LETTER **Communicated by Dina Obeid**

Identifying and Localizing Multiple Objects Using Artificial Ventral and Dorsal Cortical Visual Pathways

Zhixian Han
han594@purdue.edu
Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907, U.S.A.

Anne Sereno
asereno@purdue.edu
Department of Psychological Sciences and Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN 47907, U.S.A.

In our previous study (Han & Sereno, 2022a), we found that two artificial cortical visual pathways trained for either identity or space actively retain information about both identity and space independently and differently. We also found that this independently and differently retained information about identity and space in two separate pathways may be necessary to accurately and optimally recognize and localize objects. One limitation of our previous study was that there was only one object in each visual image, whereas in reality, there may be multiple objects in a scene. In this study, we find we are able to generalize our findings to object recognition and localization tasks where multiple objects are present in each visual image. We constrain the binding problem by training the identity network pathway to report the identities of objects in a given order according to the relative spatial relationships between the objects, given that most visual cortical areas including high-level ventral steam areas retain spatial information. Under these conditions, we find that the artificial neural networks with two pathways for identity and space have better performance in multiple-objects recognition and localization tasks (higher average testing accuracy, lower testing accuracy variance, less training time) than the artificial neural networks with a single pathway. We also find that the required number of training samples and the required training time increase quickly, and potentially exponentially, when the number of objects in each image increases, and we suggest that binding information from multiple objects simultaneously within any network (cortical area) induces conflict or competition and may be part of the reason why our brain has limited attentional and visual working memory capacities.

Neural Computation **35, 249–275** (2023)
https://doi.org/10.1162/neco_a_01559

1 Introduction

According to many neuropsychological, lesion, and anatomical studies, the human visual system has two major distinct cortical pathways (Felleman & Essen, 1991; Mishkin, Ungerleider, & Macko, 1983; Ungerleider & Mishkin, 1982). The ventral pathway is concerned with object identity (Logothetis & Sheinberg, 1996) and the dorsal pathway with spatial cognition (Colby & Goldberg, 1999). However, some recent studies argued that representations associated with shape and location processing are present in both visual streams (Konen & Kastner, 2008; Lehky & Sereno, 2007; Sereno & Lehky, 2011; Sereno, Lehky, & Sereno, 2020). In a previous study using artificial neural networks (Han & Sereno, 2022a), we showed that the two cortical visual pathways for identity and space actively retained information about both identity and space independently and differently. We also showed that this independent and different retained information about identity and space in the two modeled pathways was necessary to accurately and optimally recognize and localize objects. One limitation of our previous study was that there was only one object in each visual image, whereas in reality, there may be multiple objects in a scene.

In our current study, we try to generalize our models to multiple objects' recognition and localization tasks. One of the difficulties of dealing with images with multiple objects is the binding problem, where the representation of multiple objects with independent feature sets can lose information about which features belong to which objects (Markov, Utochkin, & Brady, 2021). Given that our previous study showed that the identity pathway actively retained information about space, we wanted to test whether it may be possible to constrain the binding problem if we take advantage of this information (i.e., the spatial information in the identity network pathway). Our previous study also showed that the kinds of information that the network actively retains depended on the tasks or goals that were used for training the network. In our current study, we trained the identity network pathway by asking it to report the identities of the objects in a certain order that depends on the relative spatial relationships between objects in the image. As a result, the identity network pathway would actively retain information about the relative spatial relationships between objects. Asking the identity network to retain relative spatial relationships is plausible because previous physiological work has shown that cells in high-level ventral areas retain spatial information of objects, including retinotopic spatial information (Op De Beeck & Vogels, 2000; Sereno & Lehky, 2011) and angle of gaze spatial information (Sereno, Sereno, & Lehky, 2014), as well as spatial relationship among object parts (Yamane, Tsunoda, Matsumoto, Phillips, & Tanifuji, 2006), information needed for scene recognition (where the objects are part of a larger scene) and object recognition, respectively. Furthermore, even fMRI studies, with their poorer spatial resolution, have also demonstrated that much of human neocortex contains topological maps of sensory

surfaces (Sereno, Sood, & Huang, 2022). In our prior work (Han & Sereno, 2022a), we showed that the simulated ventral pathway needed information about the relative spatial relationships between object parts to recognize the identity of the whole object (see also the discussion of the spatial relation of the faucet and basin of a sink in Figure 10b, in Sereno et al., 2020). Additionally, preliminary modeling results (Han & Sereno, 2022b) suggest that information about the relative spatial relationships between objects is able to constrain the binding problem when we combine the outputs of the identity network pathway and the spatial network pathway and process them together using a two-pathway neural network.

Previous studies have used artificial neural networks trained with supervised learning, self-supervised learning, or unsupervised learning to simulate the ventral and dorsal cortical visual pathways in the brain (Yamins et al., 2014; Kriegeskorte, 2015; Dobs, Martinez, Kell, & Kanwisher, 2022; Konkle & Alvarez, 2022; Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021; Zhuang et al., 2022). Many of these previous studies found that artificial convolutional neural network models could successfully produce brain-like neural responses or even predict neural responses in the biological visual cortex. However, the main goal of our study is to gain a better understanding of the consequences of brain structure or segregated streams of processing using computational modeling rather than identifying the specific response features that are similar to the real neural responses of ventral and dorsal cortical pathways.

In our study, feedforward convolutional neural networks were used to simulate the two cortical visual pathways. All neural networks in our study were trained using supervised learning. When modeling the two cortical visual pathways, it is assumed that the two pathways use the same structure for simplicity and control. We trained the two neural networks separately using multiple-objects recognition tasks and multiple-objects localization tasks, respectively, so that the trained neural networks will be able to model the ventral and dorsal pathways, respectively. We used stochastic gradient descent with backpropagation to update the weights in the neural networks during training. Stochastic gradient descent with backpropagation is currently the best method for updating connection weights between neurons in artificial neural networks, and some have argued that the brain might be able to implement backpropagation-like effective synaptic updates (Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020; Whittington & Bogacz, 2019).

Black and white images consisting of different kinds of tops, pants, and shoes (Xiao, Rasul, & Vollgraf, 2017) were used as the objects in the images (see Figure 1). Multiple objects were put in front of a black background at one of the nine possible locations. These images with multiple objects were used as visual inputs to the neural networks (see section 2 for details).

One artificial neural network, $Network_{identity}$, was trained to do an identity task (to identify whether the objects are tops, pants, or shoes). Another

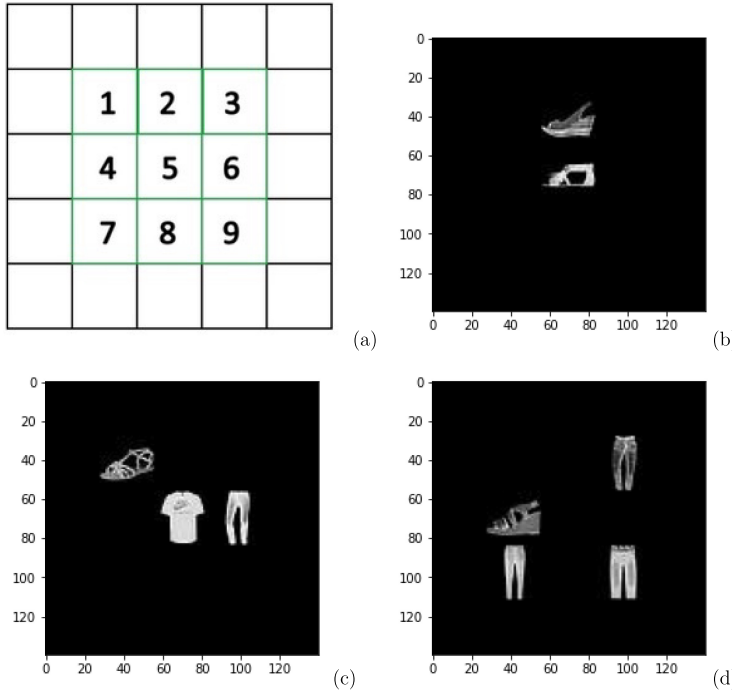


Figure 1: Possible object locations and examples of input images. (a) Nine possible locations of the objects. (b) An example of an input image with two objects. (c) An example of an input image with three objects. (d) An example of an input image with four objects.

artificial neural network, $Network_{location}$, was trained to do a localization task (to determine the locations of the objects). $Network_{identity}$ was used to model the ventral pathway, whereas $Network_{location}$ was used to model the dorsal pathway. These two networks were used to simulate the functions of ventral and dorsal cortical visual pathways in the brain. $Network_{identity}$ and $Network_{location}$ were trained independently to serve as the two pathways in $Network_{two\ pathways}$. The goal of $Network_{two\ pathways}$ is to recognize and localize multiple objects in the image at the same time. For comparison, another neural network, $Network_{one\ pathway}$, was also trained to recognize and localize multiple objects in the image at the same time. The sizes of $Network_{one\ pathway}$ and $Network_{two\ pathways}$ are equal. The difference is that $Network_{one\ pathway}$ has only one pathway, and all the training occurs as a single network (the two pathways in $Network_{two\ pathways}$ are trained as independent networks).

According to our simulation results, $Network_{two\ pathways}$ was able to outperform $Network_{one\ pathway}$ in almost all experimental conditions (different

numbers of objects in each image, different numbers of training samples). $Network_{two\ pathways}$ was able to achieve significantly higher average testing accuracy, had smaller testing accuracy variance, and required fewer training epochs and training time. However, the required training samples and training time increased quickly when the number of objects in each image increased. As a result, neither of the two networks was able to efficiently achieve high testing accuracies when there were four or more objects in the image. Though it may be a limitation of our models, this phenomenon may agree with the experimental evidence that shows our brain has a limited attention and working memory capacity for many cognitive processes, such as the processes involved in visual perception tasks, digital span tasks, and reading span tasks (Isbell, Fukuda, Neville, & Vogel, 2015; Miller, 1956; Daneman & Carpenter, 1980). Our models were not able to achieve high performance when there were four or more objects in the image because the binding problem became more difficult as the number of objects increased. Therefore, we suggest that capacity limits may be in part a consequence of the binding problem.

Similar to our previous study, our modeling is proof of the computational concept and better understanding of the effects of different organizational schemes more than an accurate model of the real human brain. Multiple-objects recognition and localization tasks are very important in both cognitive neuroscience and computer science. Our models may be able to help people get a better understanding of the computational costs and benefits of brain organization. Our models may also be able to provide insights about how to find better, more efficient, and more biologically plausible multiple-objects recognition and localization algorithms.

2 Methods

2.1 Objects. Black and white images of different kinds of tops, pants, and shoes obtained from the data set Fashion-MNIST were used as the objects in the object recognition and localization tasks (Xiao et al., 2017). There are 62 kinds of tops, 66 kinds of pants, and 58 kinds of shoes. Each object image was embedded in a black background and presented at different locations. There may be two, three, or four objects in each black background image. These object images with black background were used as visual inputs. Some examples of these input images are shown in Figure 1.

These black and white images were used to train, validate, and test the neural networks: two-thirds of the total number of images were used for training, one-sixth of the total number of images were used for validating, and one-sixth of the total number of images were used for testing.

2.2 Object Locations. Object image locations are shown and explained in Figure 1. The objects were put at different locations in a 140 $\times$ 140 (pixels) black square background. Specifically, each object image could have nine

possible locations (see Figure 1a). The objects in the same visual image are always at different locations, and they never overlap with each other.

2.3 Neural Networks. Feedforward convolutional neural networks were used to build brain networks to model the two cortical visual pathways in the brain. Each neural network consists of several hidden layers, including the convolutional layers, the pooling layers, and the fully connected dense layers. ReLU activation function was used at each layer except the final output layer, in which a softmax activation function was used. These neural networks were implemented using TensorFlow and were trained using supervised learning, the cross-entropy loss function, and stochastic gradient descent with backpropagation.

Our primary goal in this study is not optimizing each artificial neural network to achieve the highest performance. It is trying to compare the performance of one-pathway and two-pathway artificial neural network architectures when they have the same hyperparameter settings. In our previous study (Han & Sereno, 2022a), we repeated some simulations with some different hyperparameter settings (e.g., number of layers, number of filters, filter sizes) in the artificial neural networks. We found that our findings do not depend on specific hyperparameter settings of the artificial neural networks. Therefore, in our current study, we choose similar hyperparameter settings that were used in our previous study.

A batch size of 256 and the Adam optimization method were used while training. The initial learning rate of Adam optimization was 0.001. The other hyperparameters are specified in Figures 2, 3, and 4. We applied 30% random dropout to all the dense layers in all neural networks during training for regularization. All networks were trained with enough epochs to ensure that all of them had reached the highest possible validation accuracy at the end of training.

The structure of $Network_{identity}$ and $Network_{location}$ is shown in Figure 2. These two neural networks share the same structure, and the only difference between them was in their final output layers. Both networks take the same set of images as inputs. However, they were trained to do different tasks, so their output layers have different sizes. These two neural networks were trained to serve as the two pathways in $Network_{two\ pathways}$ for simultaneous multiple-objects recognition and localization.

$Network_{identity}$ was trained to determine the identities of the objects (tops, pants, or shoes) and report their identities according to their relative locations. Specifically, it was trained to report the identities according to this order: it should report the identity of the object at the top of the image first. If two objects are at the same horizontal line, then it should report the identity of the object on the left first. For example, when it receives the input image shown in Figure 1c, it should report the identities of all the objects in this order: shoe, top, pant. This information was represented in the output layer of the network using one-hot encoding. Note that the specific order

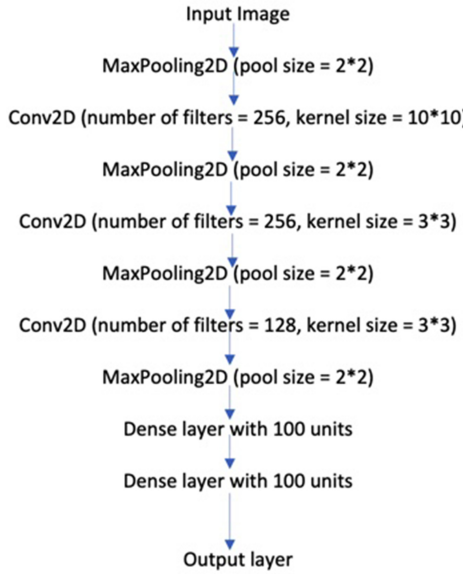


Figure 2: The structure of $Network_{identity}$ and $Network_{location}$.

described here is just an assumption without loss of generality: any particular (but consistent) spatial report would suffice. In general, the only requirement is that the one-hot vector representation in the final output layer of $Network_{identity}$ is determined by both the identities of the objects and the spatial relationships between the objects.

$Network_{location}$ was trained to determine the locations of the objects. It should report the locations of all the objects in the image regardless of their identities. For example, when it receives the input image shown in Figure 1c, it should report the locations of all the objects: locations 1, 5, and 6. This information was also represented in the output layer of the network using one-hot encoding.

The structures of $Network_{one\ pathway}$ and $Network_{two\ pathways}$ are shown in Figures 3 and 4. The sizes of $Network_{one\ pathway}$ and $Network_{two\ pathways}$ are designed to be equal, which means they have the same number of layers, the same total number of kernels in each convolutional layer, and the same total number of units in each dense layer. We chose to keep the number of units the same because our brain has a limited number of neurons, but the number of connections (the number of parameters) in our brain is more flexible. The only difference between $Network_{one\ pathway}$ and $Network_{two\ pathways}$ is their architectures. $Network_{one\ pathway}$ was trained to determine the identities and locations of all the objects in each image simultaneously using only one pathway. It took images as inputs, and the output layer

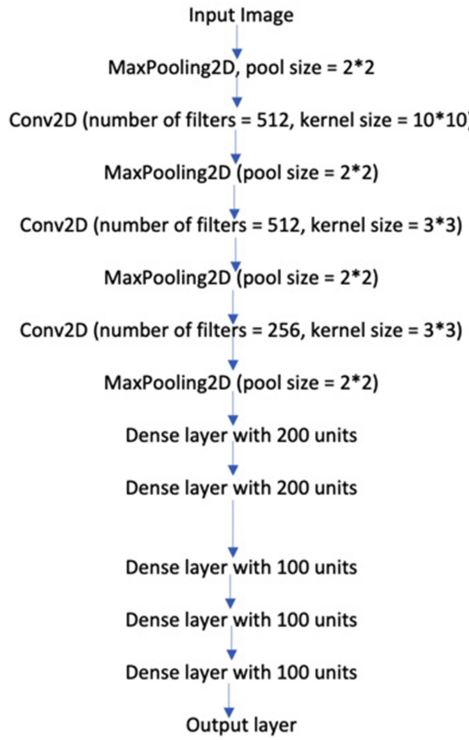


Figure 3: The structure of $Network_{one\ pathway}$.

reported the identities and locations of all the objects using one-hot encoding. $Network_{two\ pathways}$ was trained to determine the identities and locations of all the objects in each image simultaneously by processing the input image in two pathways and then combining them together. It took the images as inputs and sent this information into two pathways. The independently trained $Network_{identity}$ and $Network_{location}$ (excluding their one-hot encoded output layers) were used as the two pathways in $Network_{two\ pathways}$, which processed the input images with the $Network_{identity}$ pathway and the $Network_{location}$ pathways separately. Then the network concatenated the final layers of the two pathways together and processed the information jointly with some additional common dense layers. After the two pathways had been independently trained and their weights fixed, the common dense layers in $Network_{two\ pathways}$ were trained to report the identities and locations of all the objects using one-hot encoding.

Each network was trained five times, and testing accuracies were obtained for each of the five training sessions. The testing accuracies were obtained by dividing the number of correct classifications by the total number

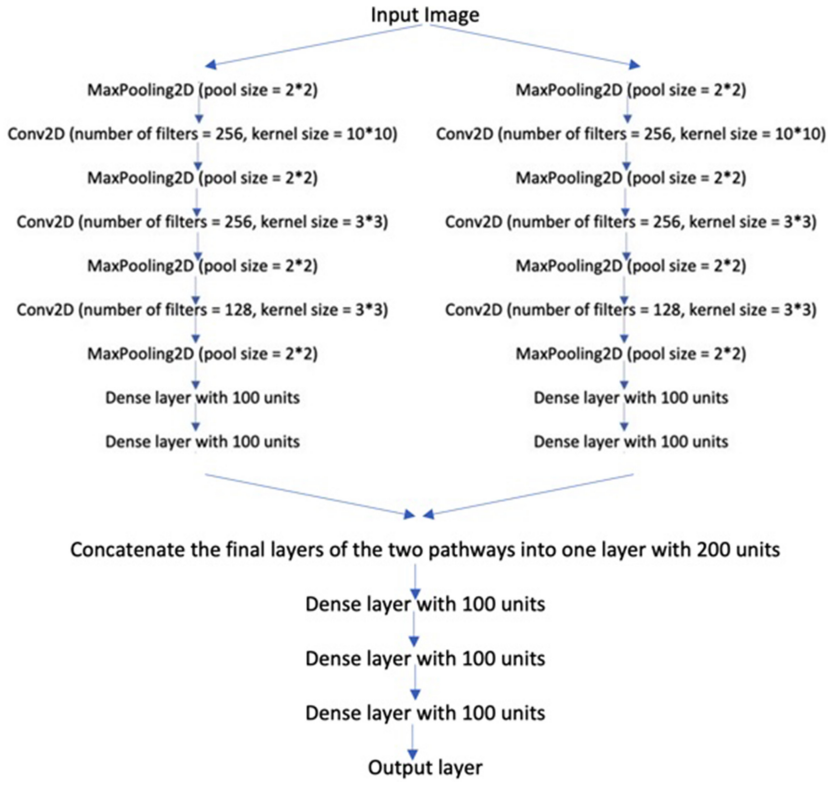


Figure 4: The structure of $Network_{two\ pathways}$.

of testing samples during the testing session. The accuracies that are used to compare different networks in this letter are always referring to the testing accuracies. Welch's two-sample t-tests were used to compare network accuracies and determine the significance of the differences.

3 Results

We performed training, validation, and testing multiple times for each network with network weights randomly initialized differently each time. When obtaining the accuracies in each experimental setting, the networks were always trained five times, and five testing accuracies were obtained for each condition after training. Welch's two-sample t-tests were used to compare different testing accuracies and determine the significance of the differences. The difference between testing accuracies is considered to be significant if the corresponding p-value <0.05. The average testing

Table 1: Average Testing Accuracies in Percentage (%) ± Standard Deviations (%) for Different Networks When There Are Two Objects in Each Image.

	Number of Samples			
Network	600	2400	6000	12,000
$Network_{identity}$ (chance level ≈ 11.1)	81.8 ± 2.2	98.4 ± 0.2	99.9 ± 0.1	100.0 ± 0.0
$Network_{location}$ (chance level ≈ 1.2)	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
$Network_{two\ pathways}$ (chance level ≈ 0.1)	62.2 ± 2.2	98.2 ± 0.1	99.9 ± 0.1	100.0 ± 0.0
$Network_{one\ pathway}$ (chance level ≈ 0.1)	61.6 ± 5.0	97.8 ± 0.5	99.9 ± 0.1	100.0 ± 0.0

Notes: The row heading are the names of the networks. The column heading are the total number of samples for training, validation, and testing. $Network_{identity}$ was trained to report identities of all objects according to their relative locations. $Network_{location}$ was trained to determine locations of all objects. $Network_{two\ pathways}$ and $Network_{one\ pathway}$ were trained to determine the identity and location of each object in the image. The chance-level accuracies (%) are reported next to the network names in the table.

accuracies of different neural networks for input images with two objects, three objects, and four objects are shown in Tables 1, 2, and 3, respectively.

3.1 Two Objects. According to the results shown in Table 1, the testing accuracies of $Network_{location}$ were always 100% for different total numbers of samples. It may be because it is very easy to determine the locations of all the objects when there are only two objects in each image. The testing accuracies of all the other neural networks increased when the total number of samples increased. Though the difference between $Network_{one\ pathway}$ and $Network_{two\ pathways}$ average testing accuracies was small and not significant ($p > 0.05$), the standard deviations of $Network_{two\ pathways}$ were smaller than or equal to the standard deviations of $Network_{one\ pathway}$. These results indicate that the performance of $Network_{two\ pathways}$ was more stable than $Network_{one\ pathway}$.

3.2 Three Objects. According to the results shown in Table 2, the testing accuracy of $Network_{location}$ was 100% when the total number of samples used was 2400 or more. It may be because it is relatively easy to determine the locations of all the objects when there are three objects in each image. The testing accuracies of all the other neural networks increased when the total number of samples increased. The difference between $Network_{one\ pathway}$ and $Network_{two\ pathways}$ average testing accuracies was significant ($p < 0.05$), and the standard deviations of $Network_{two\ pathways}$ accuracies were smaller than the standard deviations of $Network_{one\ pathway}$ accuracies. These results indicate that the performance of $Network_{two\ pathways}$ was higher than $Network_{one\ pathway}$ and the performance of $Network_{two\ pathways}$ was more stable than $Network_{one\ pathway}$.

Table 2: Average Testing Accuracies in Percentage (%) ± Standard Deviations (%) for Different Networks When There Are Three Objects in Each Image.

	Number of Samples			
Network	600	2400	6000	12,000
$Network_{identity}$ (chance level ≈ 3.7)	35.6 ± 2.6	93.8 ± 0.8	99.6 ± 0.3	99.9 ± 0.1
$Network_{location}$ (chance level ≈ 0.1)	97.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
$Network_{two\ pathways}$ (chance level ≈ 0.0)	16.8 ± 1.5	47.6 ± 0.3	85.4 ± 0.1	96.9 ± 0.0
$Network_{one\ pathway}$ (chance level ≈ 0.0)	12.2 ± 2.6	29.6 ± 3.4	63.7 ± 10.5	91.3 ± 2.5

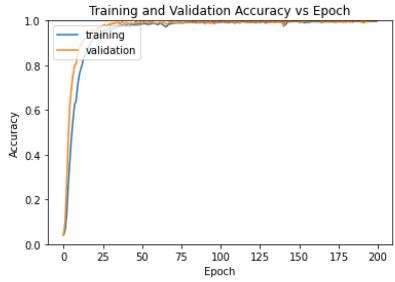
Notes: The row headings are the names of the networks. The column headings are the total number of samples for training, validation, and testing. $Network_{identity}$ was trained to report identities of all objects according to their relative locations. $Network_{location}$ was trained to determine locations of all objects. $Network_{two\ pathways}$ and $Network_{one\ pathway}$ were trained to determine the identity and location of each object in the image. The chance-level accuracies (%) are reported next to the network names in the table.

Table 3: Average Testing Accuracies in Percentage (%) ± Standard Deviations (%) for Different Networks When There Are Four Objects in Each Image.

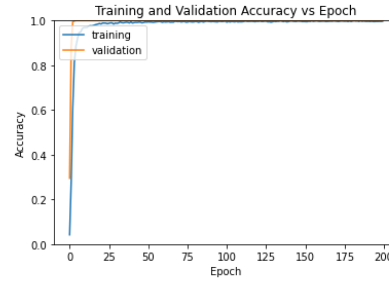
	Number of Samples			
Network	600	2400	6000	12,000
$Network_{identity}$ (chance level ≈ 1.2)	14.0 ± 3.0	69.2 ± 1.1	98.7 ± 0.8	99.8 ± 0.1
$Network_{location}$ (chance level ≈ 0.0)	99.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
$Network_{two\ pathways}$ (chance level ≈ 0.0)	6.2 ± 0.8	15.0 ± 0.5	31.4 ± 0.1	54.8 ± 0.1
$Network_{one\ pathway}$ (chance level ≈ 0.0)	3.6 ± 1.5	3.8 ± 1.1	5.1 ± 0.8	14.8 ± 1.3

Notes: The row headings are the names of the networks. The column headings are the total number of samples for training, validation, and testing. $Network_{identity}$ was trained to report identities of all objects according to their relative locations. $Network_{location}$ was trained to determine locations of all objects. $Network_{two\ pathways}$ and $Network_{one\ pathway}$ were trained to determine the identity and location of each object in the image. The chance-level accuracies (%) are reported next to the network names in the table.

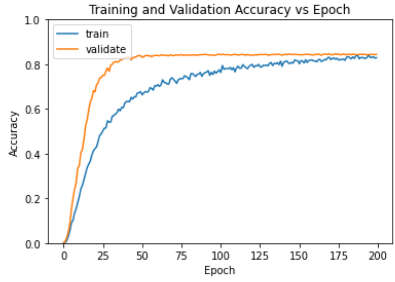
3.3 Four Objects. According to the results shown in Table 3, the testing accuracy of $Network_{location}$ was 100% when the total number of samples used was 2400 or more. It may be because it is still relatively easy to determine the locations of all the objects when there are four objects in each image. The testing accuracies of all the other neural networks increased when the total number of samples increased. The difference between the $Network_{one\ pathway}$ and $Network_{two\ pathways}$ average testing accuracies was significant ($p < 0.05$), and the standard deviations of $Network_{two\ pathways}$ accuracies were smaller than the standard deviations of $Network_{one\ pathway}$ accuracies. These results indicate that the performance of $Network_{two\ pathways}$ was higher than $Network_{one\ pathway}$, and the performance of $Network_{two\ pathways}$ was more stable than $Network_{one\ pathway}$.



(a) Identity pathway in $Network_{two\ pathways}$



(b) Location pathway in $Network_{two\ pathways}$



(c) Common dense layers in $Network_{two\ pathways}$

(d) $Network_{one\ pathway}$

Figure 5: The training and validation curves of $Network_{one\ pathway}$ and $Network_{two\ pathways}$ with 6000 total samples and three objects in each image. (a) The training and validation curves when training the identity pathway in $Network_{two\ pathways}$. (b) The training and validation curves when training the location pathway in $Network_{two\ pathways}$. (c) The training and validation curves when training the common dense layers in $Network_{two\ pathways}$. When training the common dense layers, the two pathways had been trained and their weights were fixed. (d) The training and validation curves of $Network_{one\ pathway}$ (note that the scale of the x-axis is greatly expanded in this panel: 1500 epochs, compared to all other panels, 200 epochs).

3.4 The Number of Epochs and Time Required for Training $Network_{two\ pathways}$ and $Network_{one\ pathway}$. According to the training and validation curves shown in Figure 5, $Network_{two\ pathways}$ could achieve the highest validation accuracy with fewer epochs. With three objects per image and 6000 samples in total, training $Network_{two\ pathways}$ required only around 100 epochs in total (see Figures 5a, 5b, and 5c), whereas training $Network_{one\ pathway}$ required around 1500 epochs (see Figure 5d; note that the scale of the x-axis is greatly expanded in this panel, 1500 epochs, compared to all other panels, 200 epochs). When determining the number of epochs required to train $Network_{two\ pathways}$, the number of epochs for training the two pathways independently (in Figure 5a, around 40 epochs; in Figure 5b,

around 10 epochs), and the number of epochs for training the common dense layers (in Figure 5c, around 50 epochs) were added together to get the number of epochs in total, around 100. The number of epochs required for training is an estimation. It is the estimated number of epochs until the validation accuracy no longer increases.

In addition, we measured the time spent during each training epoch according to TensorFlow logs, which are automatically output when using the model.fit command, and indicate the amount of time it took to train each epoch. For $Network_{two\ pathways}$, it took around 1 second per epoch to train $Network_{identity}$, around 1 second per epoch to train $Network_{location}$, and less than 1 second per epoch to train common dense layers. For $Network_{one\ pathway}$, each training epoch took around 3 seconds. Therefore, each training epoch in training $Network_{two\ pathways}$ (around or less than 1 s) always took much less time than each training epoch in training $Network_{one\ pathway}$, which was around 3 s). This finding may be because there are fewer weight parameters that need to be updated in $Network_{two\ pathways}$ during training.

Because training $Network_{two\ pathways}$ required fewer epochs and each training epoch took less time, the total training time required for $Network_{two\ pathways}$ was shorter. These results about required number of training epochs and training time were obtained using three-object images and 6000 samples. Similar results were also found with different numbers of objects and different numbers of samples.

3.5 Comparing the Performance of $Network_{two\ pathways}$ and $Network_{one\ pathway}$. We compared the performance of $Network_{two\ pathways}$ and $Network_{one\ pathway}$ when there were different numbers of objects in each image and when different total number of samples were used. The comparison is shown in Figure 6. $Network_{two\ pathways}$ had significantly better performance (higher average testing accuracy with smaller variance) than $Network_{one\ pathway}$ in almost all conditions. The number of samples required for both networks to reach high testing accuracy (>90%) increased when the number of objects in each image increased. With two objects in each image, 2400 samples were sufficient for $Network_{two\ pathways}$ to reach a high testing accuracy (>90%). With three objects in each image, $Network_{two\ pathways}$ required 12,000 samples to achieve a similar high testing accuracy (>90%). With four objects in each image, $Network_{two\ pathways}$ could reach only a much lower testing accuracy (54.8 ± 0.1)% with 12,000 samples. As shown in Table 4, $Network_{two\ pathways}$ reached a similar high testing accuracy (>90%) with 60,000 samples. However, the training was slow and inefficient when 12,000 or more samples were used for training, validation, and testing. Training was even slower and more inefficient when we increased the total number of samples further. The same is true and even worse for $Network_{one\ pathway}$. Therefore, we did not try to train the neural networks with more samples for all conditions (but see section 3.10). In summary, the required number of training samples and the required training time for

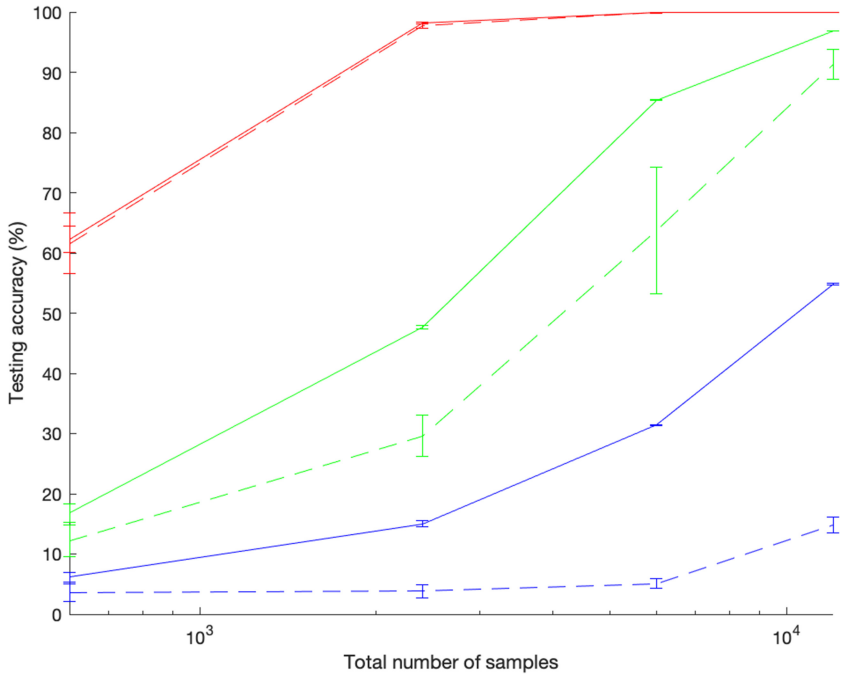


Figure 6: The testing accuracies of $Network_{two\ pathways}$ (solid lines) and $Network_{one\ pathway}$ (dashed lines) for two objects (red), three objects (green), and four objects (blue) images when the total numbers of samples used for training, validation, and testing were 600, 2400, 6000, and 12,000.

high network performance increased quickly when the number of objects in each image increased.

3.6 Compare the Performance of $Network_{one\ pathway}$ with or without Pretraining. The $Network_{one\ pathway}$ results reported above were obtained without any pretraining. In order to find out whether pretraining could improve the performance of $Network_{one\ pathway}$, we conducted additional simulations to test the performance of $Network_{one\ pathway}$ with pretraining. With three objects per image and 6000 samples in total, we pretrained $Network_{one\ pathway}$ with the identity task first and the location task later; the testing accuracy of $Network_{one\ pathway}$ on the multiple-objects recognition and localization task was (12.2 ± 8.9)%. We also pretrained $Network_{one\ pathway}$ in the other order, with the location task first and the identity task later. The testing accuracy of $Network_{one\ pathway}$ in this case was (11.6 ± 10.2)%. The accuracies of $Network_{one\ pathway}$ in both pretraining cases were significantly lower than the accuracy of $Network_{one\ pathway}$ without pretraining (63.7 ± 10.5)%.

Table 4: Average Testing Accuracies in Percentage (%) ± Standard Deviations (%) for $Network_{two\ pathways}$ When There Are Two, Three, or Four Objects in Each Image.

	1200	2400	6000	12,000	30,000	60,000
Two objects	82.2 ± 0.6[1]	98.2 ± 0.1[2]	99.9 ± 0.1	100.0 ± 0.0	NA[3]	NA[3]
Three objects	NA[3]	47.6 ± 0.3	85.4 ± 0.1[1]	96.9 ± 0.0[2]	NA[3]	NA[3]
Four objects	NA[3]	15.0 ± 0.5	31.4 ± 0.1	54.8 ± 0.1	85.7 ± 0.1[1]	98.1 ± 0.0[2]

Notes: The row headings are the numbers of objects. The column headings are the total number of samples for training, validation, and testing.
[1]The accuracies are between 82.2% and 85.7%.
[2]The accuracies are between 96.9% and 98.2%.
[3]The data for simulations that were not conducted.

3.7 Possible Reasons for the Underperformance of $Network_{one\ pathway}$. In order to help elucidate a possible reason for the underperformance of $Network_{one\ pathway}$, we used a decoder to decode information from the second-to-last layer activities of the trained $Network_{one\ pathway}$. The decoder was a multilayer perceptron with three hidden dense layers and 100 units in each hidden layer. ReLU activation function was used at each layer in the decoder except the final output layer, in which a softmax activation function was used. We used the second-to-last layer activities of the trained $Network_{one\ pathway}$ as inputs to the decoder and trained the decoder to do either the identity task or the location task with three objects per image and 6000 samples in total. The decoding accuracy for the identity task was $(66.7 \pm 3.0)\%$, and the decoding accuracy for the location task was $(91.2 \pm 1.0)\%$. The decoding accuracy for the identity task was much lower than that for the location task. In addition, the decoding accuracy for the identity task $(66.7 \pm 3.0)\%$ was very close to the accuracy of $Network_{one\ pathway}$ on the object recognition and localization tasks $(63.7 \pm 10.5)\%$ in the same condition.

3.8 The Contribution of Each Pathway in the Performance of $Network_{two\ pathways}$. In order to examine the contribution of each pathway in the performance of $Network_{two\ pathways}$, we tested the performance of $Network_{two\ pathways}$ after removing the identity pathway or location pathway, using three objects per image and 6000 samples in total. After removing the location pathway and keeping only the identity pathway, the testing accuracy of $Network_{two\ pathways}$ was $(37.0 \pm 1.9)\%$. After removing the identity pathway and keeping only the location pathway, the testing accuracy of $Network_{two\ pathways}$ was $(4.6 \pm 0.5)\%$. The accuracies of $Network_{two\ pathways}$ in both cases were significantly lower than the accuracy of $Network_{two\ pathways}$ when both pathways were present $(85.4 \pm 0.1)\%$. The accuracy of $Network_{two\ pathways}$ decreased more when the identity pathway was removed.

3.9 The Contribution of Spatial Relation Information in $Network_{identity}$ in Constraining the Binding Problem. In order to constrain the binding problem, we asked $Network_{identity}$ to learn spatial relation information by training it to report objects' identities in a certain order. In order to test whether the spatial relation information in $Network_{identity}$ contributed to performance, we trained another identity network, $Network'_{identity}$, and asked this network to report the identities of all the objects regardless of the spatial relations between these objects. Then we used the trained $Network'_{identity}$ to be the identity pathway in $Network'_{two\ pathways}$. It turned out that the testing accuracy of this $Network'_{two\ pathways}$ was (76.7 ± 1.0)% when there were three objects per image and 6000 samples in total, which was significantly lower than the testing accuracy of the original $Network_{two\ pathways}$ (85.4 ± 0.1)% in the same condition. In addition, $Network'_{two\ pathways}$ required more training epochs (around 1000 epochs) than the original $Network_{two\ pathways}$ (around 100 epochs).

For comparison, we also trained $Network'_{two\ pathways}$ and the original $Network_{two\ pathways}$ to do an object recognition and localization task without binding. We only asked the networks to report all objects' identities and locations, but we did not require them to bind each object's identity with its location. The chance level accuracy for this task without binding and the original task with binding is approximately 0%. We trained them with three objects per image and 6000 samples in total. The testing accuracy of $Network'_{two\ pathways}$ for this task was (98.2 ± 0.0)%, very close to the testing accuracy of $Network_{two\ pathways}$ on the same task (98.3 ± 0.0)%. In addition, both $Network'_{two\ pathways}$ and $Network_{two\ pathways}$ required around 100 epochs to train.

3.10 The Accuracy of $Network_{two\ pathways}$ When the Number of Objects Increased. According to the results shown in Tables 1, 2, and 3, and Figure 6, the accuracy of $Network_{two\ pathways}$ decreased when the number of objects increased with the same number of training samples. However, the required number of samples is still unclear for $Network_{two\ pathways}$ to reach a similar level of high accuracy when there are different numbers of objects.

It is hard to accurately estimate the required number of samples for $Network_{two\ pathways}$ to reach the same accuracy in different conditions. Therefore, in order to address this question, we conducted additional simulations to estimate the required number of samples for $Network_{two\ pathways}$ to reach similar high accuracies when there are two, three, or four objects. The results, shown in Table 4, suggest that $Network_{two\ pathways}$ required 1200, 6000, or 30,000 samples to reach a similar relatively high accuracy (between 82.2% and 85.7%) for two, three, or four objects, respectively. Additionally, these results also suggest that $Network_{two\ pathways}$ required 2400, 12,000, or 60,000 samples to reach a similar but even higher accuracy level (between 96.9% and 98.2%) for two, three, or four objects, respectively.

Table 5: Average Testing Accuracies in Percentage (%) ± Standard Deviations (%) for the Modified $Network_{two\ pathways}$ When There Are Two, Three, or Four Objects in Each Image.

	Remove One Convolutional Layer			Add One Convolutional Layer		
	1200	6000	30,000	1200	6000	30,000
Two objects	80.0 ± 0.7	NA	NA	85.5 ± 0.6	NA	NA
Three objects	NA	84.6 ± 0.3	NA	NA	85.6 ± 0.1	NA
Four objects	NA	NA	85.6 ± 0.1	NA	NA	85.8 ± 0.0

Notes: The row headings are the numbers of objects. The column headings are the total number of samples for training, validation, and testing. The data in the first three columns were obtained using modified $Network_{two\ pathways}$ with one convolutional layer removed in each pathway. The data in the second three columns were obtained using modified $Network_{two\ pathways}$ with one additional convolutional layer added in each pathway. The data for simulations that were not conducted are labeled "NA."

In order to test whether our results are robust to hyperparameter changes, we repeated some simulations with different numbers of convolutional layers. We either increased or decreased the number of these layers in each pathway of $Network_{two\ pathways}$. We increased the number of these layers by adding one additional convolutional layer in each pathway of the original $Network_{two\ pathways}$. The additional layer is the fourth convolutional layer, and it has the same number of filters and the same kernel sizes as the third convolutional layer in each pathway of the original $Network_{two\ pathways}$. We reduced the number of convolutional layers by removing the second convolutional layer in each pathway of the original $Network_{two\ pathways}$. We repeated simulations using these modified $Network_{two\ pathways}$ under different conditions, and the results are shown in Table 5.

3.11 Decreased $Network_{two\ pathways}$ Accuracy When the Number of Objects Increased: Role of Binding? It is possible that the decreased performance of $Network_{two\ pathways}$ with increasing numbers of objects was partly caused by a binding limitation. However, because the testing accuracies of $Network_{identity}$ also decreased when the number of objects increased, it is also possible that the decreased performance of $Network_{two\ pathways}$ was caused merely by the decreased performance of the identity pathway, and not by a binding limitation.

To test this hypothesis, we trained the same original $Network_{two\ pathways}$ to do the object recognition and localization task either with or without binding. For the case without binding, we only asked the final combined network to report all objects' identities and all objects' locations, but we did not require the network to bind each object's identity with its location. We trained each $Network_{two\ pathways}$ with three objects per image and 6000 samples in total. The chance-level accuracy for the task without binding and the original task with binding is approximately 0%. The

testing accuracy of $Network_{two\ pathways}$ for the task without binding was (98.3 ± 0.0)%, which was significantly higher than the testing accuracy of $Network_{two\ pathways}$ with the original task that required binding (85.4 ± 0.1)%. We also trained $Network_{two\ pathways}$ with four objects per image and 6000 samples in total. The chance-level accuracy for the task without binding and the original task with binding is still approximately 0%. The testing accuracy of $Network_{two\ pathways}$ for this task without binding was (89.3 ± 0.3)%, which was also significantly higher than the testing accuracy of $Network_{two\ pathways}$ with the original task that required binding (31.4 ± 0.1)%.

4 Discussion

One of the limitations of our previous study modeling the two cortical visual pathways was that there was only one object in each input image. Here, we sought to test whether our findings could be generalized to multiple-object recognition and localization tasks. In our current study, we found that the artificial neural networks with two pathways for identity and space have better performance in multiple-objects recognition and localization tasks (higher average testing accuracy, lower testing accuracy variance, less training time) than the artificial neural networks with a single pathway. Additionally, we found that the required number of training samples and the required training time increased quickly, and potentially exponentially, when the number of objects in each image increased. We also showed that the spatial relation information required in the training of our $Network_{identity}$ to constrain the binding problem was critical and increased the performance of $Network_{two\ pathways}$. Finally, we showed that testing accuracies of $Network_{two\ pathways}$ increased after training to do an object recognition and localization task without binding, suggesting that binding limited performance and may be a reason that our brain has limited attentional and visual working memory capacities.

4.1 The Performance of $Network_{two\ pathways}$ Was Significantly Better Than $Network_{one\ pathway}$. According to our simulation results, $Network_{two\ pathways}$ was able to outperform $Network_{one\ pathway}$ in almost all conditions. These are fair comparisons because the two networks have equal sizes, and we trained every network with enough epochs until it had reached the highest possible validation accuracy. $Network_{two\ pathways}$ was able to achieve higher average testing accuracy and lower testing accuracy variance in most conditions. Further, $Network_{two\ pathways}$ was able to reach the highest validation accuracy with less total training time. Therefore, our simulation results suggest that two separate pathways are advantageous in order to process the same visual information in different ways so that the network could have better performance (higher average testing accuracy, lower testing accuracy variance, less training time) in multiple-objects recognition and localization tasks.

We compared the performance of $Network_{one\ pathway}$ with or without pretraining. In the case of pretraining, we pretrained $Network_{one\ pathway}$ with the identity task and the location task. We found that the accuracy of $Network_{one\ pathway}$ with pretraining was significantly lower than the accuracy of $Network_{one\ pathway}$ without pretraining. It may be because pretraining caused $Network_{one\ pathway}$ to be more likely to get stuck in local minima. These findings show that pretraining $Network_{one\ pathway}$ with the identity task and the location task cannot improve the performance of this network. Therefore, the better performance of $Network_{two\ pathways}$ cannot be explained by pretraining. A one-pathway neural network is not optimal or efficient for learning these two different specializations (multiple-objects recognition and localization).

According to Dobs et al. (2022), a one-pathway neural network may be sufficient for learning object recognition and face recognition tasks. However, in our study, the difference between the object recognition task and the object localization task is larger than the difference between the object recognition task and the face recognition task. As a result, it is likely more difficult for a one-pathway neural network to find a common feature space to solve both the object recognition task and the object localization task. Further, we show here that the performance of a one-pathway neural network is impaired in multiple-objects recognition and localization tasks compared to a two-pathway neural network (lower average testing accuracy, higher testing accuracy variance, more training time).

4.2 Possible Reasons for the Underperformance of $Network_{one\ pathway}$. We used a decoder to decode identity and location information from the second to last layer of the trained $Network_{one\ pathway}$ with three objects per image and 6000 samples in total. According to our simulations, the decoding accuracy for the identity task was much lower than the decoding accuracy for the location task. It is important to note that with three objects per image and 6000 samples in total, $Network_{identity}$ and $Network_{location}$ have very similar accuracies on the identity and location tasks, respectively (see Table 2). Further, the identity and location tasks also have similar chance-level accuracies that are close to 0. Therefore, the findings of much lower decoding accuracy for the identity task suggest that $Network_{one\ pathway}$ learned less identity information than location information. In addition, the decoding accuracy for the identity task was very close to the accuracy of $Network_{one\ pathway}$ on the object recognition and localization tasks. Therefore, these results suggest that one reason for the underperformance of $Network_{one\ pathway}$ was that it was not able to learn enough identity information.

4.3 The Contribution of Each Pathway in the Performance of $Network_{two\ pathways}$. According to our simulations, the accuracy of $Network_{two\ pathways}$ decreased significantly when the identity pathway

or the location pathway was removed. In addition, the accuracy of $Network_{two\ pathways}$ decreased significantly more when the identity pathway was removed. Therefore, both pathways contributed to the performance of $Network_{two\ pathways}$, and the identity pathway contributed more. It is possible the identity pathway contributed more because it included both identity information and spatial relation information, whereas the location pathway contained only spatial information. It is also likely the relative contributions of each pathway may change with different task goals or task design.

4.4 The Contribution of Spatial Relation Information in $Network_{identity}$ in Constraining the Binding Problem. Previously, we showed that networks trained either for identity or location retained spatial information (Han & Sereno, 2022a). In some visual perception tasks, the goal of the task may require coordination of the information from these separated pathways (e.g., reaching for the object that is edible when multiple objects are present). In these cases, processing information independently and differently using multiple separate pathways may cause a binding problem (Treisman, 1996). We suggested that the binding problem may be lessened if we could take advantage of the spatial information contained in the identity network and object identity information in the spatial network. Therefore, in our current study, we assume that the ventral pathway has access to the relative spatial information of objects and try to constrain the binding problem in the following way. We trained $Network_{identity}$ by asking it to report the identities of all the objects in each image in a certain order that depends on the spatial relations between these objects. Note that we can choose any consistent order as long as the one-hot vector representation in the final output layer of $Network_{identity}$ is determined by both the identities of the objects and the spatial relationships between the objects. Because the information retained by the neural networks depends on the training task (Han & Sereno, 2022a), this task would make $Network_{identity}$ not only actively retain identities of the objects but also actively retain relative spatial relationships between the objects. We trained $Network_{location}$ by asking it to report the locations of all the objects in the image regardless of their identities. Then we used these trained networks as the two separate pathways in $Network_{two\ pathways}$. Therefore, $Network_{two\ pathways}$ should be able to bind the identity of each object with its location by combining the identity information in $Network_{identity}$, the relative spatial relation information in $Network_{identity}$, and the absolute location information without identities in $Network_{location}$. Our simulation results indicate that it is possible to constrain the binding problem in this way.

In order to evaluate the effectiveness of this method, we trained another identity network, $Network'_{identity}$, and asked this network to report the identities of all the objects regardless of the spatial relations between these objects. Then we used the trained $Network'_{identity}$ to be the identity pathway in

$Network'_{two\ pathways}$. According to our simulation results, the performance of $Network'_{two\ pathways}$ was significantly lower than the testing accuracy of the original $Network_{two\ pathways}$ in the same condition. These results suggest that the spatial relation information in $Network_{identity}$ improved the performance of the two-pathway neural network. However, these findings do not establish whether this improvement was because the spatial relation information in $Network_{identity}$ was important for constraining the binding problem. For comparison, we also trained $Network'_{two\ pathways}$ (identity network trained with no spatial relation information) and the original $Network_{two\ pathways}$ to do an object recognition and localization task without binding. When binding is not required in the task, the differences in performance between the two networks disappeared. The performance of $Network'_{two\ pathways}$ for this task was almost the same as the performance of $Network_{two\ pathways}$. Thus, these findings suggest that the spatial relation information retained in the identity network is not important if the task does not require binding, but when the task requires binding, it is critical in constraining the binding problem.

An important assumption that we made was that the identity network should be able to report object identities according to the relative spatial relations between objects. This assumption is biologically plausible because the ventral cortical visual pathway may have different neural representations when the relative spatial relations between the same set of objects are different (Yamane et al., 2006). Sereno & Lehky (2011) report additional experimental evidence where they showed not only that the majority of cells in late stages of the ventral pathway were spatially selective but also that it was possible to decode object location from a small population of cells. Further, they demonstrated that the recovered spatial representation was topologically correct. Topologically correct spatial information indicates that the information about relative spatial relations between objects is retained in the ventral pathway.

4.5 The Accuracy of $Network_{two\ pathways}$ When the Number of Objects Increased. According to Table 4, $Network_{two\ pathways}$ required 1200, 6000, or 30,000 samples to reach a similar relatively high accuracy (between 82.2% and 85.7%) for two, three, or four objects, respectively. Also, Table 4 shows that $Network_{two\ pathways}$ required 2400, 12,000, or 60,000 samples to reach a similar but even higher accuracy (between 96.9% and 98.2%) for two, three, or four objects, respectively. Though these required number of samples for $Network_{two\ pathways}$ to reach a similar high level of performance are just estimates, these results suggest that the required number of samples increases by around five times when the number of objects increases from two to three. The required number of samples increases by another five times when the number of objects increases from three to four. Training time for the same $Network_{two\ pathways}$ is roughly proportional to the number of training

samples, so training time rises in a similar way. It indicates that the required number of samples and the required training time for $Network_{two\ pathways}$ to reach a certain high accuracy increases quickly, and potentially exponentially, when the number of objects increases.

In addition, the results shown in Tables 4 and 5 suggest that the accuracies of the original $Network_{two\ pathways}$ and the modified $Network_{two\ pathways}$ with different numbers of convolutional layers in each pathway were very similar when they were trained under the same conditions. The modified $Network_{two\ pathways}$ with different numbers of convolutional layers in each pathway still required 1200, 6000, or 30,000 samples to reach a similar relatively high accuracy (between 80.0% and 85.8%) for two, three, or four objects, respectively. Together, these findings suggest that our results are robust to hyperparameter changes.

4.6 The Limitation of $Network_{two\ pathways}$ When the Number of Objects Increases: Role of Binding? According to our simulations, our $Network_{two\ pathways}$ model could not recognize and localize four or more objects efficiently. With four or more objects in each image, training with $Network_{two\ pathways}$ was very slow and required many more training samples if we wanted to achieve a high testing accuracy (>90%).

It is possible that the decreased performance of $Network_{two\ pathways}$ with larger numbers of objects was caused by a decreased performance of the identity pathway, not by a binding limitation. We conducted additional simulations to test this hypothesis. According to our simulations, the testing accuracy of $Network_{two\ pathways}$ for the object recognition and localization tasks that did not require binding was significantly higher than the testing accuracy of $Network_{two\ pathways}$ for the original task that required binding for each of the three- and four-object conditions. Therefore, the decreased performance of $Network_{two\ pathways}$ with larger numbers of objects cannot be fully explained with a decreased performance of the identity pathway. The binding problem increased the difficulty of the task and caused the model performance to decrease.

Though this increased difficulty with tasks that require binding may be a limitation of our model, it may agree with the computational properties of the biological brain. Our biological brain is also not good at recognizing and localizing four or more objects in the scene at the same time. According to Isbell et al. (2015), the capacity limit of visual working memory is about three simple objects in healthy young adults. Visual working memory capacity may be considered as the maximum number of objects that our brain could recognize and localize at the same time. This agrees with our findings. Our $Network_{two\ pathways}$ was also only able to achieve high testing accuracy (>90%) within a reasonable training time and training samples if the number of objects in each image was less than or equal to three. Therefore, we argue that our visual working memory capacity may be limited in part by the binding problem. Other kinds of attention and working memory

capacity may also be limited by the binding problem. Many memory span tasks that are used to measure working memory capacity require people to remember the occurrence of information in space and time (Tulving, 1972; Nairne, 2015). Combining the information with its occurrence in space and time is also often referred to as a binding problem, and some researchers have proposed that working memory is a system for building, maintaining, and updating different kinds of bindings (Oberauer, 2009).

If the brain needs to recognize and localize four or more objects, then we speculate that it could only do it sequentially. For example, Quirk, Adam, and Vogel (2020) found that the human visual working memory capacity increased for both simple and real-world objects when encoding times were longer. Their experimental tasks required participants to recognize and localize multiple objects during the encoding time and make a response about which object appeared at a certain location during the testing time. In our opinion, one possible explanation of their findings is that the brain is able to bind more objects' identities and locations sequentially when the encoding time is longer. As a result, the brain can recognize and localize more objects, so the visual working memory capacity may appear to increase.

Experimental evidence also suggests that visual working memory continues to develop throughout adolescence, and it does not reach adult levels even in 16-year-old participants (Isbell et al., 2015). It may be because multiple-objects recognition and localization requires a lot of training samples and training time. It also agrees with our findings because our $Network_{two\ pathways}$ model also required a relatively large number of training samples and training time to achieve a high testing accuracy (>90%) with three objects in each image. There are many possible combinations of the same information in different ways. Furthermore, in the real world, learning conditions are complex and changing. For example, context itself may alter the meaning of the same information (an empty pot on a stove, with the stove on versus off). Finally, our contexts and environments are themselves changing over time. Thus, there should be some improvements in working memory capacities with training time on context-appropriate sets. However, visual working memory capacity cannot be improved indefinitely through training, likely because our life and experience (training samples) are limited, as well as the fact that the greatest developmental benefits of the human brain occur before adulthood.

If our visual working memory capacity is limited in part by the binding problem, then we speculate that the measured visual working memory capacity may increase if we ask the participants to report all objects' identities regardless of their locations, and/or ask them to report all objects' locations regardless of their identities. In addition, human visual working memory capacity should be continuously developing in a long period (from infancy to adulthood) and be dependent in part on stimulus and context-appropriate training. Human visual working memory capacity should increase relatively quickly at the beginning when the capacity is low and

increase relatively slower later when the capacity is high and the individual is nearing nervous system limits. There may not be a hard limit for the maximum possible visual working memory capacity. However, there may be a soft limit because the difficulty of increasing the visual working memory capacity further should increase quickly, and potentially exponentially, when the capacity increases.

Though both our two-pathway model and human brains are not good at recognizing and localizing four or more objects at the same time, more research is needed to find out how different computational resources, hyperparameter settings, and learning types could affect this limit. Our current study suggests that the human visual working memory capacity may be limited in part by the binding problem, but our study does not suggest this limit must be three or any other specific critical number. According to some previous studies, human visual working memory capacity varies across individuals and groups (Luck & Vogel, 2013). The individual differences in working memory capacity may be caused by different hyperparameter settings in human brains (e.g., number of neurons, number of connections between neurons). Some previous studies argue that it would be biologically expensive for the brain to have a larger working memory capacity (Cowan, 2010). In addition, we have shown that the required training time and the number of samples for high neural network performance increased quickly as the number of objects increased. These findings suggest that the individual differences in terms of visual working memory capacity may exist but may not be very large. Therefore, there seem to be fairly standard numbers (three or four) for the limit of human visual working memory capacity (Luck & Vogel, 2013).

Using a computational modeling approach, we aimed to better understand whether the presence of the two separate cortical visual pathways in the brain is important for object recognition and localization when there are multiple objects in the scene. Our simulations using convolutional neural networks used simple tasks, and we ignored a lot of details of real biological neural networks. These simplifications are necessary to make direct computational comparisons possible. Our claims concern whether there could be a computational advantage for retaining information independently and differently in multiple pathways and whether this computational advantage could increase the network performance in multiple-objects recognition and localization tasks. A previous simulation study of one-pathway and two-pathway artificial neural networks compared model simulations with actual neural representations in the brain (Bakhtiari et al., 2021). They reported that their two-pathway artificial neural network models could produce better matches to the representations in the mouse ventral and dorsal visual streams than their one-pathway artificial neural network models. Though our intent with simulations was to explore the computational consequences of multiple streams architecture rather than emulate physiological conditions of the brain, interestingly our findings generally agree with

this study, which considered actual neural activities in the brain. The brain is a complex organ, sometimes described as the last frontier, and it is clear that computational approaches can play a key role, including in elucidating consequences of its organization and function.

5 Conclusion

In summary, our simulations show that our models are able to accurately and simultaneously recognize and localize multiple objects in a scene. Furthermore, we show that the artificial neural networks with two pathways for identity and space have significantly better performance (higher average testing accuracy, lower testing accuracy variance, less training time) than the artificial neural networks with a single pathway in multiple-objects recognition and localization tasks. We also find that the required number of training samples and the required training time increased quickly, and potentially exponentially, when the number of objects in each image increased. The simulations suggest that the difficulty of binding identity and spatial information increases quickly, and potentially exponentially, when the number of objects increases. We suggest that binding information from multiple segregated pathways may be a reason that our brain has a limited visual working memory capacity. Given that attention and working memory require binding information with space or time, it is possible that many attentional and working memory capacities are also limited by similar binding problems.

Acknowledgments

This work was partially supported by funds from Purdue University to A.S.

References

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In M. Ranzato, A. Beygelzimer, K. Nguyen, P. S. Liang, J. W. Vaughan, & Y. Dauphin (Eds.), *Advances in neural information processing systems*, *34* (pp. 25164–25178) Curran.

Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, *22*, 319–349. 10.1146/annurev.neuro.22.1.319

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. 10.1177/0963721409359277

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. 10.1016/S0022-5371(80)90312-6

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, *8*(11), 1–11. 10.1126/sciadv.abl8913

Felleman, D., & Essen, D. V. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47. 10.1093/cercor/1.1.1

Han, Z., & Sereno, A. (2022a). Modeling the ventral and dorsal cortical visual pathways using artificial neural networks. *Neural Computation*, *34*(1), 138–171. 10.1162/neco_a_01456

Han, Z., & Sereno, A. (2022b). Identifying and localizing multiple objects using artificial ventral and dorsal visual cortical pathways. *Computational and Mathematical Models in Vision*. Retrieved from https://docs.lib.purdue.edu/modvis/2022/session01/3/

Isbell, E., Fukuda, K., Neville, H. J., & Vogel, E. K. (2015). Visual working memory continues to develop through adolescence. *Frontiers in Psychology*, *6*, 1–10. 10.3389/fpsyg.2015.00696

Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*, *11*(2), 224–231. 10.1038/nn2036

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communication*, *13*(493), 1–12. 10.1038/s41467-022-28091-4

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. 10.1146/annurev-vision-082114-035447

Lehky, S. R., & Sereno, A. B. (2007). Comparison of shape encoding in primate dorsal and ventral visual pathways. *Journal of Neurology, Neurosurgery, and Psychiatry*, *97*, 307–319. 10.1152/jn.00168.2006

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*, 335–346. 10.1038/s41583-020-0277-3

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621. 10.1146/annurev.ne.19.030196.003045

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400. 10.1016/j.tics.2013.06.006

Markov, Y. A., Utochkin, I. S., & Brady, T. F. (2021). Real-world objects are not stored in holistic representations in visual working memory. *Journal of Vision*, *21*(3), 1–24. 10.1167/jov.21.3.18

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. 10.1037/h0043158

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. 10.1016/0166-2236(83)90190-X

Nairne, J. S. (2015). The three "ws" of episodic memory: What, when, and where. *American Journal of Psychology*, *128*(2), 267–279. 10.5406/amerjpsyc.128.2.0267

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation*, *51*, 45–100. 10.1016/S0079-7421(09)51002-X

Op De Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferiortemporal neurons. *Journal of Comparative Neurology*, *426*, 505–518. 10.1002/1096-9861(20001030)426:4⟨505::aid-cne1⟩3.0.co;2-m

Quirk, C., Adam, K. C., & Vogel, E. K. (2020). No evidence for an object working memory capacity benefit with extended viewing time. *eNeuro*, *7*(5), 1–13. 10.1523/ENEURO.0150-20.2020

Sereno, A. B., & Lehky, S. R. (2011). Population coding of visual space: Comparison of spatial representations in dorsal and ventral pathways. *Frontiers in Computational Neuroscience*, *4*(159), 1–16. 10.3389/fncom.2010.00159

Sereno, A. B., Lehky, S. R., & Sereno, M. E. (2020). Representation of shape, space, and attention in monkey cortex. *Cortex*, *122*, 40–60. 10.1016/j.cortex.2019.06.005

Sereno, A. B., Sereno, M. E., & Lehky, S. R. (2014). Recovering stimulus locations using populations of eye-position modulated neurons in dorsal and ventral visual streams of non-human primates. *Frontiers in Integrative Neuroscience*, *8*, 1–20. 10.3389/fnint.2014.00028

Sereno, M. I., Sood, M. R., & Huang, R.-S. (2022). Topological maps and brain computations from low to high. *Frontiers in Systems Neuroscience*, *16*, 1–22. 10.3389/fnsys.2022.787737

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, *6*(2), 171–178. 10.1016/S0959-4388(96)80070-5

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). Academic Press.

Ungerleider L. G., & Mishkin, M. (1982). Two cortical visual systems. In M. Goodale, D. J. Ingle, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). MIT Press.

Whittington, J. C., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, *23*(3), 235–250. 10.1016/j.tics.2018.12.005

Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms*. arXiv:1708.07747

Yamane, Y., Tsunoda, K., Matsumoto, M., Phillips, A. N., & Tanifuji, M. (2006). Representation of the spatial relationship among object parts by neurons in macaque inferotemporal cortex. *Journal of Neurophysiology*, *96*, 3147–3156. 10.1152/jn.01224.2005

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, *111*(23), 8619–8624. 10.1073/pnas.1403112111

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2022). Unsupervised neural network models of the ventral visual stream. *PNAS*, *118*(3), 1–11. 10.1073/pnas.2014196118

Received June 20, 2022; accepted October 9, 2022.