# Modeling Attention and Binding in the Brain through Bidirectional Recurrent Gating

**Saeed Salehi[1], Jordan Lei[2, ‡], Ari S. Benjamin[3, ‡], Klaus-Robert Müller[1, 4, 5, 6], and Konrad P. Kording[7, 8, *]**

[1]Machine Learning Group, Technical University of Berlin, Germany
[2]New York University, New York, NY, USA
[3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[4]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany
[5]Department of Artificial Intelligence, Korea University, Seoul, Korea
[6]Max Planck Institute for Informatics, Saarbrücken, Germany
[7]Department of Bioengineering, Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA
[8]CIFAR LMB Program
[‡]work done while at University of Pennsylvania
[*]corresponding author: *kording at upenn dot edu*

## ABSTRACT

Attention is a key component of the visual system, essential for perception, learning, and memory. Attention can also be seen as a solution to the binding problem: concurrent attention to all parts of an entity allows separating it from the rest. However, the rich models of attention in computational neuroscience are generally not scaled to real-world problems and there are thus many behavioral and neural phenomena that current models cannot explain. Here, we propose a bidirectional recurrent model of attention that is inspired by modern neural networks for image segmentation. It conceptualizes recurrent connections as a multi-stage internal gating process where bottom-up connections transmit features, while top-down and lateral connections transmit attentional gating signals. Our model can recognize and segment simple stimuli such as digits as well as objects in natural images and is able to be prompted with object labels, attributes or locations. It can learn to perform a range of behavioral findings, such as object binding, selective attention, inhibition of return, and visual search. It also replicates a variety of neural findings, including increased activity for attended objects, features, and locations, attention-invariant tuning, and relatively late onset attention. Most importantly, our proposed model unifies decades of cognitive and neurophysiological findings of visual attention into a single principled architecture. Our results highlight that the ability to selectively and dynamically focus on specific parts of stimulus streams can help artificial neural networks to better generalize and align with human brains.

## Introduction

*Attention* is a mechanism by which the brain selects a meaningful subset of incoming stimuli for perception, learning, or memory [Carrasco, 2011; Hommel et al., 2019; Itti et al., 2005; Lindsay, 2020; Maunsell, 2015; Moore and Zirnsak, 2017; Tsotsos, 2021]. By focusing on a subset of the world, attention promises a more localized, sparse, and less noisy signal in response to sensory input [Posner et al., 1980]. For learning, focusing on the relevant subset of stimuli (i.e., a target) leads to better generalization [Rutishauser et al., 2004; Walther et al., 2005]. In the case of memory, sparser and cleaner signals could be reused more often in the stimulus set (e.g., objects or object parts) and thus be more efficiently stored, recalled, and compositionally utilized [Kruijne et al., 2021; Olivers and Roelfsema, 2020]. Hence, attention plays an indispensable role in neural information processing and understanding attention is critical to advancing our comprehension of learning, memory and perception.

What makes attention difficult to model is the widespread nature of the various phenomena it includes. Attention appears on different scales and through different mechanisms [Buschman and Kastner, 2015; Itti et al., 2005; Maunsell, 2015; Moore and Zirnsak, 2017; Posner, 2016]. One frequently studied form is spotlight attention, where focus is directed towards a region in the environment, enhancing perception and processing of information in that region [Itti and Koch, 2001; Wolfe and Horowitz, 2004]. Behavioral experiments show that perception in such attention foci is faster and more reliable [Chun and Wolfe, 2005; Posner, 1980]. Another frequently studied form is feature-based attention where, as with space, knowledge about features of relevant stimuli speeds up perception and lowers error rates [Bichot et al., 2015; Treisman, 1998; Treisman and Gelade, 1980]. A third form of attention, arguably related to feature-based attention, is object-based attention, where attention is bound to an object as a whole, including its spatial and visual features [Baldauf and Desimone, 2014; Behrmann et al., 1998; Scholl,

2001; Treisman, 1998]. A fundamental question here is how an object that is represented by activity across millions of neurons and many cortical regions can be perceived and attended to as a single object [Von der Malsburg, 1999]. This is known as the binding problem, asking how the brain can integrate and merge all of the activities relating to the same entity and how to separate them from all the other entities [Feldman, 2013; Robertson, 2003; Singer, 2001; Treisman, 1996; Von der Malsburg, 1995]. Hence, understanding object-based attention is a crucial step in solving the feature binding problem [Roelfsema, 2023; Treisman, 1998]. Thus, neuroscientists and psychologists have long studied the relationship between attention and binding and sought to build a unifying model of these phenomena.

Despite their theoretical elegance, typical models of biological attention lack one or more elements that are crucial for building a comprehensive model of object-based attention [Cavanagh et al., 2023; Posner, 2023]. The first element is the top-down process, where activity in higher-level brain areas modulates the neural activity in lower-level cortices [Paneri and Gregoriou, 2017; Thiele and Bellgrove, 2018]. The second element is lateral processes and communication [Han et al., 2005; Roelfsema et al., 1998]. A third element is recurrence, where activity travels in all directions simultaneously and involves repeated processing of information over time [Hochreiter and Schmidhuber, 1997]. Indeed, there is strong experimental evidence for the importance of recurrence in binding and attention [Kar et al., 2019; Kietzmann et al., 2019; O'Reilly et al., 2013; Roelfsema, 2023; van Bergen and Kriegeskorte, 2020], and has led to considerable interest in recurrent models for attention [Kubilius et al., 2019; Larsen and Druckmann, 2022; Linsley et al., 2020; Mnih et al., 2014]. For all the richness of the existing literature, we are still lacking a model that integrates all of the aforementioned core elements, is capable of handling natural stimuli, and at the same time describes the many known findings in the field.

Attention has also been studied and utilized in machine learning, albeit often under different interpretations of the concept. Among them, the transformer architecture stands out as the most popular model bearing the name of attention. Transformers [Vaswani et al., 2017] use self-attention to dynamically create a sparse and weighted combination of input feature representations. But despite their immense success and ability to learn powerful representations [Caron et al., 2021; Dosovitskiy et al., 2020; Kirillov et al., 2023], transformers lack the three key elements of attention: lateral, top-down, and recurrent activities, which restricts their biological plausibility and limits their ability to solve tasks that require understanding of multiscale features to produce precise local binding. Similarly, other deep learning models of attention either lack recurrency [Kaul et al., 2019], or top-down control and internal gating [J. Ba et al., 2014; Locatello et al., 2020; Mnih et al., 2014; Nayebi et al., 2018], or use decoupled mechanisms for feature learning and attention [Stollenga et al., 2014]. Nonetheless, deep learning has been remarkably successful in object recognition, segmentation, and representation learning [Bengio et al., 2013; LeCun et al., 2015; Ronneberger et al., 2015] and thus promises new approaches for neuroscientists to build effective and generalizable models of visual attention [Richards et al., 2019].

Here, we propose a deep learning model of biological attention, by means of *bottom-up*, *top-down*, and *lateral connections* operating through a *bidirectional recurrent gating* mechanism. Our model is inspired by the emerging understanding of object-based attention, while borrowing elements and techniques from state-of-the-art machine learning. We base our model on the U-Net architecture [Ronneberger et al., 2015], while incorporating *residual connections* [He et al., 2016], *divisive normalization* [J. L. Ba et al., 2016; Reynolds and Heeger, 2009], attention-driven *neuromodulation*, and task prompts as explicit input. We show that our model can learn to perform a broad range of attention tasks, including but not limited to: object recognition, segmentation, spatial binding, inhibition of return, border ownership, and top-down visual search. Additionally, we illustrate the importance of recurrent attention in solving feature-binding problems using a pathological dataset and and its extension to real-world scenarios. We later scale up the model to handle attention tasks on complex natural stimuli, such as those from the COCO dataset. Finally, we demonstrate how our model describes known neurophysiological phenomena, including attention-invariant tuning, attention-mediated neuromodulation, and late-onset attention, effectively unifying a broad range of behavioral and neural phenomena related to object-based attention.

## Results

We aimed to develop a biologically motivated model that could simultaneously handle real-world naturalistic stimuli and solve benchmark tasks commonly used by cognitive scientists and neuroscientists to study vision and attention in animals and humans. Our architecture consists of a feedforward contracting *Feature* pathway that hierarchically processes and extracts learned features, alongside a top-down expansive *Attention* pathway that combines lateral and top-down connections to implement task- and stimulus-dependent bidirectional recurrent gating mechanism (Fig. 1 a, see Supplementary materials). The lateral connections form a long-range bidirectional recurrent network, while the dense recurrent layers provide the model with working memory and thus, enabling it to perform memory-dependent tasks and behaviors such as inhibition of return. Since recurrency is central to our architecture, the model expects a sequence of images as its input stimulus. Correspondingly, it outputs a sequence of attention maps and label predictions, generated sequentially in each iteration. Besides the input stimuli, the network can also accept additional inputs, such as the task index for multitask learning and prompts for top-down tasks. Altogether, our model can effectively learn to perform multiple tasks on the same architecture (Fig. 1 b).

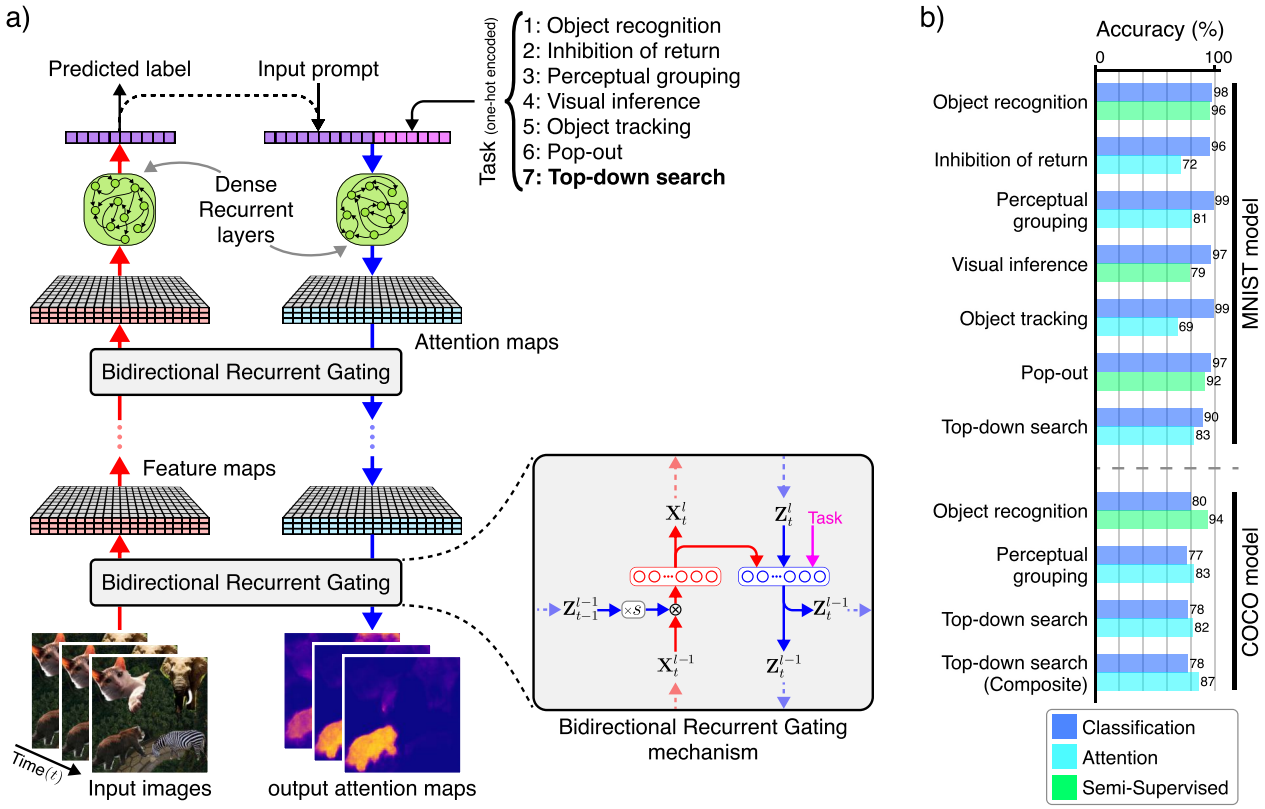


**Figure 1. Model architecture and multi-task learning. a)** The core block of our model is bidirectional recurrent gating. In this figure, we use color-coding to denote the bottom-up and top-down information streams and processes, with red representing bottom-up and blue representing top-down. Subscripts indicate the corresponding time (i.e., iteration) and superscripts denote the corresponding layer. For our experiments, we use a network constructed by stacking and weaving the main computational block, incorporating common computer vision elements for processing and learning. The network comprises of two main pathways: the *Feature* pathway, which hierarchically extracts learned feature representations (shown in red); the *Attention* pathway, which combines top-down information, the task signal, and feature maps to generate attention maps (shown in blue). The attention maps consequently and multiplicatively modulate the feature maps of the next iteration. The two pathways meet at the bottleneck, which incorporates the dense recurrent and linear layers and outputs estimated label. In addition to the bottleneck, the two pathways communicate feature maps and attention maps through the lateral connections. **b)** The architecture enables effective multi-task learning. Here we show the classification and attention accuracy for two distinct models. For some tasks, we use semi-supervised training, meaning that for some tasks we do not provide target attention maps (i.e., segmentation signal) during training.

We use stochastic gradient descent [Rumelhart et al., 1986] to train the model on a combination of two objectives: (a) the network should be able to classify objects, for which we use Cross-Entropy loss, and (b) the network should be able to segment objects in the input layer, for which we use Mean-Squared-Error loss. Although at the first glance these two goals seem insufficient for learning complex tasks, we will demonstrate that our model can leverages the multitask paradigm to effectively learn semi-supervised tasks and attend to the correct visual features. During inference, we measure the accuracy, both for classification and attention maps, and pixel error for the final iteration (see supplementary materials).

## Behavioral results and multitask learning

Visual attention encompasses an extensive range of behavioral findings [Itti et al., 2005], including: a) object recognition in a cluttered scene, b) spatial cueing of attention and perceptual grouping, c) feature saliency and pop-out, d) top-down control, e) inhibition of return, and f) pre-attentive features, to name a few. In what follows, we design a series of multitask experiments to examine whether our network is able to replicate these known behavioral findings. The first experiment (arguably using toy data) is performed on compositions of handwritten digits from the MNIST dataset [LeCun, 1998] with various aspects extended to enable a broad set of tests. The second experiment is conducted on a rather complex set of stimuli based on the COCO dataset [Lin et al., 2014], though on fewer tasks.

Object recognition is arguably one of the primary objectives of the visual system, which requires separating the target object from the background and foreground and classifying it correctly [Behrmann et al., 1998; DiCarlo et al., 2012]. The foreground often obscures parts of the object, leading to partial visibility of the object and ambiguity of the edges [Fang et al., 2009; Super et al., 2010]. Moreover, the visual system tends to maintain a stable internal representation of the object even when the foreground or background moves. To test whether the model can solve these problems, we designed an object recognition and permanence task, where a digit is placed on top of a background and partially occluded by a foreground (Fig. 2 a). Both the background and foreground are randomly generated correlated noise that move slowly across the image, while the object remains stationary. Training for this task includes only the target label, requiring the network to learn to uncover the correct attention map without explicit supervision. Our model achieves 98% classification accuracy and pixel error of 0.01 in recovering the correct attention map in the final iteration on the test dataset. This demonstrates that the model can effectively solve object recognition, segmentation and permanence tasks in the presence of significant levels of correlated noise and occlusion.

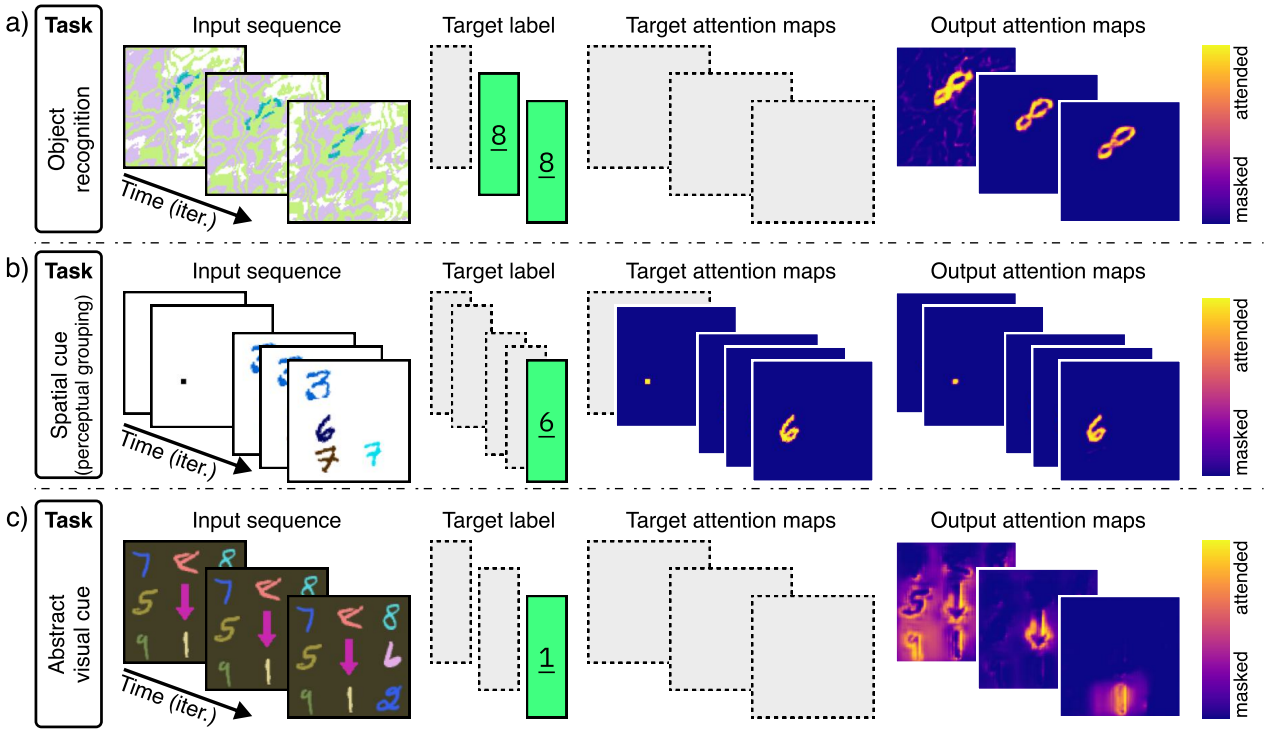


**Figure 2. Multitask training on MNIST composites (Part 1).** Results for a single model trained on multiple tasks simultaneously. For each task, from left to right: the input sequence, target labels, target attention masks, and output (estimated) attention maps are shown. For the target label and target attention map, the gray dashed boxes imply that the target is NOT provided during training. **a)** *object recognition*, **b)** *spatial cueing*, and **c)** *visual inference*. Colors of task c input sequences have been inverted for better visualization.

Spotlight attention is a well studied model system for attention [Chun and Wolfe, 2005; Itti and Koch, 2001]. A common element in cognitive visual experiments is to use a cue to spatially orient the attention to a region and task the subject to fixate on the cue and follow the instructions. Such cues could be also used to guide perceptual grouping [Herzog, 2018; Roelfsema, 2006]. Therefore, we tested our model to see if it could also learn where to direct its attention using spatial cues (Fig. 2 b). This was crucial as we intended to use a similar approach in the switchbox attention task to specify the target and distractor. Here, we trained the network to attend to a dot in the first image in the sequence for two iteration. The sequence is then followed by a random arrangement of few digits, and the network must attend to the digit that is contiguous with the previously shown cue. Our model achieves 99% classification accuracy and pixel error of 0.002 in recovering the correct attention map in the final iteration on the test dataset. These results demonstrate that our model possesses spotlight attention capabilities comparable to human subjects in such tasks.

Apart from spatial cues, we also use abstract visual cues to direct our attention to an object in the scene. The most prominent example would be pointing with the index finger. We have learned to first look at the finger, recognize the direction in which it is pointing towards, and then to orient our gaze and attention to the object in line with the finger's direction. Therefore, we trained our model on such task, where 8 digits are arranged around an arrow and the goal is to locate and classify the target

digit. During training, only information about the target digit, to which the arrow is pointing at, is provided (Fig. 2 c). It is important to note that since NO attention map for the arrow is given, the model has to learn about the arrow only through the input image and the target digit. Our model achieves 97% in classification accuracy and pixel error of 0.016 in recovering the correct attention map in the final iteration on the test dataset. Interestingly, we can see from the attention maps that the model has learned to attend to the arrow first, and infer the location of the target digit. In addition, the model is using its knowledge acquired from other tasks that digits are separate entities from the arrows, and therefore avoids learning the combination of all the arrow directions and digits, resulting in a more robust behavior. Thus, the system can learn about and attend to an abstractly cued location and use multi-task paradigm for better generalization.

Salient features in a scene are thought to be the main driving factors in bottom-up attention [Itti et al., 1998; Tsotsos et al., 1995; Wolfe, 2005], although the neural basis of this phenomenon is still debated [Einhäuser and König, 2003; Nothdurft, 2006; Parkhurst et al., 2002]. We have implemented a task where the model is presented with a grid of 9 digits, where all but one are of the same instance and color ((Fig. 3 a)). The model is trained to return the class of the digit that is different from the rest, without knowing the target attention map. The network learns to detect and classify the salient digit with 97% classification accuracy and pixel error of 0.024 in recovering the correct attention map, thus indicating that our network can learn and perform saliency-dependent pop-out tasks.

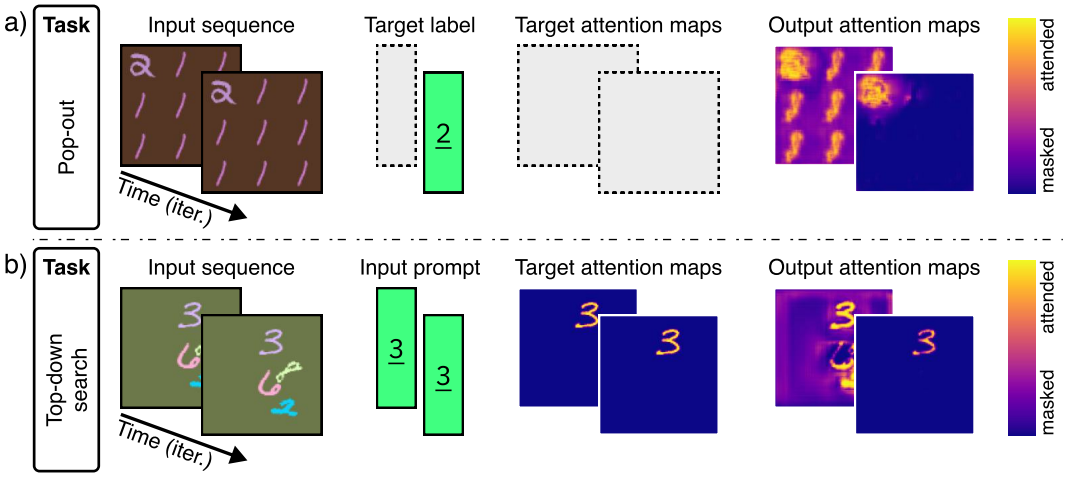


**Figure 3. Multitask training on MNIST composites (Part 2).** Results for a single model trained on multiple tasks simultaneously. For each task, from left to right: the input sequence, target labels or input prompts, target attention masks, and output (estimated) attention maps are shown. For the target label and target attention map, the gray dashed boxes imply that the target is NOT provided during training. **a)** *pop-out saliency*, and **b** *top-down visual search*.

So far, attention has been guided by the input visual stimulus mediated by the task. However, it is often the case that we are given an explicit non-visual cue or description (i.e., prompt) of what to search for in a cluttered scene. Since the prompt is the class or other distinct attributes of the object, the desired output of such tasks is the location of the prompted object [Wolfe, 2010]. Despite the many kinds of visual search, here we focus on visual search through top-down attention. To demonstrate how our model can perform top-down search, we use an input composition of digits and an input class prompt that is fed to the model from the bottleneck (i.e., top) (Fig. 3 b). The model is trained to attend to the prompted digit in the image. Our model achieves 90% classification accuracy and pixel error of 0.002 in recovering the correct attention map in the final iteration. Although here we restricted our results to only the class prompt, we have also successfully experimented with other search attributes such as color and texture (see see Supplementary materials).

A model of attention should be able to attend to an object as it moves through the scene. Spatial attention and awareness of the object's location has been shown to be an important part of the binding process [Robertson, 2003]. Here, we will train the network to tracking and classify a moving digit (Fig 4 a). The task is to fixate on and classify a stationary single digit for two iterations, and then to track the target digit as it begins to move through the scene with multiple distractor digits. Training is fully supervised with target labels and masks provided during training, and achieves 99% classification accuracy and pixel error of 0.009 in recovering the correct attention map on the test dataset. Although the network was trained for only 7-iterations of tracking, it can track the target digit for a longer period of time during inference, indicating that the network can learn to track an object through space and time.

When humans attend to multiple objects in a visual scene, they typically do so sequentially, moving from one object to the next [Klein and Ivanoff, 2008]. A large body of experimental evidence supports the idea of inhibition of return (IOR), whereby

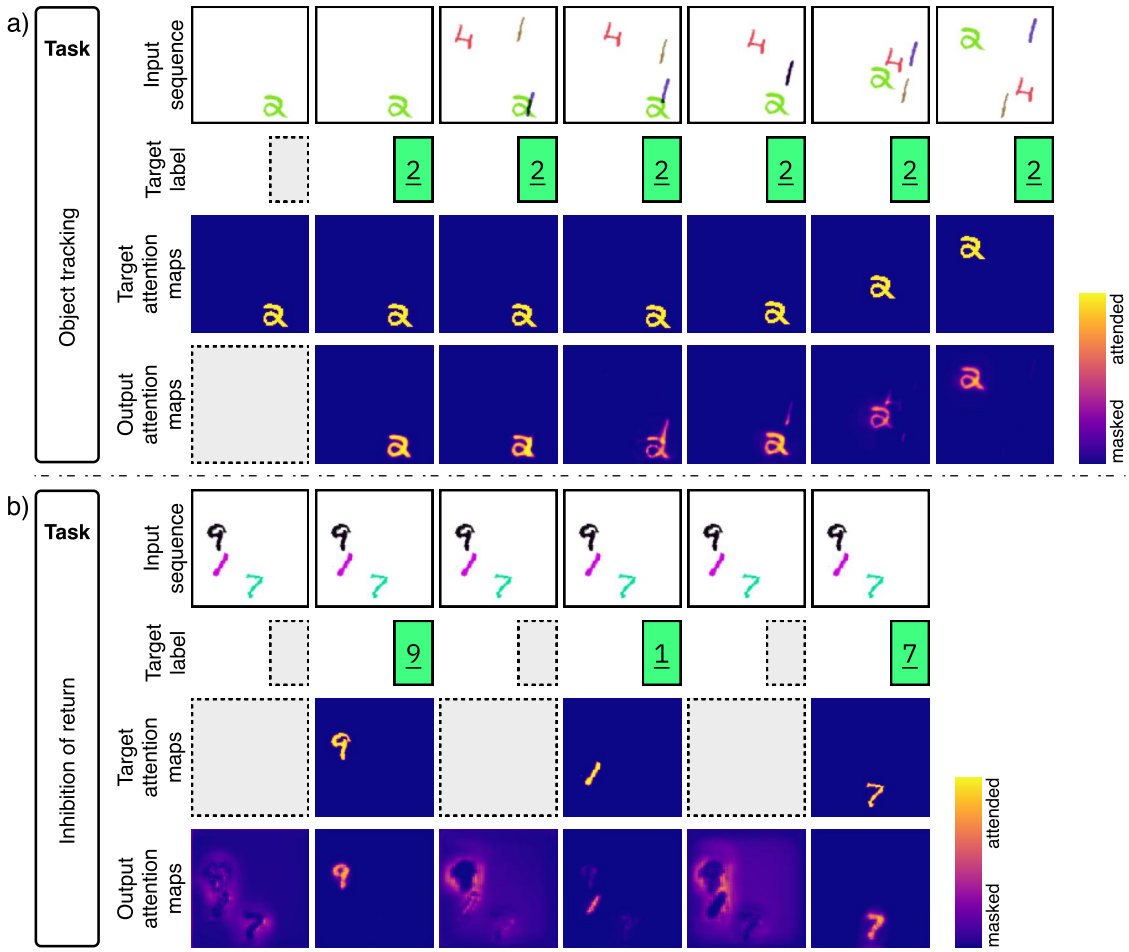


**Figure 4. Multitask training on MNIST composites (Part 3).** Results for a single model trained on multiple tasks simultaneously. For each task, from left to right: the input sequence, target labels, target attention masks, and output (estimated) attention maps are shown. For the target label and target attention map, the gray dashed boxes imply that the target is NOT provided during training. The colors of the input sequences have been inverted for better visualization. **a)** *spatial binding*, and **b)** *inhibition Of Return (IOR)*

recently attended regions of a visual input are eventually inhibited, allowing the subject to move from one object to the next [Itti and Koch, 2001; Klein, 2000; Kwak and Egeth, 1992; Posner and Cohen, 1984]. Without inhibition of return, a subject will remain fixated on the most salient object. Here, we demonstrate that our model can learn to perform IOR on a multi-digit composite using only implicit training (Fig. 4 b). Our experiment is as follows: A sequence of identical compositions of $n$ augmented digits from the MNIST dataset is fed into the model. The network learns to arbitrarily select, attend to and classify any of the digits in $k$ iterations. It must then move on to the next digit, and avoid returning to previously visited digits. We trained the network fully supervised simultaneously on $n = 1, 2,$ and 3 objects with $k = 2$ iterations per digit. Thus, the network can readily cycle through objects in a multi-object scene. Our model achieves 96% accuracy in classifying all the digits in the scene and pixel error of 0.0042 in recovering the correct attention for every digit.

### *Bregman's illusion*

Shared perceptual illusions between humans and artificial neural networks can inform us about the similarity of the underlying processes that are otherwise inaccessible to us [Benjamin et al., 2019]. Here, we want to investigate whether the model can account for well-known perceptual illusions for which its behavior is not explicitly predefined. Researchers have long known about the existence of perceptual illusions, such as the Bregman illusion, in which occlusion can provide cues that influence object perception. In the Bregman illusion, a stimulus such as a letter is occluded by a "splotchy" distractor (i.e., occlusion) in the foreground. When the distractor is removed, leaving an empty region, the stimulus appears fragmented and difficult to identify. However, when the distractor is added, the fragments are contextualized as objects occluded by a foreground distractor, making the entire scene more interpretable. The Bregman illusion is a classic demonstration of border

ownership in Gestalt psychology and perception [Bregman, 2017; Qiu et al., 2007; Williford and Heydt, 2013].

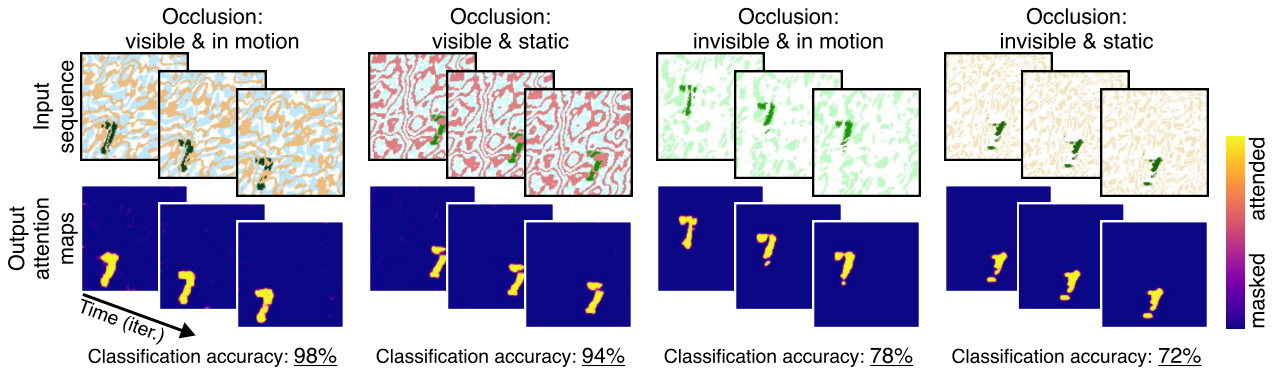


**Figure 5. Bregman's illusion.** Bregman's illusion is commonly used to demonstrate the concept of border ownership. The idea of the illusion is that presence of visible occlusion can inform us about the target boundaries, and by hiding the distractor from the scene we are also removing information about the target object. In this figure, we show that our MNIST model does produce similar behavior, under which the classification accuracy drops if the foreground distractor (i.e., occlusion) is invisible.

To investigate this illusion, we start with the aforementioned model trained on the MNIST tasks. It is important to note that we did not explicitly train or fine-tune the model on this illusion. To assess the emergence of Bregman's illusion, we only adjust the stimuli used in the object recognition task. In the original task, the foreground and background were moving and the model could slowly build a better representation of the object by integrating the incoming information. Therefore we expect the classification and attention accuracy to suffer when the foreground is made stationary. The question is whether removing the foreground and its overlap with the target object would result in a change in classification and attention accuracy. Our analysis shows that the network does indeed perform better at recognition task when the occlusion visible in the input image rather than when it is invisible, which is consistent with the experimental findings of Bregman's illusion (Fig. 5).

Although the MNIST dataset provides us with the freedom to design diverse tasks to evaluate the range of our model, the resulting compositions are far from those we encounter in nature. We chose the COCO dataset because it contains fairly complex stimuli but also the true segmentation maps. Unfortunately, the COCO dataset does not give us the flexibility of MNIST to explore the limits of the architecture, as it was curated primarily for scene-object segmentation. Nevertheless, we trained a single model on object recognition, top-down visual search, and spotlight attention (Fig. 6) on the animal subset of COCO. Since most images contain only a single animal, or multiple animals of the same species, we also used a composition of four distinct animals with natural background images taken from the BG-20k dataset [Li et al., 2022] for the top-down visual search (Fig. 6 c). For the object recognition task, we again used the BG-20k dataset, with a single animal randomly placed on the image, while the background moves slightly in each iteration (Fig. 6 a).

## Feature Attention and masking

One challenge in object recognition is the presence of spurious low-level features in the scene. For example, most waterbirds are photographed near a lake or sea, which adds a blue background as a feature to the images. Studies in both human and artificial networks show that pre-attentive features like texture and color can lead to incorrect feature integration and misclassification [Geirhos et al., 2020; Geirhos et al., 2018; Lapuschkin et al., 2019]. Attention can help separate spurious and true correlations, by masking the irrelevant features and amplifying the informative attributes. To evaluate this theory, we use the CelebA dataset which contains images of celebrity faces including 40 binary attributes such as hair color and sex per image [Liu et al., 2015]. The images are cropped such that the facial landmarks are mostly approximately aligned across the dataset. Among the attributes in this dataset, there are some highly correlated features; namely, most celebrities with blonde hair color are labeled as female (Fig. 7 a). Therefore a naively trained neural network could use the blonde hair color as a low-level proxy for sex [Sagawa et al., 2019]. This makes the CelebA an interesting dataset to test whether our model can learn to attend to the right features for classification.

To this end, we train our model on CelebA for a binary sex classification on sequences of length two, with NO target mask provided and without using class-weights to battle the class-imbalance. The results suggest that the network does not learn to remove the hair information through the feature path, but rather uses attention to integrate the relevant correct features while masking out the hair to reach high performance (Fig. 7 b). Interestingly, the attention maps show that the network has learned to mask the hair, which includes the low-level spurious features, and to attend to the face which contains the features expected for sex classification (Fig. 7 c and d). This could have important implications for how networks should learn to deal with

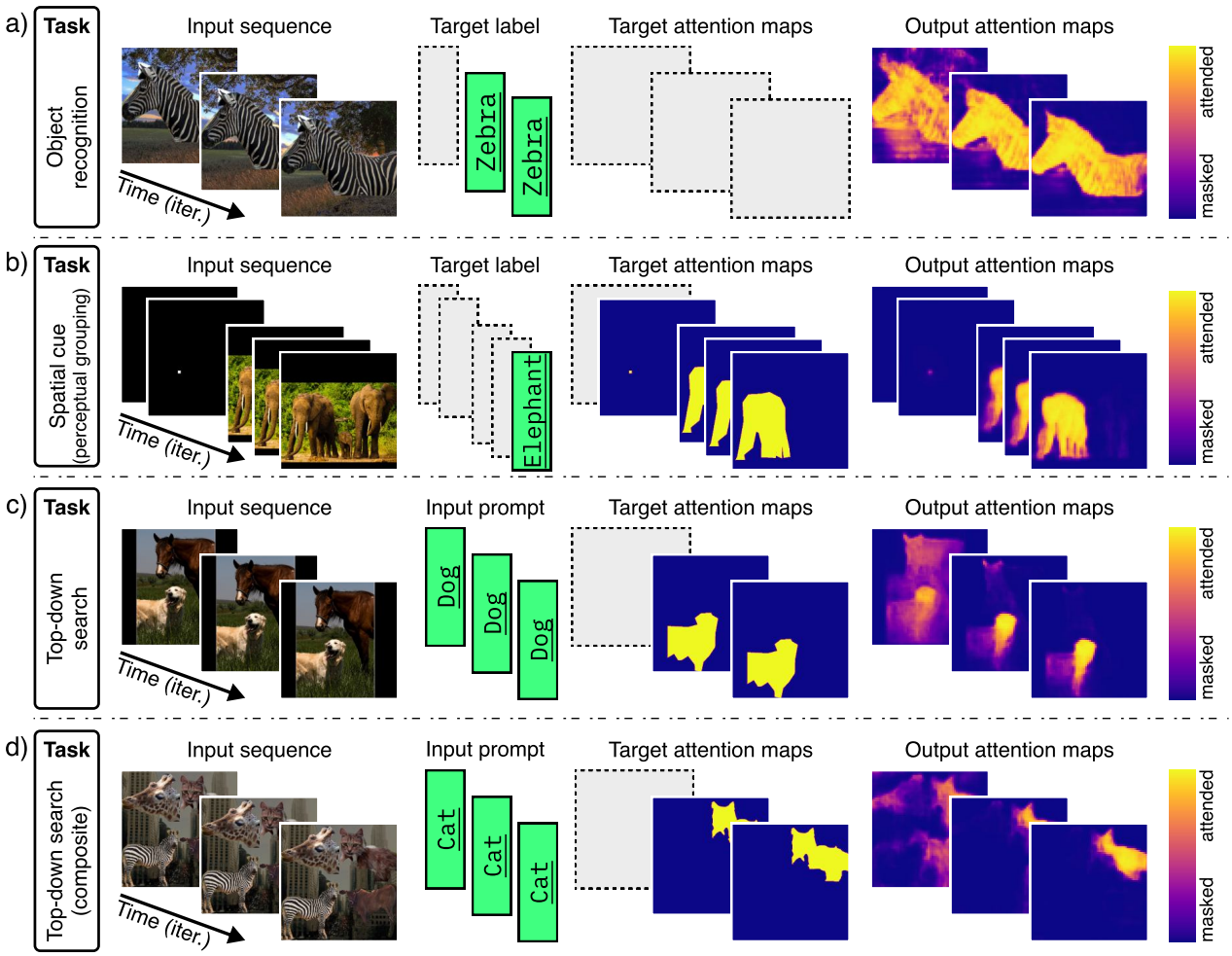


**Figure 6. multitask training on COCO** Results for a single model trained on multiple tasks simultaneously. For each task, from left to right: the input sequence, target labels, target attention masks, output (estimated) attention maps, and classification and attention accuracy on the test data are shown. For the target label and target attention map, the gray dashed boxes imply that the target is NOT provided during training. The colors for the input sequences have been inverted to ease the visualization. **a)** *object recognition*, **b** *cued perceptual grouping*, **c)** *top-down visual search*, and **d)** *top-down visual search on compositional images*.

spurious correlations, since removing the spurious features has been shown to be problematic [Khani and Liang, 2021].

## Neurophysiological Results

### *Object based attention*

Inspired by the work of Roelfsema et al. [Roelfsema et al., 1998], we tested our model's ability to perform object-based attention using switch boxes. A switch box is a technique used to evaluate whether a subject can infer the existence of an object from a visually ambiguous scene. Here, an object is defined as a continuous and regular curve. In the original switch box experiment, two curves were drawn that could either intersect or not. A cue was presented to determine the curve of interest and the subject was asked to trace the object with saccades. In our version of the switchbox experiment, we use contiguity as a cue for the model to decide which of the objects to attend to. The task is as follows: a cue is shown to the model for 2 iterations, followed by 3 iterations of two continuous and smooth curves randomly drawn from a third-order Bezier curve generator. One of the two curves is the target object, overlapping the previously shown cue, while the other curve is the distractor. Finally, the two curves disappear and two dots appear. The task is to attend to the dot that overlaps with the target object (Fig. 8).

An advantage of this experiment is that for any combination of two curves, we can use the initial cue to switch the target and distractor lines, giving us the control we require for the experiment. Not only is our model able to learn and perform the task correctly; i.e., it can learn contiguity, continuity, and regularity, but also its neural activity is consistent with the findings

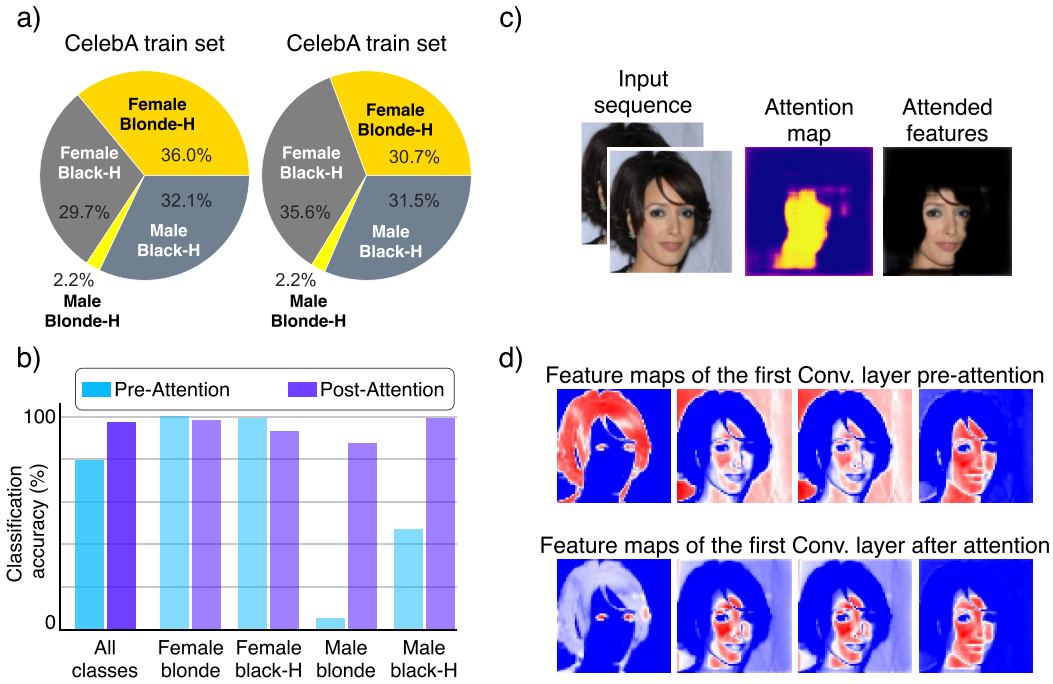


**Figure 7. Feature Attention and masking. a)** CelebA dataset contains a strong spurious correlation between hair color and sex. This spurious feature could be exploited by naive classifiers as a shortcut for sex. **b)** Classification accuracy per class and feature. Pre-attention results indicate that our model has not learned to correctly classify the expected sexes through the feature path alone, but rather uses recurrent attention to integrate the correct features. **c)** Purely trained on sex classification, our model has learned to attend to the facial features and ignore the hair and rest of the input. **d)** Feature maps in the forward path before and after attention.

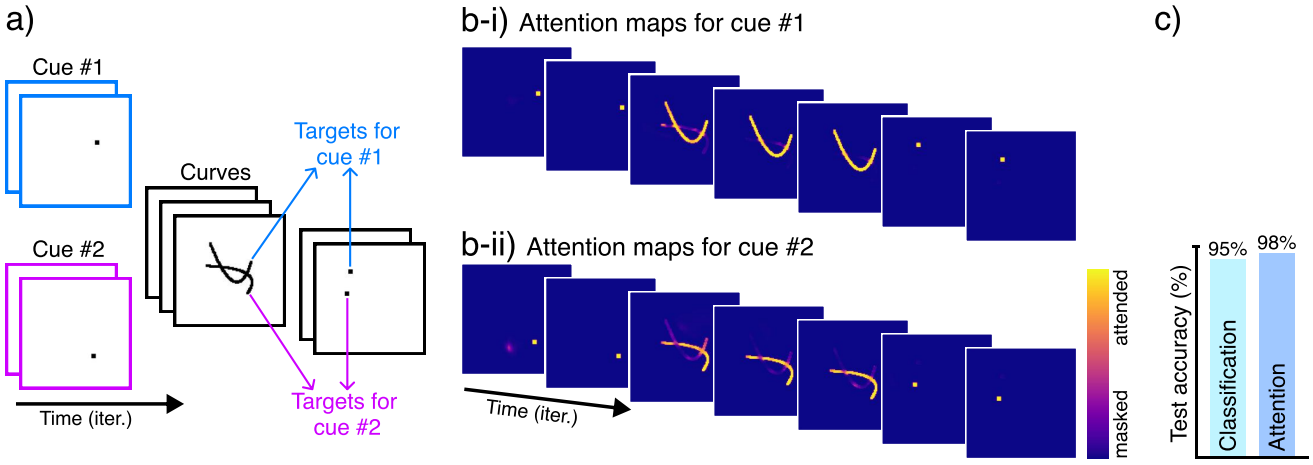


**Figure 8. Object-based attention. a)** The input sequence for the switchbox experiment consists of 3 phases: i) fixation (i.e., visual cueing) for 2 iterations; ii) stimulus (i.e., curve tracing) for 3 iterations; iii) saccade (i.e., attending to the target decision point) for 2 iterations. For any two curves and decision points, we can use two distinct cues to dictate which curve and decision point are the targets. **b-i)** and **b-ii)**) estimated attention maps for two sequences that share the same curves and decision points but different cues. **c)** Classification accuracy for the correct decision point and attention map accuracy for the target curve.

from the primary visual cortex of the macaque monkey performing a similar experiment [Roelfsema et al., 1998] (Fig. 9). We can show that attention positively modulates the activity of neurons with the target line in their receptive field, compared to when the same object is the distractor (i.e. unattended).

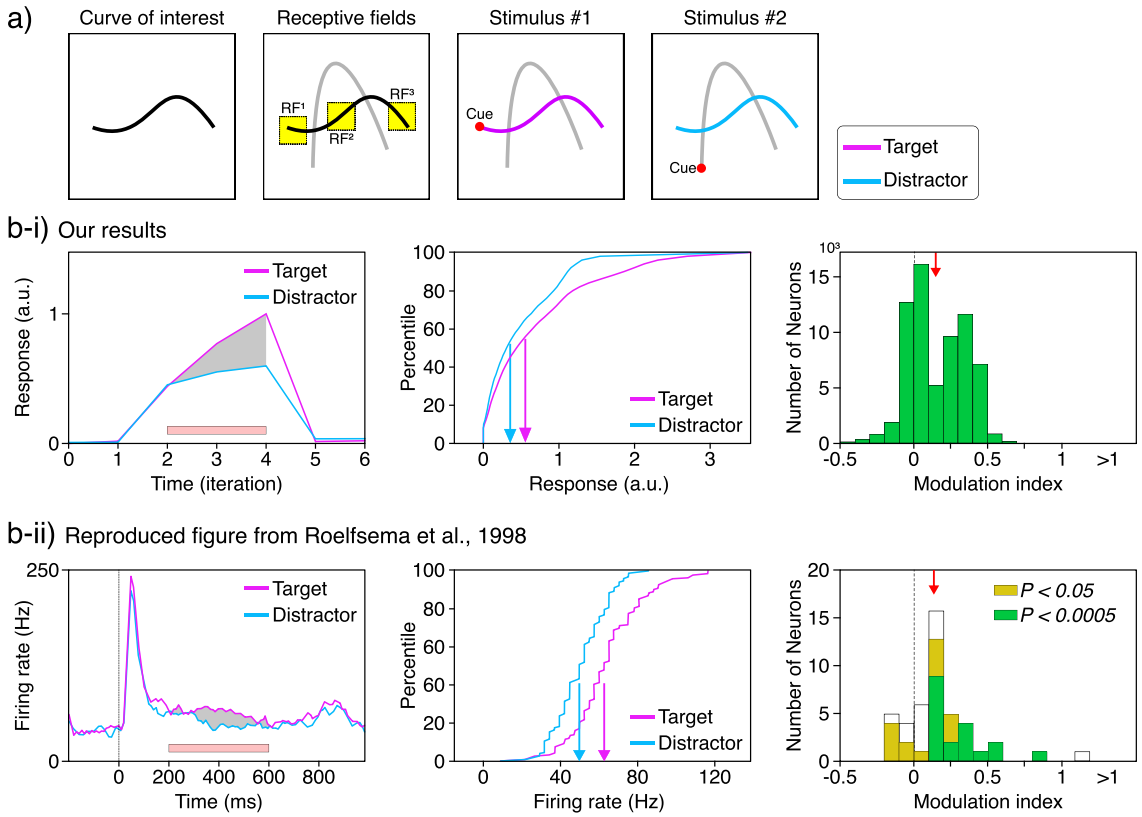


**Figure 9. Attentional Neuromodulation. a)** For a given curve of interest, we find all the neurons that their receptive fields only includes the curve of interest and reject those with overlap. We can then use the visual cue to mark the curve as target object (stimulus #1) or as distractor object (stimulus #2). **b)** Activity of neurons is enhanced through attention. Here we compare the results of **b-i)** neurons in layer two of our model for the switchbox experiment with **b-ii)** results from V1 neurons in macaque monkeys performing a similar task [Roelfsema et al., 1998]. The left graphs show neural response through time for the target (i.e., attended) versus the distractor (i.e., masked) curves. The pink bar marks the curve tracing window, used for the statistical analysis. The middle graphs demonstrate the response distributions of neurons for the target and distractor scenarios. The right graphs depict the histogram of response modulation index, showing significant positive modulation of neuronal activity. The red arrow indicates the median modulation index. The graphs are made in similar style to those in [Roelfsema et al., 1998].

### *Attention-invariant tuning*

Tuning curves are often used to depict how a neuron changes its activity (i.e., firing rate) in response to a change in a single attribute of its input. A common tuning curve for neurons in early visual areas which are sensitive to stimuli similar to 2-dimensional Gabor wavelets, is the orientation tuning. To extract the tuning curves, the experiments are designed so that the neuron's activity is mostly driven by input from early visual areas and is free of recurrent or higher-level signals. However, here we are interested in how the neuron changes its tuning in the presence of attention. Neurophysiological studies suggest that when a subject is attending to an object, neurons that are associated with the object's receptive field show a stronger activation than when the subject is not attending to it [Lee and Maunsell, 2010; Maunsell, 2015; McAdams and Maunsell, 1999; Treue and Trujillo, 1999]. Furthermore, [McAdams and Maunsell, 1999] showed that the change in the activation is multiplicative, meaning that other attributes such as width and preferred orientation are statistically unaffected, suggesting that neural tuning is attention-invariant. Like biological neurons, neurons in the early layers of an artificial neural network trained on naturalistic images exhibit tuning curves that resemble Gabor wavelets. In the deeper layers, however, the learned filters become increasingly complex, resembling the tuning curves of neurons in the ventral visual pathway (Fig. 10 a). Here we aim to test whether the neurons in our model also exhibit attention-invariant tuning.

We start with the model trained on the switchbox task and follow the method used by [McAdams and Maunsell, 1999]. As input, we use the composition of two distant bars (i.e., short straight lines). We use visual cues to switch attention between the two bars, creating two distinct states, attended and unattended, for the target bar. We then extract neurons that have the

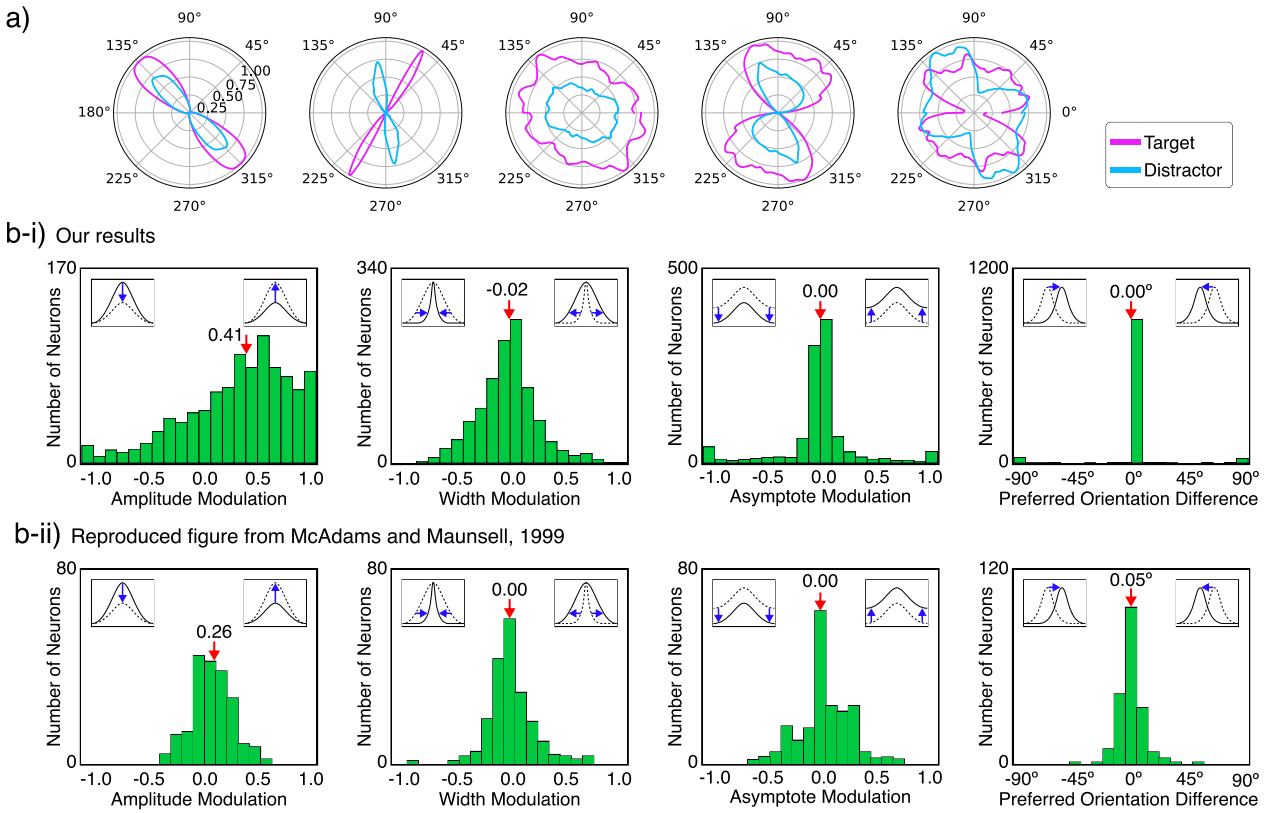


**Figure 10. Attention-invariant tuning. a)** Orientation tuning curves of five neurons in different layers of our model trained on the switchbox task. For the analysis of the attention-invariant turning, we only consider neurons with Gaussian-like tuning curves (e.g. the two plots on the left) and reject the rest. **b)** Attention increases the response amplitude but has little to no effect on the width, asymptote, and preferred orientation of the tuning curves. **b-i)** our results from the penultimate feature layer in our model compared to **b-ii)** the findings from V4 cortical neurons in macaque monkeys [McAdams and Maunsell, 1999]. The red arrows and values show the median. The graphs are made in similar style to those in [McAdams and Maunsell, 1999].

target bar in their receptive field. Furthermore, we will only consider those neurons that have a bell-curve (i.e., Gaussian like) orientation tuning. Finally, we can show that attention affects the neural activation in earlier layers of our network through multiplicative scaling without changing the overall shape of their tuning curves (Fig. 10 b).

## Discussion

We have demonstrated that a substantial portion of our understanding of visual attention can, in principle, be explained by bidirectional recurrent gating mechanism. We have introduced a unifying model that integrates multiple attentional phenomena observed in the cognitive and neural sciences. Our model performs similarly to humans, both qualitatively and quantitatively, on a range of canonical visual tasks, including object recognition, top-down search, pop-out effects, inhibition of return, and visual inference. Additionally, it replicates various neurophysiological findings, such as attention-mediated neuromodulation and attention-invariant tuning. Computationally, we showed how our model can correctly mask spurious features in the images, leading to more robust predictions. This work shows that bidirectional recurrent gating is a powerful mechanism for modeling and understanding attention.

Our model, naturally, has its limitations in describing neural and behavioral effects. For example, it does not account for overt attention mechanisms such as saccades and foveation, which are crucial for visual attention in humans [Itti et al., 2005]. Fortunately, effective models of these mechanisms exist and could, in principle, be integrated into our framework [Mnih et al., 2014]. Additionally, our model does not explicitly address aspects of neural computation such as non-stationarity, time-varying dynamics, and oscillations (beyond cycling through objects). However, given the recurrent nature of our network, we believe it can be adapted to incorporate these dynamics. Clearly, while our model describes a range of effects, there are still many findings it cannot yet quantitatively account for.

Next, our model uses Backpropagation Through Time (BPTT) to update its weights. While BPTT is a powerful tool in

machine learning, there is no direct evidence that backpropagation is used as an updating mechanism in the brain [Lillicrap et al., 2020], especially not in a BPTT setting where information would need to travel backward in time. Nevertheless, we argue that our mechanism provides a key inductive bias enabling the network to perform well on multiple attention tasks, although it currently relies on BPTT. However, computationally, all we need is gradient descent and many approximations may exist. Biologically plausible approximations of gradient descent have extensively been studied [Bengio et al., 2015; Richards and Kording, 2023]. In fact, backpropagation has been used to learn brain-like representations for visual tasks [Cadena et al., 2019; Khaligh-Razavi and Kriegeskorte, 2014; Kriegeskorte et al., 2008; Kubilius et al., 2019; Rajalingham et al., 2018]. Similarly, while weight sharing in convolutional layers has limited biological plausibility, despite being inspired by simple and complex cells in the visual cortex [Fukushima, 1988]. Recent models have proposed augmentations or wake-sleep cycles to train networks resembling emergent weight sharing [Pogodin et al., 2021]. In other words, our biologically motivated mechanism facilitates modeling attention in artificial neural networks through backpropagation, and we do not consider the biological plausibility of deep learning methods to be a limitation of the mechanism.

We have demonstrated that the proposed multitask learning paradigm can reduce the necessity of full supervision in some tasks. However, so far, our system still requires supervised signals. How do humans obtain supervised information about segmentation? Arguably, they do so by observing moving objects. The movement of these objects effectively separates them from the background, providing a signal that could be considered largely equivalent to supervised segmentation information. Similarly, additional strong cues exist for other aspects we use in supervised learning. For instance, abstract knowledge about object identity is often encoded across different modalities; for example, the presence of a horse is associated with a distinct smell. While our approach relies on supervised signals, we maintain that exploring what cues and learning algorithms the brain actually employs is a separate, albeit relevant and intriguing, research question.

Despite the drawbacks, our model of object-based attention reasonably explains data with the ability to solve real-world problems, and can efficiently learn to perform complex behavioral tasks. More importantly, we realized this with a scale-able, modular, and stable architecture that is built upon basic neural network components. Although marginally requiring more computation compared to feedforward convolutional neural networks, attention and recurrency are far too important to ignore in models of visual cognition [Lindsay and Miller, 2018]. Therefore, we argue that our model would provide a pragmatic and effective alternative to existing CNN architectures commonly used to model the visual system [Lindsay, 2021]. This would also enable design of more complex experiments that are closer to real-world behavior [Matusz et al., 2019].

Our network brings recurrency and gating, two fundamental computation schemes ubiquitous in cortex, to better model the biological visual system. Combined with other elements like hierarchical feature representation learning, the system has the right structure to model neurophysiological data. Furthermore, the multitask paradigm opens the door to design concurrent tasks that could bring about emergent behaviors at both behavioral and neurophysiological levels.

Importantly, our model provides a sufficiently complex baseline model that can be used to develop better models of object-based attention in the brain. Our results state the positive impact of attention and multi-task learning on generalization and robustness [Geirhos et al., 2020; Lapuschkin et al., 2019; Sinz et al., 2019]. Our model of object-based attention motivates new insights into our understanding of decision-making in a broad range of visual tasks, including credit assignments, binding, and memorability. This promises to be a starting point for further research in vision, learning, and binding that combines real-world tasks with cognitive modeling. Simultaneously, with the recent rise of interest in object-based attention and binding in the field of machine learning [Greff et al., 2020; Locatello et al., 2020], we see our network as a productive common ground between neuroscience and machine learning.

We also propose that the deep learning literature may benefit from incorporating insights from such attention modeling. By introducing attention as a primitive in the standard artificial neural network framework, we provide a flexible mechanism that allows stimuli, prompts, tasks, and outputs to jointly determine what is attended to and propagate this information through a large network. This approach fundamentally acknowledges the compositionality of the real world. Further research is needed to identify the types of tasks where incorporating human-like attention enhances network performance.

## Acknowledgements

This work was partly funded by the German Ministry for Education and Research (under refs 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18056A, 01IS18025A and 01IS18037A) and the German Research Foundation (DFG) within the RTG 2433 DAEDALUS. Furthermore, KRM was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and No. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation).

## Code and data availability

The code for this project will be made publicly available on September 13th 2024 (which incidentally is the day of the programmer) on Github: https://github.com/ssnio/bio-attention or upon request. The datasets are publicly available on their corresponding online repositories, but the scripts to generate the composites are part of the source code.

## References


Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science*, *344*(6182), 424–427.

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.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

Benjamin, A., Qiu, C., Zhang, L.-Q., Kording, K., & Stocker, A. (2019). Shared visual illusions between humans and artificial neural networks. *2019 Conference on Cognitive Computational Neuroscience*, *10*, 2019–1299.

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.

Bregman, A. S. (2017). Asking the "what for" question in auditory perception. https://api.semanticscholar.org/CorpusID:141511323

Buschman, T. J., & Kastner, S. (2015). From behavior to neural dynamics: An integrated theory of attention. *Neuron*, *88*(1), 127–144.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, *15*(4), e1006897.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, *51*(13), 1484–1525.

Cavanagh, P., Caplovitz, G. P., Lytchenko, T. K., Maechler, M. R., Tse, P. U., & Sheinberg, D. L. (2023). The architecture of object-based attention. *Psychonomic Bulletin & Review*, *30*(5), 1643–1667.

Chun, M. M., & Wolfe, J. M. (2005). Visual attention. *Blackwell handbook of sensation and perception*, 272–310.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European journal of neuroscience*, *17*(5), 1089–1097.

Fang, F., Boyaci, H., & Kersten, D. (2009). Border ownership selectivity in human early visual cortex and its modulation by attention. *Journal of Neuroscience*, *29*(2), 460–465.

Feldman, J. (2013). The neural binding problem(s). *Cognitive Neurodynamics*, *7*, 1–11.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*(2), 119–130.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Greff, K., Van Steenkiste, S., & Schmidhuber, J. (2020). On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*.

Han, S., Jiang, Y., Mao, L., Humphreys, G. W., & Gu, H. (2005). Attentional modulation of perceptual grouping in human visual cortex: Functional mri studies. *Human Brain Mapping*, *25*(4), 424–432.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Herzog, M. H. (2018). Perceptual grouping. *Current Biology*, *28*(12), R687–R688.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.
Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics*, *81*, 2288–2303.
Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.
Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.
Itti, L., Rees, G., & Tsotsos, J. K. (2005). *Neurobiology of attention*. Elsevier.
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.
Kaul, C., Manandhar, S., & Pears, N. (2019). Focusnet: An attention-based fully convolutional network for medical image segmentation. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 455–458.
Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915.
Khani, F., & Liang, P. (2021). Removing spurious features can hurt accuracy and affect groups disproportionately. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 196–205.
Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., 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*, *116*(43), 21854–21863.
Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything.
Klein, R. M., & Ivanoff, J. (2008). Inhibition of return [revision #125033]. *Scholarpedia*, *3*(10), 3650. https://doi.org/10.4249/scholarpedia.3650
Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.
Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.
Kruijne, W., Bohte, S. M., Roelfsema, P. R., & Olivers, C. N. (2021). Flexible working memory through selective gating and attentional tagging. *Neural Computation*, *33*(1), 1–40.
Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent anns. *arXiv preprint arXiv:1909.06161*.
Kwak, H.-W., & Egeth, H. (1992). Consequences of allocating attention to locations and to other attributes. *Perception & Psychophysics*, *51*(5), 455–464.
Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1096.
Larsen, B. W., & Druckmann, S. (2022). Towards a more general understanding of the algorithmic utility of recurrent connections. *PLoS Comput Biol*, *18*(6).
LeCun, Y. (1998). The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.
Lee, J., & Maunsell, J. H. (2010). Attentional modulation of mt neurons with single or multiple stimuli in their receptive fields. *Journal of Neuroscience*, *30*(8), 3058–3066.
Li, J., Zhang, J., Maybank, S. J., & Tao, D. (2022). Bridging composite and real: Towards end-to-end deep image matting. *International Journal of Computer Vision*, *130*(2), 246–266.
Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*(6), 335–346.
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision*, 740–755.
Lindsay, G. W. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, *14*, 29.
Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031.
Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, *7*, e38105.
Linsley, D., Karkada Ashok, A., Govindarajan, L. N., Liu, R., & Serre, T. (2020). Stable and expressive recurrent vision models. *Advances in Neural Information Processing Systems*, *33*, 10456–10467.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of International Conference on Computer Vision (ICCV)*.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., & Kipf, T. (2020). Object-centric learning with slot attention. *Advances in neural information processing systems*, *33*, 11525–11538.

Matusz, P. J., Dikker, S., Huth, A. G., & Perrodin, C. (2019). Are we ready for real-world neuroscience?

Maunsell, J. H. (2015). Neuronal mechanisms of visual attention. *Annual review of vision science*, *1*(1), 373–391.

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience*, *19*(1), 431–441.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*.

Moore, T., & Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annual Review of Psychology*, *68*, 47–72.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *Advances in neural information processing systems*, *31*.

Nothdurft, H.-C. (2006). Salience and target selection in visual search. *Visual Cognition*, *14*(4-8), 514–542.

Olivers, C. N., & Roelfsema, P. R. (2020). Attention for action in visual working memory. *Cortex*, *131*, 179–194.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. *Frontiers in Psychology*, *4*, 124.

Paneri, S., & Gregoriou, G. G. (2017). Top-down control of visual attention by the prefrontal cortex. functional specialization and long-range interactions. *Frontiers in neuroscience*, *11*, 545.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, *42*(1), 107–123.

Pogodin, R., Mehta, Y., Lillicrap, T., & Latham, P. E. (2021). Towards biologically plausible convolutional networks. *Advances in Neural Information Processing Systems*, *34*, 13924–13936.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

Posner, M. I. (2016). Orienting of attention: Then and now. *Quarterly journal of experimental psychology*, *69*(10), 1864–1875.

Posner, M. I. (2023). The evolution and future development of attention networks. *Journal of Intelligence*, *11*(6), 98.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance X: Control of Language Processes*, *32*, 531–556.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of experimental psychology: General*, *109*(2), 160.

Qiu, F. T., Sugihara, T., & Von Der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature neuroscience*, *10*(11), 1492–1499.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770.

Richards, B. A., & Kording, K. P. (2023). The study of plasticity has always been about gradients. *The Journal of Physiology*, *601*(15), 3141–3149.

Robertson, L. C. (2003). Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, *4*(2), 93–102.

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.*, *29*(1), 203–227.

Roelfsema, P. R. (2023). Solving the binding problem: Assemblies form when neurons enhance their firing rate—they don't need to oscillate or synchronize. *Neuron*, *111*(7), 1003–1019.

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

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*, 234–241.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, *2*, II–II.

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.
Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1-2), 1–46.
Singer, W. (2001). Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, *929*(1), 123–146.
Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*(6), 967–979. https://doi.org/https://doi.org/10.1016/j.neuron.2019.08.034
Stollenga, M., Masci, J., Gomez, F., & Schmidhuber, J. (2014). Deep networks with internal selective attention through feedback connections. *arXiv preprint arXiv:1407.3068*.
Super, H., Romeo, A., & Keil, M. (2010). Feed-forward segmentation of figure-ground and assignment of border-ownership. *PloS one*, *5*(5), e10705.
Thiele, A., & Bellgrove, M. A. (2018). Neuromodulation of attention. *Neuron*, *97*(4), 769–785.
Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, *6*(2), 171–178. https://doi.org/https://doi.org/10.1016/S0959-4388(96)80070-5
Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1295–1306.
Treisman, A., & Gelade, G. A. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136. https://api.semanticscholar.org/CorpusID:353246
Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579.
Tsotsos, J. K. (2021). *A computational perspective on visual attention*. MIT Press.
Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*(1-2), 507–545.
van Bergen, R. S., & Kriegeskorte, N. (2020). Going in circles is the way forward: The role of recurrence in visual inference. *Current Opinion in Neurobiology*, *65*, 176–193.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.
Von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current opinion in neurobiology*, *5*(4), 520–526.
Von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, *24*(1), 95–104.
Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Harrow, J., Psomopoulos, F. E., & Tosatto, S. C. (2021). Dome: Recommendations for supervised machine learning validation in biology. *Nature methods*, *18*(10), 1122–1127.
Walther, D., Rutishauser, U., Koch, C., & Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, *100*(1-2), 41–63.
Williford, J. R., & Heydt, R. v. d. (2013). Border-ownership coding [revision #186629]. *Scholarpedia*, *8*(10), 30040. https://doi.org/10.4249/scholarpedia.30040
Wolfe, J. M. (2005). Guidance of visual search by preattentive information. *Neurobiology of attention*, 101–104.
Wolfe, J. M. (2010). Visual search. *Current biology*, *20*(8), R346–R349.
Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, *5*(6), 495–501.
Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

# Supplements / Appendix

## Validation

We followed the DOME recommendations and framework presented by [Walsh et al., 2021] with respect to assessment and reproducibility.

### *Data*

The data samples are task-specific and are generated by augmenting and composing images from publicly available datasets: MNIST [LeCun, 1998], COCO [Lin et al., 2014], and CelebA [Liu et al., 2015]. For MNIST and COCO, we use different image augmentation to improve the generalization. Regarding the train-test sets, the CelebA dataset comes with three distinct sets of: training, validation, and test datasets. But the MNIST and COCO datasets do not include an available labeled test dataset and hence we use cross validation on the train set. In all experiments, the validation sets are only used for inference to find the best trained model and the test dataset is used for final reporting and visualizations. The data samples for the SwitchBox experiment were created randomly using Bezier curves.

### *Optimization*

The models are trained on a single NVIDIA A100 GPU. We use Adam algorithm as the gradient descent optimizer with weight decay regularization. We perform grid search for finding a learning rate that yields better validation error. Our loss function could be written as the composition of three weighted objectives:

$$\mathbf{L} = \sum_{k=1}^{m} \sum_{i=1}^{n} \left[ \alpha_i^k \, MSE(\hat{M}_i^k, \, M_i^k) + \beta_i^k \, CrossEntropy(\hat{y}_i^k, \, y_i^k) \right] + \lambda \|\mathbf{W}\|^2 \tag{1}$$

where subscript $i$ is the iteration and superscript $k$ denote the task index. The first term inside the sum is the Mean-Squared Error (MSE) loss for the estimated $\hat{M}_i^1$ and true attention maps $M_i$, if available. Next is cross-entropy as the classification loss of output labels. For a given experiment, we control for the existence and importance of each term in the iteration $i$ with the hyper-weights $\alpha_i$, $\beta_i$, and $\gamma_i$. The last term is the $L_2$ regularization term for the learnable weights $\mathbf{W}$, the only explicit regularization included in the loss function. For the COCO model, we also used class-weights to improve the performance.

### *Model*

We use distinct architectures for each experiment. Each network is scaled and equipped differently depending on the task and data complexity. For example, we used a simpler architecture for the CelebA task. For MNIST, we used RNN layers as working memory which is necessary for the IOR task. For COCO we used a scaled up version of MNIST model with, but did not include the RNN layers.

The network is built by stacking multiple modules of bidirectional recurrent gating blocks that could have different "processing and learning layers" 11. The first iteration always starts with a flat attention map (i.e. pre-attention). As the image passes through the feature path, the model extract higher level features and builds a richer representation of the input. The model also stores the feature maps for each layer to be used later by the attention path. Finally, the feature path ends (if necessary) with sequential RNN layers that act as working memory. The output of the last RNN layer is used for label classification, but we do not consider the first label prediction for classification error since the first feature pass is unaware of the task and is the results of flat attention. The attention path starts with the output of the forward path concatenated with the task encoding. Note that for the top-down tasks, the forward output is discarded and the one-hot encoded prompt is fed into model instead. After attention RNN layers, the resulted activity is concatenated with the feature maps from the corresponding forward layer and scaled according to the task-embedding. This signal now contains information about the task, the feature representations, and latent information and can be used by the consequent layers (namely transposed-convolution, normalization, and non-linearity) to generate the attention maps for corresponding forward layer. Other than the bottle neck, the concatenated features and attention maps, form a tensor of shape (channels, height, width). The task embedding modulates the channels and not in the pixel level. In the attention path we use hyperbolic tangent non-linearity (Tanh) to get attention maps that range between negative and positive one (i.e. (-1, 1)). This process is repeated for every layer in the attention path until the attention map reaches the bottom of the model, which concludes the first iteration. The next iterations is similar to the first iteration except the inputs to each layer are now modulated by the scaled attention maps from the previous iteration. We use a multiplicative scale of 0.5 to allow partial attention and avoid losing all the distractor information, as they might be relevant for later iterations [Maunsell, 2015]. The details of architectures are available in the code repository.

### *Evaluation*

The code and trained models are all made available for validation and further investigation (please see Code Availability section). For classification, we use Cross-Entropy loss. The MNIST dataset is sufficiently balanced, but for COCO we used class-weights to combat the imbalanced class distributions. For CelebA experiment, although the dataset is not balanced, we did not use class-weights to rather emphasize on the networks ability in learning without prior knowledge of class distribution and spurious correlations.

For segmentation, we use mean-squared loss for training, but we use attention accuracy and pixel error for reporting as they are more intuitive and easier to interpret.

**Attention accuracy** Attention accuracy measures the normalized accuracy of a pixel being correctly attended or masked (i.e., the target and estimated pixel value having the same sign). We use normalization to correct the target to scene area ratio, leading to chance level to always be at 50% (i.e., the whole scene being attended or the whole scene being masked). The measure is similar to the Intersection over Union (IoU) commonly used in computer vision [Rezatofighi et al., 2019]. The following equation explains the measure:

$$AttentionAccuracy = 100\% \, \frac{1}{h \times w} \sum_{i,j}^{h,w} \sigma(\hat{m}_{ij} \cdot m_{ij}) \quad \text{where:} \quad \sigma(a,b) = \begin{cases} 1 & \text{if} \quad Sign(a) = Sign(b) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

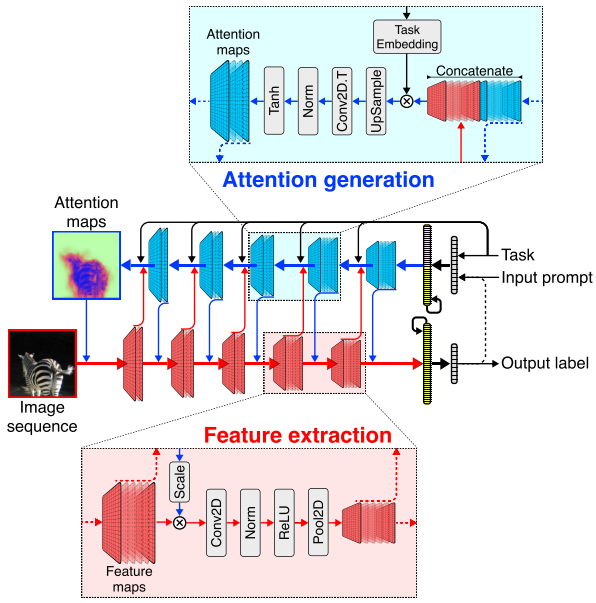


**Figure 11.** Detailed architecture used for the MNIST experience. The elements of the two pathways are convolution (Conv2D) and transposed-convolution (Conv2D.T), nonlinear activation functions (ReLU and Tanh), pooling and up-sampling, and divisive normalization (Norm). The feature map $\mathbf{X}_t^l$ of block $l$ at iteration $t$ is multiplicatively modulated by the attention map $\mathbf{Z}_{t-1}^l$ from the previous iteration before being processed through the forward block $l+1$. The output of block $l+1$ is passed to the corresponding block in the attention path, concatenated with the attention map $l+1$ before modulation by the task-embedding. The attention block then creates the attention map $\mathbf{Z}_{t-1}^l$ for the next iteration

where $\hat{m}_{ij}$ and $m_{ij}$ denote the $ij-th$ pixel of estimated and true attention maps respectively, while $h$ and $w$ are the height and width of the attention map.

**Pixel Error** Pixel Error denotes the mean squared euclidean distance between estimated and true attention maps.

**Modulation index** We use the Modulation Index (MI) similar to those described in [Roelfsema et al., 1998] and [McAdams and Maunsell, 1999]:

$$ModulationIndex = \frac{T-D}{T+D} \tag{3}$$

where $T$ and $D$ denote the averaged response (i.e. activity) of neurons to the target and distractor curves respectively.

## Multi-modal top-down visual search

We also trained a separate model based on our architecture on a top-down visual search for class and color using the FashionMNIST dataset [Xiao et al., 2017]. The input prompt, therefore, includes both the target class and color to search for.

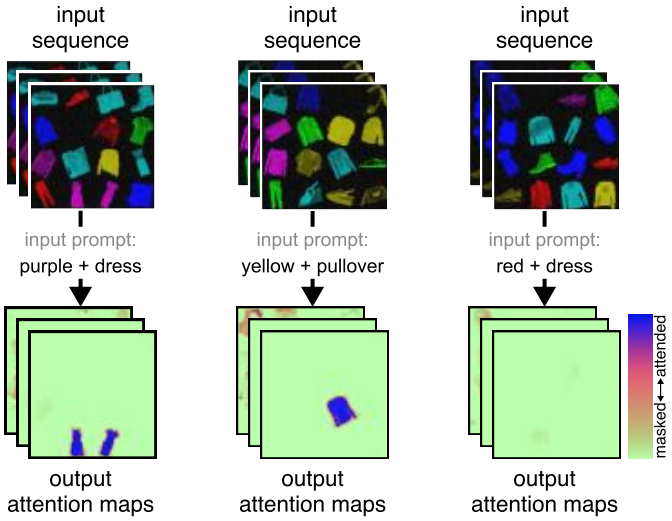


**Figure 12.** **Top-down visual search**