# How the brain learns to parse images using an attentional, incremental grouping process

Sami Mollard[1*], Sander M. Bohte[2,3], Pieter R. Roelfsema[1,4,5,6]

Affiliations:
[1]Department of Vision & Cognition, Netherlands Institute for Neuroscience, Meibergdreef 47, 1105 BA, Amsterdam, The Netherlands.
[2]Machine Learning Group, Centrum Wiskunde & Informatica, Science Park 123, 1098XG, Amsterdam, The Netherlands
[3]Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1090GE 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
[5]Department of Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research, VU University, Amsterdam, The Netherlands.
[6]Department of Psychiatry, Academic Medical Center, Amsterdam, The Netherlands

* Corresponding author: Sami Mollard

**Email:** s.mollard@nin.knaw.nl

**Abstract**

Natural scenes usually contain a vast number of objects that need to be segmented and segregated from each other and from the background to guide behaviour. In the visual brain, object-based attention is the process by which image fragments belonging to the same objects are grouped together. The curve-tracing task is a special case of a perceptual grouping task that tests our ability to group image elements of an elongated curve. The task consists in determining which image elements belong to the same curve, and in the brain, neurons spread an enhanced activity level over the representation of the relevant curve. A previous "growth-cone model of attention" accounted for the scale invariance of tracing by proposing that the enhanced activity propagates at multiple levels of the visual cortical hierarchy. However, the precise neuronal circuitry for learning and implementing scale-invariant tracing remains unknown. We propose a new recurrent architecture for the scale-invariant labelling of curves and objects. The architecture is composed of a feedforward pathway that dynamically selects the right scale and prevents the spilling over of the enhanced activity to other curves, and a recurrent pathway for tag spreading that involves horizontal and feedback interactions, mediated by a disinhibitory loop involving VIP and SOM interneurons. We trained the network with curves up to seven pixels long using reinforcement learning and a learning rule local in time and space and we found that it generalized to curves of any length and to spatially extended objects. The network chose the appropriate scale and switched to higher or lower scales as dictated by the distance between curves, just has as been observed in human psychophysics and in the visual cortex of monkeys. Our work provide a mechanistic account of the learning of scale-invariant perceptual grouping in the brain.

**Significance Statement**

Objects are labelled and grouped in the visual cortex via a tag of enhanced activity. If the scale-invariant dynamics of propagations of this tag are well characterised, it remains unknown what neural architectures and learning rules can produce those dynamics. This work is the first to propose a neural architecture trained with reward that give rises to the same dynamics observed in monkeys' visual cortex or human reaction times, shedding light on the mechanisms of multiscale object-based attention in the visual cortex.

# Main Text

## Introduction

Natural scenes usually contain a vast number of objects that need to be segmented and segregated from each other and from the background in order to guide behaviour. In the visual brain, elements of spatially extended objects are added to a growing object representation by the propagation of a tag of enhanced activity (1), a process known as "object-based attention" in psychology, which is a highly serial process.

In the first half of the 20th century, Gestalt psychologists described several rules to determine what was grouped with what during visual perception (2, 3). Such rules include for example the law of connectedness, following which connected elements are grouped together, or the law of good continuation, where aligned contours are assigned to the same object. First theories of perception assumed that those Gestalt rules were applied preatentively and in parallel across the visual field (4), but further experiments revealed that, in certain cases, the grouping of image elements together is a serial process. This is the case for curve tracing tasks for example. In these tasks, participants have to report if two elements belong to the same curve or not while keeping their eye on a fixation point (Fig. 1A). As the target curve is not traced by the participant following it with its eyes, the tracing of the curve is a mental operation. Jolicoeur et al. showed

that the time needed to trace the curve was proportional to the length of the curve to trace, indicating a serial process (5). Later, Roelfsema et al., 1998 recorded the activity of neurons in the visual cortex of monkeys performing a curve tracing task (1). Neurons whose receptive field fell on the curve that needed to be traced (the target curve) showed an enhanced activity compared to neurons whose receptive field fell on a distractor curve (Fig. 1A,B). Fig.1B shows the activity of one neuron when its receptive field fell either on the target curve (orange) or the distractor curve (blue). After 130ms the activity of the neuron is enhanced when its receptive field is on the target curve. It was also shown that this enhancement appeared later when neurons where further from the beginning of the curve (6), showing that object grouping was performed by the spread of a tag of enhanced activity over the representation of the target object, and confirming the serial nature of this spreading, from the cue toward the target (Fig. 1A). To unify those seemingly disparate findings that the same gestalt rules can be applied in parallel in certain situation and sequentially in others, it was later proposed that grouping happens in two phases (7, 8). In the first phase, dominated by feedforward processing, neurons tuned to different features, like colour or orientation, become active if the corresponding feature is present in their receptive field. The sets of active neurons after this phase forms the 'base-representation'. This mechanism is in effect when gestalt rules are applied preatentively and in parallel. During the second, recurrent processing phase, neurons in the base representation belonging to the same objects are grouped together incrementally, via feedback and horizontal connections. This phase is more flexible but takes more time.

Further work on curve tracing in human and monkeys showed that this iterative grouping process depends not only on the length of the curve to trace, but also on the distance between the target and the distractor curves: the tracing speed is lower when the curves are close, and higher when they are further apart (6, 9). This characteristic of the grouping process can be explained if it occurs at multiple levels of the visual hierarchy. When the stimulus is closer to the observer, the length of the curves, measured in degrees of visual angle, increases. However, the distance between the curves also increases and the grouping process can occur more swiftly through neurons in higher visual areas, which have bigger receptive fields, without attention spilling from the target to the distractor curve, such that the reaction time remains the same. This also implies that, when two curves are close to each other, the spreading occurs at a low-level of the visual hierarchy and takes time, while when the two curves are further apart, the spreading can occur faster. Those predictions were confirmed by Pooresmaeili et al. (6): they recorded the activity of neurons in the visual cortex of monkeys solving a curve tracing task. Critically, the stimuli shown presented a bottleneck where the two curves would become close to each other (Fig. 1C). By measuring the time needed for the attentional signal to spread to neurons before and after the bottleneck, they showed that a "growth-cone" model of attention, where the attentional tag spreads at the highest level such that only one curve is in the receptive field of the neurons, best explained results. Later, Jeurissen et al. extended those results from simple curves to 2-D images (10): reaction times of humans participants asked to report if two dots were on the same object or not were best explained by a growth-cone model of attentional spreading.

While the growth-cone model explains the dynamics of attentional spreading in object-based attention, it is a descriptive model, and it remains unknown what neural circuitry and learning rule can learn to group objects by dynamically choosing the appropriate scale depending on the visual input. We thus examined how a recurrent neural network with four spatial scales could learn to trace curves in a scale-invariant way using only, as for monkeys, reward, and with a biologically plausible learning rule.

We endowed the network's architecture with two innovations. The first one, following (11), is the presence of two distinct group of neurons (Fig. 2A). Neurons in the feedforward group only receive feedforward input; they are responsible for base-representation and dynamic, flexible scale selection. Neurons in the recurrent group receive feedforward, but also feedback and horizontal connections. They are responsible for recurrent processing and tag propagation. This segregation of neurons in two group is coherent with experimental results showing that in the visual cortex some neurons are influenced by curve-tracing while others are not (12–15). Neurons sensitive only to feedforward inputs are found in layers 4 and 6, while neurons active during the

recurrent processing phase are more prevalent in layers 2, 3 and 5 (12–14). The feedforward group of our model could thus correspond to layers 4 and 6 of the visual cortex, while the recurrent group would correspond to layers 2, 3 and 5. Critically, neurons in the recurrent group in our model where multiplicatively gated by neurons in the recurrent group. This ensured the stability of the grouping process: a neuron in the recurrent group can only be active if the corresponding neuron in the feedforward group is active (i.e. if it is part of the base representation). This enable the tag of enhanced activity to spread on the target object without accidentally tagging the distractor object.

The second innovation is a disinhibitory loop mediating feedback and horizontal connections (Fig. 2B). It was shown that in mice visual cortex, disinhibitory circuits contribute to solve figure-ground task, known to rely on iterative perceptual processes similar to the ones used in curve-tracing tasks (16). Modelling work has also shown how disinhibition and perceptual attention could be related (17). In our work, disinhibitory circuit facilitate learning by making the network both stable and expressive, and ensure that the enhancement of activity can spread over the full length of the target object.

The goal of our work was threefold. We asked whether the network 1- would learn to indicate if a cue and a target belonged to the same objects, 2- If it solved the task by propagating a tag of enhanced activity, 3- if this tag was propagated at multiple level of the hierarchy.

We trained the network on curve-tracing tasks for which electrophysiological recording of monkeys performing the same task exist (6). We used RELEARNN (11, 18), a biologically plausible learning rule that determines how synapses modify their connection strength using information both local in space and time. After having been trained to trace curves of up to 7 pixels, the network could generalize to arbitrary long curves. We found that the activation of units of the network became similar to that of neurons in the visual cortex of monkeys: the network learnt to propagate enhanced activity along curves. After training on a curve-tracing task, the network could also perform an object-tracing task and propagate attention over 2D, spatially extended shapes with minimal fine-tuning. We found that our networks reaction time could predict human-reaction times as well as the growth-cone model of attention. Our model shows how a recurrent neural network trained with a biologically plausible learning rule can learn to propagate enhanced activity at different levels to group objects.

# Methods

## Tasks

We trained the networks on a curve-tracing task for which electrophysiological recordings of monkeys visual cortex are available (6) (Fig. 1A). In this task, monkeys started the trial by directing their gaze to a fixation point. After 300 milliseconds, a stimulus consisting of two curves appeared. Monkeys then had to mentally trace the curve that was connected to the fixation point (target curve), while keeping their eyes fixated on the fixation point. When the fixation point disappeared, the monkeys had to make an eye movement toward the blue circle at the other end of the target curve. In the version of the task tailored for the network, the whole stimulus was presented at once.

After successful training on the curve-tracing task, networks were also tested on an object-grouping task for which human psychophysics results are available (10, 19) (Fig. 1B). In this task, the stimulus consisted of two objects, either outlined or in a naturalistic setting. Participants had to report whether a cue appeared on the same or a different object as the fixation point, while their reaction time was recorded.

We note that, to solve both tasks, the agent must learn to group elements connected to each other. In the curve-tracing task this grouping process takes place on a curve along one dimension, while in the object-grouping task the grouping occurs in two dimensions.

In the version of the tasks tailored for the neural network, we used a grid with a depth of three features, one for each colour the pixels could take (red, blue and green). For training speed purpose the networks were trained on a 108x108 pixels grid. To probe the networks on curves

long enough to see the effect of scale selection, and on object with high enough resolution, results were obtained on a 144x144 pixels grid for the curve tracing task, and a 594x594 pixels grid for the object tracing task. We were able to change the input size between training and testing time because we used weight-sharing (see Model). We probed the capacity of the network to select the blue pixel either at the other end of the curve connected to the red pixel (curve-tracing task), or on the same object as the red pixel (object-grouping task).

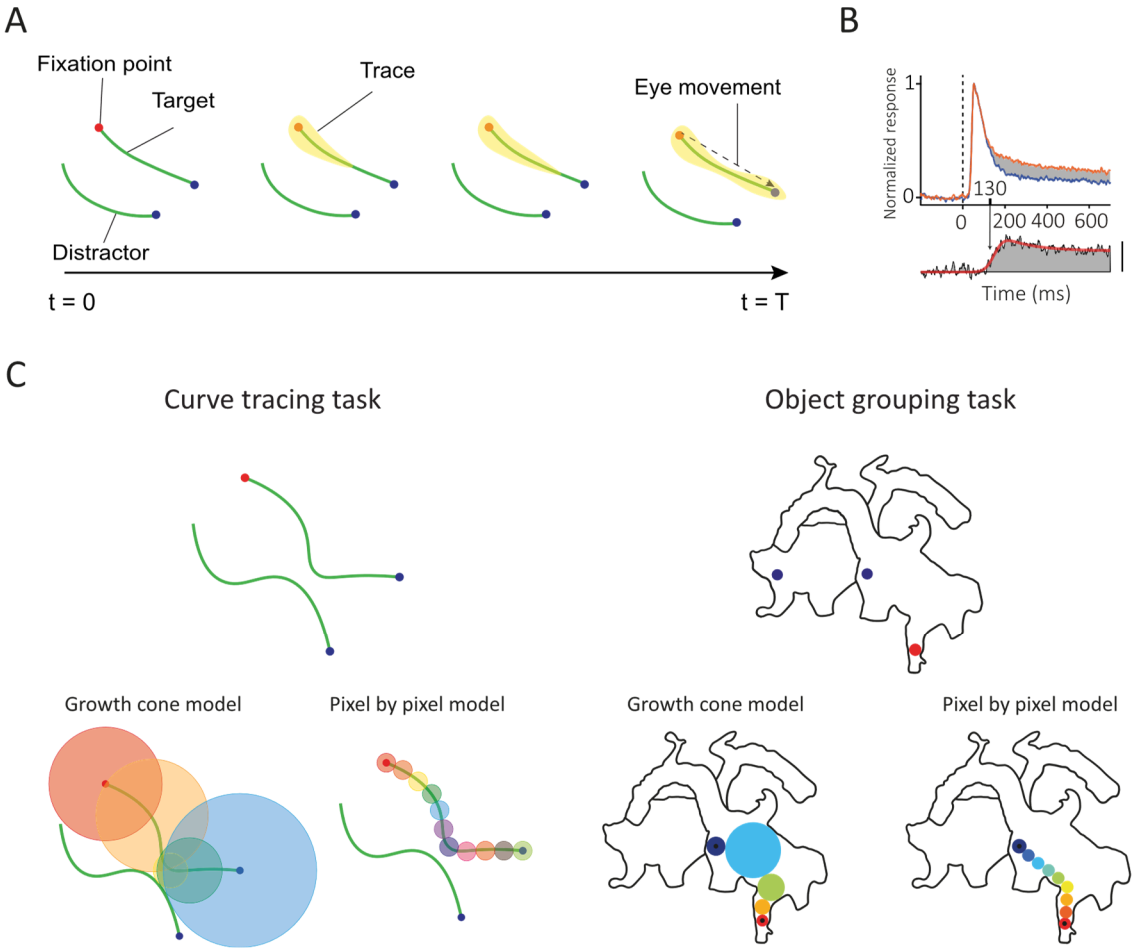



**Figure 1. Tasks used to probe the dynamics of object-based attention. A.** In the curve tracing task, the agent has to make an eye movement toward the blue dot connected to the red fixation point. The representation of the target curve in the visual cortex is enhanced because extra neuronal activity spreads over this curve (yellow). **B.** Recordings in monkey's visual cortex show that in the curve tracing task, the target curve is identified by a tag of enhanced activity of the neurons whose receptive field fall on the target curve (orange line) versus the distractor curve (blue line). **C**. Curve tracing and object grouping tasks. In the curve tracing task, a growth-cone model of attention explains the dynamics of neural activity by positing that a tag of enhanced activity can propagate at multiple scales of the visual hierarchy, depending on the distance between the curves. In the pixel-by-pixel model however, the assumption is that the tag spreads through horizontal connections at one scale only, making the speed of propagation constant. The object grouping task is analogous as the curve-tracing task, but the agent has to spread activity not over a one-dimensional line, but a two-dimensional shape.

## Model

To solve those tasks, we developed a neural network with a recurrent architecture and four spatial scales, where neurons in each layer could belong to one of two groups: the feedforward group and the recurrent group (Fig. 2A) .

The role of the feedforward group is twofold. First, it permits dynamical scale selection. Neurons in the feedforward group are trained to activate only when the stimulus in their receptive field is not ambiguous (see training). If a neuron is activated even in the presence of an ambiguous stimulus in its receptive field, the tag of enhanced activity risks to "spill" to the distractor object. For the curve-tracing task, we consider that the stimulus inside a receptive field is not ambiguous when all the pixels are connected and colinear (Fig. 2C). For the object-tracing task, the stimulus inside a receptive field is not ambiguous if it doesn't contain a boundary (Fig. 2C). Upon presentation of the stimulus, feedforward activity flows in the network and, being activated if the stimulus in their receptive field is not ambiguous, neurons in the feedforward group encode a multiscale base-representation (Fig. 2D). The feedforward group also has a gating role : neurons in the feedforward group multiplicatively gate neurons in the recurrent group (see Eq. 4). Thus if a neurons in the feedforward group is inactive because the stimulus in its receptive field is ambiguous, the corresponding neuron in the recurrent group will also be inactive. This ensures that the tag of enhanced activity can't propagate over a distractor object.

Neurons in the recurrent group learn to spread enhanced neuronal activity over the representation of the target object, to label it as one coherent item.

To train using RELEARNN (18), recurrent neural networks need to reach a stable state, which is not guaranteed with nonlinearities such as ReLU functions. Using squashing nonlinearities like sigmoid however can lead to a loss of expressivity (20). To circumvent this issue, Mollard et al. used a ReLU-like function with a decreasing slope after a threshold (11). While this made the network both stable and expressive, the difference of activities between pixel on the target and distractor curve was decreasing the further pixels were from the fixation point, making the task unsolvable for long curves. Here we implemented tracing as a disinhibitory process, to ensure that the network would be stable and expressive for arbitrarily long curves (Fig. 2B). Feedback and horizontal connections activate Vasoactive intestinal peptide-expressing (VIP) interneurons that have a disinhibitory effect on their corresponding pyramidal neurons, via inhibition of Somatostatin-expressing (SOM) interneurons (16, 21). Hence, tracing causes the incremental disinhibition of pyramidal units that represent the target object, which is thus serially labeled with enhanced activity. The activity of each pyramidal neurons can't be greater than the feedforward activity they would get if there was no SOM inhibition, preventing the network to diverge and ensuring it reaches a stable state.

The activity $\boldsymbol{X^l}$ of units in the feedforward group and $\boldsymbol{Y^l}(t)$ of units in the recurrent group in layer $l$ and time $t$ was determined by:

$$\boldsymbol{X^l} = \sigma\left(V \otimes \boldsymbol{X^{l-1}}\right) \tag{1}$$

$$\boldsymbol{VIP^l}(t) = H \otimes \boldsymbol{Y^l}(t-1) + FB \otimes \boldsymbol{Y^{l+1}}(t-1) \tag{2}$$

$$\boldsymbol{SOM^l}(t) = \sigma\left(\boldsymbol{SOM_0} - \boldsymbol{VIP^l}(t)\right) \tag{3}$$

$$\boldsymbol{Y^l}(t) = \sigma\left(\phi\left(\boldsymbol{X^l}\right) \odot \left(FF \otimes \boldsymbol{Y^{l-1}}(t) - \boldsymbol{SOM^l}(t)\right)\right) \tag{4}$$

Where FF and FB represent feedforward and feedback weights between scales, H horizontal weights within a scale, $SOM_0$ is the default activation of SOM neurons in the absence of feedback, $\sigma$ is a ReLU activation function and $\phi$ a gating function (see methods). $\odot$ represents the element-wise product between recurrent and feedforward elements and $\otimes$ indicates

convolution between weights and neurons. Colors in the equations correspond to units and connections in Fig. 2A and Fig. 2B.

To reduce computation time, we used weight sharing, despite the absence of weight sharing in the brain. In previous work, we showed that networks with a similar structure trained on a curve-tracing task could learn the task both with or without weight-sharing, but the number of trials needed to learn the task without weight-sharing was 7 times greater than with weight-sharing (11). Our results are thus likely to generalize to biologically plausible learning without weight sharing.

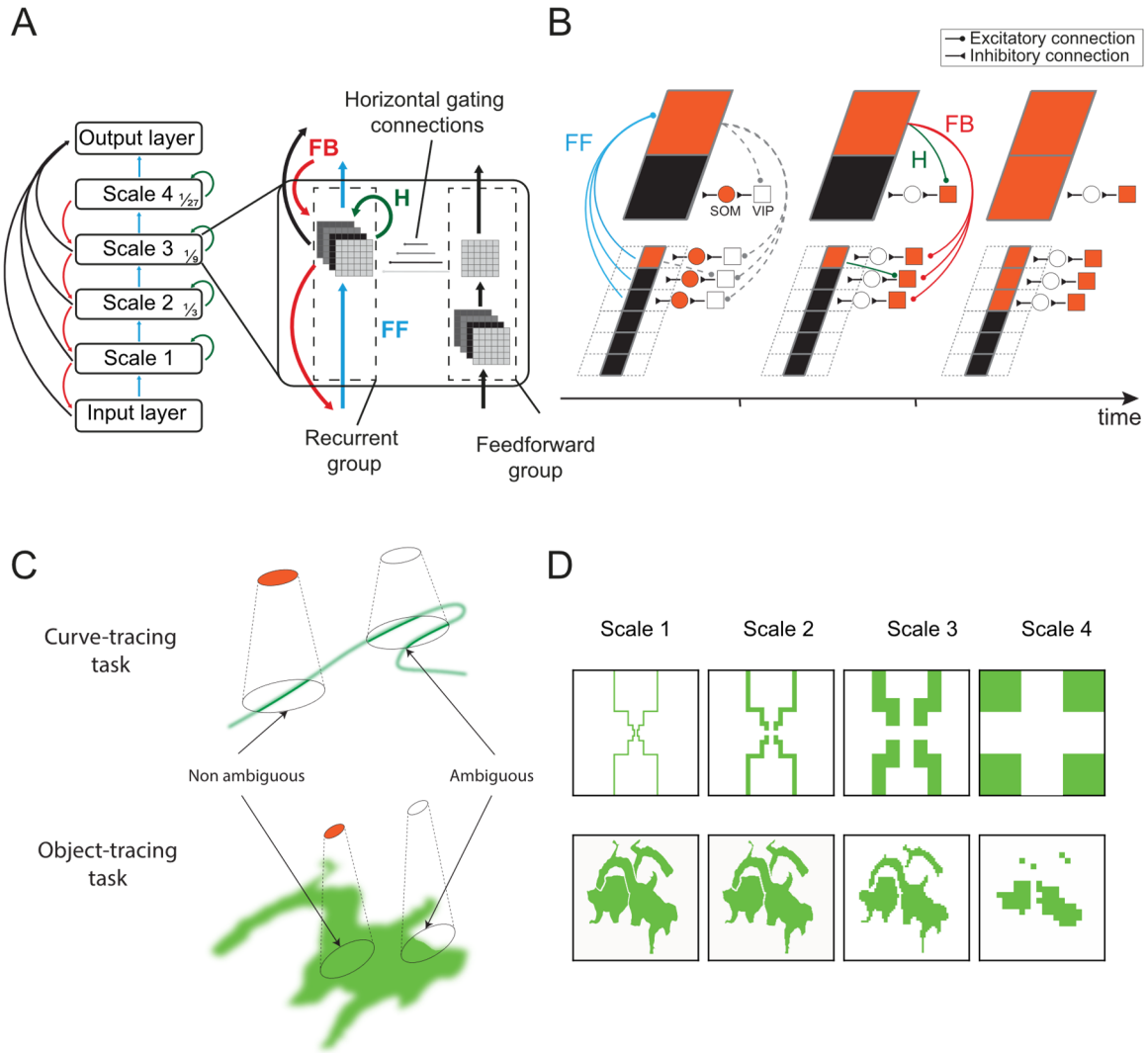



**Figure 2. Model. A.** The network incorporates four spatial scales. Neurons can belong either to the feedforward group, responsible for scale selection and backbone representation, or to the recurrent group, responsible for attentional modulation. Neurons in the recurrent group are gated by neurons in the feedforward group. **B.** Grouping as a disinhibitory process. When the attentional tag reaches a pyramidal neuron (orange), feedback and horizontal connections activate nearby VIP interneurons that have a disinhibitory effect on their corresponding pyramidal neurons, via the inhibition of SOM interneurons. The activity of those pyramidal neurons increases: the attentional tag has propagated. Grouping thus consists in the incremental disinhibition of pyramidal units that represent the target object, which is therefore serially labelled with enhanced activity.. **C.** Scale selection in the feedforward group. For the curve tracing task, neurons in the feedforward group will be active (orange) only if all the elements in their receptive field are connected and colinear. If that is not the case, the stimulus in the receptive field is ambiguous, and the tag could spill from the target to the distractor curve. In the object-tracing task, neurons in the feedforward group will be active (orange) only if their receptive field fall on an homogeneous image region. If that is not the case, the tag could spill outside of the attended object. **D.** Representation learnt by the feedforward group for the curve tracing task (top) and the object tracing task (bottom). At higher scale, the representation is coarser, less detailed and presents gaps, but the receptive field sizes are bigger and propagation can happen faster.

## Training

Training of the network was accomplished in two steps.

First, we pretrained two feedforward networks, one for the curve-tracing task and one for the object-grouping task (see Methods). The curve-tracing feedforward network was trained to classify colinear and non-colinear elements in a supervised fashion: if all the active elements inside a unit's receptive field are colinear and connected, this unit was trained to become active (Fig. 2C). Similarly, units in the object-grouping feedforward network were trained with supervised learning to become active if all the elements in their receptive field were active (Fig. 2C). We generated two synthetic datasets with randomly generated curves or objects and the corresponding labels, as shown in Fig. 2C, and trained the feedforward networks with supervised learning on those two datasets. When monkeys or human are first presented with the curve-tracing or the object-tracing tasks, they come with pre-existing knowledge and inductive bias that they form through developmental and learning processes (22–26). They already know how to segment and group objects. We argue that the pretraining of the feedforward network would correspond to this already formed knowledge that humans and monkeys are equipped with when they learn the curve and object-tracing tasks for the first time. Since our objective is to model the learning of those tasks, and not the inductive bias that helps learn them, we used backpropagation and supervised learning to train the feedforward network.

After pretraining, weights in the feedforward group were frozen while weights in the recurrent group were trained to trace curves using reinforcement learning, analogous to the way monkeys or humans learnt the tasks. Weights in the recurrent group were updated using RELEARNN, the local learning rule inspired by the Almeida-Pineda algorithm (7, 11, 18, 20), see methods.

To avoid the recurrent network getting stuck in a local minimum, we used a curriculum for training, a strategy also used to train monkeys. The network was first presented with curves that could be 3 pixels long. Once the network achieved 85% accuracy during a test phase with fixed weight and no exploration, we added one pixel to the curve and repeated the operation until curves were 7 pixels long.

# Results

## Curve-tracing task

We trained 5 networks on the curve-tracing task on a 108x108 pixels grid, and they reached our convergence criterion in an average of 23 200 trials.

The curves seen during training were drawn randomly at each trial. As the number of curves grows exponentially with their length, the number of curves possible exceeds the number of trials needed to learn the task and there is little chance that the network could memorize all the input presented. Instead, the network learns an abstract, general grouping rule that can be applied irrespective of the characteristics of the stimulus shown (shape, length of the curve). To validate this, we tested the network at the end of training on randomly generated 30-pixels-long curves. Not only were those curves never seen during training, they were also 4 times longer than the curves the models were trained on. The networks achieved perfect accuracy on 30-pixel-long curves.

To isolate the effect of the disinhibitory loop, we trained networks with no disinhibition, where neurons activate their neighbour directly. Networks were either unstable if we used Relu activation functions, or not expressive enough if we used squashing non linearities: the further a pixel was from the fixation point, the weaker the tag, and the network could not learn the task for longer curves, this effect was also observed in shallower networks (11). The architectural inclusion of a disinhibition loop thus enabled the network to be both expressive and stable for the learning of the curve-tracing task.

When monkeys perform the curve-tracing task, the latency of the attentional modulation of neurons whose receptive field fall on the target curve is best explained by a growth-cone model

of attention. In this model, the tracing speed depends not only on the distance of the receptive field from the fixation point, but also on the distance to the distractor curve. To determine the dynamics of attentional modulation in our artificial neural network model, we generated 15 test stimuli, with curves alternatively getting closer or further away from each other, and we sought to determine if those dynamics were better explained by a pixel-by-pixel model or a growth-cone model. Fig. 3A shows one of those test stimuli, with superimposed in orange the neural dynamics in the recurrent group across time. When curves are far apart, tagging proceeds faster (between steps 20 and 25 for instance), but when the curves are closer or if the curvature is higher, tagging is slower (between steps 5 and 10 for instance). We emphasize that the pixel-by-pixel and the growth-cone models are descriptive models that can be applied to monkeys or artificial neural networks neural data, while our artificial neural network is a mechanistic model with dynamics that can be compared to the one recorded in monkey's visual cortex.

To determine the latency of the attentional modulation at each point on the target curve, we looked at the normalized difference of activity between neurons whose receptive field fall on the target curve, and those whose receptive field fall on the distractor curve. The latency of the attentional modulation was defined as the timestep when this normalized difference reached 30% of its maximum. Fig. 2B shows the normalized difference for different groups of neurons whose receptive fields fall along an example curve. When the curves are far apart, like for the red, orange, green and blue groups, neurons with bigger receptive fields (in this case 3x3 receptive fields) can spread enhanced activity and the tag propagates faster. However when the two curves are closer, this is not possible anymore and the spreading has to happen at a lower level, in this case a pixel-by-pixel level.

To investigate which of the pixel-by-pixel model or the grouping model accounted better for the latency data we used a regression analysis. As can be seen on Fig. 3C, the growth-cone model fitted the dynamics of the artificial neural network much better than the pixel-by-pixel model, indicating that grouping speed in our network happens at different scale, depending on the curvature of the target curve or the distance to the distractor curve, in line with what has been observed in the visual cortex of monkeys (Fig. 3D).

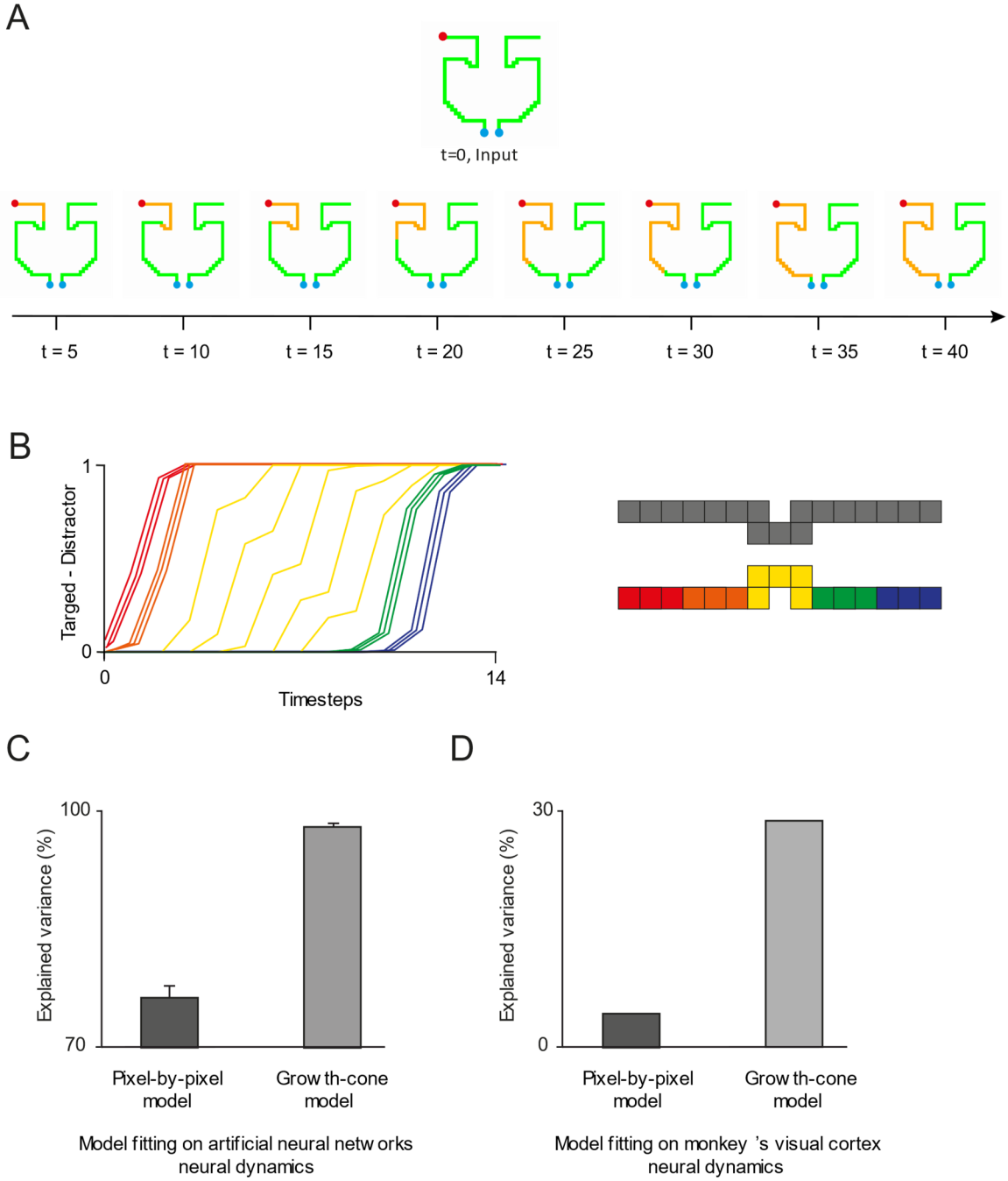



**Figure 3. Performance of the network on the curve-tracing task.**. **A.** Example stimuli (green) with superimposed neural dynamics in the recurrent group in orange. When curves are far apart, tagging proceeds faster (between steps 20 and 25 for instance), but when the curves are closer or if the curvature is higher, tagging is slower (between steps 5 and 10 for instance). **B.** Left, normalized response enhancement for the target curve. Each curve is normalized by its maximum over time. When the curves are far apart, activity can propagate at a higher level (red, orange, green and blue group), and is fast, but when get closer together, activity propagates at a slower pace, one pixel at a time (yellow). Right, example stimulus corresponding to the activity curves on the left. Colours of the pixels correspond to the colours of the response curves **C.** Following (3), we fitted neural dynamics in the artificial neural networks to the ones predicted by the pixel-by-pixel and the growth-cone model. Error bars, 95%-confidence interval **D.** Similarly as what has been observed in monkey's visual cortex, the growth-cone model of attention explains the dynamics of attentional spreading in our artificial neural network better than the pixel-by-pixel model.

## Object-tracing task

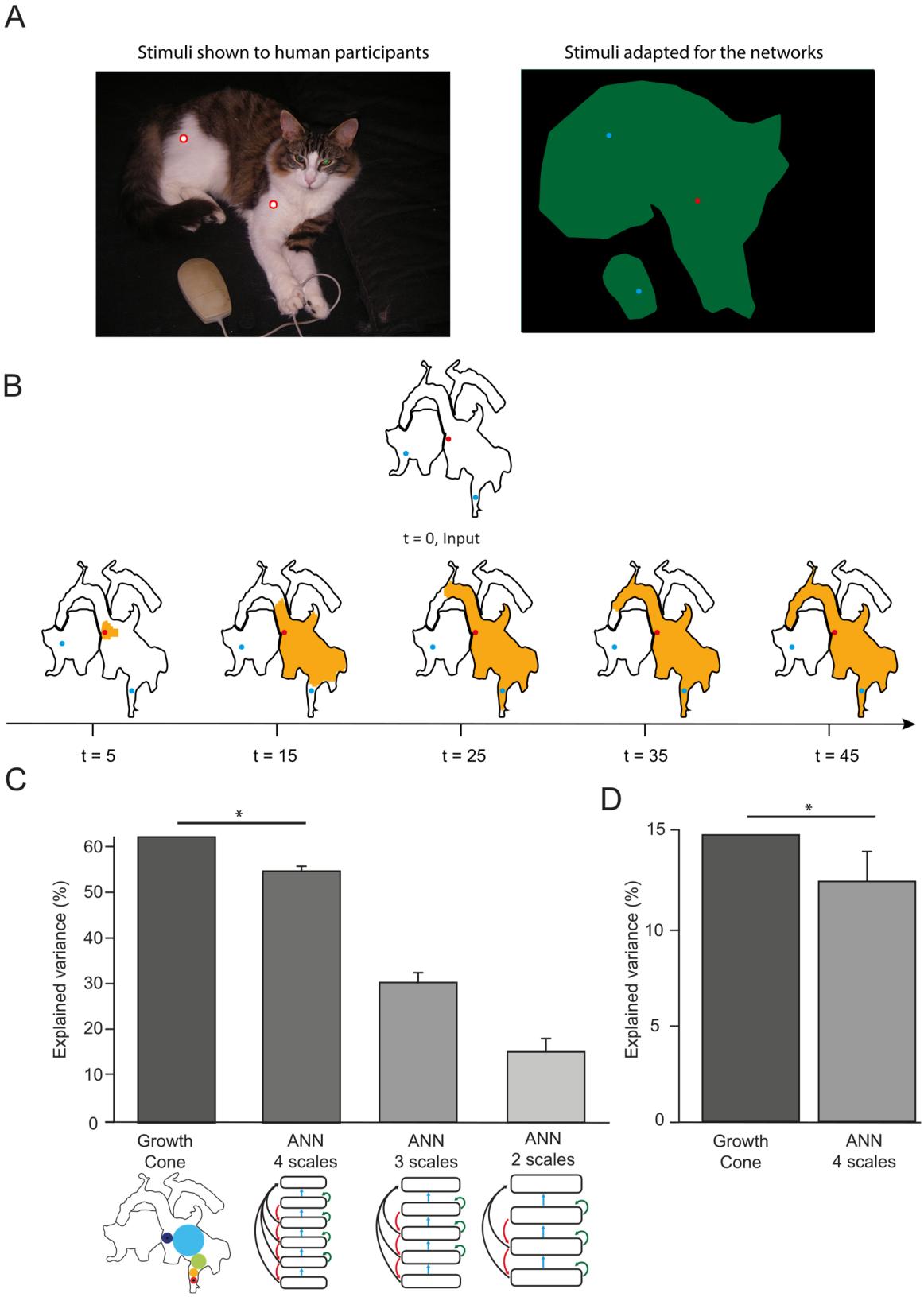
After training networks on the curve-tracing task, we probed their ability to group elements over two dimensional, spatially extended objects. For this, we replaced the curve-tracing feedforward network by the object-grouping one, while we left the weights in the recurrent group unchanged. Interestingly, out of the ten networks that we trained on the curve-tracing task, all were able to learn the object-grouping task in less than 100 trials, while still being able to perform the curve-tracing task with no loss of performance. This result suggests that the spread of attention over one dimensional objects or spatially extended ones can be achieved by the same mechanisms and the same architecture.

To determine the dynamics of object-based attention in human, Jeurissen et al. devised the “two-dot tasks”: an image containing two objects (two monkeys for instance) was presented to human participants (10). The fixation point was inside one of the objects, and a cue could be either inside the same object, or inside a different object. The goal of the participants was to determine if the two dots were on the same or different objects. Jeurissen et al. probed participants with four categories of stimuli: scrambled shapes, outlines, cartoons, and pictures on a white background. For the scrambled shapes and outline categories, Jeurissen et al. found that a growth-cone model of attention better explained human reactions time than a pixel-by-pixel model (10). The stimuli used by Jeurissen et al. were not naturalistic, as both the target and distractor objects were either outlined or presented on a white background. Subsequent work by Adeli et al. (19) collected human reaction times for a more challenging version of the two-dot task where the stimuli consisted of natural images.

Given that a growth-cone model better explained our artificial neural networks neural dynamics, we set to determine how well reaction times in our artificial neural networks could predict human reaction time. We tested our networks on the scrambled stimuli from Jeurissen et al., and on the natural stimuli from Adeli et al. For the latter, we added a preprocessing step where we first outlined the target and distractor object, and removed all textures (Fig. 4A). We then determined the reaction time of the networks as the timestep when the neurons in the output layer whose receptive field falls on the target reached 90% of their maximum value.

We found that while the growth-cone model explained 62% of the variance of human reactions time in Jeurissen et al. dataset, our artificial neural network explained 55% of it ($p < 0.05$ Student's t-test) (Fig. 4C). We tested a similar architecture with a lower number of scales and observed that the more scales used in the network, the better its explanatory power. Similarly, even though the growth-cone model assumes an infinite number of scales while our model uses only four, our trained networks nearly matched the predictive performance of the growth-cone model in the Adeli et al. dataset (Fig. 4D).

Those results point to the fact that our networks architecture and learning rule remarkably accounts for dynamics of object-based attention in humans and provide a mechanistic account of the growth-cone model in a learning setting.


A
Stimuli shown to human participants
Stimuli adapted for the networks
B
t = 0, Input
t = 5
t = 15
t = 25
t = 35
t = 45
C
Explained variance (%)
0
10
20
30
40
50
60
*
Growth Cone
ANN 4 scales
ANN 3 scales
ANN 2 scales
D
Explained variance (%)
0
5
10
15
*
Growth Cone
ANN 4 scales

362

**Figure 4. Object-tracing task. A.** We adapted stimuli shown to human participants for the networks. After outlining the shape of the target and distractor objects, we removed all textures such that objects where green on a black background. **B.** Example stimulus with superimposed neural dynamics in the recurrent group (orange). Note that when homogeneous images regions are large, tagging goes faster (between steps 5 and 15 for instance), but when they are narrower, tagging takes more time (between steps 25 and 35 for instance). **C.** Explained variance in human reaction times by different models. We fitted human reactions times obtained from Jeurissen et al. (10) with reactions times predicted by the growth-cone model and our artificial neural network model using either 2, 3 or 4 scales. If the growth-cone model gives the best fit ($p < 0.05$, Student's t-test), it assumes a continuous hierarchy, while our artificial neural network only has four levels. Error bars, 95%-confidence interval. **D.** Similarly**,** the reaction times predicted by our networks predict the reaction times collected by Adeli et al. (19) almost as well as the growth-cone model ($p < 0.05$, Student's t-test).

## Discussion

In this study, we used an optimization-base approach (27) to understand the scale-invariant propagation of object-based attention. In the visual cortex of monkeys, curve-tracing tasks are solved by the propagation of an attentional tag along the representation of the attended curve (1). The speed of tracing depends not only on the length of the curve to trace, but also on the distance between the target and distractor curves. This can be explained by the growth-cone model of attention, a descriptive model where the tag of enhanced activity can propagate at different levels of the visual hierarchy (28). Similar dynamics can be observed in human subjects. When asked to report if two dots belong to the same object or not, their reaction times can also be explained fairly accurately by the growth-cone model of attention (10). It remained unknown however what architecture and learning rule are responsible for such dynamics.

We trained artificial neural networks on a curve-tracing tasks, with rewards only, mimicking monkeys training. Our model was endowed with two innovations. First, artificial units could either receive feedforward inputs only (feedforward group), or feedforward, horizontal and feedback inputs (recurrent group). Neurons in the feedforward group represent the base-representation, the backbone on which object-based representation can take place (29), and map to neurons in the visual cortex which are mainly activated by feedforward input (12–15). The feedforward group is also responsible for dynamics scale selection in the network, as they only activate when a non-ambiguous stimulus is present in their receptive field. Neurons in the recurrent group were responsible for the propagation of a modulatory tag of enhanced activity over the representation of the target object, to label it as a coherent whole. The second innovation is the introduction of a disinhibitory loop connecting neurons in the recurrent group, through VIP and SOM interneurons. Such a disinhibitory loop stabilizes training, as the activity of neurons is bounded by their feedforward input, and the network is guaranteed to reach a stable state. The disinhibitory loop also enables the network to trace over arbitrary big objects without attenuation, as has been observed in previous work (11). After training, we compared dynamics as predicted by the growth-cone model or a simpler, pixel-by-pixel model, to the dynamics in our artificial neural networks. The growth-cone model predicted dynamics in our artificial neural networks better than the pixel-by-pixel model, as observed in monkeys' visual cortex, indicating that our networks had learnt to propagate activity in a multiscale, dynamics fashion.

After training, and despite having been trained only to trace one-dimensional curves, the networks were also able to propagate activity over 2D, spatially extended objects. For this task, we did not use the growth-cone model to predict activities in our artificial neural networks, but directly compared the ability of the growth-cone model and our networks to predict human reactions time. If the growth-cone model has ultimately a higher prediction power, it assumes infinite scales, while our artificial neural network only has 3. Moreover, the growth-cone model is merely descriptive, while our neural network offers a mechanistic account of human reaction times in an object-based attentional task.

This work is not the first aiming to understand the dynamics of multiscale object-based attention. Domijan & Marić built a neural network to trace curves, using separate feedforward and recurrent units. In their model, gabor filters are used to obtain a multiscale representation of

curves (30). Units at each level must then be above a scale-specific threshold to be activated. The presence of a distractor contour inside a given receptive field thus prevent the corresponding unit to get activated. In our work, this scale-selection process is implemented through the feedforward network, and the filters ensuring that the presence of the distractor in the receptive field inactivate the cell are learnt during training. More recently, Schmidt and Neumann (31) proposed a neural dynamical model of incremental binding. Their model can trace curves in a scale-invariant way and, through ablation studies, they showed that it was due to the interraction of different scales through feedback connections. Both of those studies however only investigated the implementation of scale-invariant tracing, and did not look into how such strategies can be learnt. Our work is the first to demonstrate how multiscale perceptual grouping can be learnt. Moreover, it unifies results obtained in monkeys through curve-tracing tasks, and in human through object-tracing tasks.

We propose that perceptual grouping is implemented in the visual cortex through disinhibitory loops between pyramidal neurons. The features belonging to the attended object are singled out by a tag of enhanced activity, and this representation grows by the iterative propagation of this tag through the removal of baseline inhibition of pyramidal neurons by SOM interneurons. Experimental evidence indicate that disinhibitory loops are essential for the recurrent processing underlying figure-ground modulation (16). Further experiments could test whether the same mechanisms is responsible for the perceptual grouping of coherent objects.

A shortcoming of our architecture is the need to pretrain a feedforward group of neurons. We argue that the features learnt in those networks during the pretraining could be learnt prior to the specific tasks we used, through unsupervised learning for instance. The need to train two feedforward groups, one for the curve-tracing task and one for the object-tracing task is also problematic. Indeed, it appears that for the curve-tracing task the scale used to propagate activity is chosen so that the receptive field of the highest activated neurons only has colinear and connected elements. In the case of the object-tracing task, the highest region with homogeneous texture is chosen. It is unclear how those two conditions could be fused together.

It has been observed in humans experiments that the growth-cone model of attention best described results when the objects over which attentional spreading was happening were only outlines of unrecognizable shapes. When realistic objects were presented, with colour, varied texture and interior contours, the growth-cone model performed poorly. It also seems that grouping speed is influenced by image recognition: perception starts with object classification, which is rapid and feedforward, and then is followed by a serial perceptual grouping phase (32). As such, future work could extend our model and train it on naturalistic stimuli to account for the dynamics of object-based attention in such a setting and understand the influence of object recognition, texture and colour on grouping speed.

## Materials and Methods

### Feedforward network training

The feedforward networks were trained with backpropagation in a supervised fashion. The binary cross entropy loss was used, with the adam optimizer (33) and a learning rate of $10e^{-3}$. Each scale was trained on the same set of stimuli but independently of the others. The input consisted of random curves on a 36x36 grid for the curve tracing task, and of random objects for the object-tracing task. The label could be either one if all the elements inside the receptive field of a neurons were colinear and connected (curve-tracing) or with a homogeneous texture (object tracing task).

### Recurrent network training

The output layer of the network is retinotopically organised, each neuron being associated with one pixel of the input grid. Their activity represents the expected reward if the corresponding pixel is selected to make an eye movement, also called Q-value (34).

During training, the action with the highest Q-value was chosen with 95% chance. In the remaining 5% trials, a random action was sampled from the Boltzmann distribution $P_B$:

$$P_B(a) = \frac{exp(q_a)}{\sum_k exp(q_k)}$$

We trained the networks using RELEARNN (18), a biologically plausible implementation of the Almeida-Pineda algorithm for credit assignment. It can be broken down into three phases. First, input neurons are clamped and activity propagates through the network until convergence to a stable state. An action was selected if activity of neurons between two consecutive timesteps was constant or after 30 timesteps. When an action is selected, an attention signal originating from the winning action is propagated through an accessory network to determine the influence of each neuron on the selected action. Critically, attentional feedback signal is present locally at each synapse, and only depends on the last stable state on the network, not the full history of activations, which makes it local in time. After the end of the attentional signal propagation, the network gets a reward r from the environment and computed a reward prediction error $\delta$ that is determined by:

$$\delta = r - Q_a$$

Where $Q_a$ is the state-action value of the selected action. $\delta$ is broadcasted in the whole network by a neuromodulatory signal (35) and thus available at all synapses. $\delta$ and the attentional feedback signal are then combined to update the weights.

**Pixel-by-pixel model**

In the pixel-by-pixel mode, attention spreads at a constant speed. In the curve-tracing task, the latency at a position A only depends on the arc length between the fixation point and the receptive field:

$$t_{pixel}^{curve}(A) = t_c + \beta_{curve} \cdot L(A)$$

In the object-grouping task, the reaction time depends on the length shortest distance between the cues, measured within the object's interior. The length of the shortest path between the cues was defined as the number of pixels connecting the two cues, where pixels were connected by their 8-neighbourhood.

$$RT_{pixel}^{object} = RT_C + \beta_{object} \cdot L_{shortest}$$

**Growth-cone model**

The growth-cone model is similar to the pixel-by-pixel one, but grouping can now occur at multiple spatial scale.

For the curve-tracing task, the grouping happens at the biggest scale such that their are only colinear and connected elements in the receptive field of the neurons.

For the object tracing task, we followed (10). Here, the model selects the appropriate scale for the growth-cone for every-pixel of the object. It corresponds to the diameter $d_{inscr}$ of the largest inscribed circle centered on that pixel and which fits within the boundaries of the object. To compute the shortest path between the cue and the fixation point, we defined an 8-connected lattice G=(V,E) with vertices (pixels) V connected by edges E where two neighbouring pixels are only connected if they do not cross the object boundary. A path is a set of connected edges $e \in$

$E$. We defined P as the set of all paths that emanate from $p_1 \in V$ and terminate at $p_2 \in V$. An edge e has length $f_{GC}(E)$

Where $f_{GC}(e) = \frac{1}{(d_{inscr}(pix_1)+d_{inscr}(pix_2))/2}$ for horizontally and vertically aligned pixels and $f_{GC}(e) = \frac{\sqrt{2}}{(d_{inscr}(pix_1)+d_{inscr}(pix_2))/2}$ for diagonally aligned pixels. RT is then estimated as follows:

$$RT_{GC} = RT_{Offset} + C \cdot p \in P_{min} \sum_{e \in p} f_{GC}(e)$$

**Human Reaction time fit**

Human reaction time (RT) analysis was done following (10). We only analysed RTs on correct trials for the "same" condition (when the fixation point and the cue were on the same object) after removing outliers if they deviated from the mean of the 1/RT distribution by more than 2.5 standard deviation. We then performed a regression analysis to test how well RTs as predicted by the growth-cone model, our artificial neural network or the pixel-by-pixel model predicted human RTs.

Human RTs were averaged across participant, giving one experimental RT for each image-cue conditions. We performed one regression against those RTs separately for each of the 5 networks. The mean explained variance of the artificial neural network models was tested against the explained variance of the growth-cone model using a one sample t test.

**Acknowledgments**

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3, Task 3.7), an ERC grant (101052963 "NUMEROUS"), an NWO Crossover grant 17619 "INTENSE", "DBI2" (a Gravitation program of the Dutch Ministry of Science) and NWO-OCENW.KLEIN.178.
We acknowledge the use of Fenix Infrastructure resources, which are partially funded from the European Union's Horizon 2020 research and innovation programme through the ICEI project under the grant agreement No. 800858.

**References**

1. P. R. Roelfsema, V. A. F. Lamme, H. Spekreijse, Object-based attention in the primary visual cortex of the macaque monkey. *Nat. 1998 3956700* **395**, 376–381 (1998).

2. K. Koffka, Principles of gestalt psychology. *Princ. Gestalt Psychol.*, 1–720 (2013).

3. M. Wertheimer, Untersuchungen zur Lehre von der Gestalt. II. *Psychol. Forsch. 1923 41* **4**, 301–350 (1923).

4. J. R. Bergen, B. Julesz, Parallel versus serial processing in rapid pattern discrimination. *Nat. 1983 3035919* **303**, 696–698 (1983).

5. P. Jolicoeur, S. Ullman, M. Mackay, Curve tracing: A possible basic operation in the perception of spatial relations. *Mem. Cognit.* **14**, 129–140 (1986).

6. A. Pooresmaeili, P. R. Roelfsema, A growth-cone model for the spread of object-based attention during contour grouping. *Curr. Biol.* **24**, 2869–2877 (2014).

7. P. R. Roelfsema, Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.* **29**, 203–227 (2006).

8. P. R. Roelfsema, R. Houtkamp, Incremental grouping of image elements in vision. *Attention, Perception, Psychophys.* **73**, 2542–2572 (2011).

9. P. A. McCormick, P. Jolicoeur, Predicting the shape of distance functions in curve tracing: Evidence for a zoom lens operator. *Mem. Cognit.* **19**, 469–486 (1991).

10. D. Jeurissen, M. W. Self, P. R. Roelfsema, Serial grouping of 2D-image regions with object-based attention in humans. *Elife* **5** (2016).

11. S. Mollard, C. Wacongne, S. M. Bohte, P. R. Roelfsema, Recurrent neural networks that learn multi-step visual routines with reinforcement learning. *bioRxiv*, 2023.07.03.547198 (2023).

12. T. Van Kerkoerle, M. W. Self, P. R. Roelfsema, Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nat. Commun. 2017 81* **8**, 1–14 (2017).

13. A. Pooresmaeili, J. Poort, A. Thiele, P. R. Roelfsema, Separable Codes for Attention and Luminance Contrast in the Primary Visual Cortex. *J. Neurosci.* **30**, 12701–12711 (2010).

14. M. W. Self, T. van Kerkoerle, H. Supèr, P. R. Roelfsema, Distinct roles of the cortical layers of area V1 in figure-ground segregation. *Curr. Biol.* **23**, 2121–2129 (2013).

15. J. D. Semedo, *et al.*, Feedforward and feedback interactions between visual cortical areas use different population activity patterns. *Nat. Commun. 2022 131* **13**, 1–14 (2022).

16. L. Kirchberger, *et al.*, The essential role of recurrent processing for figure-ground perception in mice. *Sci. Adv.* **7** (2021).

17. D. Sridharan, E. I. Knudsen, Selective disinhibition: A unified neural mechanism for predictive and post hoc attentional selection. *Vision Res.* **116**, 194–209 (2015).

18. T. Brosch, H. Neumann, P. R. Roelfsema, Reinforcement Learning of Linking and Tracing Contours in Recurrent Neural Networks. *PLoS Comput. Biol.* **11**, e1004489 (2015).

19. H. Adeli, S. Ahn, N. Kriegeskorte, G. Zelinsky, Affinity-based Attention in Self-supervised Transformers Predicts Dynamics of Object Grouping in Humans (2023) (March 21, 2024).

20. D. Linsley, A. K. Ashok, L. N. Govindarajan, R. Liu, T. Serre, “Stable and expressive recurrent vision models” (2020) (March 29, 2021).

21. H. J. Pi, *et al.*, Cortical interneurons that specialize in disinhibitory control. *Nat. 2013 5037477* **503**, 521–524 (2013).

22. L. Kiorpes, S. A. Bassin, Development of contour integration in macaque monkeys. *Vis. Neurosci.* **20**, 567–575 (2003).

23. J. Atkinson, O. Braddick, Visual segmentation of oriented textures by infants. *Behav. Brain Res.* **49**, 123–131 (1992).

24. A. Norcia, V. Sampath, C. Hou, F. Pei, Contour Integration in Human Infants. *Invest. Ophthalmol. Vis. Sci.* **43**, 3993–3993 (2002).

25. A. Needham, Infants' use of featural information in the segregation of stationary objects. *Infant Behav. Dev.* **21**, 47–76 (1998).

26. D. M. Werchan, A. G. E. Collins, M. J. Frank, D. Amso, 8-Month-Old Infants Spontaneously Learn and Generalize Hierarchical Rules. *Psychol. Sci.* **26**, 805–815 (2015).

27. B. A. Richards, *et al.*, A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).

28. A. Pooresmaeili, J. Poort, P. R. Roelfsema, Simultaneous selection by object-based attention in visual and frontal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6467–6472 (2014).

29. P. R. Roelfsema, F. P. de Lange, Early Visual Cortex as a Multiscale Cognitive Blackboard. *Annu. Rev. Vis. Sci.* **2**, 131–151 (2016).

30. D. Domijan, M. M. Mari´c, A multi-scale neurodynamic implementation of incremental grouping. *Vision Res.* **197**, 108057 (2022).

31. D. Schmid, H. Neumann, Thalamo-Cortical Interaction for Incremental Binding in Mental Contour-Tracing. *bioRxiv*, 2023.12.20.572705 (2023).

32. I. Korjoukov, *et al.*, The Time Course of Perceptual Grouping in Natural Scenes. *https://doi.org/10.1177/0956797612443832* **23**, 1482–1489 (2012).

33. D. P. Kingma, J. L. Ba, Adam: A Method for Stochastic Optimization. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* (2014) (June 6, 2023).

34. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction, Second Edition* (2018).

35. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science (80-. ).* **275**, 1593–1599 (1997).