A brain-inspired object-based attention network for multi-object recognition and visual reasoning

Hossein Adeli[1*], Seoyoung Ahn[1], Gregory J. Zelinsky[1,2]

[1]Department of Psychology, [2]Department of Computer Science
Stony Brook University
{hossein.adelijelodar, seoyoung.ahn, gregory.zelinsky}@stonybrook.edu
[*] corresponding author

Abstract

The visual system uses sequences of selective glimpses to objects to support behavioral goals, but how is this attention control learned? Here we present an encoder-decoder model inspired by the interacting bottom-up and top-down visual pathways making up the recognition-attention system in the brain. At every iteration, a new glimpse is taken from the image and is processed through the 'what' encoder, a hierarchy of feedforward, recurrent, and capsule layers, to obtain an object-centric (object-file) representation. This representation feeds to the 'where' decoder, where the evolving recurrent representation provides top-down attentional modulation to plan subsequent glimpses and impact routing in the encoder. We demonstrate how the attention mechanism significantly improves the accuracy of classifying highly overlapping digits. In a visual reasoning task requiring comparison of two objects, our model achieves near-perfect accuracy and significantly outperforms larger models in generalizing to unseen stimuli. Our work demonstrates the benefits of object-based attention mechanisms taking sequential glimpses of objects.

1 Introduction

Objects are the units through which we interact with the world and perform tasks [69, 57]. Consider the simple task of searching for an object in a cluttered environment (e.g. a paper clip in a busy drawer). This task requires one to sequentially move spatial attention to select objects and to use object-based attention to route and bind the features of these objects so that they can be recognized as a target or non-target. The role of the attention mechanism in this process is to make the objects in the sequence the figure, in order to facilitate interacting with the object and performing the task [67, 83]. Works have also suggested that object-based attention is important for achieving human-level performance in accuracy and generalization [24, 66]. However, in contrast to spatial [13, 1] or feature-based attention [49, 47], where attention studies can focus on simple, well-defined visual features (e.g., location, color, etc.), the units of selection for object-based attention [5, 17] are entities consisting of complex spatial properties (shapes) and distributed feature representations. Moreover, perception is akin to hypothesis testing, where a top-down modulation reflecting object hypotheses and goals affect object selection and bottom-up processing. Here we attempt to model a general-purpose object-based attention mechanism, a task that requires the solution to three sub-tasks: (a) solving the binding problem, thereby creating an object-centric representation, (b) capturing the interaction between a (largely) bottom-up mechanism for recognition and a top-down mechanism for attention planning, and (c) learning to sequentially sample objects.

To solve the object binding problem [76], recent deep neural networks (DNNs) have been proposed that create object-centric representations of the entities in a scene by spatially segregating the features from the background and dynamically grouping those features in spatial and representational domains [27]. This segregation and grouping relies on (bottom-up) part-whole matching and Gestalt processes interacting with (top-down) objectness priors and knowledge of object categories [27, 83, 85]. The representation of different objects are stored in separate "slots" [27, 26, 8, 48, 19], realizing the cognitive concept of "object files" [37, 25]. A recent development in this domain, Capsule Networks (CapsNets) [64, 32], attempt to represent scenes as parse trees. Capsules in different layers represent visual entities at different levels of object granularity in the image, from small object parts in the lower levels to whole objects at the top level. Information is dynamically routed between the capsules in adjacent layers, eventually settling on a global part-whole matching (i.e. scene parse tree). Capsules provide an encapsulation and grouping of object representations that has shown benefits for performing object-based and class-conditioned reasoning for downstream tasks (e.g. deflecting adversarial attacks; [61]). Object-centric representations have great potential for modeling visual perception (e.g. for predicting global shape processing behavior in humans [16]), but to our knowledge, have not yet been studied in the context of planning attention glimpses.

An object-centric perspective requires an integrated attention-recognition mechanism that is nicely embodied by the interacting structures organized along the ventral and dorsal visual pathways in the brain [77, 78]. The ventral "what" pathway is involved in feature processing and recognizing objects and scenes [20, 77]. Core object recognition refers to the rapid recognition of briefly presented objects (100ms) [81] and is believed to be carried out primarily in the initial feedforward processing pass in the ventral pathway [14]. Modeling works support this claim by showing that neural activation along this pathway during core object recognition can be predicted [9, 29, 10] using feedforward Convolutional Neural Networks (CNNs [45, 44]) trained on the same recognition tasks. However, difficult recognition tasks require recurrent and feedback connections beyond the feedforward pass [39, 38, 70, 86]. The dorsal "where" pathway is believed to be involved in the spatial prioritization of visual inputs and the guidance of actions to objects [6, 75]. Here we take inspiration from the role played by dorsal structures in providing attentional modulation of ventral pathway activity, as shown in previous work [11, 6]. The attention signal generated in the dorsal pathway prioritizes and modulates the routing of the visual inputs in the ventral pathway for the purpose of improving object classification success and better performing the task. The repetition of this process imposes seriality on behavior when confident classification decisions are needed, making visual perception a sequential process. Previous neuro-cognitive modeling work on an integrated recognition-attention mechanism [11] have been largely hand designed and applied only to simpler stimuli, limitations that could be better addressed with a learning-based approach.

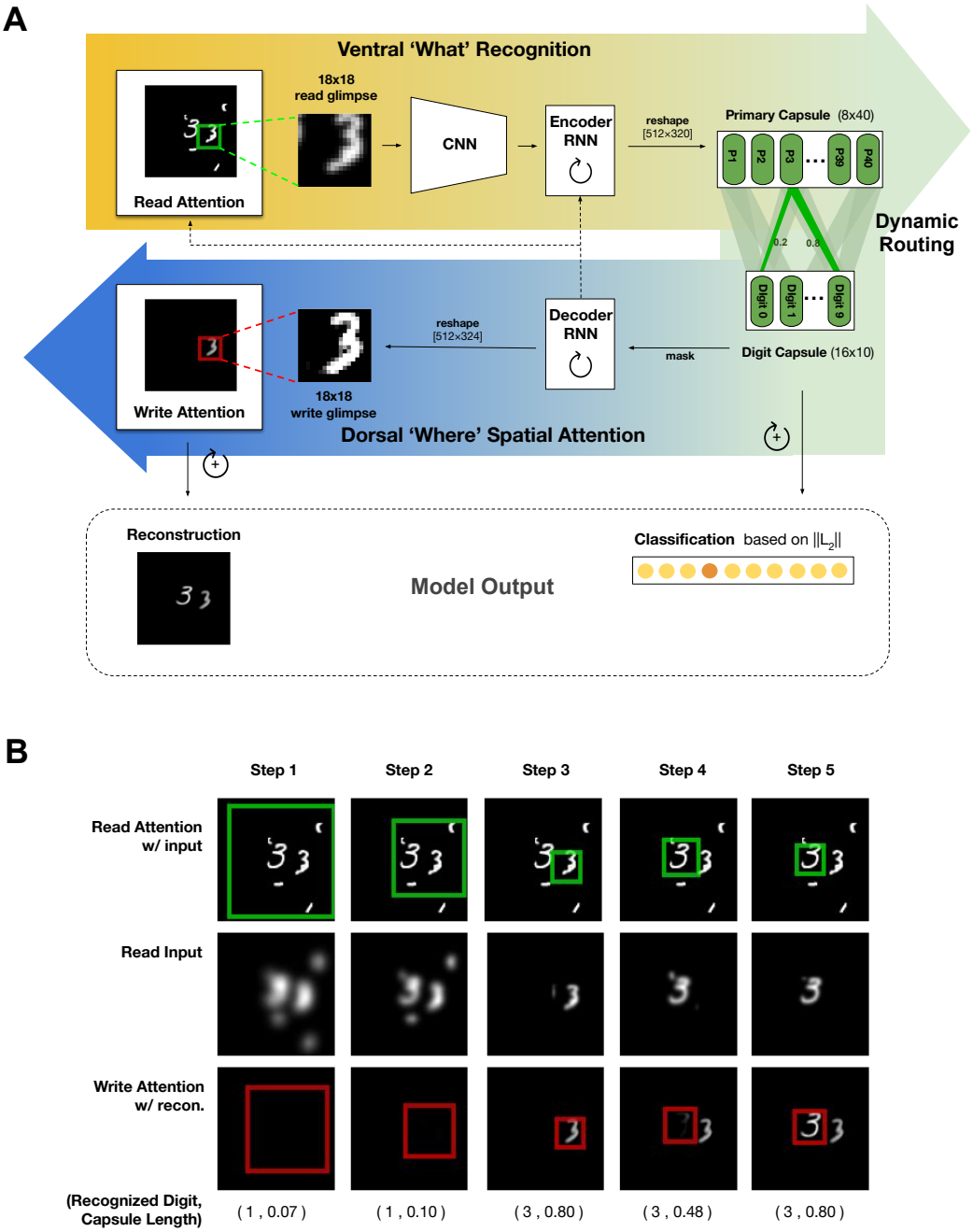

Figure 1: **A)** OCRA architecture. At each timestep a new read glimpse is taken from the image with the attention window parameters specified by the decoder RNN. The read glimpse is then processed using first a small CNN network and the features are then inputted to a recurrent layer. The primary capsules are then read from this recurrent activation using a linear mapping. The primary capsules are dynamically routed through agreement to the class capsules and the size of the these latter capsules determines the evidence for a class in the selected glimpse. The process so far maps onto the processing along the ventral pathway. The decoder takes the masked capsules and writes to the canvas using a similar attention mechanism as the read process. The sequential process goes on for multiple steps. At

the end, the cumulative size of the class capsules determines the final classification, which along the cumulative canvas are used for training. **B)** The stepwise process of moving the glimpse over a sample MultiMNIST cluttered image. The read attention windows are overlaid as green boxes on the original image in the first row. The read glimpse are shown in the second row and the bottom row shows the cumulative canvas and the location of write attention in red. The most active capsule that is routed to the decoder at each step, and its length, are provided at the bottom.

Our model of object-based attention employs the general structure of autoencoders (and more generally encoder-decoders). This class of models learn to encode the sensory input into a compact representation that captures the important aspects of the input, and then decode that representation to reconstruct the sensory input (or translate it into another modality, e.g. image to text) [42, 87]. Notably the DRAW [28] architecture introduced a sequential spatial attention mechanism to Variational Autoencoders (VAE, [42]) and showed that the model can learn to iteratively glimpse different parts of the input images and reconstruct them. Our premise is that the encoder-decoder framework loosely maps onto interactions between dorsal and ventral processing. Encoder processing can be mapped onto the ventral pathway and the decoder processing to the dorsal pathway that provides a top-down attention modulation signal for processing in the ventral pathway, with the encoding and decoding steps taken iteratively (cycle of prioritization and selection repeats). Using this framework, we present OCRA, an **O**bject-**C**entric **R**ecurrent **A**ttention model that combines recurrent glimpse-based attention and object-centric representation learning. Like a CapsNet, it performs encapsulation of features to structure the higher level representations for object processing. However, we place this structure within the aforementioned encoder-decoder model with recurrent attention, thereby enabling integration of structured information across multiple attentional glimpses. We show that capsule-based binding of object features and grouping (part-whole matching) is effective in sequential detection of multiple overlapping, distinct objects. It is also very effective in performing visual reasoning tasks (judging whether two randomly generated shapes are the same or different) and on a challenging generalization task where the model is tested on stimuli that are different from the training set.

2 Results

Object-based attention mechanisms endow a model with the ability to segregate and represent different objects effectively in order to perform a task. To test this in OCRA we used multi-object recognition and visual reasoning tasks. First, the model's ability to segregate the figure objects from the background clutter is tested on MultiMNIST-cluttered dataset[3]. Second, we test the model's recognition performance when objects are highly overlapped thus occluded using the MultiMNIST task[64]. We also examine the model's performance on learning visual reasoning using a similar paradigm as the task 1 in the Synthetic Visual Reasoning Test (SVRT) [21], where the model must detect whether two randomly generated objects in a given scene are the same or different. All model accuracies presented in this section are averages of 5 runs (see 4.5 for details of training and hyperparameters selection for different experiments). For recognition tasks, the accuracies are measured on the image level, meaning that the response is correct only if both objects in an image are recognized correctly.

The architecture for OCRA is shown in Fig. 1. Building on the DRAW architecture [28], the "attention window" in OCRA is a grid of filters applied to a small or a large area of the image. However, because the number of filters covering the attention window remains constant, as the window gets bigger it samples increasingly low-resolution information, creating a tradeoff between the size of the attention window and the resolution of information extracted. This property is aligned with "zoom lens" theories of human attention [18, 52], that similarly propose a variable-sized attention process that can be broadly or narrowly allocated to an input, but one that is constrained by the same trade off (See Fig. 4 for a visual illustration of the attention mechanism). The original DRAW model [28] was formulated as an autoregressive VAE, as it was trained for stepwise self-supervised reconstruction.

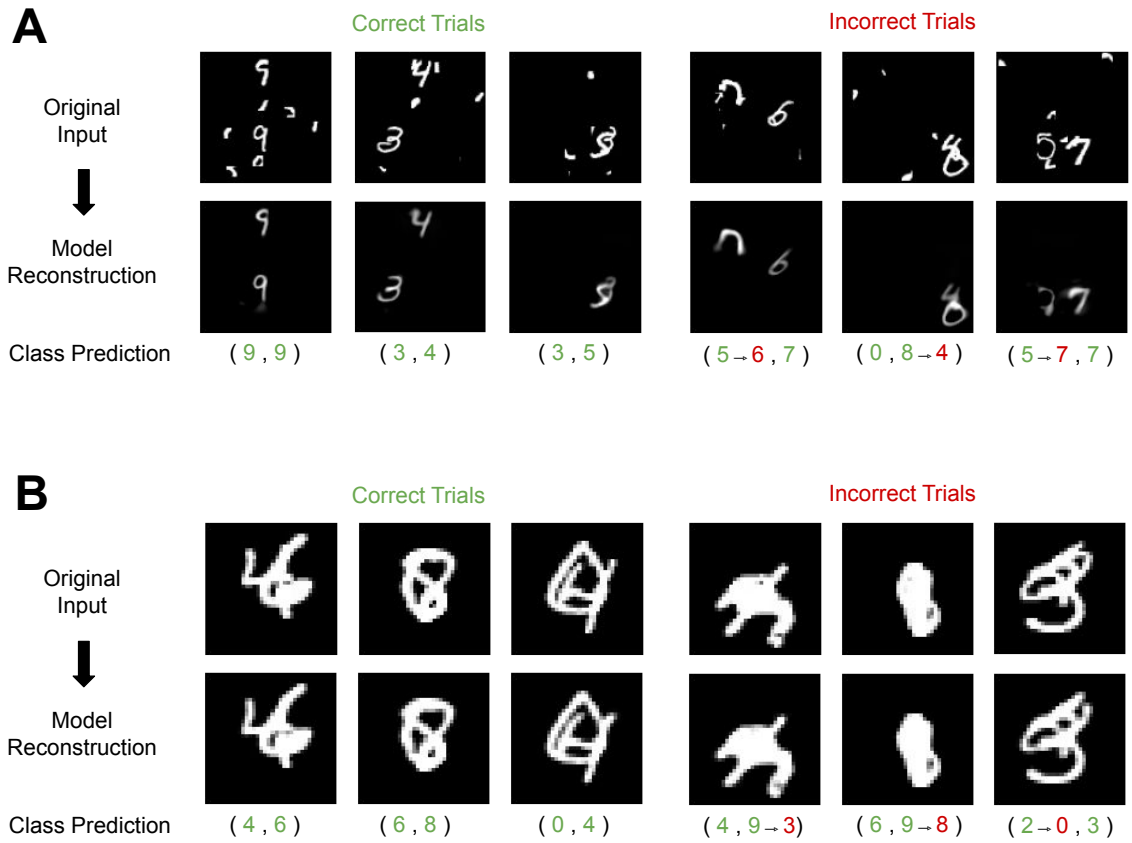

Figure 2: Sample classification and reconstruction predictions for OCRA on **A)** MultiMNIST cluttered and **B)** MultiMNIST tasks. The green digits show the ground truth and correct model predictions and erroneous predictions from the model are in red.

In contrast, our formulation uses a deterministic encoder-decoder approach and is trained to predict both the category classification based on object-centric capsules, and image reconstruction based on decoder output. Refer to the method section 4 for the detailed description of the model architecture and mechanisms.

2.1 Object in Clutter

Recognizing objects in noisy environments requires grouping and binding the object features and segregating them from the noisy background or other objects. We tested how well our model can learn to do this using the the MultiMNIST cluttered task. The stimuli for this task are generated by placing two random digits and object-like distractor pieces on a canvas (see 4.1).

Fig. 1B shows OCRA performing detection, classification and reconstruction of the digits in five timesteps for a sample image from this task. The top row shows the size and extent of the attention window for each step. The initial glimpse (left column) is set to cover the image and be centred to get a gist of the whole scene. The second row shows the read glimpse in the image dimension but note that the selected glimpse is always the same size given the constant number of filters, as discussed earlier. For this task, we defined a reconstruction mask that is the averaged sum of all the read glimpses

Table 1: MultiMNIST classification accuracy

Model	Model size	Accuracy
CapsNet (Sabour et al., 2018)	11.36M	87.3% (±0.16)
OCRA-3glimpse	3.87M	92.76% (±0.11)
OCRA-10glimpse	3.87M	94.92% (±0.17)

converted into image dimensions (average of Fig. 1B middle row). We integrate this mask into our loss function by multiplying it by the input image before comparing it to the model reconstruction. This mask effectively focuses the loss so that the model is accountable for reconstructing only the areas where it had glimpsed, allowing the model to be selective with its glimpses and write operations. The most active capsule that is routed to the decoder at each step, and its length, are provided at the bottom row. Utilizing the reconstruction mask, the model, with only classification supervision, learns that the best strategy is to move its glimpse to digits and to write to the canvas only when it is confident of the digit classification. Fig. 1B shows how the model's object-centric representation affects its attention planning and reconstruction, clearly selecting objects and ignoring the visual noise. Fig. 2 shows more examples of model predictions for both recognition and reconstruction, with correct responses on the left side and errors on the right. Most errors by OCRA on this task are due to a digit overlapping with the other digit or noise pieces in ways that change their appearance from the underlying ground truth. Fig. 5 shows the multi-step process of the model for these images. The model also successfully disengages from one object and moves on to glimpse and recognize the second object which captures the fact that humans recognize objects serially, with attention needing to shift from one object to another, but something that is not dealt with when the models are tested only on images with a single object.

OCRA's performance (%94 accuracy) on this task is comparable to other glimpse based models [51, 3], however, OCRA is much easier to train given its differentiable attention mechanism in contrast to reinforcement learning-based mechanisms employed in earlier work. Another notable difference is the use of the context network in [3] where the model inputs the whole image in a separate pathway to plan attention selection, thereby separating it from the recognition pathway. OCRA has one pathway and strings together glimpses from multiple steps to have an integrated recognition and attention planning mechanism. Our model does this through its "zoom lens" attention processes that can switch between local and global processing as needed. The early glimpses in the model, taken from the whole image, while being low resolution gists (Fig. 1, middle row, left), allow the model to plan its future glimpses. Later glimpses are focused on individual objects to mediate the recognition and reconstruction processes.

2.2 Overlapping Objects

Objects commonly occlude each other, and to study the attention-recognition system under these conditions we used the MultiMNIST task. This task involves recognizing digits with on average 80 percent overlap between the two bounding boxes. Fig. 2B shows some samples of the model making correct and incorrect predictions and the resulting reconstructions. Table 1 shows the OCRA accuracy (whether both objects in the image are correctly classified) for this task compared to the CapsNet model [64] where the task was originally introduced. Our model with 3 glimpses significantly outperforms the CapsNet model while having only a third of the number of parameters. We saw a clear effect of the number of timesteps in our model, with the error rate dropping from 3 timesteps to 10 timesteps.

In order to better understand the model dynamics, we start from the OCRA-3glimpse model and ablate different model components; the object-centric representation, the sequential glimpse mechanism and the recurrent processing; to measure their impact on model performance. Results are shown in Table 2.

The effect of recurrent attention mechanisms: In the first ablation experiment we asked how well a recurrent model using object-centric representation would perform without the ability to sample glimpses. OCRA-Recurrent (Table 2) performs multi-step processing on the input image using the recurrence in its encoder and decoder RNNs but lacks the ability to glimpse at specific locations. Therefore the model receives the entire image as input at each processing step for which it requires a bigger number of parameters (36×36 pixel input in contrast to 18×18 pixel glimpse). We then trained this model using three timesteps to make it comparable to OCRA-3glimpse. As shown in Table 2, this model performs worse than OCRA-3glimpse highlighting the important role of glimpse mechanism in our model performance. We then asked what if we removed the recurrent attention mechanism completely from our model. What is left is a model that makes one feedforward pass with the full resolution image as its input, binds features for either objects in separate category capsules, and then feeds them to the decoder (without masking the object capsules) to reconstruct the whole image at once. This is the Feedfoward model in Table 2. Similar to the OCRA-recurrent, having the model input the whole image in full resolution requires a bigger backbone and a larger number of parameters. This model shows higher error rate compared to the OCRA-recurrent supporting previous work on how recurrent dynamics can assist recognition tasks high degrees of occlusion [70, 86]. This is hypothesized to be due to recurrence providing more computational depth [54, 79] and leveraging contextualized iterative computations [79]. Taken together, the results show that while recurrent computation can be effective for challenging recognition tasks with high degrees of occlusion, when it is paired with an attention glimpse mechanism more compact and better performing models can be built. [68]

The role of object-centric representation: We performed another ablation study to examine the impact of object-centric representation on model performance. We replaced the capsule architecture architecture with two fully connect layers (with the same size as two capsule layers; 320 and 160 units) and a classification readout. The new model still iteratively processes the image and the classification scores are a linear readout from the second fully connected layer representation at each time step, which are then all combined across timesteps to make the final classification decision. As shown in Table 2, the new model performed worse than the model with object-centric representation showing the effect of encapsulation of information for performing this task. However, the effect was relatively smaller compared to the ones from ablating the recurrent attention. This suggests that the models can to some extent compensate for removal of the capsule architecture on this task. This could be due to the fact the the model has a glimpse mechanism and it can route information globally, thereby potentially reducing the benefit derived from feature encapsulation on this task.

2.3 Visual Reasoning

The ability to reason over visual entities is an important cognitive capability in humans and other animals [55] with perhaps the ability to judge whether two objects are the same or different being the most fundamental form of visual reasoning. Fleuret et al. [21] proposed the Synthetic Visual Reasoning Test (SVRT) as a benchmark for testing this capability in AI systems. Specifically task 1 in this benchmark (see examples in Fig. 3A) tests models ability on judging whether two randomly generated objects in an image are the the same or different. While behavioral results showed humans are able to perform this task easily, the early CNN models applied to this task failed to reach very high

Table 2: Ablation study results on MultiMNIST classification

Model	# of Glimpses	Capsules	Model size	Accuracy
OCRA (proposed)	3	Yes	(3.87M)	92.76% (±0.11)
OCRA-Recurrent	No	Yes	(8.58M)	91.02% (±0.14)
OCRA-Feedforward	No	Yes	(6.47M)	89.37% (±0.10)
OCRA-no Capsules	3	No	(3.87M)	91.96% (±0.13)

accuracy exposing a shortcoming [71, 40]. However, later works have found that deeper CNN models (with more layers and residual connections) such as ResNet [30] family of models can perform this task to a near perfect accuracy (>99%) while shallower models such as AlexNet [43] stay at chance level [50, 22], identifying computational depth as an important factor in learning this task.

Interestingly, some alternative approaches can achieve high levels of accuracy on this task with smaller networks. For example, siamese networks where two networks with shared parameters are applied to two object in the scene, are also able to learn this task [40]. Also if the objects are presented in different channels of the input image the models can preform this task better [72], These approaches are in a way pre-attending (binding) through feeding the objects separately to each network (the case with siamese networks) or by providing them in separate channels, by-passing the binding problem. These findings have led to the hypothesis that the main difficulty in performing this task is due to the the models not having a notion of object entities that are segregated from each other and and whose feature are bound for each object. However, both of those approaches are ad-hoc and unrealistic as the objects do not come pre-segmented with feature binding solved. As argued by [71**?**] the more plausible approach is to in fact model the the object-based attention process and detect and bind the features of the objects through sequential processing (also see [63]. On top that, the computational depth can be achieved through recurrent sequential processing rather than higher number of layers. Our model is designed exactly for this.

For applying OCRA to this task, we tied the response to the task of the same-different judgment to the capsule slot that would reconstruct the second object. As shown in Fig. 3B top row, the model starts with a more global glimpse but then moves its attention window, first to one object and then to the other. One object capsule would be representing the first object. This object-centric representation is then maintained in the encoder and decoder RNNs across timesteps. When the model glimpses the second object, the recurrent representation would bias the model to route the second object information to one of the two judgment capsule slots. This biasing is realized by the feedback connection that we added to OCRA for this task from the decoder RNN to the encoder RNN. Therefore the encoder RNN takes feedfoward, recurrent and feedback inputs to update its activity as each timestep. The model is trained with the same classification and reconstruction losses. Over training the model learns to route and reconstruct the second glimpsed object to the first classification capsule if the object have the same features and to the second classification capsule if the features are different from the one that is maintained in the recurrent connection. Note that we do not have explicit working memory modules in our model but the connection loops and the local recurrent connections do realize that function by maintaining the features of the first glimpsed object. We also hypothesized an additional capsule to represent the global scene gist and temporary object representation (that is likely to be formed when the glimpse moves between objects), which makes the the total number of capsules used in the representational bottleneck four. Fig. 3B bottom row shows the cumulative write canvas with the location of the write attention during the 10 timesteps. The gradual spreading of the reconstruction shown (bottom row) is consistent with the spreading of attention within an object hypothesized by object-based models of attention[36, 17].

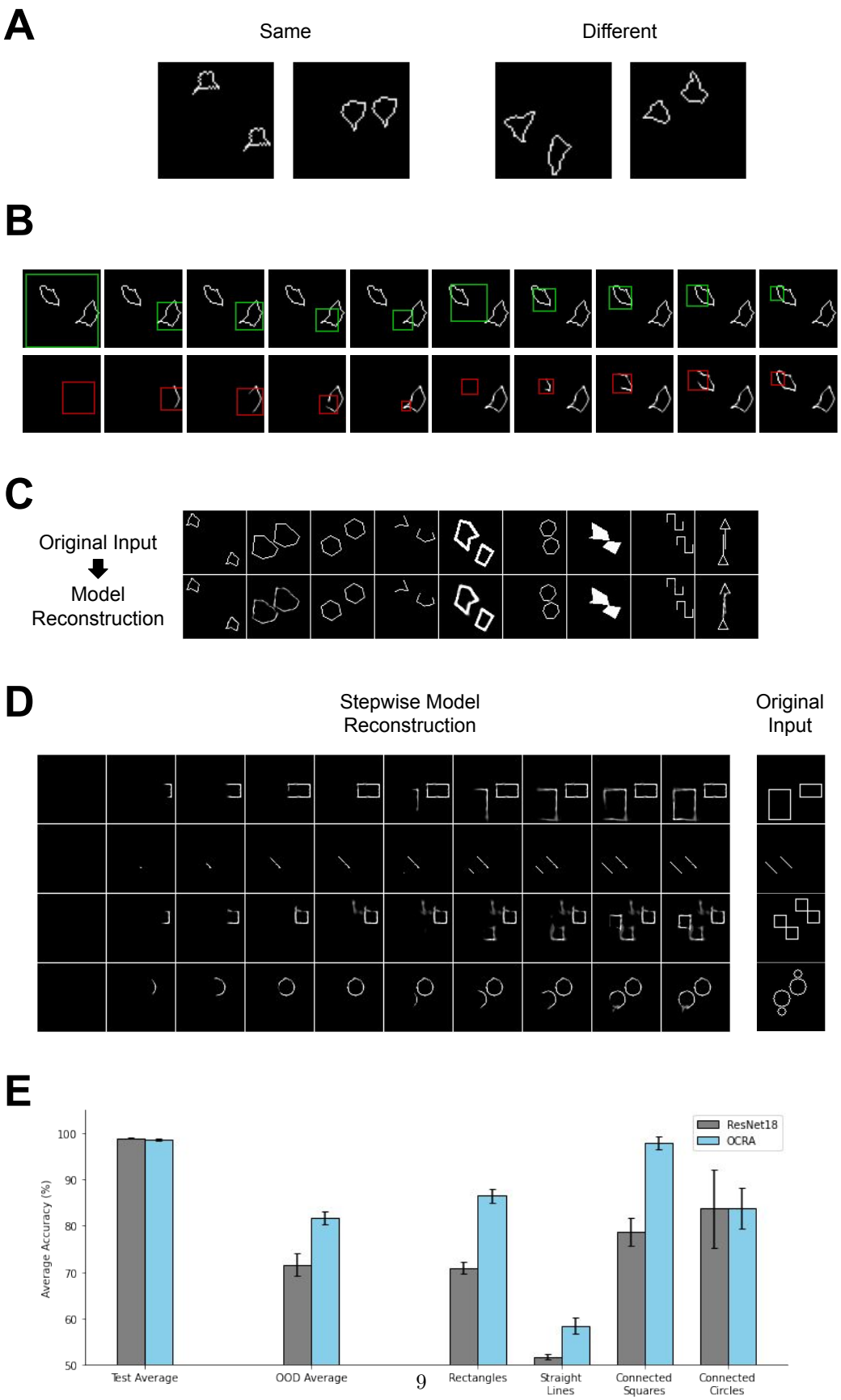
A
Same
Different
B
C
Original Input
Model Reconstruction
D
Stepwise Model Reconstruction
Original Input
E
ResNet18
OCRA
Average Accuracy (%)
100
90
80
70
60
50
Test Average
OOD Average
Rectangles
Straight Lines
Connected Squares
Connected Circles

Figure 3: OCRA performing the visual reasoning tasks. **A)** Sample images from SVRT task 1 for same and different responses. **B)** The sequential processing on a sample trial. The top row shows the model's read attention window in green at each time step. The bottom row shows the cumulative canvas with write attention window in red. **C)** Sample images for training OCRA on the OOD generalization task on the top row with one example for each shape type. The bottom row shows trained model's reconstructions for each sample. **D)** The sequential processing of four OOD sample images with a different OOD shape in each row. The original images are shown in on the right and the stepwise cumulative canvas is shown on the left for each sample illustrating the object-based behavior. **E)** The model accuracies for both ResNet18 (gray) and OCRA (cyan) on the original test set and the OOD test set on average and broken down by OOD shape type. Both model achieve near perfect accuracy on the original test set but OCRA significantly outperforms the ResNet18 on the generalization task. Error bards show the standard error.

Table 3 shows the model accuracies on this task. We also verified the results from earlier work for deeper models by applying the ResNet18 model. OCRA achieved the same near perfect level of accuracy as the ResNet model on this task. We then performed two ablation studies to examine the effect of the feedback connection and the object-centric representation bottleneck. The model without the capsules is constructed in a similar way as the previous section. The capsule layers were replaces by fully connected layers and a classification readout. For the no feedback model, we just removed the decoder RNN-encoder RNN feedback connection. We found ablating either of these components from our model significantly impacted the accuracy (Table 3). This shows that to perform this task, it is important for OCRA to leverage both the object-centric representation and the feedback mechanism for top-down biasing to modulate bottom-up information routing.

Generalization to out-of-distribution stimuli: As discussed earlier, deeper CNNs can solve the same-different task in the SVRT benchmark, however, there is still a big weakness in them as revealed by [59]. In this work, the authors argued that the true test of a model on whether they have learned the concept of "sameness" is to show that they can generalize to other displays where the pixel distribution of the generated objects are not the same as the ones in the training set. The models would then have to generalize to out of distribution (OOD) stimuli. We tested OCRA and ResNet18 on the dataset and a task adopted and modified from [59]. The training set includes a collection of nine different types of stimuli as shown in Fig. 3B top row. This was done to make sure that the training set is diverse and can allow the models to learn a rich feature representation for performing this task. The model was then not only tested on the original test set, but also on the OOD generalization test set, a collection of four types of stimuli with different shapes from the training set (examples shown in Fig. 3D right column).

As shown in Fig. 3E, OCRA and the ResNet model both achieved near perfect accuracy on the original test set on this task. Fig. 3B bottom row shows examples of OCRA reconstructions for some test samples (See Fig. 6 for the cumulative canvas at each timestamp for these samples). On the OOD generalization test OCRA significantly outperformed the ResNet model; average and four separate test shape accuracies are shown in Fig. 3E. This result confirms the previous work [59] showing that ResNet models critically struggle with same-different judgment generalization. For OCRA, the sequential

Table 3: Ablation study results on visual reasoning

Model	Accuracy
ResNet18	99.1% ($\pm$0.43)
OCRA	99.14% ($\pm$0.38)
OCRA-no Capsules	91.38% ($\pm$2.79)
OCRA-no Feedback	92.10% ($\pm$1.81)

object-based processing of the image generalizes to the OOD samples, as evident in the cumulative canvas for four examples shown in Fig. 3D. We believe this is the advantage that OCRA has over the competing models and can outperform them despite being much smaller in size. Note that the reconstruction qualities for OOD images are not as good as the original test images showing how the model still struggles with representing these stimuli and that there is room for improving the model.

3 Discussion

Object-based attention endows humans with the ability to perform an unparalleled diversity of tasks, and we built OCRA motivated by what we believe are its essential components. The object-centric representations allow the model to work in the space of object entities rather than pixels. The sequential processing and its inherent recurrent computations with cognitively plausible (zoom-lens) glimpse mechanism can provide more processing depth with smaller networks. The two pathway architecture can provide an interacting bottom-up an top-down processes where the feedforward processing is enriched and controlled by current top-down hypotheses. The results show that the synthesis of these approaches can lead to models with improved robustness and generalizability under challenging conditions that benefit from a mechanism for parsing the scene to entities and a recurrent process for iterative attentional sampling. We also observed qualitative hallmarks of object-based attention (such as the gradual filling in of the objects) in the model behavior [36].

While many works in the intersection of machine learning and cog-neuro fields have focused on convolutional and recurrent neural networks, encoder-decoder methods have the potential to be very useful to model behavior and brain responses. The generative component in these models is suitable for modeling different top-down modulations and feedback processing. To date, these models have been used for modeling top-down feedback in the ventral pathway [74, 2] in predictive coding [34], mental imagery [7] and continual learning [80]. These models can also be used for representation learning in general where they are trained using self-supervised methods to generate the visual input. The bottleneck representation can then be used as a more compact representation of the image which through using appropriate constraints can be made disentangled and interpretable [73, 31]. In this work we used this framework to address the relative paucity of work in modeling the dorsal pathway (but see [4] for a dual-stream architecture interpretation) compared to the ventral pathway, as its role is more complicated and varied across tasks (but see [65, 62] for recent work showing functional diversity beyond object recognition in the ventral pathway as well).

Previous work on modeling top-down attention focused on building "soft"-attention models that capture certain aspects of feature-based attention [13, 49] where top-down modulations can weight the incoming representation based on task relevance [87, 53]. In this vein, transformer-based attention mechanisms [82] have recently been used to create models with an integrated process of soft attention sampling and recognition [89, 35] leading to better performance on adversarial images. In these models the ongoing state representation can produce queries that will be matched to the keys from the bottom-up processing to create attention weightings to sample the bottom-up values to update the state representation. We believe that combining these approaches with the proposed glimpse based mechanism will help build a more robust model, where the model not only learns an optimal policy for moving glimpses but also the top-down modulation can bias feature processing along the processing hierarchy depending on the current object hypotheses.

Visual perception is the process of carving up the visual world into meaningful objects and entities. Discovering objects and representing them in ways that are conducive to performing downstream tasks has become an important area of research for relating deep learning (connectionist) approaches with human like symbolic and object-based reasoning [23, 27]. In these models, the first stage of generating object-based representations by segmenting the object in the scene is then followed by a second stage of modeling the interaction between the objects either through a key-query attention mechanism (similar to transformers) [25, 15] or graph neural networks [60]. The object-centric models provide the necessary representational capabilities to realize object-files and model many object-based

behavior and brain activities such as untethering of the object representation from the visual input pixels [57, 25]. These approaches can allow the models to work and learn on object level and should be particularly useful in modeling different aspects of object-based attention [58, 84, 46, 24, 12]. In return, we believe that taking an object-centric perspective can facilitate bringing ideas and inductive biases from object-based attention literate and more generally cognitive science to build more robust and generalization AI systems.

4 Materials and Methods

4.1 Stimuli generation

MultiMNIST Cluttered dataset: We generated the MultiMNIST Cluttered dataset, similar to the cluttered translated MNIST dataset from [51]. For each image, 2 digits and 6 digit-like clutter pieces are placed in random locations on a 100×100 blank canvas. All digits were sampled from the original MNIST dataset[45] and the two digits in each image could be from the same or different categories. Clutter pieces were generated from other MNIST images by randomly cropping 8×8 patches. We generated 180K images for training and 30K for testing, ensuring to maintain the same MNIST training/testing separation.

MultiMNIST Dataset: We generated the MultiMNIST dataset following the method from [64]. Each image in this dataset contains two overlapping digits sampled from different classes of the MNIST hand-written digits dataset[45] (size 28×28 pixels). After the two digits are overlaid, each is shifted randomly up to 4 pixels in horizontal and vertical directions, resulting in images of size 36×36 pixels with on average 80% overlap between the digit bounding boxes. We generated 3M images for training, and 500K images for testing, and ensured that the training/testing sets were generated from the corresponding MNIST sets (i.e., the training set in MultiMNIST was only generated from the training set in MNIST).

SVRT task 1: We generated the SVRT stimuli for task 1 directly using the code from [21]. We generated 60K images for training and validation and 10K images for testing. The images are 64x64 pixels. The stimuli for the generalization task were generated using the code modified from [59]. For each of the nine types we generated 6K images for training (54K in total) and 1.2k for testing (10.8K in total). We also generated 5.4k images for the four OOD shape types that were only used for testing.

4.2 OCRA Architecture

The OCRA architecture is shown in Fig. 1. An overview of the OCRA components, loss functions with pseudocodes and implementation details are provided in this section. A pytorch[56] implementation of OCRA with additional details and results are provided in these repositories: github.com/hosseinadeli/OCRA and github.com/hosseinadeli/OCRA_samediff.

Pseudocode 1 OCRA Architecture Overview

Initialize encoder and decoder RNN hidden states, h_0^{enc}, h_0^{dec}, to zero
Initialize classification score vector s_0 and reconstruction $canvas_0$ to zero
for t in $timesteps = 1, 2, \ldots$ **do**
    Get read glimpse: $g_t \leftarrow READ(x, h_{t-1}^{dec})$ ▷ read operation in Pseudocode 2
    Apply two convolutional layers with max pooling: $g_t^{conv} \leftarrow CNN(g_t)$
    Update encoder RNN's hidden representations: $h_t^{enc} \leftarrow RNN^{enc}(h_{t-1}^{enc}, g_t^{conv})$
    Compute object capsules through dynamic routing: $d_t \leftarrow CAPSULE(h_t^{enc})$ ▷ capsule operation in Pseudocode 4
    Accumulate classification score for digit j: $s_t^j \leftarrow s_{t-1}^j + \|d_t^j\|$
    Mask with zeros all but the digit capsule with max classification score: $d_t^{mask} \leftarrow mask(d_t)$
    Update decoder RNN's hidden representations: $h_t^{dec} \leftarrow RNN^{dec}(h_{t-1}^{dec}, d_t^{mask})$
    Reconstruct image with write attention: $w_t \leftarrow WRITE(h_t^{dec})$ ▷ write operation in Pseudocode 3
    Update reconstruction canvas: $canvas_t = canvas_{t-1} + w_t$
end for

Read and Write Attention: At each timestep a glimpse, g_t, is sampled through applying a grid of N $\times$ N Gaussian filters on the input image x. We set the glimpse size to 18×18 for our experiments, with a sample glimpse shown in Fig. 1 left. The Gaussian filters are generated using four parameters: $g_X, g_Y, \delta, \sigma^2$, which specify the center coordinates of the attention window, the distance between equally spaced Gaussian filters in the grid, and the variance of the filters, respectively. All of these parameters are computed via a linear transformation of the previous step decoder RNN (Recurrent Neural Network) output h_{t-1}^{dec} using a weight matrix W_{read}, which makes the attention mechanism fully-differentiable. A similar procedure applies to the *write* attention operation. The decoder RNN output h_t^{dec} is linearly transformed into an M $\times$ M write patch w_t (set to 18 $\times$ 18 in our experiments), which is then multiplied by the Gaussian filters to reconstruct the written parts in the original image size (Fig. 1 right). The Gaussian filters used for the *write* operation differed from those used for the *read* operation, and were computed from four parameters obtained from a separate linear transformation, $W_{write_attention}$, of the decoder RNN output h_t^{dec}. Detailed algorithms for Read and Write attention operations are provided in Pseudocode 2 and 3 in the Supplementary along with an illustration of the attention mechanisms (Fig. 4).

Pseudocode 2 Read Attention

h_{t-1}^{dec} = previous decoder hidden representation
N= the size of read attention glimpse; W, H= (image width, height)
procedure READATTENTION(x, h_{t-1}^{dec}) ▷ Get a N×N read glimpse g_t by applying horizontal and vertical Gaussian filterbank matrices F_X, F_Y to the image
    Get attention grid parameters: $g_X, g_Y, \log\delta, \log\sigma^2 \leftarrow W_{read_attention}(h_{t-1}^{dec})$
    Scale attention grid centers: $g_X \leftarrow \frac{A+1}{2}(g_X + 1)$, $g_Y \leftarrow \frac{A+1}{2}(g_Y + 1)$
    Scale attention grid stride: $\delta \leftarrow \frac{max(A,B)-1}{N-1}\delta$
    Get x, y locations of each grid point i, j:
        $\mu_X^i = g_X + (i - N/2 - 0.5)\delta$
        $\mu_Y^i = g_Y + (j - N/2 - 0.5)\delta$
    Compute Gaussian filterbank matrices (Z_X and Z_Y are normalization constants):
        $F_X[i, w] = \frac{1}{Z_X}\exp(-\frac{(w-\mu_X^i)^2}{2\sigma^2})$
        $F_Y[j, h] = \frac{1}{Z_Y}\exp(-\frac{(h-\mu_X^j)^2}{2\sigma^2})$
    Get a read glimpse g_t: $g_t \leftarrow F_Y x F_X^T$
end procedure

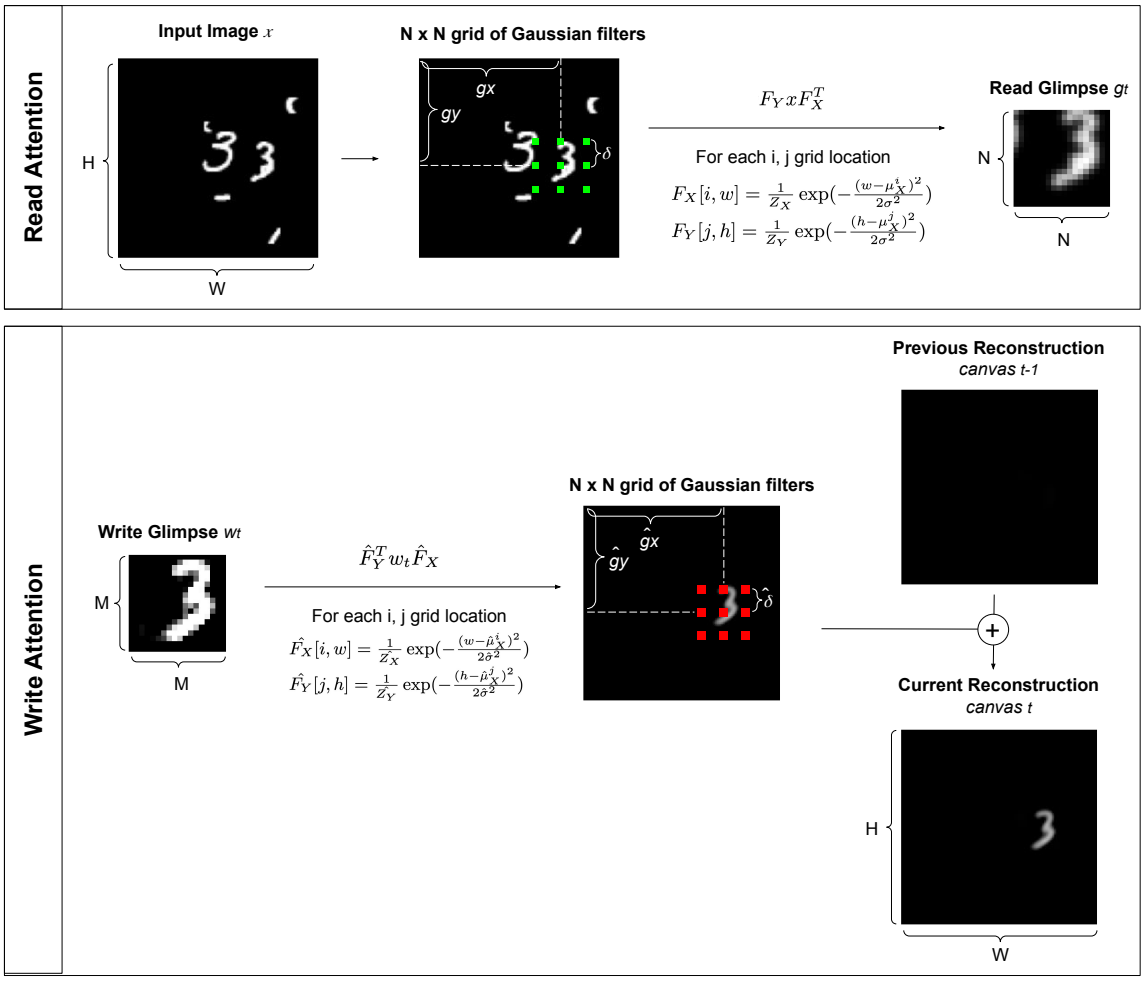

Figure 4: Visual illustrations of read and write attention mechanisms

Pseudocode 3 Write Attention

h_t^{dec} = current decoder hidden representation
M= the size of write attention glimpse, W, H= (image width, height)
procedure WRITEATTENTION(h_t^{dec}) ▷ Reconstruct a M×M write glimpse w_t to the original image size by applying transposed Gaussian filterbank matrices $\hat{F_X}, \hat{F_Y}$
    Get a write glimpse: $w_t \leftarrow W_{write}(h_t^{dec})$
    Get attention grid parameters: $\hat{g_X}, \hat{g_Y}, \log\hat{\delta}, \log\hat{\sigma}^2 \leftarrow W_{write_attention}(h_t^{dec})$
    Scale attention grid centers: $\hat{g_X} \leftarrow \frac{A+1}{2}(g_X+1)$, $\hat{g_Y} \leftarrow \frac{A+1}{2}(g_Y+1)$
    Scale attention grid stride: $\hat{\delta} \leftarrow \frac{max(A,B)-1}{M-1}\hat{\delta}$
    Get x, y locations of each grid point i, j:
        $\hat{\mu}_X^i = \hat{g_X} + (i - M/2 - 0.5)\hat{\delta}$
        $\hat{\mu}_Y^i = \hat{g_Y} + (j - M/2 - 0.5)\hat{\delta}$
    Compute Gaussian filterbank matrices ($\hat{Z_X}$ and $\hat{Z_Y}$ are normalization constants):
        $\hat{F_X}[i,w] = \frac{1}{\hat{Z_X}}\exp(-\frac{(w-\hat{\mu}_X^i)^2}{2\hat{\sigma}^2})$
        $\hat{F_Y}[j,h] = \frac{1}{\hat{Z_Y}}\exp(-\frac{(h-\hat{\mu}_X^j)^2}{2\hat{\sigma}^2})$
    Convert the write glimpse into the original image size: $w_t \leftarrow \hat{F}_Y^T w_t \hat{F}_X$
end procedure

Encoder: After a glimpse is selected from the input image by the read attention operation, it is processed first using a two-layer convolutional neural network (CNN) with 32 filters in each layer. Kernel sizes were set to 5 and 3 respectively for the first and the second layers. Each convolutional layer is followed by max pooling with a kernel size of 2 and a stride of 2, and Rectified Linear Units (ReLU) were used for non-linear activation functions. Given the glimpse size of 18 $\times$ 18, the resulting 32 feature maps are of size 4 $\times$ 4. The feature maps, g_t^{conv}, are reshaped (to a vector of size 512) and used as input to the encoder Recurrent Neural Network (RNN), along with the encoder RNN hidden state from the previous step; h_{t-1}^{enc}. We used LSTM[33] units (size 512) for the recurrent layers in our model.

Latent Capsule Representations and Dynamic Routing: We use a vector implementation of capsules[64] where the length of the vector represents the existence of the visual entity and the orientation characterizes its visual properties. The primary level capsules are generated through a linear read out of the encoder RNN; h_t^{enc}. These capsules are meant to represent lower-level visual entities ("parts") that belong to one of the higher-level capsules in the object capsule layer ("whole"). To find this part-whole relationship, we used the dynamic routing algorithm proposed by [64]. Dynamic routing is an iterative process where the assignments of parts to whole objects (coupling coefficients) are progressively determined by agreement between the two capsules (measured by the dot product between the two vector representations). Each primary level capsule (i) provides a prediction for each object level capsule (j). These predictions are then combined using the coupling coefficients (cij) to compute the object level capsule. Then the agreement (dot product) between the object level capsules and the predictions from each primary level capsule impacts the coupling coefficients for the next routing step.For example, if the prediction for a digit capsule j from a primary capsule i, ($\hat{p}_t^{j|i} \leftarrow W_t^{ij} p_t^i$), highly agrees with the computed digit capsule ($\sum_i c_t^{ij} \hat{p}_t^{j|i}$), the coupling coefficient c_t^{ij} is enhanced so that more information is routed from primary capsule i to object capsule j. Coupling coefficients are normalized across the class capsule dimension following the max-min normalization[88] as in Eq. 1. Lower- and upper-bounds for normalization, lb and ub, were set to 0.01 and 1.0. This routing procedure iterates three times. We used this method instead of the softmax normalization in [64] because we observed the latter method would not differentiate between the coupling coefficients. In our experiments we used 40 primary level capsules, each a vector of size 8. The object capsules are vectors of size 16 and there are 10 of them corresponding to the 10 digit categories for the multi-object recognition task and 4 of them for the visual reasoning task. For the object level capsules, we use a squash function (the Supplementary Eq. 2) to ensure that its vector length is within the range of 0 to 1. For the mutli-object recognition task, these would represent the probability of a digit being present in the glimpse at each step. Once the routing is completed, we compute the vector length (L2 norm) of each object capsule to obtain classification scores. The final digit classification is predicted based on the scores accumulated over all timesteps. For the visual reasoning task, two capsules (among four total) were designated to be the response capsules and the cumulative length of these judgment capsules were used for same-different response. The Algorithms are provided in Pseudocode 4.

$$c_t^{ij} = lb + (ub - lb)\frac{c_t^{ij} - min(c_t^{ij})}{max(c_t^{ij}) - min(c_t^{ij})} \tag{1}$$

$$d_t^j = \frac{\|v_t^j\|^2}{1 + \|v_t^j\|^2}\frac{v_t^j}{\|v_t^j\|} \tag{2}$$

Pseudocode 4 Capsule Representation and Dynamic Routing

h_t^{enc} = current encoder representation
procedure CAPSULE(h_t^{enc})
  Apply linear transform on h_t^{enc} to create primary capsule: $p_t \leftarrow W_p(h_t^{enc})$
  For all object capsule j and primary capsule i initialize b_t^{ij} to zero
  for r in $routings = 1, 2, \ldots$ **do**
    For all object capsule j compute prediction from all primary capsule i: $\hat{p}_t^{j|i} \leftarrow W_t^{ij} p_t^i$
    For all object capsule j: $d_t^j \leftarrow squash(\sum_i c_t^{ij} \hat{p}_t^{j|i})$ ▷ squash function in Eq. 2
    For all object capsule j and primary capsule i: $b_t^{ij} \leftarrow b_t^{ij} + \hat{p}_t^{j|i} \cdot d_t^j$
    Max-Min Normalize: $c_t^{ij} \leftarrow maxmin(b_t^{ij})$ ▷ maxmin function in Eq.1
  end for
  For all object capsule j: $d_t^j \leftarrow squash(\sum_i c_t^{ij} \hat{p}_t^{j|i})$
end procedure

Decoder: The object capsules provide a structured representation that can be used for decoding and glimpse selection. We first mask the object capsules so that only the vector representation from the most active capsule is forwarded to the decoder RNN, which also inputs the hidden state from the previous step, h_{t-1}^{dec}. Because the decoder maintains through recurrence the ongoing and evolving object-based representation of the image, it is best suited to determine the next read glimpse location (as discussed earlier). The state of the decoder RNN is also used through two linear operations to determine what and where to write in the canvas to reconstruct the image.

4.3 Loss Function

OCRA outputs object classification scores (cumulative capsule lengths) and image reconstruction (cumulative write canvas). Losses are computed for each output and combined with a weighting as in Eq. 3. For reconstruction loss, we simply computed the mean squared differences in pixel intensities between the input image and the model's reconstruction. For classification, we used margin loss (Eq. 4). For each class capsule j, the first term is only active if the target object is present ($T_j > 0$) where minimizing the loss pushes the capsule length to be bigger than target capsule length minus a small margin (m). The second term is only active when the target capsule length is zero and in that case minimizing the loss pushes the predicted capsule length to be below a small margin (m). For all the experiments in this paper we used Adam optimizer[41]. See Supplementary **??** for a detailed explanation and the pseudocode for the loss functions.

$$Total\,Loss = \sum_{j \in class} Class\,Loss_j + \lambda_{recon} Recon\,Loss \tag{3}$$

$$\begin{aligned} Class\,Loss = \sum_{j \in class} & max(0, min(T_j, 1)) \cdot max(0, (T_j - m) - \|d_j\|)^2 + \\ & \lambda_{absent} \cdot max(0, 1 - T_j) \cdot max(0, \|d_j\| - m)^2 \end{aligned} \tag{4}$$

4.4 Measuring accuracy

When we convert the model scores to multi-object classification, we take into account the thresholds that are set in the loss function. If the model prediction for one class capsule is larger than 1.8 (conservatively selected to be slightly lower than 1.9), this signals the presence of two objects from this class in the image. If no class score is larger than this threshold, the top two highest scores are selected as the model predictions for the two objects in the image.

4.5 Training Regime and Hyperparameter selection

All the model training was done on a single GPU workstation. All the Models trained on the MultiMNIST task were stopped after 50 epochs of training (taking about 40 hours for the 10glimpse model). The models trained on the MultiMNIST cluttered task were stopped after 1000 epochs (taking about 40 hours for the 5glimpse model). The models trained on the visual reasoning task task were stopped after 100 epochs (taking about 2 hours). We used the Adam optimizer [41] for all the experiments.

Table 4: Hyperparameter Setting for MultiMNIST and MultiMNIST Cluttered tasks

Hyperparameters	MultiMNIST-(10/3)glimpse	Cluttered-(5)glimpse	Same-Different
# timesteps, t	10/3	5	10
# epoch	50	1000	100
lr	0.001	0.001	0.001
batch size	128	128	128
read glimpse size, N	18	18	18
write glimpse size, M	18	18	18
# conv1 filters	32	32	32
# conv2 filters	32	32	32
lstm size, $dim(h_{enc}), dim(h_{dec})$	512	512	512
# primary capsule	40	40	40
primary capsule dimension, $dim(p)$	8	8	8
# routings r	3	3	3
object capsule dimension, $dim(d)$	16	16	16
# background capsules	0	1	0
reconstruction loss weight, λ_{recon}	10/3	175	30
clipping final canvas to [0,1]	TRUE	FALSE	TRUE
use reconstruction mask	FALSE	TRUE	FALSE

Table 4 provides all the hyperparameters used for the three tasks. The hyperparameters were mostly the same between the tasks with the few differences. For the MultiMNIST Cluttered task, we added one background capsule and utilized a reconstruction mask, both to allow the model to focus on the main objects and ignore the background clutter. The image is larger (100×100 vs 36×36) in the MultiMNIST Cluttered task, with most of it being empty, the reconstruction loss therefore has a much smaller range compared to the other task. For this reason, and the use of the reconstruction mask, we use a much larger weight for the reconstruction loss to make it comparable to the margin loss. For this task we added a background capsule to the 10 class capsules, similar in approach to [61]). This gives the model the choice to dynamically route background noise in its attention windows to a non-class capsule, thereby allowing the model to exclude the noise from its object representations.

For the MultiMNIST task, we clip the cumulative reconstruction canvas to be between [0,1] before comparing it to the input image. This allows the model to overlap different segments in the multi-step process of writing to the canvas without increasing the loss, improving model reconstruction given the high degree of overlap between the digits.

For the Visual Reasoning task, we used mostly similar hyperparameters. The number of class capsules were set to 4 as explained in the main text.

Acknowledgements

References

[1] Hossein Adeli, Françoise Vitu, and Gregory J Zelinsky. A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*, 37(6):1453–1467,

2017.

[2] Haider Al-Tahan and Yalda Mohsenzadeh. Reconstructing feedback representations in the ventral visual pathway with a generative adversarial autoencoder. *PLoS Computational Biology*, 17(3):e1008775, 2021.

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

[4] Shahab Bakhtiari, Patrick Mineault, Tim Lillicrap, Christopher C Pack, and Blake A Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *bioRxiv*, 2021.

[5] Daniel Baldauf and Robert Desimone. Neural mechanisms of object-based attention. *Science*, 344(6182):424–427, 2014.

[6] James W Bisley and Michael E Goldberg. Attention, intention, and priority in the parietal lobe. *Annual review of neuroscience*, 33:1–21, 2010.

[7] Jesse L Breedlove, Ghislain St-Yves, Cheryl A Olman, and Thomas Naselaris. Generative feedback explains distinct brain activity codes for seen and mental images. *Current Biology*, 30(12):2211–2224, 2020.

[8] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[9] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

[10] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016.

[11] Gustavo Deco and Edmund T Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44(6):621–642, 2004.

[12] Antoine Dedieu, Rajeev V Rikhye, Miguel Lázaro-Gredilla, and Dileep George. Learning attention-controllable border-ownership for objectness inference and binding. *bioRxiv*, pages 2020–12, 2021.

[13] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

[14] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[15] David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. *Advances in Neural Information Processing Systems*, 34, 2021.

[16] Adrien Doerig, Lynn Schmittwilken, Bilge Sayim, Mauro Manassi, and Michael H Herzog. Capsule networks as recurrent models of grouping and segmentation. *PLoS computational biology*, 16(7):e1008017, 2020.

[17] Matthias Ekman, Pieter R Roelfsema, and Floris P de Lange. Object selection by automatic spreading of top-down attentional signals in v1. *Journal of Neuroscience*, 40(48):9250–9259, 2020.

[18] Charles W Eriksen and James D St James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & psychophysics*, 40(4):225–240, 1986.

[19] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *arXiv preprint arXiv:1603.08575*, 2016.

[20] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[21] François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.

[22] Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16, 2021.

[23] Marta Garnelo and Murray Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019.

[24] Dileep George, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan, Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 358(6368):eaag2612, 2017.

[25] Anirudh Goyal, Alex Lamb, Phanideep Gampa, Philippe Beaudoin, Sergey Levine, Charles Blundell, Yoshua Bengio, and Michael Mozer. Object files and schemata: Factorizing declarative and procedural knowledge in dynamical systems. *arXiv preprint arXiv:2006.16225*, 2020.

[26] Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

[27] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

[28] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471. PMLR, 2015.

[29] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):1–14, 2021.

[32] Geoffrey E Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with em routing. In *International conference on learning representations*, 2018.

[33] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[34] Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, Tan Nguyen, Doris Y Tsao, and Anima Anandkumar. Neural networks with recurrent generative feedback. *arXiv preprint arXiv:2007.09200*, 2020.

[35] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

[36] Danique Jeurissen, Matthew W Self, and Pieter R Roelfsema. Serial grouping of 2d-image regions with object-based attention in humans. *Elife*, 5:e14320, 2016.

[37] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992.

[38] Kohitij Kar and James J DiCarlo. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, 109(1):164–176, 2021.

[39] Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019.

[40] Junkyung Kim, Matthew Ricci, and Thomas Serre. Not-so-clevr: learning same–different relations strains feedforward neural networks. *Interface focus*, 8(4):20180011, 2018.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[44] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[45] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[46] Jordan Lei, Ari S Benjamin, and Konrad P Kording. Object based attention through internal gating. *arXiv preprint arXiv:2106.04540*, 2021.

[47] Grace W Lindsay and Kenneth D Miller. How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, 7:e38105, 2018.

[48] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*, 2020.

[49] John HR Maunsell and Stefan Treue. Feature-based attention in visual cortex. *Trends in neurosciences*, 29(6):317–322, 2006.

[50] Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Solving the same-different task with convolutional neural networks. *Pattern Recognition Letters*, 143:75–80, 2021.

[51] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*, 2014.

[52] Notger G Müller, Oliver A Bartelt, Tobias H Donner, Arno Villringer, and Stephan A Brandt. A physiological correlate of the "zoom lens" of visual attention. *Journal of Neuroscience*, 23(9):3561–3565, 2003.

[53] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

[54] Aran Nayebi, Javier Sagastuy-Brena, Daniel M Bear, Kohitij Kar, Jonas Kubilius, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel LK Yamins. Goal-driven recurrent neural network models of the ventral visual stream. *bioRxiv*, 2021.

[55] David L Oden, Roger KR Thompson, and David Premack. Infant chimpanzees spontaneously perceive both concrete and abstract same/different relations. *Child development*, 61(3):621–631, 1990.

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[57] Benjamin Peters and Nikolaus Kriegeskorte. Capturing the objects of vision with neural networks. *Nature Human Behaviour*, pages 1–18, 2021.

[58] Arezoo Pooresmaeili and Pieter R Roelfsema. A growth-cone model for the spread of object-based attention during contour grouping. *Current Biology*, 24(24):2869–2877, 2014.

[59] Guillermo Puebla and Jeffrey Bowers. Can deep convolutional neural networks support relational reasoning in the same-different task? *bioRxiv*, 2021.

[60] Haozhi Qi, Xiaolong Wang, Deepak Pathak, Yi Ma, and Jitendra Malik. Learning long-term visual dynamics with region proposal interaction networks. *arXiv preprint arXiv:2008.02265*, 2020.

[61] Yao Qin, Nicholas Frosst, Colin Raffel, Garrison Cottrell, and Geoffrey Hinton. Deflecting adversarial attacks. *arXiv preprint arXiv:2002.07405*, 2020.

[62] Hamidreza Ramezanpour and Mazyar Fallah. The role of temporal cortex in the control of attention. *arXiv preprint arXiv:2106.10344*, 2021.

[63] Matthew Ricci, Rémi Cadène, and Thomas Serre. Same-different conceptualization: a machine vision perspective. *Current Opinion in Behavioral Sciences*, 37:47–55, 2021.

[64] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*, 2017.

[65] Ilaria Sani, Heiko Stemmann, Bradley Caron, Daniel Bullock, Torsten Stemmler, Manfred Fahle, Franco Pestilli, and Winrich A Freiwald. The human endogenous attentional control network includes a ventro-temporal cortical node. *Nature communications*, 12(1):1–16, 2021.

[66] Brian J Scholl. Objects and attention: The state of the art. *Cognition*, 80(1-2):1–46, 2001.

[67] Brian J Scholl and Zenon W Pylyshyn. Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive psychology*, 38(2):259–290, 1999.

[68] Avi Schwarzschild, Arjun Gupta, Amin Ghiasi, Micah Goldblum, and Tom Goldstein. The uncanny similarity of recurrence and depth. *arXiv preprint arXiv:2102.11011*, 2021.

[69] Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

[70] Courtney J Spoerer, Patrick McClure, and Nikolaus Kriegeskorte. Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, 8:1551, 2017.

[71] Sebastian Stabinger, David Peer, Justus Piater, and Antonio Rodríguez-Sánchez. Evaluating the progress of deep learning for visual relational concepts. *Journal of Vision*, 21(11):8–8, 2021.

[72] Sebastian Stabinger, David Peer, and Antonio Rodríguez-Sánchez. Arguments for the unsuitability of convolutional neural networks for non-local tasks. *Neural Networks*, 142:171–179, 2021.

[73] Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, pages 1–16, 2021.

[74] Michele Svanera, Andrew T Morgan, Lucy S Petro, and Lars Muckli. A self-supervised deep neural network for image completion resembles early visual cortex fmri activity patterns for occluded scenes. *Journal of Vision*, 21(7):5–5, 2021.

[75] Sara M Szczepanski, Mark A Pinsk, Malia M Douglas, Sabine Kastner, and Yuri B Saalmann. Functional and structural architecture of the human dorsal frontoparietal attention network. *Proceedings of the National Academy of Sciences*, 110(39):15806–15811, 2013.

[76] Anne Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, 1996.

[77] Leslie G Ungerleider and James V Haxby. 'what'and 'where'in the human brain. *Current opinion in neurobiology*, 4(2):157–165, 1994.

[78] Leslie G Ungerleider and Luiz Pessoa. What and where pathways. *Scholarpedia*, 3(11):5342, 2008.

[79] Ruben S van Bergen and Nikolaus Kriegeskorte. Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193, 2020.

[80] Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14, 2020.

[81] Rufin VanRullen and Simon J Thorpe. The time course of visual processing: from early perception to decision-making. *Journal of cognitive neuroscience*, 13(4):454–461, 2001.

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[83] Shaun P Vecera. Toward a biased competition account of object-based segregation and attention. *Brain and Mind*, 1(3):353–384, 2000.

[84] Shaun P Vecera and Randall C O'reilly. Figure-ground organization and object recognition processes: an interactive account. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2):441, 1998.

[85] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[86] Dean Wyatte, David J Jilk, and Randall C O'Reilly. Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in psychology*, 5:674, 2014.

[87] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[88] Zhen Zhao, Ashley Kleinhans, Gursharan Sandhu, Ishan Patel, and KP Unnikrishnan. Capsule networks with max-min normalization. *arXiv preprint arXiv:1903.09662*, 2019.

[89] Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9483–9492, 2020.

A Supplementary 2

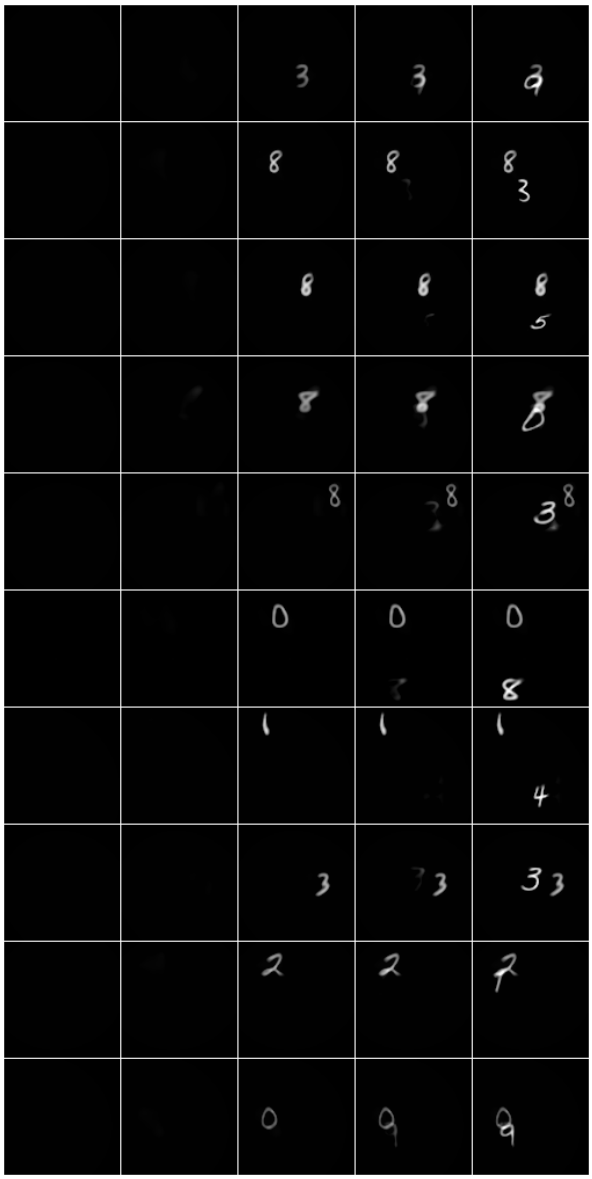

Figure 5: OCRA MultiMNIST Cluttered output with 5 glimpses showing the gradual object-based reconstruction on the cumulative canvas.

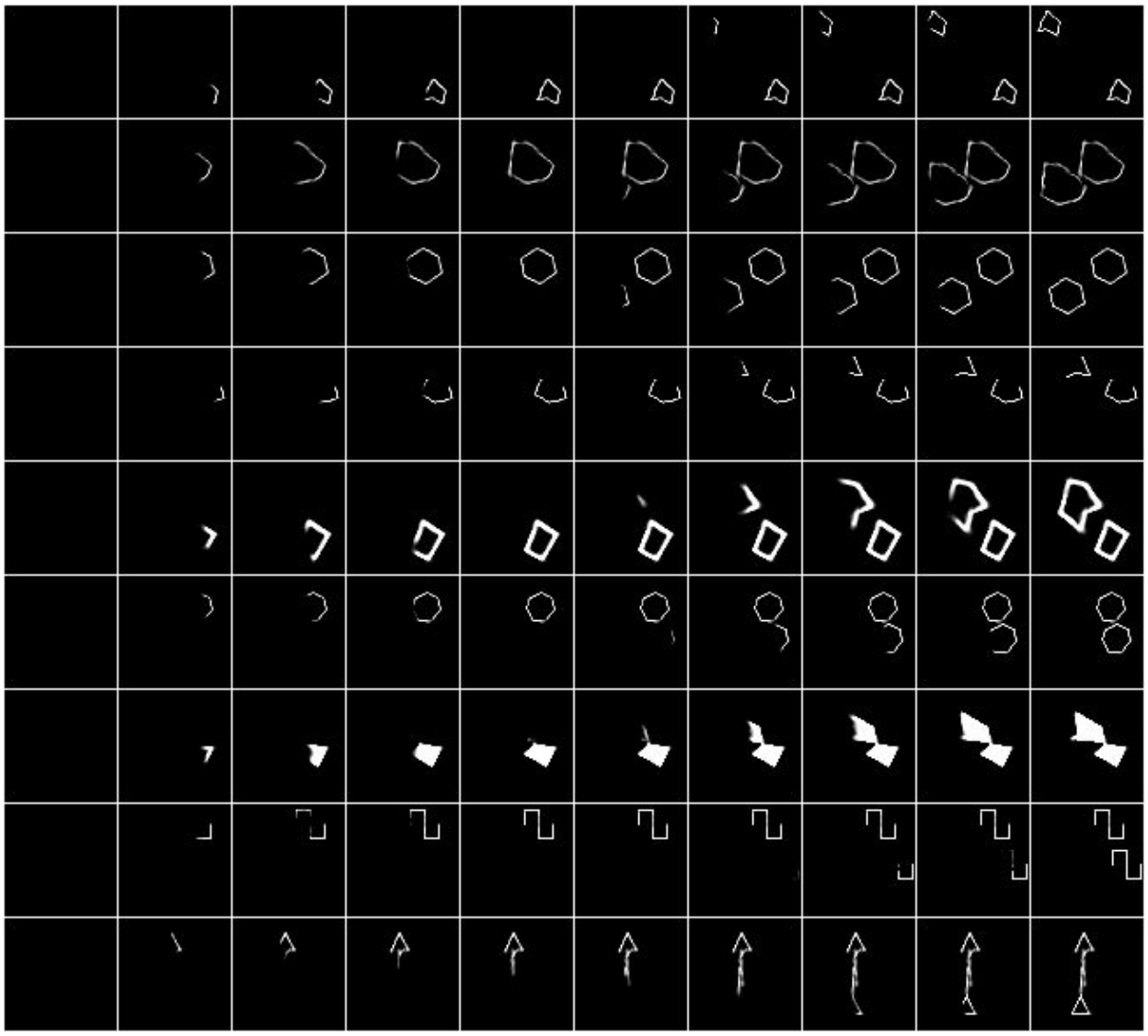

Figure 6: OCRA visual reasoning with 10 glimpses showing the gradual object-based reconstruction on the cumulative canvas.