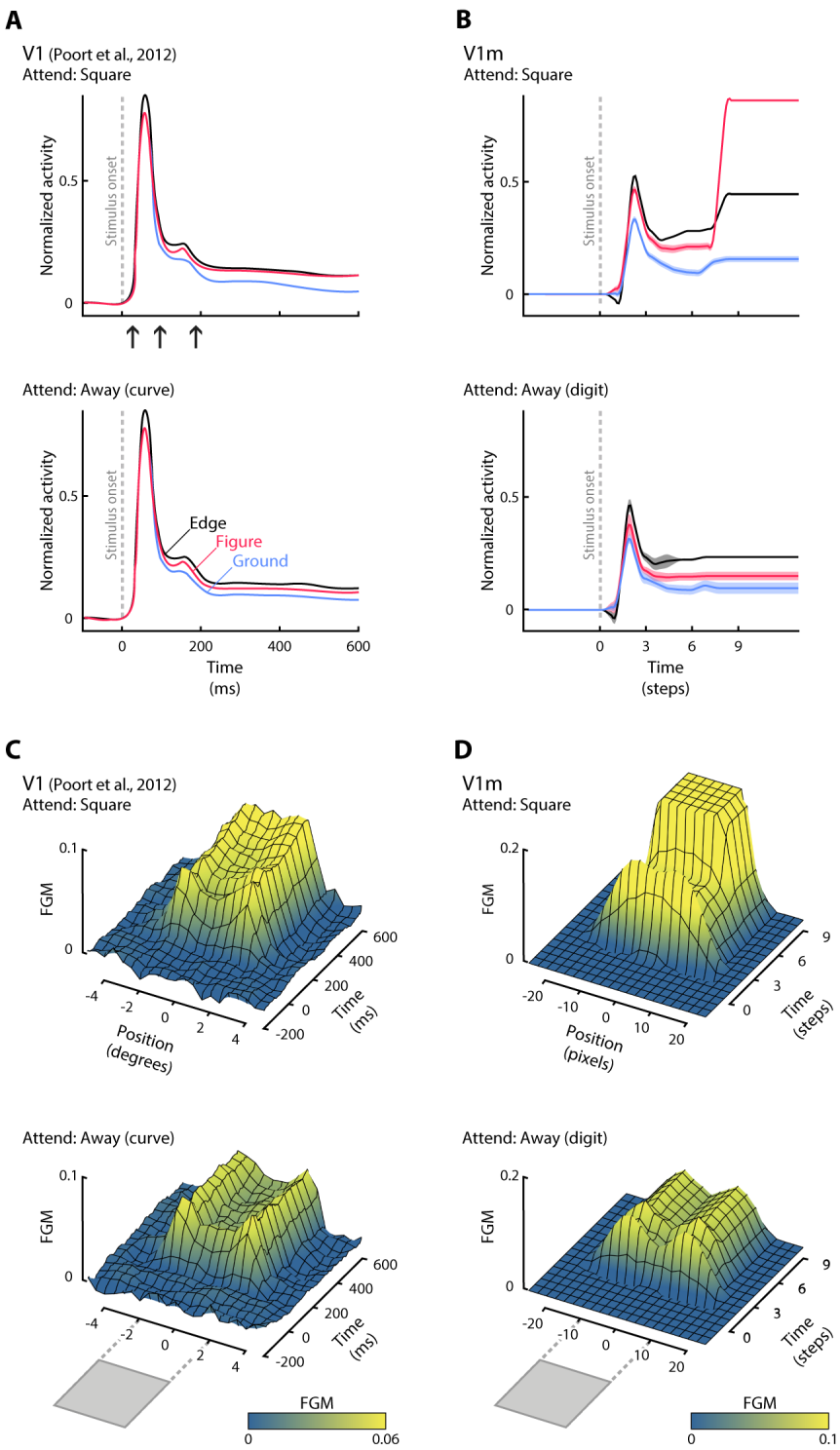

Figure 3: V1m evolves figure-ground modulation by selective attention as observed in monkey V1
(A) Activity of V1 neurons from the study of Poort et al, in which macaques were cued to either detect texture-based square figures (top) or to perform another task in which they did not attend the figures. V1 activity depended both on texture elements falling in RF (part of the figure, the ground or the edge between figure and ground) and attention. There were three different phases of activity: a first transient response (top, left arrow), a later phase in which figures elicited more activity than grounds (middle arrow) and an even later phase in which attention increased the figure response (right arrow). **(B)** Response of feedback units in V1m when attention is directed to the figure (top) or to one of the digits (bottom). The successive time steps were simulated by propagating the activity of the units one connection at a time. Dashed regions represent SEM (across 1,000 stimuli). **(C,D)** Space-time profile of figure-ground modulation (FGM: difference between figure and background response) for V1 neurons when attending to the figure (**C**, top) and when attending away (**C**, bottom) and for V1m neurons (**D**, top: figure: bottom: away).

neurons. For each of the units in the three lowest areas (V1m, V2m and V4m) of our model, we computed the degree of selectivity to many stimulus properties that are common in the primate visual system (Fig. 4A-B) and we also evolved an image (image insets in Fig. 4C, D) to show the stimulus preferred by each different unit (Bashivan et al., 2019; Papale et al., 2021; Walker et al., 2019). Indeed, neurons in the visual cortex are selective for the orientation of edges inside the RF (a schematic of typical responses is shown in Fig. 4A, left). This tuning relies to a large degree on the selectivity of feedforward connections to V1 (Chung & Ferster, 1998; Jin et al., 2011; Lien & Scanziani, 2013). For each unit, we also quantified the tuning (i.e. the orientation index (OI): see Methods) using stimuli with different orientations centered on the RF of the units, and then considering the relationship between the response to the preferred orientation and that to the orthogonal orientation. Most units in our model also exhibited tuning to the orientation of image elements (Fig. 4C, D, blue).

Another common property of visual neurons is surround suppression (Fig. 4B, red), i.e. the evidence that elements in the surround of the RF can suppress the activity of a neuron. This property is the result of the influence of the activity of both nearby neurons, in the same cortical region, and top-down feedback from neurons in higher regions (Angelucci et al., 2017; Bijanzadeh et al., 2018). Some of the units in our model showed surround suppression (Fig. 4C-E, red), that we quantified considering the relationship between the response to a stimulus of the preferred size, centered on the RF of each unit, and the response to a stimulus covering the whole image (surround suppression index: SSI).

A property of visual neurons that has been more directly linked to figure-ground perception and to the role of top-down feedback is the coding of border-ownership (Qiu et al., 2007; Williford & von der Heydt, 2016). In most natural

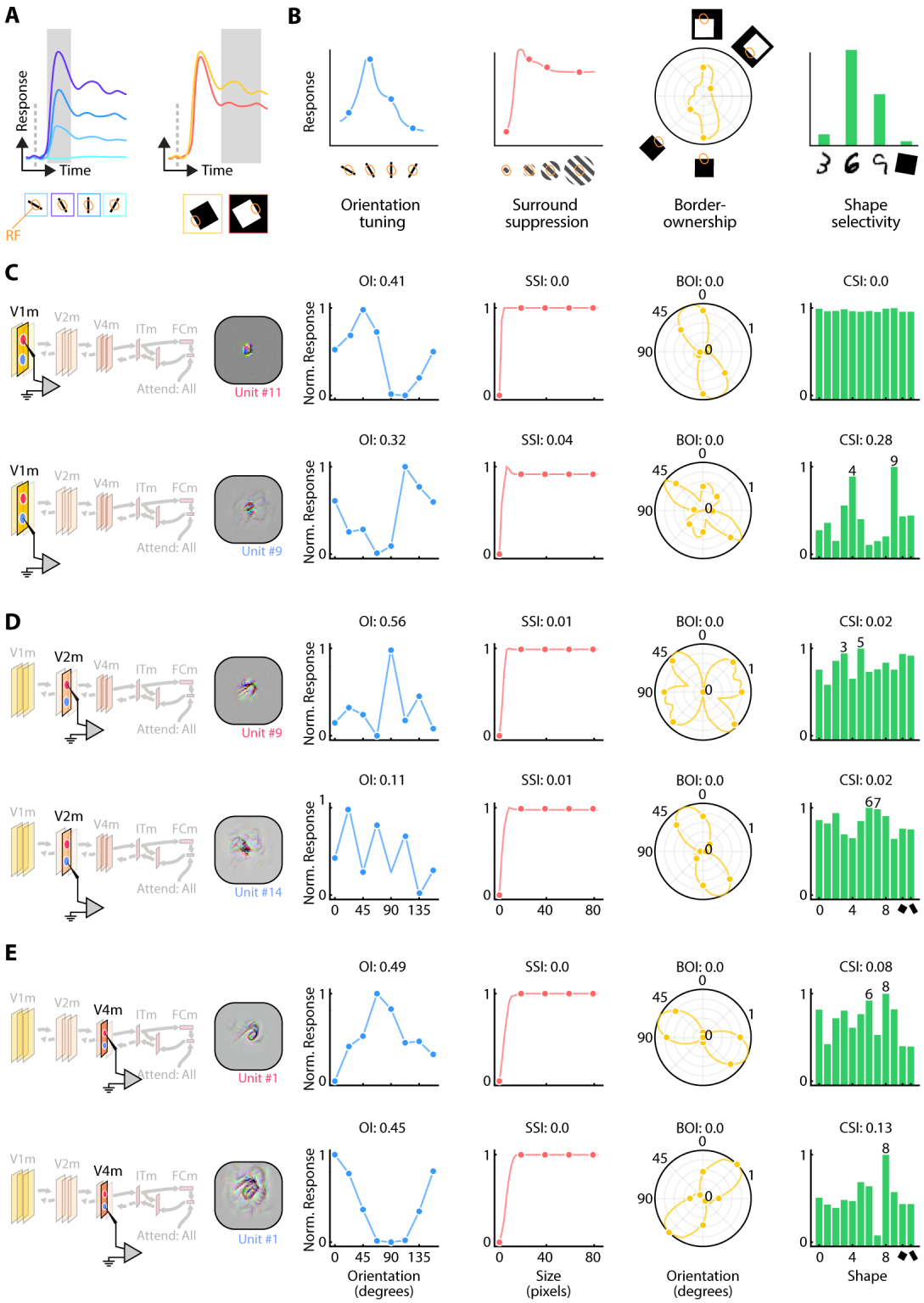

Figure 4: Artificial neurons show tuning properties similar to real neurons
(A) Schematic of typical responses of tuned selective neurons. Left: feedforward driven orientation tuning. Right: feedback driven border-ownership selectivity. **(B)** Schematics of the tuning of typical V1 neuron responses. From left to right: orientation tuning, surround suppression, border-ownership and tuning to more complex shapes. **(C)** Representative examples of the selectivity of feedforward (rows 1,3) and feedback model units (rows 2,4) in V1m and **(D)** in V2m. (OI: orientation-index: SSI: surround suppression index; BOI: border-ownership index; CSI: class-shape-selectivity index). **(E)** The impact of feedback modulation on neural selectivity is represented by a same example V4m neuron in either feedforward or feedback path.

images, boundaries belong to a figure that is seen as occluding a background. Border-ownership tuning means that neurons are selective for the figural side of boundary: a schematic of the response of a typical border-ownership neuron is shown in Fig. 4A (left) To quantify the degree of border-ownership selectivity (BOI), these responses can be averaged and plotted against the corresponding stimulus features, as shown in the schematic in Figure 4B (yellow). Despite its more direct role in segmentation, none of the units in our model showed border-ownership selectivity (Fig. 4C-E, yellow): a unite selective for border-ownership would show a different response when their RF is placed on the edge between figure and ground that is dependent on the side of the figure. However, the units in our model show an identical response when the figure is either on one side or the opposite.

Finally, we investigated the tuning to shape classes: a form of selectivity that is more common in neurons in higher regions, that tend to respond more strongly when a specific shape is present, irrespective of low-level differences such as orientation or view-point (Zoccolan et al., 2007). Some of the units in our model showed shape tuning (Fig. 4C-E, green), computed as the sparseness of the distribution of their average response to stimuli belonging to the 12 different shape classes (shape-class selectivity index: CSI).

Of note, all the tuning properties for all the units in our model can be explored in a Google Colab: https://colab.research.google.com/drive/13O0-4uYq3l1tZupVn-NLP8C-WvakajCG?usp=sharing.

The influence of feedback connections on the tuning of units at different network levels

Therefore, we tested more systematically the difference between the degree of selectivity in the feedforward and in the feedback path in each region. Figure 5 shows the population distributions of selectivity to orientation (blue), surround suppression (red) and object shape (green) in V1m to V4m in both feedforward and feedback paths. In V1m, surround suppression in the feedback pathway was stronger than in the feedforward pathway ($p < 0.01$, Bonferroni corrected, Wilcoxon signed rank test; Fig. 5A), whereas orientation tuning and shape selectivity did not differ (both ps > 0.05, Bonferroni corrected, Wilcoxon signed rank test), while the RF size increased ($p < 0.001$, Bonferroni corrected, Wilcoxon signed rank test). In V2m, surround suppression and shape-selectivity were stronger in the feedback pathway (both ps < 0.001, Bonferroni corrected, Wilcoxon signed rank test; Fig. 5B) with no difference in orientation tuning ($p > 0.05$, Bonferroni corrected, Wilcoxon signed rank test), while the RF size increased ($p < 0.001$, Bonferroni corrected, Wilcoxon signed rank test).. In V4m, shape-selectivity was stronger in the feedback pathway ($p < 0.001$, Bonferroni corrected, Wilcoxon signed rank test; Fig. 5C), orientation tuning was of similar strength (both ps > 0.05, Bonferroni corrected, Wilcoxon signed rank test), surround suppression was weak in both paths and no significant difference in the RF between the two paths was observed ($p > 0.05$, Bonferroni corrected, Wilcoxon signed rank test; note that the RF size across areas was different by design)..

In all regions, however, we found the modulation to determine a positive shift in learned weights (all ps < 0.001, Bonferroni corrected, Wilcoxon signed rank test). Finally, we found no significant correlation between different tuning properties neither in feedforward nor in feedback areas (Pearson's correlation: all ps > 0.05, Bonferroni corrected, t-test), meaning that we observed no relationship between, for instance, being selective for a particular object shape and the amount of surround suppression.

Discussion

In this work, we built a biologically inspired ANN implementing a model of the primate object processing pathway, composed of a driving feedforward path and a modulatory feedback path. We demonstrated that this simple but biologically plausible architecture can integrate selective attention to solve with high accuracy a complex object segmentation task on cluttered stimuli, relying solely on modulatory feedback.

We also compared the activity of model units to the responses of neurons in the visual cortex of monkeys, and we found that it replicated many key neurophysiological findings. Interestingly, the model was able to reproduce the spatial profile of the segmentation result, with a strong enhancement of the representation of image elements at all figural locations compared to background, even though the image elements of figure and ground were, on average, the same.

The time-course of activity of model units in V1m resembled the time-course of response of V1 neurons. Activity started with a transient response, followed by figure-ground modulation and the influence of attention was expressed at an even later time-point, because feedforward processing had to reach FCm, which then had to reach back to V1m. The delayed attentional influence is in accordance with neurophysiological evidence about comparable top-down influence (e.g., (Bichot et al., 2015, 2019; Moore & Fallah, 2004; van Vugt et al., 2018)). The attentional influence in V1m was stronger than in monkey V1. A likely reason for this difference is that our model was only trained on a binary segmentation task, whereas V1 of animals contributes to many more tasks (Roelfsema & de Lange, 2016).

Interestingly, many of the properties of neurons in the visual cortex were shared by the model, including orientation tuning, surround suppression and shape-selectivity. Furthermore, there were interesting differences between tuning in the feedforward and feedback pathways. For example, the degree of orientation tuning was similar in the two pathways, but RFs were larger in the feedback pathway, given the contribution of units in higher areas with larger RFs. It was also interesting to observe that shape selectivity was higher in the feedback pathway in V4m and V2m, indicating that the identity of objects was carried back from higher regions (ITm and FCm) through feedback connections. These results provide insight into why object

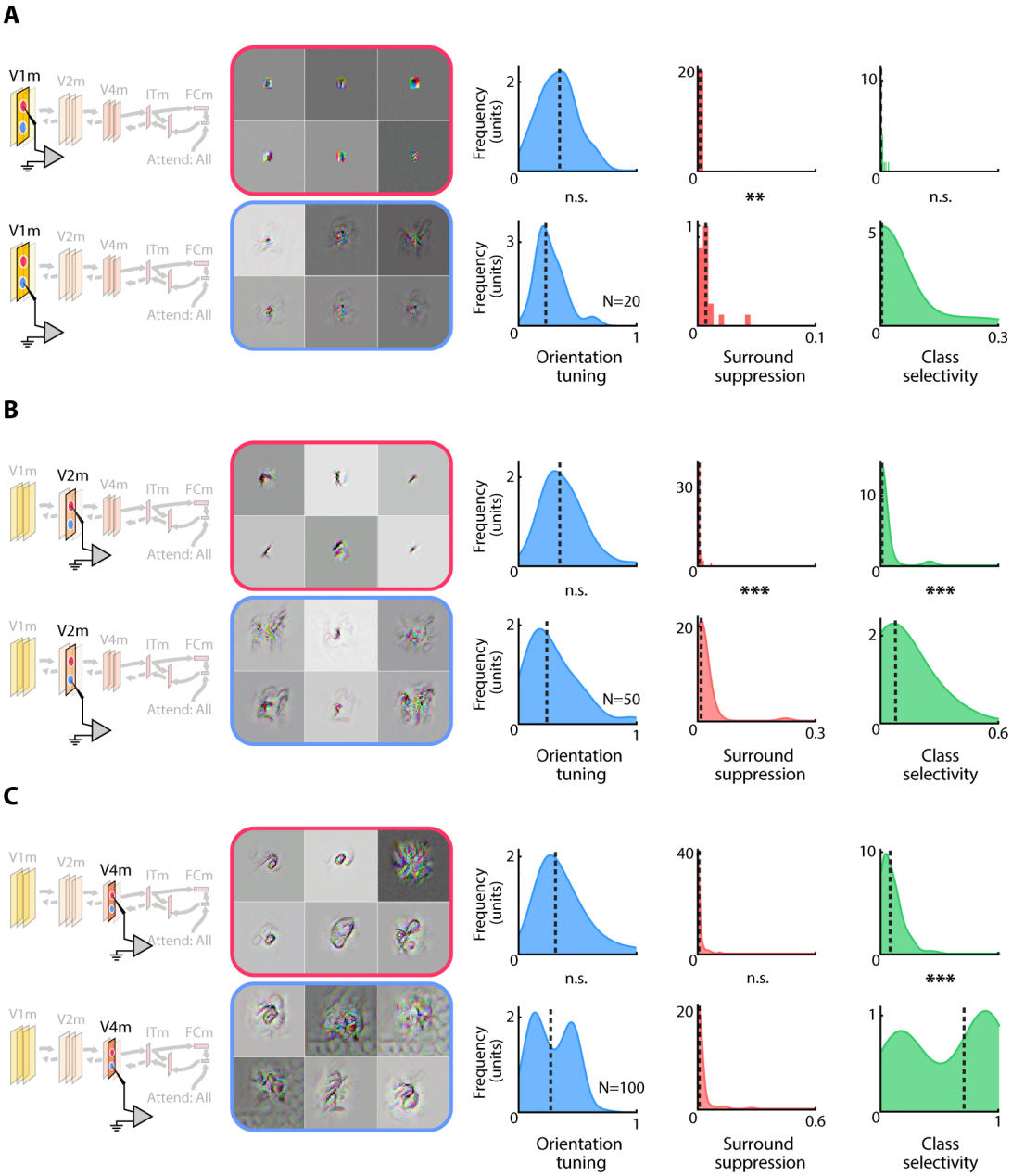

Figure 5: Modulatory feedback reshape feedforward tuning
The impact of feedback modulation on neural selectivity is represented by a same example V4m neuron in either feedforward or feedback path. Tuning of six example units and distribution of orientation tuning (blue), surround suppression (red) and shape-selectivity (green) across all units in V1m **(A)**, V2m **(B)** and V4m **(C)**. In each panel, units of the feedforward are shown in the top row and units of the feedback pathway in the bottom row. We assessed significance of differences between the feedforward and feedback pathways with Wilcoxon signed rank tests (Bonferroni corrected). n.s., $p > 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

class can be decoded from the activity of early visual neurons in the primate brain (e.g., (Muckli et al., 2015; Papale et al., 2019, 2020, 2021; Williams et al., 2008)).

Surround suppression was strongest in the feedback pathway although this property was less common than in V1 of animals (Jones et al., 2001). Again, there are multiple possible reasons for this discrepancy. For example, the model did not include horizontal connections so that surround suppression was mediated by feedback only, whereas horizontal connections in the visual cortex of animals are thought to contribute to surround suppression (Adesnik et al., 2012).

A surprising result in our analyses was the lack of border-ownership selective cells in our model. A previous study used cloned Markov random field model in a similar setting as ours (Dedieu et al., n.d.), parsing MNIST digits on a background with distracting image elements. This model using a number of computational mechanisms, including "cliques" and "query training", that do not translate as easily into neurobiological mechanisms as the model units and training curriculum used here. Interestingly, units in (Dedieu et al., n.d.) did exhibit border-ownership. Although we did not carry out simulations to explore the differences between models in detail, an interesting possibility is that

this previous study included horizontal connections and was able to disregard spurious contours from shapes that are caused by the occlusion of overlying shapes. Hence, it is reasonable to believe that border-ownership might be crucial to recognize occluded objects, a condition that we did not explore .

As a matter of fact, many complex visual tasks have been solved with ANNs that were only barely inspired by brain mechanisms. Moreover, recent work has showed that some form of object segmentation can be retrieved from internal representations in very deep feedforward models (Seijdel et al., 2020). Thus, while our model evolved brain-like mechanisms, it is possible that similar mechanisms can be observed with different, not brain inspired architectures. However, mounting evidence show that brain-like mechanisms can improve the performance of ANNs in several ways. Incorporating feedback not only leads to higher similarity with brain representations (Kietzmann et al., 2019; Kubilius et al., 2018) but increases classification accuracy and noise resistance (Jarvers & Neumann, 2019; Yan et al., 2019), while integrating also horizontal connections is critical to solve complex perceptual grouping tasks (Kim et al., 2019; Linsley et al., 2018). This evidence shows that biological plausibility can indeed lead to better artificial models, calling for an increased collaboration between the neuroscience and artificial intelligence (Doerig et al., 2022).

Previous studies implemented similar, but simpler, models by assembling hand-crafted functions derived from neurophysiological evidence, leading to a limited set of applications (Poort et al., 2012; Roelfsema et al., 2002). Our model, instead, presents a remarkable advancement with respect to these previous models, as it solved a much harder segmentation task by learning and developing internal representations similar to what observed in the primate brain, without explicitly modeling them. A more similar work to ours is represented by the model of (Lei et al., 2021), which is based on a similar architecture and that can also solve complex segmentations. Our work, however, goes one step further in testing the role of modulatory feedback, by implementing a biologically plausible form of top-down feedback and it also provides a direct comparison of the effect of these feedback on the tuning of units, thus providing novel falsifiable findings about the visual system that can be tested in future neurophysiological experiments.

Overall, we demonstrated that modulatory feedback can produce accurate object segmentations of attended shapes.

Future work, using ecological stimuli, will further test if the same architecture can also capture known behavioral (Fowlkes et al., 2007; Neri, 2014, 2017; Zuiderbaan et al., 2017) and neural (Hesse & Tsao, 2016; Papale et al., 2018; Williford & von der Heydt, 2016) signatures of object segmentation in natural vision, and further versions of our model will also improve biologically-plausibility by testing for the role of horizontal modulatory connections and using brain-like learning rules (Pozzi et al., 2020)

References

Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z. J., & Scanziani, M. (2012). A neural circuit for spatial summation in visual cortex. *Nature*, *490*(7419), 226–230. https://doi.org/10.1038/nature11526

Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., & Bressloff, P. C. (2017). Circuits and Mechanisms for Surround Modulation in Visual Cortex. *Annu Rev Neurosci*, *40*, 425–451. https://doi.org/10.1146/annurev-neuro-072116-031418

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

Behrmann, M., Zemel, R. S., & Mozer, M. C. (1998). Object-based attention and occlusion: Evidence from normal participants and a computational model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1011–1036. https://doi.org/10.1037//0096-1523.24.4.1011

Bichot, N. P., Heard, M. T., DeGennaro, E. M., & Desimone, R. (2015). A Source for Feature-Based Attention in the Prefrontal Cortex. *Neuron*, *88*(4), 832–844. https://doi.org/10.1016/j.neuron.2015.10.001

Bichot, N. P., Xu, R., Ghadooshahy, A., Williams, M. L., & Desimone, R. (2019). The role of prefrontal cortex in the control of feature attention in area V4. *Nature Communications*, *10*(1), 1–12. https://doi.org/10.1038/s41467-019-13761-7

Bijanzadeh, M., Nurminen, L., Merlin, S., Clark, A. M., & Angelucci, A. (2018). Distinct Laminar Processing of Local and Global Context in Primate Primary Visual Cortex. *Neuron*, *100*(1), 259-274 e4. https://doi.org/10.1016/j.neuron.2018.08.020

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do we know what the early visual system does? *J Neurosci*, *25*(46), 10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005

Chung, S., & Ferster, D. (1998). Strength and orientation tuning of the thalamic input to simple cells revealed by electrically evoked cortical suppression. *Neuron*, *20*(6), 1177–1189. https://doi.org/10.1016/S0896-6273(00)80498-5

Dedieu, A., Rikhye, R. v, Lázaro-Gredilla, M., & George, D. (n.d.). *Learning attention-controllable border-ownership for objectness inference and binding*. https://doi.org/10.1101/2020.12.31.424926

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., Kording, K., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2022). The neuroconnectionist research programme. *ArXiv*. http://arxiv.org/abs/2209.03718

Duncan, J. (1984). Selective Attention and the Organization of Visual Information. In *Journal of Experimental Psychology: General* (Vol. 113, Issue 4). https://psycnet.apa.org/record/1985-29839-001

Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47. http://cercor.oxfordjournals.org/content/1/1/1.1.short

Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure-ground cues are valid for natural images. *Journal of Vision*, *7*(8), 2. https://doi.org/10.1167/7.8.2

Guclu, U., & van Gerven, M. A. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J Neurosci*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hesse, J. K., & Tsao, D. Y. (2016). Consistency of Border-Ownership Cells across Artificial Stimuli, Natural Stimuli, and Stimuli with Ambiguous Contours. *J Neurosci*, *36*(44), 11338–11349. https://doi.org/10.1523/JNEUROSCI.1857-16.2016

Hong, S., Noh, H., & Han, B. (2015). Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation. *Advances in Neural Information Processing Systems*, *2015-January*, 1495–1503. http://arxiv.org/abs/1506.04924

Jarvers, C., & Neumann, H. (2019). *Incorporating Feedback in Convolutional Neural Networks*.

Jin, J., Wang, Y., Swadlow, H. A., & Alonso, J. M. (2011). Population receptive fields of on and off thalamic inputs to an orientation column in visual cortex. *Nature Neuroscience*, *14*(2), 232–240. https://doi.org/10.1038/nn.2729

Jones, H. E., Grieve, K. L., Wang, W., & Sillito, A. M. (2001). Surround Suppression in Primate V1. *Journal of Neurophysiology*, *86*(4), 2011–2028. https://doi.org/10.1152/jn.2001.86.4.2011

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. https://doi.org/10.1038/s41593-019-0392-5

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2019). Disentangling neural mechanisms for perceptual grouping. *ArXiv*. http://arxiv.org/abs/1906.01558

Kingma, D. P., & Ba, J. L. (2015, December 22). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1412.6980v9

Kramer, A. F., Weber, T. A., & Watson, S. E. (1997). Object-based attentional selection--Grouped arrays or spatially invariant representations?: Comment on Vecera and Farah (1994). *Journal of Experimental Psychology: General*, *126*(1), 3–13. https://doi.org/10.1037//0096-3445.126.1.3

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. In *bioRxiv* (p. 408385). bioRxiv. https://doi.org/10.1101/408385

LECUN, & Y. (n.d.). THE MNIST DATABASE of handwritten digits. *Http://Yann.Lecun.Com/Exdb/Mnist/*. Retrieved January 28, 2021, from https://ci.nii.ac.jp/naid/10027939599

Lei, J., Benjamin, A. S., & Kording, K. P. (2021). Object Based Attention Through Internal Gating. *ArXiv*. http://arxiv.org/abs/2106.04540

Lien, A. D., & Scanziani, M. (2013). Tuned thalamic excitation is amplified by visual cortical circuits. *Nature Neuroscience*, *16*(9), 1315–1323. https://doi.org/10.1038/nn.3488

Linsley, D., Kim, J., Veerabadran, V., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated-recurrent units. *Advances in Neural Information Processing Systems*, *2018-December*, 152–164. http://arxiv.org/abs/1805.08315

Moore, T., & Fallah, M. (2004). Microstimulation of the Frontal Eye Field and Its Effects on Covert Spatial Attention. *Journal of Neurophysiology*, *91*(1), 152–162. https://doi.org/10.1152/jn.00741.2002

Muckli, L., de Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., Goebel, R., & Yacoub, E. (2015). Contextual Feedback to Superficial Layers of V1. *Curr Biol*, *25*(20), 2690–2695. https://doi.org/10.1016/j.cub.2015.08.057

Neri, P. (2014). Semantic control of feature extraction from natural scenes. *J Neurosci*, *34*(6), 2374–2388. https://doi.org/10.1523/JNEUROSCI.1755-13.2014

Neri, P. (2017). Object segmentation controls image reconstruction from natural scenes. *PLoS Biol*, *15*(8), e1002611. https://doi.org/10.1371/journal.pbio.1002611

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, *401*(6753), 584–587. https://doi.org/10.1038/44134

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

Papale, P., Betta, M., Handjaras, G., Malfatti, G., Cecchetti, L., Rampinini, A., Pietrini, P., Ricciardi, E., Turella, L., & Leo, A. (2019). Common spatiotemporal processing of visual features shapes object representation. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-43956-3

Papale, P., Leo, A., Cecchetti, L., Handjaras, G., Kay, K. N., Pietrini, P., & Ricciardi, E. (2018). Foreground-background segmentation revealed during natural image viewing. *ENeuro*, *5*(3). https://doi.org/10.1523/ENEURO.0075-18.2018

Papale, P., Leo, A., Handjaras, G., Cecchetti, L., Pietrini, P., & Ricciardi, E. (2020). Shape coding in occipito-temporal cortex relies on object silhouette, curvature and medial-axis. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.00212.2020

Papale, P., Zuiderbaan, W., Teeuwen, R. R. M., Gilhuis, A., Self, M. W., Roelfsema, P. R., & Dumoulin, S. O. (2021). The influence of objecthood on the representation of natural images in the visual cortex. *BioRxiv*, 2021.09.21.461209. https://doi.org/10.1101/2021.09.21.461209

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *ArXiv*. http://arxiv.org/abs/1912.01703

Poort, J., Raudies, F., Wannig, A., Lamme, V. A. F., Neumann, H., & Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron*, *75*(1), 143–156. https://doi.org/10.1016/j.neuron.2012.04.032

Poort, J., Self, M. W., van Vugt, B., Malkki, H., & Roelfsema, P. R. (2016). Texture Segregation Causes Early Figure Enhancement and Later Ground Suppression in Areas V1 and V4 of Visual Cortex. *Cereb Cortex*, *26*(10), 3964–3976. https://doi.org/10.1093/cercor/bhw235

Pozzi, I., Bohté, S. M., & Roelfsema, P. R. (2020). Attention-Gated Brain Propagation: How the brain can implement reward-based error backpropagation. *Advances in Neural Information Processing Systems*.

Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience*, *10*(11), 1492–1499. https://doi.org/10.1038/nn1989

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu Rev Neurosci*, *29*, 203–227. https://doi.org/10.1146/annurev.neuro.29.051605.112939

Roelfsema, P. R., & de Lange, F. P. (2016). Early Visual Cortex as a Multiscale Cognitive Blackboard. *Annu Rev Vis Sci*, *2*, 131–151. https://doi.org/10.1146/annurev-vision-111815-114443

Roelfsema, P. R., & Houtkamp, R. (2011). Incremental grouping of image elements in vision. *Atten Percept Psychophys*, *73*(8), 2542–2572. https://doi.org/10.3758/s13414-011-0200-0

Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H., & Bosch, H. (2002). Figure - Ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, *14*(4), 525–537. https://doi.org/10.1162/08989290260045756

Seijdel, N., Tsakmakidis, N., de Haan, E. H. F., Bohte, S. M., & Scholte, H. S. (2020). Depth in convolutional neural networks solves scene segmentation. *PLOS Computational Biology*, *16*(7), e1008022. https://doi.org/10.1371/journal.pcbi.1008022

Self, M. W., Jeurissen, D., van Ham, A. F., van Vugt, B., Poort, J., & Roelfsema, P. R. (2019). The Segmentation of Proto-Objects in the Monkey Primary Visual Cortex. *Curr Biol*, *29*(6), 1019-1029 e4. https://doi.org/10.1016/j.cub.2019.02.016

Self, M. W., Kooijmans, R. N., Super, H., Lamme, V. A., & Roelfsema, P. R. (2012). Different glutamate receptors convey feedforward and recurrent processing in macaque V1. *Proc Natl Acad Sci U S A*, *109*(27), 11031–11036. https://doi.org/10.1073/pnas.1119527109

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Spitzer, H., Desimone, R., & Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science*, *240*(4850), 338–340. http://www.ncbi.nlm.nih.gov/pubmed/3353728

Ullman, S. (1995). Sequence Seeking and Counter Streams: A Computational Model for Bidirectional Information Flow in the Visual Cortex. *Cerebral Cortex*, *5*(1), 1–11. https://doi.org/10.1093/cercor/5.1.1

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37

van Essen, D. C., Anderson, C. H., & Felleman, D. J. (1992). Information processing in the primate visual system: an integrated systems perspective. *Science*, *255*(5043), 419–423. http://www.ncbi.nlm.nih.gov/pubmed/1734518

van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., & Roelfsema, P. R. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science*, *360*(6388), 537–542. https://doi.org/10.1126/science.aar7186

Vecera, S. P., & Farah, M. J. (1997). Is visual image segmentation a bottom-up or an interactive process? *Perception and Psychophysics*, *59*(8), 1280–1296. https://doi.org/10.3758/BF03214214

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*. https://doi.org/10.1038/s41593-019-0517-x

Williams, M. A., Baker, C. I., op de Beeck, H. P., Mok Shim, W., Dang, S., Triantafyllou, C., & Kanwisher, N. (2008). Feedback of visual object information to foveal retinotopic cortex. *Nature Neuroscience*, *11*(12), 1439–1445. https://doi.org/10.1038/nn.2218

Williford, J. R., & von der Heydt, R. (2016). Figure-Ground Organization in Visual Cortex for Natural Scenes. *ENeuro*, *3*(6). https://doi.org/10.1523/ENEURO.0127-16.2016

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yan, S., Fang, X., Xiao, B., Rockwell, H., Zhang, Y., & Lee, T. S. (2019). Recurrent Feedback Improves Feedforward Representations in Deep Neural Networks. *ArXiv*. http://arxiv.org/abs/1912.10489

Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *J Neurosci*, *20*(17), 6594–6611. http://www.ncbi.nlm.nih.gov/pubmed/10964965

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J Neurosci*, *27*(45), 12292–12307. https://doi.org/10.1523/JNEUROSCI.1897-07.2007

Zuiderbaan, W., van Leeuwen, J., & Dumoulin, S. O. (2017). Change Blindness Is Influenced by Both Contrast Energy and Subjective Importance within Local Regions of the Image. *Front Psychol*, *8*, 1718. https://doi.org/10.3389/fpsyg.2017.01718

Acknowledgements

We thank Jasper Poort for providing the code and pre-processed data from his experiment, and Alexander van Meegen for his early contribution to the model implementation. This work was supported by NWO ('INTENSE'), the European Union FP7 (ERC grant number 339490 'Cortic_al_gorithms'), the European Union Horizon 2020 Framework Programme for Research and Innovation under the Framework Partnership (HBP FPA agreement number 650003), the H2020 Framework Programme for Research and Innovation (Human Brain Project SGA2 grant number 785907, and HBP SGA3 grant number 945539), the European Union Horizon 2020 Future and Emerging Technologies (FET Open grant number 899287 'NeuraViPeR').