Recurrent computations for visual pattern completion

Supporting Information Appendix

Hanlin Tang[1,4*], Martin Schrimpf[2,4*], William Lotter[1,3,4*], Charlotte Moerman[4], Ana Paredes[4], Josue Ortega Caro[4], Walter Hardesty[4], David Cox[3], Gabriel Kreiman[4]

1. Supplementary Materials and Methods

Psychophysics experiments

A total of 106 volunteers (62 female, ages 18-34 y) with normal or corrected to normal vision participated in the psychophysics experiments reported in this study. All subjects gave informed consent and the studies were approved by the Institutional Review Board at Children's Hospital, Harvard Medical School. In 67 subjects, eye positions were recorded during the experiments using an infrared camera eye tracker at 500 Hz (Eyelink D1000, SR Research, Ontario, Canada). We performed a main experiment (reported in **Figure 1F-G**) and three variations (reported in **Figures 1I-J, 2**, **S1** and **S8-9**).

Backward masking. Multiple lines of evidence from behavioral (e.g. (1, 2)), physiological (e.g. (3-6)), and computational studies (e.g. (7-11)) suggest that recognition of whole isolated objects can be approximately described by rapid, largely feed-forward, mechanisms. Despite the success of these feed-forward architectures in describing the initial steps in visual recognition, each layer has limited spatial integration of its inputs. Additionally, feed-forward algorithms lack

www.pnas.org/cgi/doi/10.1073/pnas.1719397115

mechanisms to integrate temporal information or to take advantage of the rich temporal dynamics characteristic of neural circuits that allow comparing signals within and across different levels of the visual hierarchy. It has been suggested that backward masking can interrupt recurrent and top-down signals: when an image is rapidly followed by a spatially overlapping mask: the new high-contrast mask stimulus interrupts any additional, presumably recurrent, processing of the original image (3, 12-20). Thus, the psychophysical experiments tested recognition under both unmasked and backward masked conditions.

Main experiment. Both spatial and temporal integration are likely to play an important role in pattern completion mechanisms (21-27). A scheme of the experiment designed to study the spatial and temporal integration during recognition of occluded or partially visible objects is shown in **Figure 1**. Twenty-one subjects were asked to categorize images into one of 5 possible semantic groups (5-alternative forced choice) by pressing buttons on a gamepad. Stimuli consisted of contrast-normalized gray scale images of 325 objects belonging to five categories (animals, chairs, human faces, fruits, and vehicles). Each object was only presented once in each condition. Each trial was initiated by fixating on a cross for at least 500 ms. After fixation, subjects were presented with the image of an object for a variable time (25 ms, 50 ms, 75 ms, 100 ms, or 150 ms), referred to as the stimulus onset asynchrony (SOA). The image was followed by either a noise mask (**Figure 1B**) or a gray screen (**Figure 1A**), with a duration of 500 ms, after which a choice screen appeared requiring the subject to respond. We use the term "pattern completion" to indicate successful categorization of partial images in the 5-alternative forced choice task used here and we do not mean to imply that subjects are forming any mental image of the entire object, which we did not test. The noise mask was generated by scrambling the phase of the images, while retaining the spectral coefficients. The images (256 x 256 pixels) subtended approximately 5 degrees of the visual field. In approximately 15% of the trials, the objects were presented in unaltered fashion (the 'Whole' condition, **Figure 1C** left). In the other 85% of the trials, the objects were rendered partially visible by presenting visual features through Gaussian bubbles (28) (the 'Partial condition', standard deviation = 14 pixels, **Figure 1C** right). Each subject performed an initial training session to familiarize themselves with the task and the stimuli. They were presented with 40 trials of whole objects, then 80 calibration trials of occluded objects. During the calibration trials, the number of bubbles was titrated using a staircase procedure to achieve an overall task difficulty of 80% correct rate. The number of bubbles (but not their positions) was then kept constant for the rest of the experiment. Results from the familiarization and calibration phase were not included in the analyses. Despite calibrating the number of bubbles, there was a wide range of degrees of occlusion because the positions of the bubbles were randomized in every trial. Each image was only presented once in the masked condition and once in the unmasked condition.

Physiology-based psychophysics experiment. In the physiology-based psychophysics experiment (**Figure 2**, n = 33 subjects), stimuli consisted of 650 images from five categories for which we had previously recorded neural responses (see below). In the neurophysiological recordings (25), bubble positions were randomly selected in each subject and therefore each subject was presented with different images (except for the fully visible ones). The main difference between the physiology-based psychophysics experiment and the Main experiment is that here we used the exact same images that were used in the physiological recordings (see description under "Neurophysiological Recordings" below).

Occlusion experiment. In the occlusion experiment (**Figure 1I**, **Figure S1**, n=14 subjects in the partial objects experiment and n =15 subjects in the occlusion experiment), we generated occluded images that revealed the same sets of features as the partial objects, but contained an explicit occluder (**Figure 1D**) to activate amodal completion cues. The stimulus set consisted of 16 objects from 4 different categories. For comparison, we also collected performance with partial objects from this reduced stimulus set.

Novel objects experiment. The main set of experiments required categorization of images containing pictures of animals, chairs, faces, fruits and vehicles. None of the subjects involved in the psychophysics or neurophysiological measurements had had any previous exposure to the *specific pictures* in these experiments, let alone with the partial images rendered through bubbles. Yet, it can be surmised that all the subjects had had extensive previous experience with *other* images of objects from those categories, including occluded versions of other animals, chairs, faces, fruits and vehicles. In order to evaluate whether experience with occluded instances of objects from a specific category is important to recognize novel instances of partially visible objects from the same category, we conducted a new psychophysics experiment with novel objects. We used 500 unique novel objects belonging to 5 categories, all the novel objects were chosen from the Tarr Lab stimulus repository (29). An equal amount of stimuli were chosen from each category. One exemplar from each category is shown in **Figure S8A**. In the Cognitive Science community, the first three categories are known as “Fribbles” and the last two categories as “Greebles” and “Yufos” (29). In our experiments, each category was assigned a Greek letter name (**Figure S8A**) so as not to influence the subjects with potential meanings of an invented name.

The experiment followed the same protocol as the main experiment (**Figure 1**). Twenty-three new subjects (11 female, 20 to 34 years old) participated in this experiment. Since the subjects had no previous exposure to these stimuli, they underwent a short training session where they were presented with 2 fully visible exemplars from each category so that they could learn the mapping between categories and response buttons. In order to start the experiment, subjects were required to get 8 out of 10 correct responses, 5 times in a row using these practice stimuli. On average, reaching this level of accuracy required 80±40 trials. Those 2 stimuli from each category were not used in the subsequent experiments. Therefore, whenever we refer to “novel” objects, what we mean is objects from 5 categories where subjects were exposed to ~80 trials of 2 fully visible exemplars per category, different from the ones used in the psychophysics tests. This regime represented our compromise of ensuring that subjects knew which button they had to press, while at the same time keeping only minimal initial training. Importantly, this initial training only involved whole objects and subjects had no exposure to partial novel objects before the onset of the psychophysics measurements. Halfway through the experiment, we repeated 3 runs of the recognition test with the same 2 initial fully visible exemplars as a control to ensure that subjects were still performing the task correctly, and all subjects passed this control (>80% performance in just 3 consecutive runs).

During the experiment, subjects were presented with 1,000 uniquely rendered stimuli from 500 contrast-normalized gray scale novel objects, resized to 256x256 pixels, subtending approximately 5° of visual angle. All images were contrast normalized using the `histMatch` function from the SHINE toolbox (30). This function equates the luminance histogram of sets of images. For each subject, 1,000 unique renderings were obtained by applying different bubbles to the original images, resulting in a total of 23,000 different stimuli across subjects.

The SOAs and other parameters were identical to those used in the main experiment. The analyses and models for the novel object experiments follow those in the main experiment (**Figures S8B-D** are the analogs of **Figure 1F-H**, **Figure S9A** is the analog of **Figure 3A**, **Figure S9B-D** are the analogs of **Figure 4B-D**).

Neurophysiology experiments

The neurophysiological data analyzed in **Figures 2** and **3** were taken from the study by Tang *et al* (25), to which we refer for further details. Briefly, subjects were patients with pharmacologically intractable epilepsy who had intracranial electrodes implanted for clinical purposes. These electrodes record intracranial field potential signals, which represent aggregate activity from large numbers of neurons. All studies were approved by the hospital's Institutional Review Board and were carried out with the subjects' informed consent. Images of partial or whole objects were presented for 150 ms, followed by a gray screen for 650 ms. Subjects performed a five-alternative forced choice categorization task as described in **Figure 1** with the following differences: (i) the physiological experiment did not include the backward mask condition; (ii) 25 different objects were used in the physiology experiment; (iii) the SOA was fixed at 150 ms in the physiology experiment.

Bubbles were randomly positioned in each trial. In order to compare models, behavior and physiology on an image-by-image basis, we had to set up a stimulus set based on the exact images (same bubble locations) presented to a given subject in the physiology experiment. To construct the stimulus set for the physiology-based psychophysics experiment (**Figure 2**), we chose two electrodes according to the following criteria: (i) those two electrodes had to come from different physiology subjects (to ensure that the results were not merely based on any peculiar properties of one individual physiology subject), (ii) the electrodes had to respond both to whole objects and partially visible objects (to ensure a robust response where we could estimate latencies in single trials), and (iii) the electrodes had to show visual selectivity (to compare the responses to the preferred and non-preferred stimuli). The electrode selection procedure was strictly dictated by these criteria and was performed before even beginning the psychophysics experiment. We extracted the images presented during the physiological recordings in n = 650 trials for psychophysical testing. For the preferred category for each electrode, only trials where the amplitude of the elicited neural response was in the top 50th percentile were included, and trials were chosen to represent a distribution of neural response latencies. After constructing this stimulus set, we performed psychophysical experiments with n = 33 new subjects (Physiology-based psychophysics experiment) to evaluate the effect of backward masking for the exact same images for which we had physiological data.

For the physiological data, we focused on the neural latency, defined as the time of the peak in the physiological response, as shown in **Figure 2B**. These latencies were computed in single trials (see examples in **Figure 2C**). Because these neural latencies per image are defined in single trials, there are no measures of variation in the x-axis in **Figure 2F** or **Figure 3C-D**. A more extensive analysis of the physiological data, including extensive discussion of many ways of measuring neural latencies, was presented in (25).

Behavioral and neural data analysis

Masking Index. To quantify the effect of backward masking, we defined the masking index as 100%-pAUC, where pAUC is the percent area under the curve when plotting performance as a function of SOA (e.g. **Figure 2E**). To evaluate the variability in the masking index, we used a half-split reliability measure by randomly partitioning the data into two halves and computing the masking index separately in each half. **Figure S2** provides an example of such a split. Error bars in **Figure 2F** constitute half-split reliability values.

Correlation between masking index and neural latency. To determine the correlation between masking index and neural response latency, we combined data from the two recording sites by first standardizing the latency measurements (z-score, **Figure 2F**). We then used a linear regression on neural response latency with masking index, percent visibility, and recording site as predictor factors to avoid any correlations dictated by task difficulty or differences between recording sites. We used only trials from the preferred category for each recording site and reported the correlation and statistical significance in **Figure 2F**. There was no significant correlation between the masking index and neural latency when considering trials from the non-preferred category.

Correlation between model distance and neural response latency. As described below, we simulated the activity of units in several computational models in response to the same images used in the psychophysics and physiology experiments. To correlate the model responses with neural response latency, we computed the Euclidean distance between the model representation of partial and whole objects. We computed the distance between each partial object in the physiology-based psychophysics experiment stimulus set and the centroid of the whole images from the same category (distance-to-category). We then assessed significance by using a linear regression on the model distance versus neural response latency while controlling for masking index, percent visibility, and recording site as factors.

Feed-forward Models

We considered the ability to recognize partially visible images by state-of-the-art feed-forward computational models of vision (**Figure 3A, Figure S3** and **Figure S4**). First, we evaluated whether it was possible to perform recognition purely based on pixel intensities. Next, in the main text we evaluated the performance of the AlexNet model (31). AlexNet is an eight-layer deep convolutional neural network consisting of convolutional, max-pooling and fully-connected layers with a large number of weights trained in a supervised fashion for object recognition on ImageNet, a large collection of labeled images from the web (31, 32). We used a version of AlexNet trained using *caffe* (33), a deep learning library. Two layers within the AlexNet were tested: pool5 and fc7. Pool5 is the last convolutional (retinotopic) layer in the architecture. fc7 is the last layer before the classification step and is fully connected, that is, every unit in fc7 is connected to every unit in the previous layer. The number of features used to represent each object was 256x256=65536 for pixels, 9216 for pool5 and 4096 for fc7.

We also considered many other similar feed-forward models: VGG16 block5, fc1 and fc2 (25088, 4096 and 4096 features respectively) (34), VGG19 fc1 and fc2 (4096 features each) (34), layers 40 to 49 of ResNet50 (200704 to 2048 features) (35), and InceptionV3 mixed 10 layer (131072 features) (36). In all of these cases, we used models pre-trained for the ImageNet 2012 data set and randomly downsampled the number of features to 4096 as in AlexNet. Results for all of these models are shown in **Figure S4;** more layers and models can be found in the accompanying web site: http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html

Classification performance for each model was evaluated on a stimulus set consisting of 13,000 images of partial objects (generated from 325 objects from 5 categories). These were the same partial objects used to collect human performance in the main experiment (**Figure 1**). We used a support vector machine (SVM) with a linear kernel to perform classification on the features computed by each model. We used 5-fold cross-validation across the 325 objects. Each split contained 260 objects for training, and 65 objects split for validation and testing, such that each object was used exactly in one validation and testing split, and such that there was an equal number of objects from each category in each split. Decision boundaries were fit on the training set using the SVM with the C parameter determined through the validation set by considering the following possible C values: 10^{-4}, 10^{-3}, …, 10^{3}, 10^{4}. The SVM boundaries were fit using images of whole objects and tested on images of partial objects. Final performance numbers for partial objects were calculated on the full data set of 13,000 images -- that is, for each split, classification performance was evaluated on the partial objects corresponding to the objects in the test set, such that, over all splits, each partial object was evaluated exactly once.

As indicated above, all the results shown on **Figure 3A**, **Figure S3** and **Figure S4** are based on models that were trained on the ImageNet 2012 data set and then tested using our stimulus set. We also tested a model created by fine-tuning the AlexNet network. We fine-tuned AlexNet using the set of whole objects in our data set and then re-examined the model's performance under the low visibility conditions in **Figure S5**. We fine-tuned AlexNet by replacing the original 1000-way fully-connected classifier layer (fc8) trained on ImageNet with a 5-way fully-connected layer (fc8') over the categories in our dataset and performing back-propagation over the entire network. We again performed cross validation over objects, choosing final weights by monitoring validation accuracy. To be consistent with previous analysis, after fine-tuning the representation, we used an SVM classifier on the resulting fc7 activations.

To graphically display the representation of the images based on all 4096 units in the fc7 layer of the model in a 2D plot (**Figure 4C**), we used stochastic neighborhood embedding (t-SNE) (37). We note that this was done exclusively for display purposes and all the analyses, including distances, classification and correlations, are based on the model representation with all the units in the corresponding layer as described above. For each model and each image, we computed the Euclidian distance between the model's representation and the mean point across all whole objects within the corresponding category. This distance-to-category corresponds to the y-axis in **Figure 3B-C**.

Recurrent Neural Network Models

A recurrent neural network (RNN) was constructed by adding all-to-all recurrent connections to different layers of the bottom-up convolutional networks described in the previous section (for example, to the fc7 layer of AlexNet in **Figure 4A**). We first describe here the model for AlexNet; a similar procedure was followed for the other computational models. An RNN consists of a state vector that is updated according to the input at the current time step and its value at the previous time step. Denoting $\mathbf{h}_t$ as the state vector at time t and $\mathbf{x}_t$ as the input into the network at time t, the general form of the RNN update equation is $\mathbf{h}_t = f(\mathbf{W}_h \mathbf{h}_{t-1}, \mathbf{x}_t)$ where f introduces a non-linearity as defined below. In our model, $\mathbf{h}_t$ represents the fc7 feature vector at time t and $\mathbf{x}_t$ represents the feature vector for the previous layer, fc6, multiplied by the transition weight matrix $\mathbf{W}_{6\rightarrow7}$. For simplicity, the first six layers of AlexNet were kept fixed to their original feed-forward versions.

We chose the weights $\mathbf{W}_h$ by constructing a Hopfield network (38), RNN_h, as implemented in MATLAB's `newhop` function, which is a modified version of the original description by Hopfield (39). Since this implementation is based on binary unit activity, we first converted the scalar activities in $\mathbf{x}$ to {-1,+1} by mapping those values greater than 0 to +1 and all other values to -1. Depending on the specific layer and model, this binarization step in some cases led to either an increase or a decrease in performance (even before applying the attractor network dynamics); all the results shown in the Figures report the results after applying the Hopfield dynamics. The weights in RNN_h are symmetric ($W_{ij} = W_{ji}$) and are dictated by the Hebbian learning rule $W_{ij} = \frac{1}{n_p}\sum_{p=1}^{n_p} x_i^p x_j^p$ where the sum goes over the n_p patterns of whole objects to be stored (in our case n_p=325) and x_i^p represents the activity of unit i in response to pattern p. This model does not have any free parameters that depend on the partial objects and the weights are uniquely specified by the activity of the feed-forward network in response to the whole objects. After specifying $\mathbf{W}_h$, the activity in RNN_h was updated according to $\mathbf{h}_0=\mathbf{x}$ and $\mathbf{h}_t = satlins(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{b})$ for t>0 where *satlins* represents the saturating linear transfer function,

$satlins(z) = \max(\min(1, z), -1)$ and **b** introduces a constant bias term. The activity in RNN_h was simulated until convergence, defined as the first time point where there was no change in the sign of any of the features between two consecutive time points.

To evaluate whether the increase in performance obtained in the RNN_h was specific to the AlexNet architecture, we also implemented recurrent connections added onto other networks. **Figure S7** shows a comparison between performance of the VGG16 network layer fc1 (34) and a VGG16 fc1 model endowed with additional recurrent connections in the same format as used with AlexNet. We used the time steps of the Hopfield network that yielded maximal performance. The VGG16+Hopfield model also showed performance improvement with respect to the purely bottom-up VGG16 counterpart. Several additional models were tested for other layers of AlexNet, VGG16, VGG19, ResNet and InceptionV3, showing a distribution with different degrees of consistent improvement upon addition of the recurrent connectivity (shown in the accompanying web material at http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html).

We ran an additional simulation with the RNN models to evaluate the effects of backward masking (**Figure 4F**). For this purpose, we simulated the response of the feed-forward AlexNet model to the same masks used for the psychophysical experiments to determine the fc6 features for each mask image. Next, we used this mask as the fixed input $\mathbf{x}_t$ into the recurrent network, at different time points after the initial image input.

2. Supplementary Discussion

Partially visible versus occluded objects

In most of the experiments, we rendered objects partially visible by presenting them through "bubbles" (**Fig. 1C**) in an attempt to distill the basic mechanisms required for spatial integration during pattern completion. It was easier to recognize objects behind a real occluder (**Fig. 1D, S1**, (40)). The results presented here were qualitatively similar (**Fig. S1**) when using explicit occluders (**Fig. 1D**): recognition of occluded objects was also disrupted by backward masking (**Fig. 1I, S1**). As expected, performance was higher for the occlusion versus the bubbles condition.

"Unfolding" recurrent neural networks into feed-forward neural networks

Before examining computational models including recurrent connections, we analyzed bottom-up architectures and showed that they were *not* robust to extrapolating from whole objects to partial objects (**Figure 4**). However, there exist infinitely many possible bottom-up models. Hence, even though we examined state-of-the-art models that are quite successful in object recognition, the failure to account for the behavioral and physiological results in the bottom-up models examined here (as well as similar failures reported in other studies, e.g. (41, 42)) should be interpreted with caution. We do not imply that it is impossible for *any* bottom-up architecture to recognize partially visible objects. In fact, it is possible to unfold a recurrent network with a finite number of time steps into a bottom-up model by creating an additional layer for each additional time step. However, there are several advantages to performing those computations with a recurrent architecture including a drastic reduction in the number of units required as well as in the number of weights that need to be trained and the fact that such unfolding is applicable only when we know *a priori* the fixed number of computational steps required, in contrast with recurrent architectures that allow an arbitrary and variable number of computations.

Recurrent computations and "slower" integration

A related interpretation of the current findings is that more challenging tasks, such as recognizing objects from minimal pixel information, may lead to "slower processing" throughout the ventral visual stream. According to this idea, each neuron would receive weaker inputs and require a longer time for integration, leading to the longer latencies observed experimentally at the behavioral and physiological level. It seems unlikely that the current observations could be fully accounted by longer integration times at all levels of the visual hierarchy. First, all images were contrast normalized to avoid any overall intensity effects. Second, neural delays for poor visibility images were not observed in early visual areas (25). Third, the correlations between the effects of backward masking and neural delays persisted even after accounting for difficulty level (**Fig. 3**). Fourth, none of the state-of-the-art purely bottom-up computational models were able to account for human level performance (see further elaboration of this point below). These arguments rule out slower processing throughout the entire visual system due to low intensity signals in the lower visibility conditions. However, the results presented here are still compatible with the notion that the inputs to higher-level neurons in the case of partial objects could be weaker and could require further temporal integration. This possibility is consistent with the model proposed here. Because the effects of recurrent computations are delayed with respect to the bottom-up inputs, we expect that any such slow integration would have to interact with the outputs of recurrent signals.

Extensions to the proposed proof-of-concept architecture

A potential challenge with attractor network architectures is the pervasive presence of spurious attractor states, particularly prominent when the network is near capacity. Furthermore, the simple instantiation of a recurrent architecture presented here still performed below humans, particularly under very low visibility conditions. It is conceivable that more complex architectures that take into account the known lateral connections in every layer as well as top-down connections in visual cortex might improve performance even further. Additionally, future extensions will benefit from incorporating other cues that help in pattern completion such as relative positions (front/behind), segmentation, movement, source of illumination, and stereopsis, among others.

Mixed training regime

All the computational results shown in the main text and discussed thus far involve training models *exclusively* with whole objects and testing performance with images of partially visible objects. Here we discuss a “mixed training” regime where the models are trained with access to partially visible objects. As emphasized in the main text, these are weaker models since they show less extrapolation (from partially visible objects to other partially visible objects as opposed to from whole objects to partially visible objects) and they depart from the typical ways of assessing invariance to object transformations (e.g. training at one rotation and testing at other rotations). Furthermore, humans do not require this type of additional training as described in the novel object experiments reported in **Figures S8** and **S9**. Despite these caveats, the mixed training regime is interesting to explore because it seems natural to assume that, at least in some cases, humans may be exposed to both partially visible objects and their whole counterparts while learning about objects. We emphasize that we cannot directly compare models that are trained only with whole objects and models that are trained with both whole objects and partially visible ones.

We considered two different versions of RNN models that were trained to reconstruct the feature representations of the whole objects from the feature representations of the corresponding partial objects. These models were based on a mixed training regime whereby both whole objects and partial objects were used during training. The state at time *t*>0 was computed as the activation of the weighted sum of the previous state and the input form the previous layer: $\mathbf{h}_t = \mathrm{ReLU}(\mathbf{W}_h \mathbf{h}_{t-1}, \mathbf{x}_t)$ where ReLU*(z)=max(0,z)*. The loss function was the mean squared Euclidean distance between the features from the partial objects and the features from the whole objects. Specifically, the RNN was iterated for a fixed number of time steps (t_{max} = 4) after the initial feed-forward pass, keeping the input from fc6 constant. Thus, letting $\mathbf{h}^i_{t_{\max}}$ be the RNN state at the last time step for a given image *i* and ${}_{whole}\mathbf{h}^i_{t0}$ be the feed-forward feature vector of the corresponding whole image, the loss function has the form

$$E = \frac{1}{T_I}\sum_{i=1}^{T_I}\left[\frac{1}{T_u}\sum_{j=1}^{T_u}(h^i_{t_{\max}}[j] - {}_{whole}h^i_{t0}[j])^2\right]$$

where j goes over all the T_u units in fc7 and i goes over all the T_I images in the training set. The RNN was trained in a cross validated fashion (5 folds) using the same cross validation scheme as with the feed-forward models and using the RMSprop algorithm for optimization. In RNN_5, the weights of the RNN were trained with 260 objects for each fold. All of the partial objects from the psychophysics experiment for the given 260 objects, as well as one copy of the original 260 images, were used to train the RNN for the corresponding split. In the case where the input to the RNN was the original image itself, the network did not change its representation over the recurrent iterations. Given the high number of weights to be learned by the RNN as compared to the number of training examples, the RNNs overfit fairly quickly. Therefore, early stopping (10 epochs) was implemented as determined from the validation set, i.e., we used the weights at the time step where the validation error was minimal.

To evaluate the extent of extrapolation across categories, we considered an additional version, RNN_1. In RNN_1, the recurring weights were trained using objects from only one category and the model was tested using objects from the remaining 4 categories. In all RNN versions, once $\mathbf{W}_h$ was fixed, classification performance was assessed using a linear SVM, as with the feed-forward models. Specifically, the SVM boundaries were trained using the responses from the feed-forward model to the whole objects and performance was evaluated using the representation at different time steps of recurrent computation.

The RNN_5 model had 40962 recurrent weights trained on a subset of the objects from all five categories. The RNN_5 model matched or surpassed human performance (**Figure S11**). Considering all levels of visibility, the RNN_5 model performed slightly above human levels ($p=3x10^{-4}$, Chi-squared test). While the RNN_5 model can extrapolate across objects and categorize images of partial objects that it has not seen before, it does so by exploiting features that are similar for different objects within the 5 categories in the experiment. RNN_1, a model where the recurrent weights were trained using solely objects from one of the categories and performance was evaluated using objects from the remaining 4 categories, did not perform any better than the purely feed-forward architecture ($p=0.05$, Chi-squared test). Upon inspection of the fc7 representation, we observed that several of the features were sparsely represented across categories. Therefore, the recurrent weights in RNN_1 only modified a fraction of all the possible features, missing many important features to distinguish the other objects. Thus, the improvement in RNN_5 is built upon a sufficiently rich dictionary of features that are shared among objects within a category. These results show that recurrent neural networks trained with subsets of the partially visible objects can achieve human level performance, extrapolating across objects, as long as they are trained with a sufficiently rich set of features.

We also evaluated the possibility of training the bottom-up model (AlexNet) using the mixed training regime and the same loss function as with RNN_5 and RNN_1, i.e. the Euclidean distance between features of whole and occluded images. Using the fc7 representation of the AlexNet model trained with partially visible objects also led to a model that either matched or surpassed human level performance at most visibility levels (**Figure S11**). The bottom-up model in the mixed training regime showed slightly worse performance than humans at very high visibility levels, including whole objects, perhaps because of the extensive fine-tuning with partially visible objects (note performance above humans at extremely low visibility levels). Within the mixed-training regimes, the RNN_5 model slightly outperformed the bottom-up model (**Figure S11**).

A fundamental distinction between the models presented in the text, particularly RNN_h, and the models introduced here, is that the mixed training models require training with partial objects from the same categories in which they will be evaluated. Although the specific photographs of objects used in the psychophysics experiments presented here were new to the subjects, humans have extensive experience in recognizing similar objects from partial information. It should also be noted that there is a small number of partially visible images in ImageNet, albeit not with such low visibility levels as the ones explored here, and all the models considered here were pre-trained using ImageNet. Yet, the results shown in **Figures S8-S9** demonstrate that humans can recognize objects shown under low visibility conditions even when they have had no experience with partial objects of a specific category and have had only minimal experience with the corresponding whole objects.

Temporal scale for recurrent computations

The models presented here, and several discussions in the literature, schematically and conceptually separate feed-forward computations from within-layer recurrent computations. Physiological signals arising within ~150 ms after stimulus onset have been interpreted to reflect largely feed-forward processing (1, 3, 5, 8, 10, 11, 43), whereas signals arising in the following 50 to 100 ms may reflect additional recurrent computations (27, 44, 45) . This distinction is clearly an oversimplification: the dynamics of recurrent computations can very well take place quite rapidly and well within ~150 ms of stimulus onset (46). Rather than a schematic initial feed-forward path followed by recurrent signals within the last layer in discrete time steps as implemented in RNN_h, cortical computations are based on continuous time and continuous interactions between feed-forward and within-layer signals (in addition to top-down signals). A biologically plausible implementation of a multi-layered spiking network including both feed-forward and recurrent connectivity was presented in ref. (46), where the authors estimated that recurrent signaling can take place within ~15 ms of computation per layer. Those time scales are consistent with the results shown here. Recurrent signals offer dynamic flexibility in terms of the amount of computational processing. Under noisy conditions (an injected noise term added to modify the input to each layer in (46), more occlusion in our case, and generally any internal or external source of noise), the system can dynamically use more computations to solve the visual recognition challenge.

Figures 4C-F, S10, S11, and **S12** show dynamics evolving over tens of discrete recurrent time steps. The RNNh model performance and correlation with humans saturate within approximately 10-20 recurrent steps (**Fig. 4C-F**). Membrane time constants of 10-15 ms (47) and one time constant per recurrent step would necessitate hundreds of milliseconds. Instead, the behavioral and physiological delays accompanying recognition of occluded objects occur within a delay of 50 to 100 ms (**Fig. 1-2, S12**) (25, 48), which are consistent with a continuous time implementation of recurrent processing (46).

3. Supplementary Figures Legends

Figure S1: Robust performance with occluded stimuli
We measured categorization performance with masking (solid lines) or without masking (dashed lines) for (**A**) partial and (**B**) occluded stimuli on a set of 16 exemplars belonging to 4 categories (chance = 25%, dashed lines). There was no overlap between the 14 subjects that participated in (**A**) and the 15 subjects that participated in (**B**). The effect of backward masking was consistent across both types of stimuli. The black lines indicate whole objects and the gray lines indicate the partial and occluded objects. Error bars denote SEM.

Figure S2: Example half-split reliability of psychophysics data
Figure 2E in the main text reports the masking index, a measure of how much recognition of each individual image is affected by backward masking. This measure is computed by averaging performance across subjects. In order to evaluate the variability in this metric, we randomly split the data into two halves and computed the masking index for each image for each half of the data. This figure shows one such split and how well one split correlates with the other split. **Figure 2F** shows error bars defined by computing standard deviations of the masking indices from 100 such random splits.

Figure S3: Bottom-up models can recognize minimally occluded images
A. Extension to **Figure 3A** showing that bottom-up models successfully recognize objects when more information is available (**Figure 3A** showed visibility values up to 35% whereas this figure extends visibility up to 100%). The format and conventions are the same as those in **Figure 3A**. The black dotted line shows interpolated human performance between the psychophysics experimental values measured at 35% and 100% visibility levels.

(**B**) Stochastic neighborhood embedding dimensionality reduction (t-SNE, **Methods**) to visualize the fc7 representation in the AlexNet model for whole objects (open circles) and partial objects (closed circles). Different categories are separable in this space, but the boundaries learned on whole objects did not generalize to the space of partial objects. The black arrow shows a schematic example of model distance definition, from an image of a partial face (green circle) to the average face centroid (black cross).

Figure S4: All of the purely feed-forward models tested were impaired under low visibility conditions

The human, AlexNet-pool5 and AlexNet-fc curves are the same ones shown in **Figure 3A** and are reproduced here for comparison purposes. This figure shows performance for several other models: VGG16-fc2, VGG19-fc2, ResNet50-flatten, inceptionV3-mixed10, VGG16-block5 (see text for references). In all cases, these models were pre-trained to optimize performance under ImageNet 2012 and there was no additional training (see also **Figure S5**). An expanded version of this figure with many other layers and models can be found on our web site: http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html

Figure S5: Fine-tuning did not improve performance under heavy occlusion

The human and fc7 curves are the same ones shown in **Figure 3A** and are reproduced here for comparison purposes. The pre-trained AlexNet network used in the text was fine tuned using back-propagation with the set of *whole* images from the psychophysics experiment (in contrast with the pre-trained Alexnet network which was trained using the Imagenet 2012 data set). The fine-tuning involved all layers (**Methods**).

Figure S6: Correlation between RNN_h model and human performance for individual objects as a function of time

At each time step in the recurrent neural network model (RNN_h), the scatter plots show the relationship between the model's performance on individual partial exemplar objects and human performance. Each dot is an individual exemplar object. In **Figure 4E** we report the average correlation coefficient across all categories.

Figure S7: Adding recurrent connectivity to VGG16 also improved performance

This Figure parallels the results shown in **Figure 4B** for AlexNet, here using the VGG16 network, implemented in keras (**Methods**). The results shown here are based on using 4096 units from the fc1 layer. The red curve (vgg16-fc1) corresponds to the original model without any recurrent connections. The implementation of the RNN$_h$ model here (VGG16-fc1-Hopfield) is similar to the one in **Figure 4B**, except that here we use the VGG16 fc1 activations instead of the AlexNet fc7 activations. An expanded version of this figure with similar results for several other layers and models can be found on our web site: http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html

Figure S8: Robust recognition of *novel* objects under low visibility conditions

A. Single exemplar from each of the 5 novel object categories (**Methods**). (**B-C**) Behavioral performance for the unmasked (**B**) and masked (**C**) trials. The experiment was identical to the one in **Figure 1** and the format of this figure follows that in **Figure 1F-G**. The colors denote different SOAs. Error bars=SEM. Dashed line = chance level (20%). Bin size=2.5%. Note the discontinuity in the x-axis to report performance for whole objects (100% visibility). (**D**) Average recognition performance as a function of the stimulus onset asynchrony (SOA) for partial objects (same data and conventions as **B-C**, excluding 100% visibility). Error bars=SEM. Performance was significantly degraded by masking (solid) compared to the unmasked trials (dotted) (p<0.0001, Chi-squared test, d.f.=4).

Figure S9: The performance of feed-forward and recurrent computational models for *novel* objects was similar to those for known object categories

A. Performance of feed-forward computational models (format as in **Figure 3A**) for novel objects.

B. Performance of the recurrent neural network RNN_h (format as in **Figure 4B**) for novel objects.

C. Temporal evolution of the feature representation for RNN_h (format as in **Figure 4C**). The colors and greek letters denote the five object categories (see examples in **Figure S8A**).

D. Performance of RNN_h as a functon of recurrent time for novel objects (format as in **Figure 4D**).

Figure S10: Side-by-side comparison of neurophysiological signals, psychophysics and computational model

A. Adaptation of Figure 6C from Tang et al 2014. This figure shows the dynamics of decoding object information for whole objects and (black) and partial objects (gray) from neurophysiological recordings as a function of time post stimulus onset (see Tang et al 2014 for details.

B. Reproduction of **Figure 1H** (behavior).

C. Reproduction of **Figure 4F** (RNN_h model).

Above each subplot, the experiment schematic highlights that part **A** involves no masking and fixed SOA = 150 ms whereas parts **B** and **C** involve masking and variable SOAs. The inset in part **C** directly overlays the results of the RNN_h model in part **C** onto the results of the psychophysics experiment in part **B.** In order to create this plot, we mapped 0 time steps to 25ms, 256 time steps to 150 ms and linearly interpolated the time steps in between.

Figure S11: Mixed training regimes.

A. This figure follows the format of **Fig3A**, **4B** and **S3, S4, S5, S7, S9A-B**. The black line shows human performance and is copied from **Fig. 3A.** The green and blue lines show the recurrent model (RNN_5) and bottom-up model (AlexNet fc7), respectively, trained in a mixed regime that included the occluded objects with visibility levels within the gray rectangle (the same ones used to evaluate human psychophysics performance). In the RNN5 model, there were ~16 million weights trained (all-to-all in the fc7 layer) whereas in the Alexnet fc7 model, there were ~60 million weights trained (all the weights across layers in the Alexnet model). Cross-validated test performance is shown here as well as in the other figures throughout the manuscript. As noted in the text, we emphasize that this figure involves a different training regime from the ones in the previous figures and therefore one cannot directly compare performance with the previous figures.

B. This figure follows the format of **Fig. 4E**. The green and blue bars show the correlation between human and model for the recurrent model and bottom-up model, respectively, both trained using occluded objects. The gray rectangle shows human-human correlation, see **Fig. 4E** for details..

Figure S12: Image-by-image comparison between RNNh model performance and human performance in the masked condition

Expanding on **Figure 4E**, this figure shows the correlation coefficient between human recognition performance in the masked condition (**Figure 1B**) at a given SOA (y-axis) and RNN$_h$ model performance at a given time step (x-axis). The top row shows the unmasked condition (**Figure 1A**). In this figure, there is no mask for the model (see **Figure 4F** for model performance with a mask). The computation of the correlation coefficient follows the same procedure illustrated in **Figure S6** and **4E**. The color scale for the correlation coefficient is shown on the right. As an upper bound and as shown in **Figure 4E**, the correlation coefficient between different human subjects was 0.41 for the unmasked condition. The yellow boxes highlight the highest correlation for a given SOA value.

4. Author contributions

Conceptualization: HT, BL, MS, DC, GK

Physiology experiment design: HT, GK

Physiological data collection and analyses: HT

Psychophysics experiment design: HT, BL, MS, CM, GK

Psychophysics data collection: HT, BL, MS, AP, JO, WH, CM

Computational models: HT, BL, MS, DC, CM, GK

Resources: DC, GK

Manuscript writing: HT, BL, MS, GK

5. Data availability

All relevant data and code (including image databases, behavioral measurements, physiological measurements and computational algorithms) are publicly available through the lab's website and through the lab's GitHub page: http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html

6. References

1. Kirchner H & Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision research* 46(11):1762-1776.
2. Potter M & Levy E (1969) Recognition memory for a rapid sequence of pictures. *Journal of experimental psychology* 81(1):10-15.
3. Keysers C, Xiao DK, Foldiak P, & Perret DI (2001) The speed of sight. *Journal of Cognitive Neuroscience* 13(1):90-101.
4. Hung CP, Kreiman G, Poggio T, & DiCarlo JJ (2005) Fast Read-out of Object Identity from Macaque Inferior Temporal Cortex. *Science* 310:863-866.
5. Liu H, Agam Y, Madsen JR, & Kreiman G (2009) Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron* 62(2):281-290.
6. Tovee M & Rolls E (1995) Information encoding in short firing rate epochs by single neurosn in the primate temporal visual cortex. *Visual Cognition* 2(1):35-58.
7. Pinto N, Doukhan D, DiCarlo JJ, & Cox DD (2009) A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol* 5(11):e1000579.
8. Riesenhuber M & Poggio T (1999) Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2(11):1019-1025.
9. Wallis G & Rolls ET (1997) Invariant face and object recognition in the visual system. *PROGRESS IN NEUROBIOLOGY* 51(2):167-194.
10. Yamins DL*, et al.* (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 111(23):8619-8624.
11. Serre T*, et al.* (2007) A quantitative theory of immediate visual recognition. *Progress In Brain Research* 165C:33-56.
12. Breitmeyer B & Ogmen H (2006) *Visual Masking: Time Slices through Conscious and Unconscious Vision* (Oxford University Press, New York).

13. Bridgeman B (1980) Temporal response characteristics of cells in monkey striate cortex measured with metacontrast masking and brightness discrimination. *Brain Res* 196(2):347-364.
14. Macknik SL & Livingstone MS (1998) Neuronal correlates of visibility and invisibility in the primate visual system. *Nature neuroscience* 1(2):144-149.
15. Lamme VA, Zipser K, & Spekreijse H (2002) Masking interrupts figure-ground signals in V1. *J Cogn Neurosci* 14(7):1044-1053.
16. Kovacs G, Vogels R, & Orban GA (1995) Cortical correlate of pattern backward masking. *Proceedings of the National Academy of Sciences* 92(12):5587-5591.
17. Rolls ET, Tovee MJ, & Panzeri S (1999) The neurophysiology of backward visual masking: information analysis. *Journal of Cognitive Neuroscience* 11(3):300-311.
18. Keysers C & Perrett DI (2002) Visual masking and RSVP reveal neural competition. *Trends Cogn Sci* 6(3):120-125.
19. Enns JT & Di Lollo V (2000) What's new in visual masking? *Trends Cogn Sci* 4(9):345-352.
20. Thompson KG & Schall JD (1999) The detection of visual signals by macaque frontal eye field during masking. *Nature neuroscience* 2(3):283-288.
21. Kellman PJ, Guttman S, & Wickens T (2001) Geometric and neural models of object perception. *From framents to objects: Segmentation and grouping in vision*, eds Shipley TF & Kellman PJ (Elsevier Science Publishers, Oxford, UK).
22. Murray RF, Sekuler AB, & Bennett PJ (2001) Time course of amodal completion revealed by a shape discrimination task. *Psychon Bull Rev* 8(4):713-720.
23. Kosai Y, El-Shamayleh Y, Fyall AM, & Pasupathy A (2014) The role of visual area V4 in the discrimination of partially occluded shapes. *Journal of Neuroscience* 34(25):8570-8584.
24. Nakayama K, He Z, & Shimojo S (1995) Visual surface representation: a critical link between lower-level and higher-level vision. *Visual cognition*, eds Kosslyn S & Osherson D (The MIT press, Cambridge), Vol 2.
25. Tang H*, et al.* (2014) Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* 83:736-748.
26. Johnson JS & Olshausen BA (2005) The recognition of partially visible natural objects in the presence and absence of their occluders. *Vision research* 45(25-26):3262-3276.
27. Lee TS (2003) Computations in the early visual cortex. *J Physiol Paris* 97(2-3):121-139.
28. Gosselin F & Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. *Vision research* 41(17):2261-2271.
29. Williams P (1998) Representational organization of multiple exemplars of object categories.
30. Willenbockel V*, et al.* (2010) Controlling low-level image properties: the SHINE toolbox. *Behav Res Methods* 42(3):671-684.
31. Krizhevsky A, Sutskever I, & Hinton G (2012) ImageNet Classification with Deep Convolutional Neural Networks. in *NIPS* (Montreal).

32. Russakovsky O*, et al.* (2014) ImageNet Large Scale Visual Recognition Challenge. in *CVPR* (arXiv:1409.0575, 2014).
33. Yangqing J*, et al.* (2014) Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv*:1408.5093.
34. Simonyan K & Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv* 1409.1556.
35. He K, Zhang X, Ren S, & Sun J (2015) Deep residual learning for image recognition. *arXiv* 1512.03385.
36. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, & Wojna Z (2015) Rethinking the inception architecture for computer vision. *arXiv* 1512.005673v3.
37. van der Maaten L & Hinton G (2008) Visualizing High-Dimensional Data Using t-SNE. *J. Machine Learning Res.* 9:2579-2605.
38. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *PNAS* 79:2554-2558.
39. Li J, Michel A, & Porod W (1989) Analysis and synthesis of a class of neural networks: linear systems operating on a closed hypercube. *IEEE Transactions on Circuits and Systems* 36(11):1405-1422.
40. Bregman AL (1981) *Asking the "what for" question in auditory perception* (Erlbaum, Hillsdale, NJ) p 19.
41. Pepik B, Benenson R, Ritschel T, & Schiele B (2015) What is holding back convnets for detection? 1508.
42. Spoerer CJ, McClure P, & Kriegeskorte N (2017) Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. *Frontiers in psychology* 8:1551.
43. DiCarlo JJ & Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11(8):333-341.
44. Lamme VA & Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci* 23(11):571-579.
45. Gilbert CD & Li W (2013) Top-down influences on visual processing. *Nat Rev Neurosci* 14(5):350-363.
46. Panzeri S, Rolls ET, Battaglia F, & Lavis R (2001) Speed of feedforward and recurrent processing in multilayer networks of integrate-and-fire neurons. *Network* 12(4):423-440.
47. Koch C (1999) *Biophysics of Computation* (Oxford University Press, New York).
48. Fyall AM, El-Shamayleh Y, Choi H, Shea-Brown E, & Pasupathy A (2017) Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. *eLife* 6.

Supplementary Figure 1

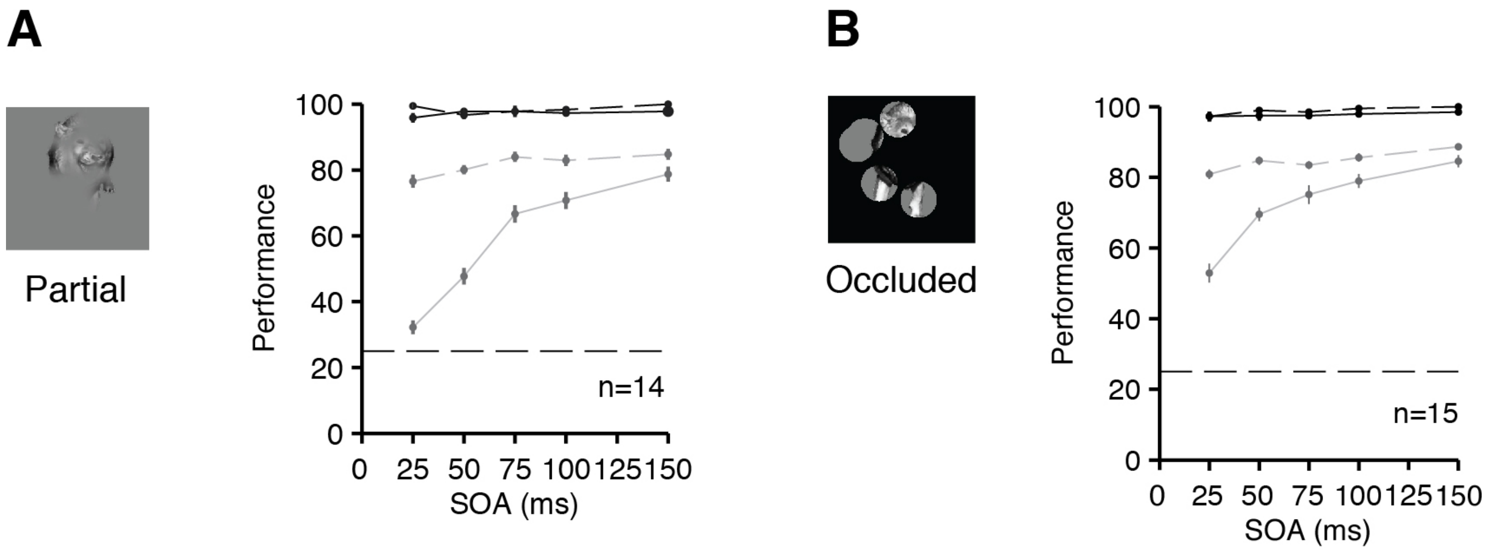

Figure S1: Robust performance with occluded stimuli

We measured categorization performance with masking (solid lines) or without masking (dashed lines) for (**A**) partial and (**B**) occluded stimuli on a set of 16 exemplars belonging to 4 categories (chance = 25%, dashed lines). There was no overlap between the 14 subjects that participated in (**A**) and the 15 subjects that participated in (**B**). The effect of backward masking was consistent across both types of stimuli. The black lines indicate whole objects and the gray lines indicate the partial and occluded objects. Error bars denote SEM.

Supplementary Figure 2

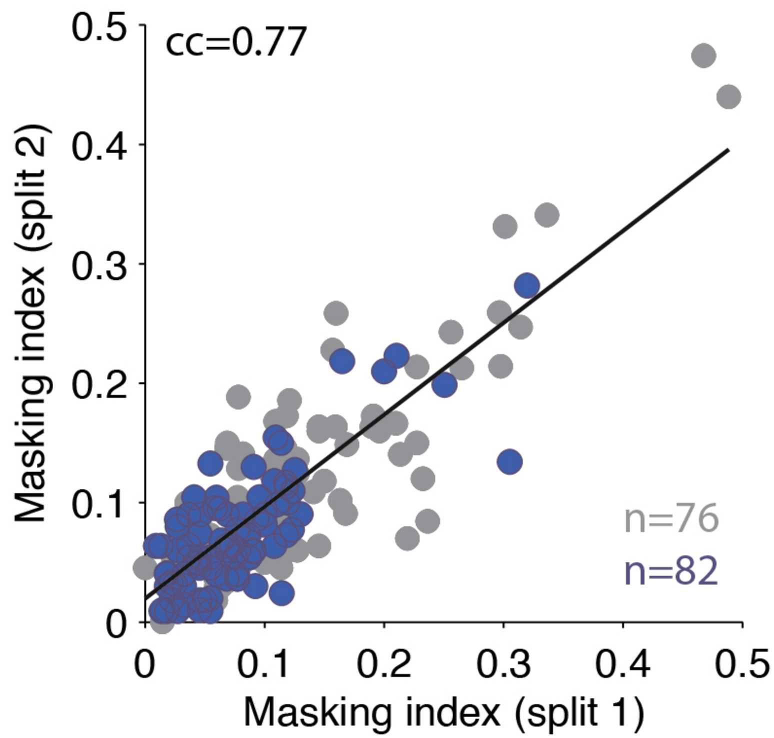

Figure S2: Example half-split reliability of psychophysics data

Figure 2E in the main text reports the masking index, a measure of how much recognition of each individual image is affected by backward masking. This measure is computed by averaging performance across subjects. In order to evaluate the variability in this metric, we randomly split the data into two halves and computed the masking index for each image for each half of the data. This figure shows one such split and how well one split correlates with the other split. **Figure 2F** shows error bars defined by computing standard deviations of the masking indices from 100 such random splits.

Supplementary Figure 3

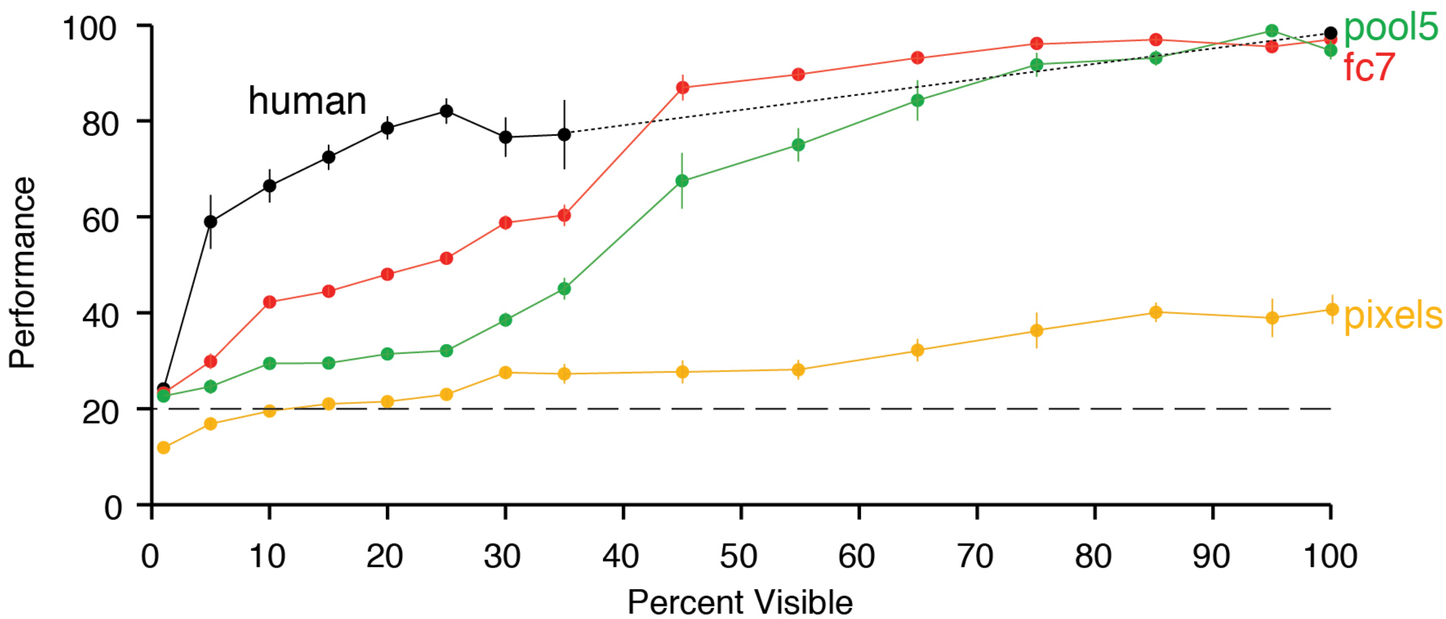

Figure S3: Bottom-up models can recognize minimally occluded images

Extension to **Fig. 3A** showing that bottom-up models successfully recognize objects when more information is available (**Fig. 3A** showed visibility values up to 35% whereas this figure extends visibility up to 100%). The format and conventions are the same as those in **Fig. 3A**. The black dotted line shows interpolated human performance between the psychophysics experimental values measured at 35% and 100% visibility levels.

Supplementary Figure 4

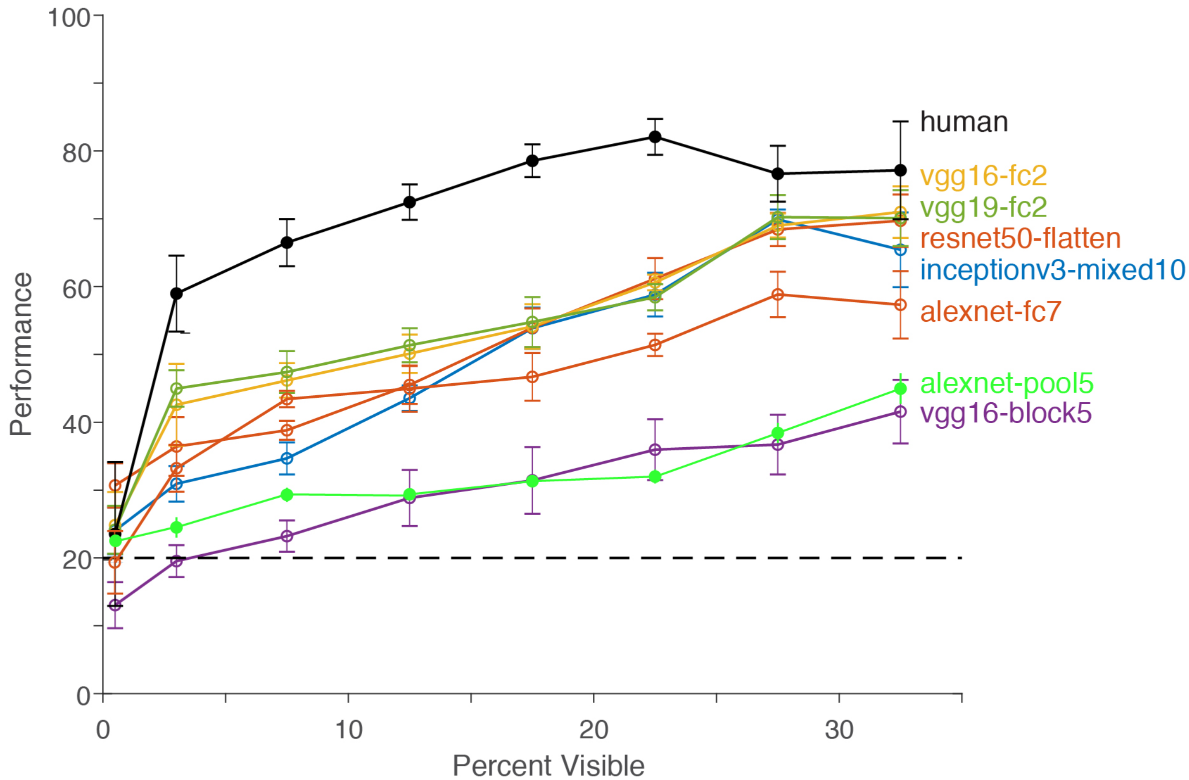

Figure S4: All of the purely feed-forward models tested were impaired under low visibility conditions

The human, AlexNet-pool5 and AlexNet-fc curves are the same ones shown in **Figure 3A** and are reproduced here for comparison purposes. This figure shows performance for several other models: VGG16-fc2, VGG19-fc2, ResNet50-flatten, inceptionV3-mixed10, VGG16-block5 (see text for references). In all cases, these models were pre-trained to optimize performance under ImageNet 2012 and there was no additional training (see also **Figure S5** for fine tuning results). An expanded version of this figure with many other layers and models can be found on our web site: http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html

Supplementary Figure 5

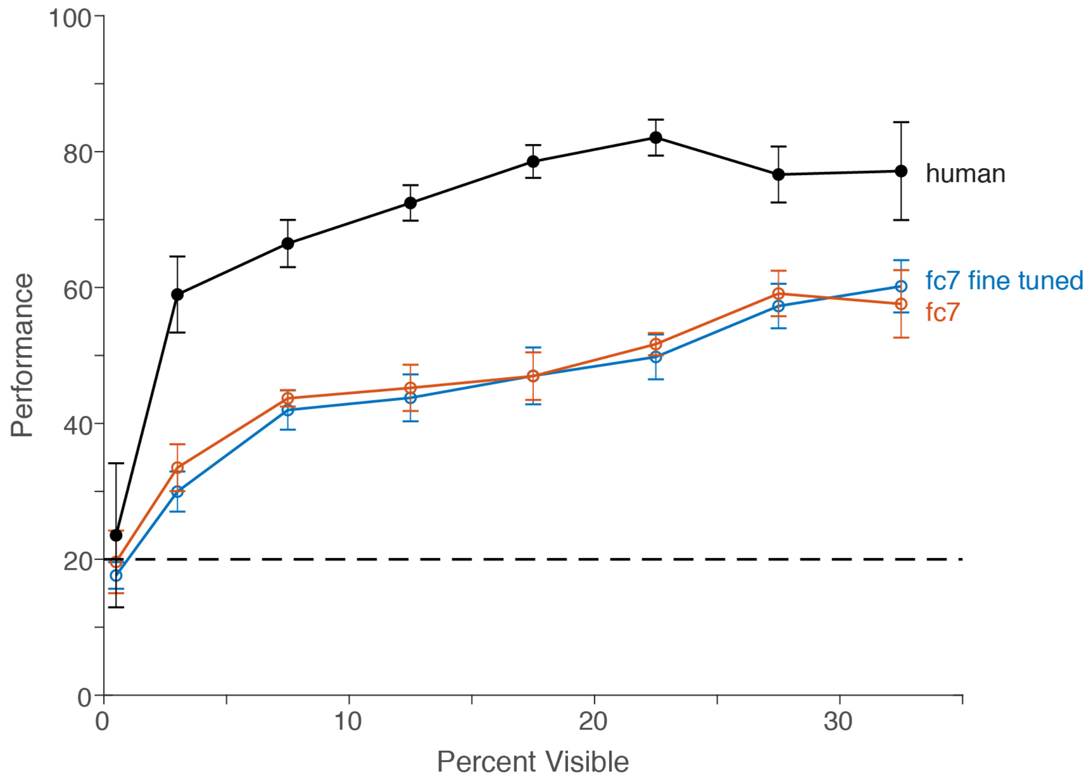

Figure S5: Fine-tuning did not improve performance under heavy occlusion

The human and fc7 curves are the same ones shown in **Figure 3A** and are reproduced here for comparison purposes. The pre-trained AlexNet network used in the text was fine tuned using back-propagation with the set of *whole* images from the psychophysics experiment (in contrast with the pre-trained Alexnet network which was trained using the Imagenet 2012 data set). The fine-tuning involved all layers (**Methods**).

Supplementary Figure 6

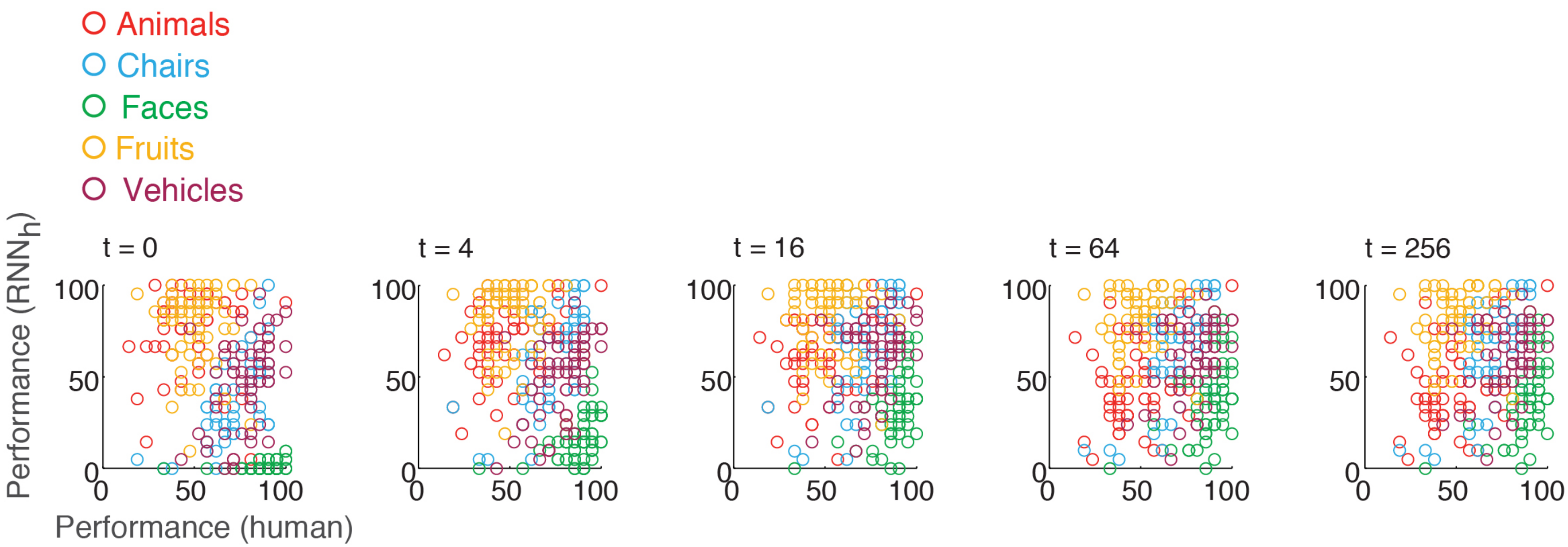

Figure S6: Correlation between RNN$_h$ model and human performance for individual objects as a function of time
At each time step in the recurrent neural network model (RNN$_h$), the scatter plots show the relationship between the model's performance on individual partial exemplar objects and human performance. Each dot is an individual exemplar object. In **Fig. 4E** we report the average correlation coefficient across all categories.

Supplementary Figure 7

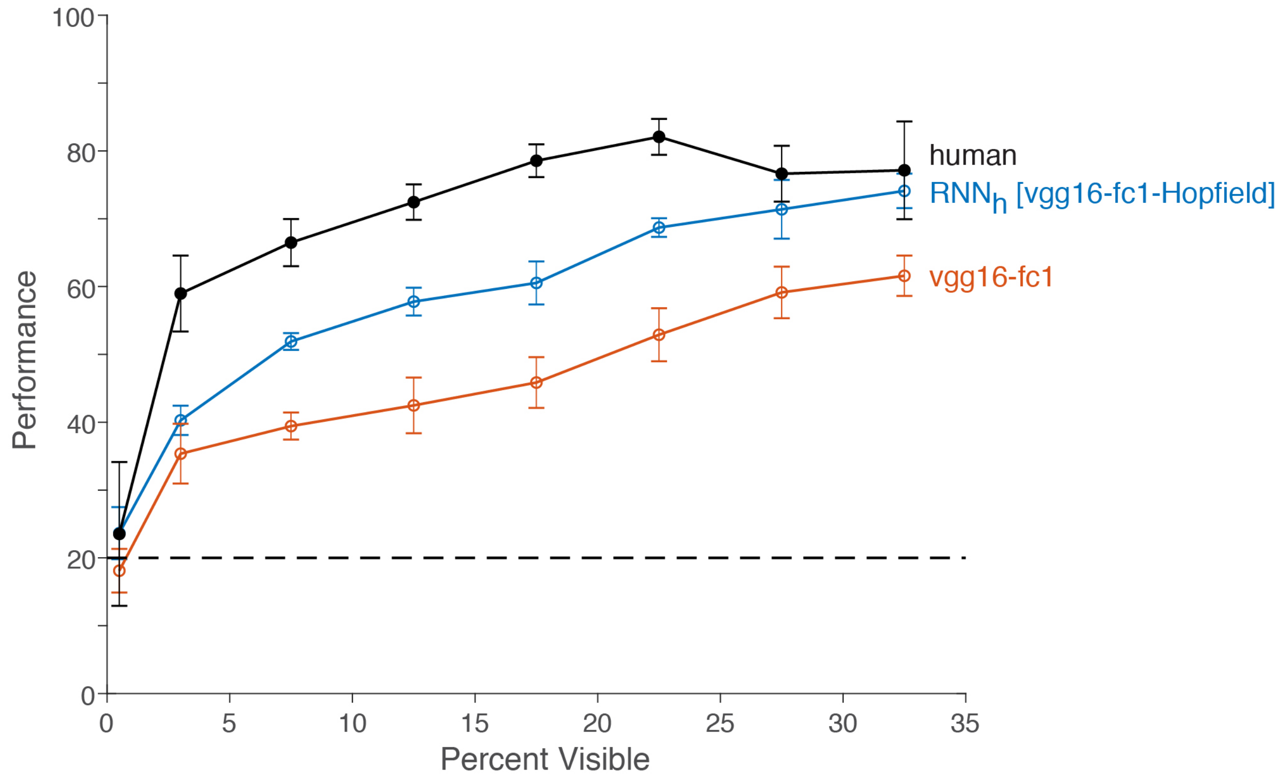

Figure S7: Adding recurrent connectivity to VGG16 also improved performance

This Figure parallels the results shown in **Figure 4B** for AlexNet, here using the VGG16 network, implemented in keras (**Methods**). The results shown here are based on using 4096 units from the fc1 layer. The red curve (vgg16-fc1) corresponds to the original model without any recurrent connections. The implementation of the RNN_h model here (VGG16-fc1-Hopfield) is similar to the one in **Figure 4B**, except that here we use the VGG16 fc1 activations instead of the AlexNet fc7 activations. An expanded version of this figure with similar results for several other layers and models can be found on our web site: http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html

Supplementary Figure 8

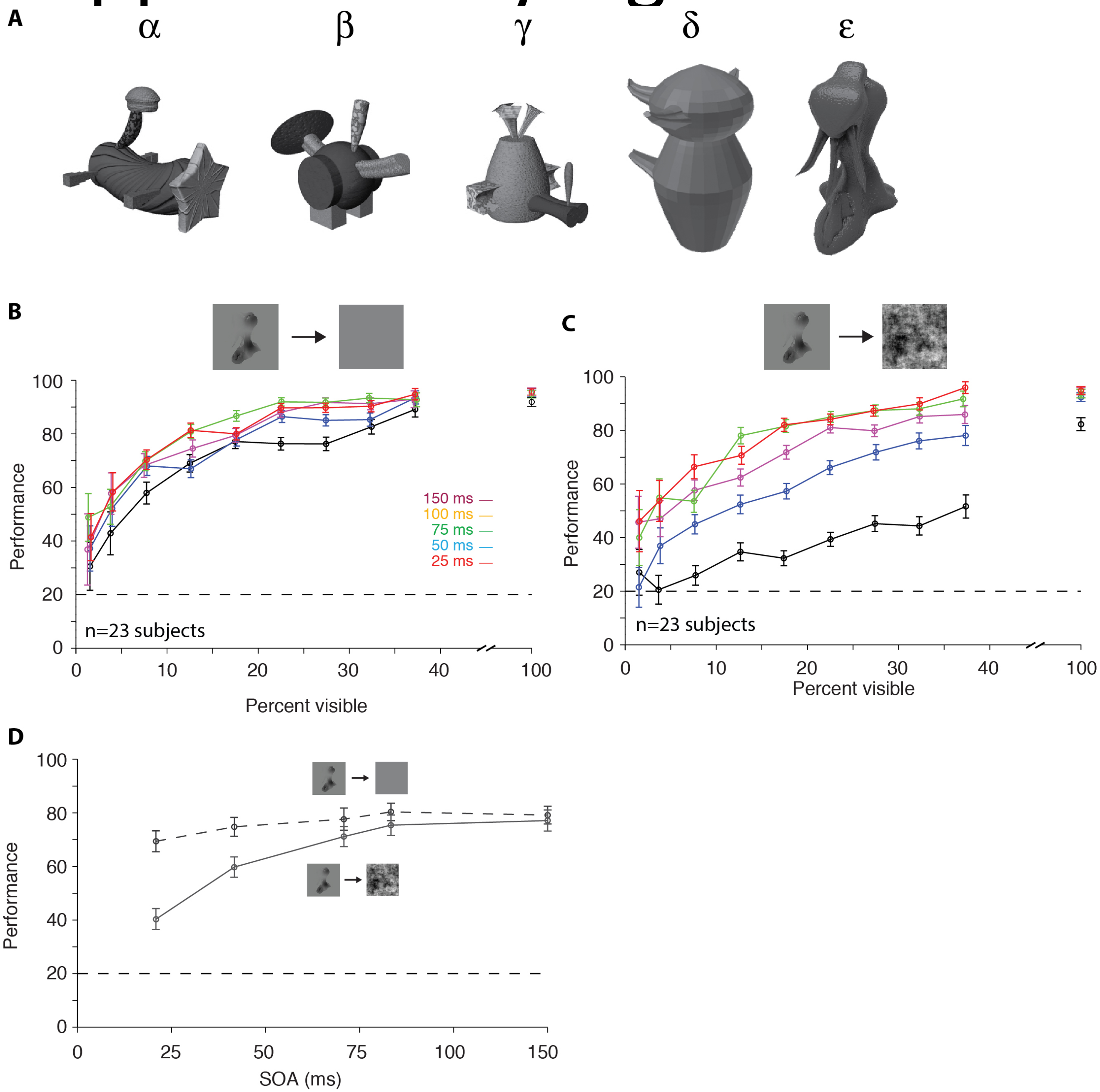

Figure S8: Robust recognition of *novel* objects under low visibility conditions

A. Single exemplar from each of the 5 novel object categories (**Methods**). (**B-C**) Behavioral performance for the unmasked (**B**) and masked (**C**) trials. The experiment was identical to the one in **Figure 1** and the format of this figure follows that in **Figure 1F-G**. The colors denote different SOAs. Error bars=SEM. Dashed line = chance level (20%). Bin size=2.5%. Note the discontinuity in the x-axis to report performance for whole objects (100% visibility). (**D**) Average recognition performance as a function of the stimulus onset asynchrony (SOA) for partial objects (same data and conventions as **B-C**, excluding 100% visibility). Error bars=SEM. Performance was significantly degraded by masking (solid) compared to the unmasked trials (dotted) (p<0.0001, Chi-squared test, d.f.=4).

Supplementary Figure 9

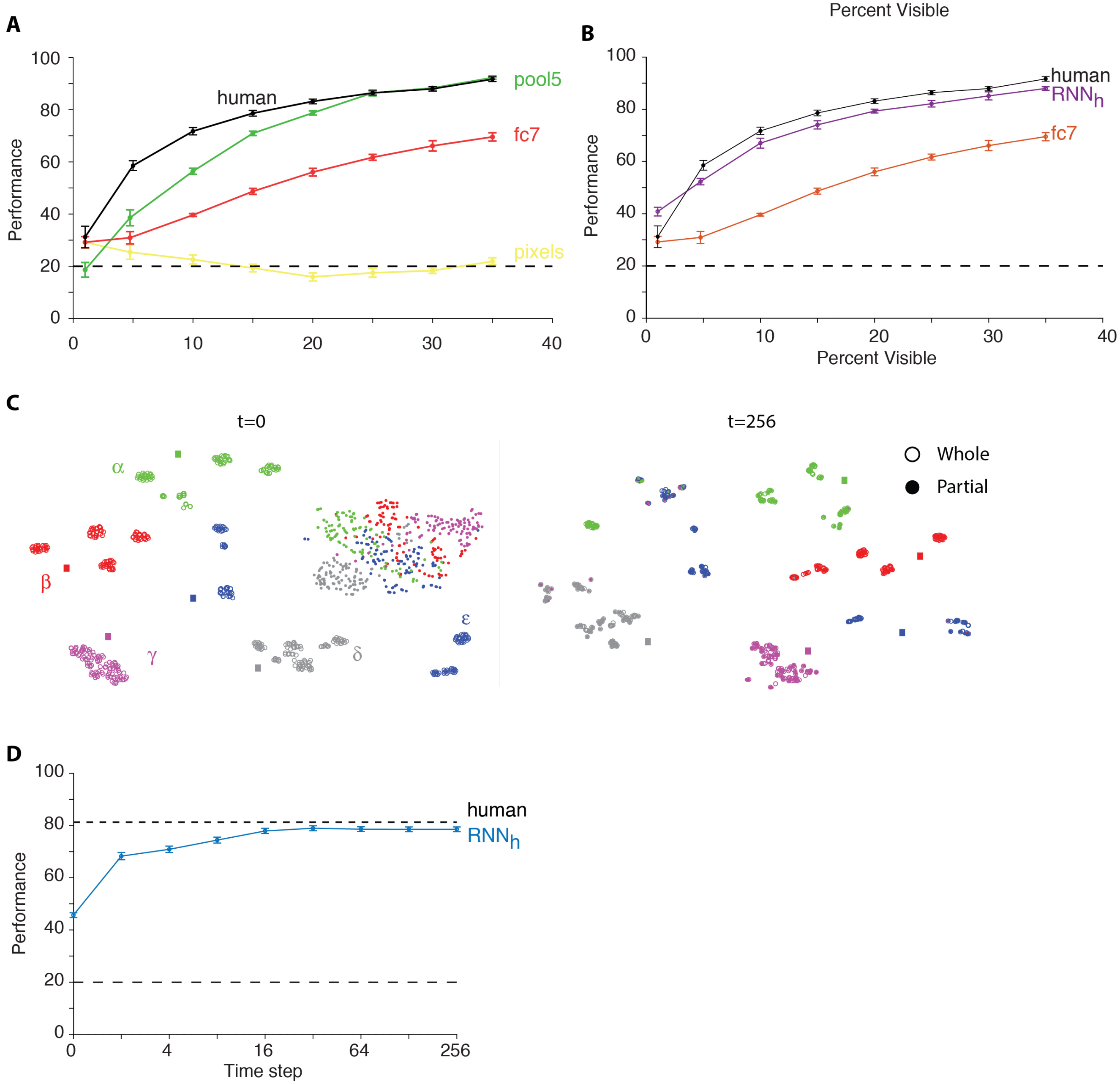

Figure S9: The performance of feed-forward and recurrent computational models for *novel* objects was similar to those for known object categories

A. Performance of feed-forward computational models (format as in **Figure 3A**) for novel objects.

B. Performance of the recurrent neural network RNN_h (format as in **Figure 4B**) for novel objects.

C. Temporal evolution of the feature representation for RNN_h (format as in **Figure 4C**). The colors and greek letters denote the five object categories (see examples in **Figure S8A**).

D. Performance of RNN_h as a functon of recurrent time for novel objects (format as in **Figure 4D**).

Supplementary Figure 10

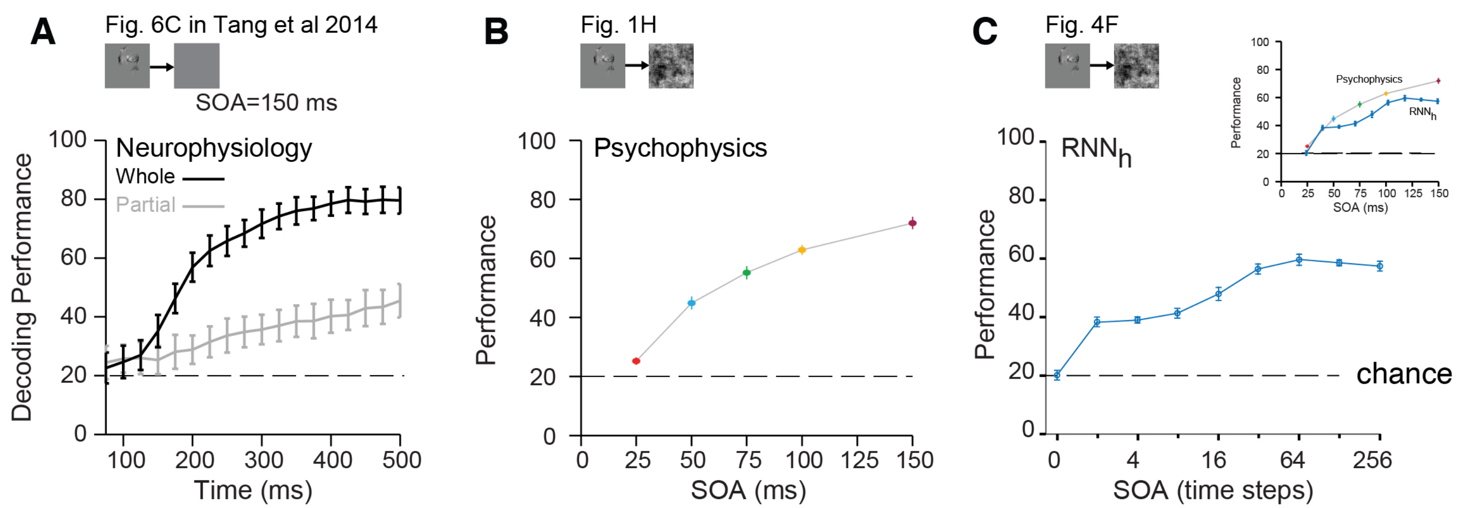

Figure S10: Side-by-side comparison of neurophysiological signals, psychophysics and computational model

A. Reproduction of Figure 6C from Tang et al 2014. This figure shows the dynamics of decoding object information for whole objects and (black) and partial objects (gray) from neurophysiological recordings as a function of time post stimulus onset (see Tang et al 2014 for details.

B. Reproduction of **Figure 1H** (behavior).

C. Reproduction of **Figure 4F** (RNN_h model).

Above each subplot, the experiment schematic highlights that **A** involves no masking and fixed SOA = 150 ms whereas **B, C** involve masking and variable SOAs. The inset in part **C** directly overlays the results of the RNN_h model in **C** onto the results of the psychophysics experiment in **B**. In order to create this plot, we mapped 0 time steps to 25ms, 256 time steps to 150 ms and linearly interpolated the time steps in between.

Supplementary Figure 11

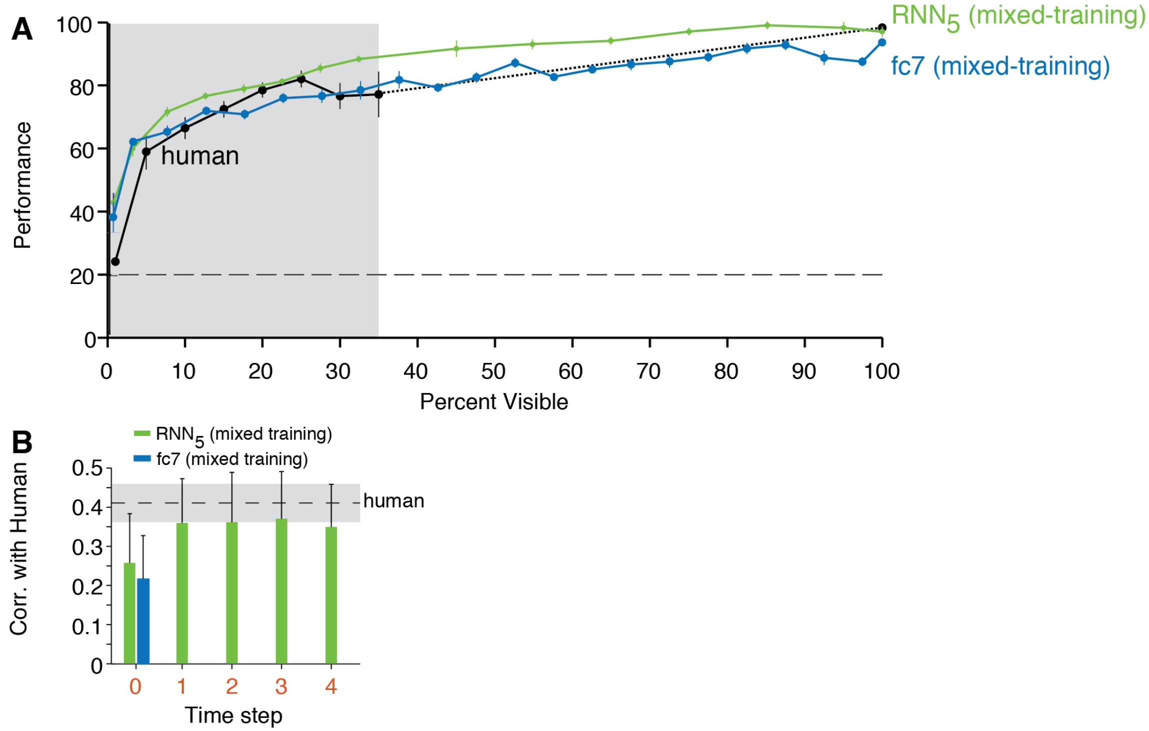

Figure S11: Mixed training regimes.

A. This figure follows the format of **Fig3A**, **4B** and **S3A, S4, S5, S7, S9A-B**. The black line shows human performance and is copied from **Fig. 3A** for comparison purposes. The green and blue lines show the recurrent model (RNN_5) and bottom-up model (AlexNet fc7), respectively, trained in a mixed regime that included the occluded objects with visibility levels within the gray rectangle (the same ones used to evaluate human psychophysics performance). In the RNN_5 model, there were ~16 million weights trained (all-to-all in the fc7 layer) whereas in the Alexnet fc7 model, there were ~60 million weights trained (all the weights across layers in the Alexnet model). Cross-validated test performance is shown here as well as in the other figures throughout the manuscript. As noted in the text, we emphasize that this figure involves a different training regime from the ones in the previous figures (here the models are trained with occluded objects) and, therefore, one cannot directly compare performance in this figure with the previous figures.

B. This figure follows the format of **Fig. 4E**. The green and blue bars show the correlation between human and model for the recurrent model and bottom-up model, respectively, both trained using occluded objects. The gray rectangle shows human-human correlation, see **Fig. 4E** for details..

Supplementary Figure 12

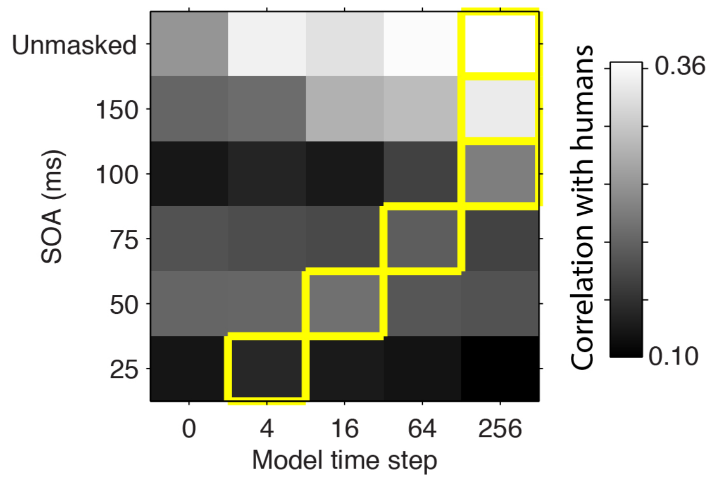

Figure S12: Image-by-image comparison between RNNh model performance and human performance in the masked condition

Expanding on **Figure 4E,** this figure shows the correlation coefficient between human recognition performance in the masked condition (**Figure 1B**) at a given SOA (y-axis) and RNNh model performance at a given time step (x-axis). The top row shows the unmasked condition (**Figure 1A**). In this figure, there is no mask for the model (see **Figure 4F** for model performance with a mask). The computation of the correlation coefficient follows the same procedure illustrated in **Figure S6** and **4E**. The color scale for the correlation coefficient is shown on the right. As an upper bound and as shown in **Figure 4E**, the correlation coefficient between different human subjects was 0.41 for the unmasked condition. The yellow boxes highlight the highest correlation for a given SOA value.