1001	Recurrent computations for visual pattern completion
1002	Supporting Information Appendix
1003	
1004	Hanlin Tang ^{1,4*} , Martin Schrimpf ^{2,4*} , William Lotter ^{1,3,4*} , Charlotte Moerman ⁴ , Ana
1005	Paredes ⁴ , Josue Ortega Caro ⁴ , Walter Hardesty ⁴ , David Cox ³ , Gabriel Kreiman ^{4 ×}
1006	
1007	1. Supplementary Materials and Methods
1008	2. <u>Supplementary Discussion</u>
1009	3. <u>Supplementary Figures Legends</u>
1010	4. Author contributions
1011	5. <u>Data availability</u>
1012	6. <u>References</u>
1013	
1014	1. Supplementary Materials and Methods
1015	Psychophysics experiments
1016	A total of 106 volunteers (62 female, ages 18-34 y) with normal or corrected to
1017	normal vision participated in the psychophysics experiments reported in this study.
1018	All subjects gave informed consent and the studies were approved by the
1019	Institutional Review Board at Children's Hospital, Harvard Medical School. In 67
1020	subjects, eye positions were recorded during the experiments using an infrared
1021	camera eye tracker at 500 Hz (Eyelink D1000, SR Research, Ontario, Canada). We
1022	performed a main experiment (reported in Figure 1F-G) and three variations
1023	(reported in Figures 1I-J, 2 , S1 and S8-9).
1024	
1025	Backward masking. Multiple lines of evidence from behavioral (e.g. (1, 2)),
1026	physiological (e.g. (3-6)), and computational studies (e.g. (7-11)) suggest that
1027	recognition of whole isolated objects can be approximately described by rapid,
1028	largely feed-forward, mechanisms. Despite the success of these feed-forward
1029	architectures in describing the initial steps in visual recognition, each layer has
1030	limited spatial integration of its inputs. Additionally, feed-forward algorithms lack

1031 mechanisms to integrate temporal information or to take advantage of the rich 1032 temporal dynamics characteristic of neural circuits that allow comparing signals 1033 within and across different levels of the visual hierarchy. It has been suggested that 1034 backward masking can interrupt recurrent and top-down signals: when an image is 1035 rapidly followed by a spatially overlapping mask: the new high-contrast mask 1036 stimulus interrupts any additional, presumably recurrent, processing of the original 1037 image (3, 12-20). Thus, the psychophysical experiments tested recognition under 1038 both unmasked and backward masked conditions.

1039

1040 *Main experiment*. Both spatial and temporal integration are likely to play an 1041 important role in pattern completion mechanisms (21-27). A scheme of the 1042 experiment designed to study the spatial and temporal integration during recognition of occluded or partially visible objects is shown in Figure 1. Twenty-one 1043 1044 subjects were asked to categorize images into one of 5 possible semantic groups (5-1045 alternative forced choice) by pressing buttons on a gamepad. Stimuli consisted of 1046 contrast-normalized gray scale images of 325 objects belonging to five categories 1047 (animals, chairs, human faces, fruits, and vehicles). Each object was only presented 1048 once in each condition. Each trial was initiated by fixating on a cross for at least 500 1049 ms. After fixation, subjects were presented with the image of an object for a variable 1050 time (25 ms, 50 ms, 75 ms, 100 ms, or 150 ms), referred to as the stimulus onset 1051 asynchrony (SOA). The image was followed by either a noise mask (**Figure 1B**) or a 1052 grav screen (Figure 1A), with a duration of 500 ms, after which a choice screen 1053 appeared requiring the subject to respond. We use the term "pattern completion" to 1054 indicate successful categorization of partial images in the 5-alternative forced choice 1055 task used here and we do not mean to imply that subjects are forming any mental 1056 image of the entire object, which we did not test. The noise mask was generated by 1057 scrambling the phase of the images, while retaining the spectral coefficients. The 1058 images (256 x 256 pixels) subtended approximately 5 degrees of the visual field. In 1059 approximately 15% of the trials, the objects were presented in unaltered fashion 1060 (the 'Whole' condition, **Figure 1C** left). In the other 85% of the trials, the objects 1061 were rendered partially visible by presenting visual features through Gaussian

1062 bubbles (28) (the 'Partial condition', standard deviation = 14 pixels, Figure 1C 1063 right). Each subject performed an initial training session to familiarize themselves 1064 with the task and the stimuli. They were presented with 40 trials of whole objects. 1065 then 80 calibration trials of occluded objects. During the calibration trials, the 1066 number of bubbles was titrated using a staircase procedure to achieve an overall 1067 task difficulty of 80% correct rate. The number of bubbles (but not their positions) 1068 was then kept constant for the rest of the experiment. Results from the 1069 familiarization and calibration phase were not included in the analyses. Despite 1070 calibrating the number of bubbles, there was a wide range of degrees of occlusion 1071 because the positions of the bubbles were randomized in every trial. Each image 1072 was only presented once in the masked condition and once in the unmasked 1073 condition.

1074

1075 *Physiology-based psychophysics experiment.* In the physiology-based psychophysics 1076 experiment (Figure 2, n = 33 subjects), stimuli consisted of 650 images from five 1077 categories for which we had previously recorded neural responses (see below). In 1078 the neurophysiological recordings (25), bubble positions were randomly selected in 1079 each subject and therefore each subject was presented with different images (except 1080 for the fully visible ones). The main difference between the physiology-based 1081 psychophysics experiment and the Main experiment is that here we used the exact 1082 same images that were used in the physiological recordings (see description under 1083 "Neurophysiological Recordings" below).

1084

1085 *Occlusion experiment*. In the occlusion experiment (**Figure 1I**, **Figure S1**, n=14

1086 subjects in the partial objects experiment and n =15 subjects in the occlusion

1087 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

1089 amodal completion cues. The stimulus set consisted of 16 objects from 4 different

1090 categories. For comparison, we also collected performance with partial objects from

1091 this reduced stimulus set.

1092

1093 *Novel objects experiment.* The main set of experiments required categorization of 1094 images containing pictures of animals, chairs, faces, fruits and vehicles. None of the 1095 subjects involved in the psychophysics or neurophysiological measurements had 1096 had any previous exposure to the *specific pictures* in these experiments, let alone 1097 with the partial images rendered through bubbles. Yet, it can be surmised that all 1098 the subjects had had extensive previous experience with *other* images of objects 1099 from those categories, including occluded versions of other animals, chairs, faces, 1100 fruits and vehicles. In order to evaluate whether experience with occluded instances 1101 of objects from a specific category is important to recognize novel instances of 1102 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 1103 1104 categories, all the novel objects were chosen from the Tarr Lab stimulus repository 1105 (29). An equal amount of stimuli were chosen from each category. One exemplar 1106 from each category is shown in **Figure S8A**. In the Cognitive Science community, the 1107 first three categories are known as "Fribbles" and the last two categories as 1108 "Greebles" and "Yufos" (29). In our experiments, each category was assigned a Greek 1109 letter name (**Figure S8A**) so as not to influence the subjects with potential meanings 1110 of an invented name.

1111 The experiment followed the same protocol as the main experiment (**Figure** 1112 **1**). Twenty-three new subjects (11 female, 20 to 34 years old) participated in this 1113 experiment. Since the subjects had no previous exposure to these stimuli, they 1114 underwent a short training session where they were presented with 2 fully visible 1115 exemplars from each category so that they could learn the mapping between 1116 categories and response buttons. In order to start the experiment, subjects were 1117 required to get 8 out of 10 correct responses, 5 times in a row using these practice 1118 stimuli. On average, reaching this level of accuracy required 80±40 trials. Those 2 1119 stimuli from each category were not used in the subsequent experiments. Therefore, 1120 whenever we refer to "novel" objects, what we mean is objects from 5 categories 1121 where subjects were exposed to ~ 80 trials of 2 fully visible exemplars per category, 1122 different from the ones used in the psychophysics tests. This regime represented 1123 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).

1131During the experiment, subjects were presented with 1,000 uniquely1132rendered stimuli from 500 contrast-normalized gray scale novel objects, resized to1133256x256 pixels, subtending approximately 5° of visual angle. All images were1134contrast normalized using the histMatch function from the SHINE toolbox (30).1135This function equates the luminance histogram of sets of images. For each subject,11361,000 unique renderings were obtained by applying different bubbles to the original1137images, resulting in a total of 23,000 different stimuli across subjects.

1138 The SOAs and other parameters were identical to those used in the main 1139 experiment. The analyses and models for the novel object experiments follow those 1140 in the main experiment (**Figures S8B-D** are the analogs of **Figure 1F-H**, **Figure S9A** 1141 is the analog of **Figure 3A**, **Figure S9B-D** are the analogs of **Figure 4B-D**).

1142

1143 Neurophysiology experiments

1144 The neurophysiological data analyzed in **Figures 2** and **3** were taken from 1145 the study by Tang et al (25), to which we refer for further details. Briefly, subjects 1146 were patients with pharmacologically intractable epilepsy who had intracranial electrodes implanted for clinical purposes. These electrodes record intracranial field 1147 1148 potential signals, which represent aggregate activity from large numbers of neurons. 1149 All studies were approved by the hospital's Institutional Review Board and were 1150 carried out with the subjects' informed consent. Images of partial or whole objects 1151 were presented for 150 ms, followed by a gray screen for 650 ms. Subjects 1152 performed a five-alternative forced choice categorization task as described in 1153 **Figure 1** with the following differences: (i) the physiological experiment did not 1154 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 physiologyexperiment.

1157 Bubbles were randomly positioned in each trial. In order to compare models, 1158 behavior and physiology on an image-by-image basis, we had to set up a stimulus 1159 set based on the exact images (same bubble locations) presented to a given subject 1160 in the physiology experiment. To construct the stimulus set for the physiology-1161 based psychophysics experiment (Figure 2), we chose two electrodes according to 1162 the following criteria: (i) those two electrodes had to come from different 1163 physiology subjects (to ensure that the results were not merely based on any 1164 peculiar properties of one individual physiology subject), (ii) the electrodes had to 1165 respond both to whole objects and partially visible objects (to ensure a robust 1166 response where we could estimate latencies in single trials), and (iii) the electrodes 1167 had to show visual selectivity (to compare the responses to the preferred and non-1168 preferred stimuli). The electrode selection procedure was strictly dictated by these 1169 criteria and was performed before even beginning the psychophysics experiment. 1170 We extracted the images presented during the physiological recordings in n = 6501171 trials for psychophysical testing. For the preferred category for each electrode, only 1172 trials where the amplitude of the elicited neural response was in the top 50th 1173 percentile were included, and trials were chosen to represent a distribution of 1174 neural response latencies. After constructing this stimulus set, we performed 1175 psychophysical experiments with n = 33 new subjects (Physiology-based 1176 psychophysics experiment) to evaluate the effect of backward masking for the exact 1177 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).

1185

1186 Behavioral and neural data analysis

- 1187 *Masking Index.* To quantify the effect of backward masking, we defined the masking
- 1188 index as 100%-pAUC, where pAUC is the percent area under the curve when
- 1189 plotting performance as a function of SOA (e.g. **Figure 2E**). To evaluate the
- 1190 variability in the masking index, we used a half-split reliability measure by
- 1191 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
- 1193 **Figure 2F** constitute half-split reliability values.
- 1194

1195 *Correlation between masking index and neural latency.* To determine the correlation

between masking index and neural response latency, we combined data from thetwo recording sites by first standardizing the latency measurements (z-score,

Figure 2F). We then used a linear regression on neural response latency with

- 1199 masking index, percent visibility, and recording site as predictor factors to avoid any
- 1200 correlations dictated by task difficulty or differences between recording sites.
- 1201 We used only trials from the preferred category for each recording site and reported
- 1202 the correlation and statistical significance in **Figure 2F**. There was no significant
- 1203 correlation between the masking index and neural latency when considering trials
- 1204 from the non-preferred category.
- 1205

1206 *Correlation between model distance and neural response latency.* As described below. 1207 we simulated the activity of units in several computational models in response to 1208 the same images used in the psychophysics and physiology experiments. To 1209 correlate the model responses with neural response latency, we computed the 1210 Euclidean distance between the model representation of partial and whole objects. 1211 We computed the distance between each partial object in the physiology-based 1212 psychophysics experiment stimulus set and the centroid of the whole images from 1213 the same category (distance-to-category). We then assessed significance by using a 1214 linear regression on the model distance versus neural response latency while 1215 controlling for masking index, percent visibility, and recording site as factors. 1216

1217 Feed-forward Models

1218 We considered the ability to recognize partially visible images by state-of-1219 the-art feed-forward computational models of vision (Figure 3A, Figure S3 and 1220 Figure S4). First, we evaluated whether it was possible to perform recognition 1221 purely based on pixel intensities. Next, in the main text we evaluated the 1222 performance of the AlexNet model (31). AlexNet is an eight-layer deep convolutional 1223 neural network consisting of convolutional, max-pooling and fully-connected layers 1224 with a large number of weights trained in a supervised fashion for object 1225 recognition on ImageNet, a large collection of labeled images from the web (31, 32). 1226 We used a version of AlexNet trained using *caffe* (33), a deep learning library. Two 1227 layers within the AlexNet were tested: pool5 and fc7. Pool5 is the last convolutional 1228 (retinotopic) layer in the architecture. fc7 is the last layer before the classification 1229 step and is fully connected, that is, every unit in fc7 is connected to every unit in the 1230 previous layer. The number of features used to represent each object was 1231 256x256=65536 for pixels, 9216 for pool5 and 4096 for fc7.

1232 We also considered many other similar feed-forward models: VGG16 block5, 1233 fc1 and fc2 (25088, 4096 and 4096 features respectively) (34), VGG19 fc1 and fc2 1234 (4096 features each) (34), layers 40 to 49 of ResNet50 (200704 to 2048 features) 1235 (35), and InceptionV3 mixed 10 layer (131072 features) (36). In all of these cases, 1236 we used models pre-trained for the ImageNet 2012 data set and randomly 1237 downsampled the number of features to 4096 as in AlexNet. Results for all of these 1238 models are shown in Figure S4; more layers and models can be found in the 1239 accompanying web site:

1240 <u>http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html</u>

1241 Classification performance for each model was evaluated on a stimulus set 1242 consisting of 13,000 images of partial objects (generated from 325 objects from 5 1243 categories). These were the same partial objects used to collect human performance 1244 in the main experiment (**Figure 1**). We used a support vector machine (SVM) with a 1245 linear kernel to perform classification on the features computed by each model. We 1246 used 5-fold cross-validation across the 325 objects. Each split contained 260 objects 1247 for training, and 65 objects split for validation and testing, such that each object was 1248 used exactly in one validation and testing split, and such that there was an equal 1249 number of objects from each category in each split. Decision boundaries were fit on 1250 the training set using the SVM with the C parameter determined through the 1251 validation set by considering the following possible C values: 10⁻⁴, 10⁻³, ..., 10³, 10⁴. 1252 The SVM boundaries were fit using images of whole objects and tested on images of 1253 partial objects. Final performance numbers for partial objects were calculated on 1254 the full data set of 13,000 images -- that is, for each split, classification performance 1255 was evaluated on the partial objects corresponding to the objects in the test set, 1256 such that, over all splits, each partial object was evaluated exactly once.

1257 As indicated above, all the results shown on Figure 3A, Figure S3 and 1258 **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-1259 1260 tuning the AlexNet network. We fine-tuned AlexNet using the set of whole objects in 1261 our data set and then re-examined the model's performance under the low visibility 1262 conditions in **Figure S5**. We fine-tuned AlexNet by replacing the original 1000-way 1263 fully-connected classifier layer (fc8) trained on ImageNet with a 5-way fully-1264 connected layer (fc8') over the categories in our dataset and performing back-1265 propagation over the entire network. We again performed cross validation over 1266 objects, choosing final weights by monitoring validation accuracy. To be consistent 1267 with previous analysis, after fine-tuning the representation, we used an SVM 1268 classifier on the resulting fc7 activations.

1269 To graphically display the representation of the images based on all 4096 1270 units in the fc7 layer of the model in a 2D plot (Figure 4C), we used stochastic 1271 neighborhood embedding (t-SNE) (37). We note that this was done exclusively for 1272 display purposes and all the analyses, including distances, classification and 1273 correlations, are based on the model representation with all the units in the 1274 corresponding layer as described above. For each model and each image, we 1275 computed the Euclidian distance between the model's representation and the mean 1276 point across all whole objects within the corresponding category. This distance-to-1277 category corresponds to the y-axis in Figure 3B-C.

1278

1279 Recurrent Neural Network Models

1280 A recurrent neural network (RNN) was constructed by adding all-to-all 1281 recurrent connections to different layers of the bottom-up convolutional networks 1282 described in the previous section (for example, to the fc7 layer of AlexNet in Figure 1283 **4A**). We first describe here the model for AlexNet; a similar procedure was followed 1284 for the other computational models. An RNN consists of a state vector that is 1285 updated according to the input at the current time step and its value at the previous 1286 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 $\mathbf{h}_{t} = f(\mathbf{W}_{h}\mathbf{h}_{t-1},\mathbf{x}_{t})$ 1287 1288 where f introduces a non-linearity as defined below. In our model, \mathbf{h}_t represents the 1289 fc7 feature vector at time t and \mathbf{x}_t represents the feature vector for the previous 1290 layer, fc6, multiplied by the transition weight matrix $W_{6,7}$. For simplicity, the first 1291 six layers of AlexNet were kept fixed to their original feed-forward versions.

1292 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 1293 1294 original description by Hopfield (39). Since this implementation is based on binary 1295 unit activity, we first converted the scalar activities in \mathbf{x} to $\{-1,+1\}$ by mapping those 1296 values greater than 0 to +1 and all other values to -1. Depending on the specific layer 1297 and model, this binarization step in some cases led to either an increase or a 1298 decrease in performance (even before applying the attractor network dynamics); all 1299 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 1300

1301 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

1302 whole objects to be stored (in our case n_p =325) and x_i^p represents the activity of 1303 unit *i* in response to pattern *p*. This model does not have any free parameters that 1304 depend on the partial objects and the weights are uniquely specified by the activity 1305 of the feed-forward network in response to the whole objects. After specifying \mathbf{W}_{h} , 1306 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 1307 t>0 where *satlins* represents the saturating linear transfer function,

1308 satlins(z) = max(min(1,z),-1) and **b** introduces a constant bias term. The activity in 1309 RNN_h was simulated until convergence, defined as the first time point where there 1310 was no change in the sign of any of the features between two consecutive time 1311 points.

1312 To evaluate whether the increase in performance obtained in the RNN_b was 1313 specific to the AlexNet architecture, we also implemented recurrent connections 1314 added onto other networks. Figure S7 shows a comparison between performance of 1315 the VGG16 network layer fc1 (34) and a VGG16 fc1 model endowed with additional 1316 recurrent connections in the same format as used with AlexNet. We used the time 1317 steps of the Hopfield network that yielded maximal performance. The 1318 VGG16+Hopfield model also showed performance improvement with respect to the 1319 purely bottom-up VGG16 counterpart. Several additional models were tested for 1320 other layers of AlexNet, VGG16, VGG19, ResNet and InceptionV3, showing a 1321 distribution with different degrees of consistent improvement upon addition of the 1322 recurrent connectivity (shown in the accompanying web material at 1323 http://klab.tch.harvard.edu/resources/Tangetal RecurrentComputations.html). 1324 We ran an additional simulation with the RNN models to evaluate the effects 1325 of backward masking (Figure 4F). For this purpose, we simulated the response of 1326 the feed-forward AlexNet model to the same masks used for the psychophysical 1327 experiments to determine the fc6 features for each mask image. Next, we used this 1328 mask as the fixed input \mathbf{x}_t into the recurrent network, at different time points after 1329

1330

1331 2. Supplementary Discussion

the initial image input.

1332

Partially visible versus occluded objects 1333

1334 In most of the experiments, we rendered objects partially visible by 1335 presenting them through "bubbles" (Fig. 1C) in an attempt to distill the basic 1336 mechanisms required for spatial integration during pattern completion. It was 1337 easier to recognize objects behind a real occluder (Fig. 1D, S1, (40)). The results 1338 presented here were qualitatively similar (**Fig. S1**) when using explicit occluders

1339 (Fig. 1D): recognition of occluded objects was also disrupted by backward masking

1340 (**Fig. 1I, S1**). As expected, performance was higher for the occlusion versus the

1341 bubbles condition.

1342

1343 **"Unfolding" recurrent neural networks into feed-forward neural networks**

1344 Before examining computational models including recurrent connections, we 1345 analyzed bottom-up architectures and showed that they were *not* robust to 1346 extrapolating from whole objects to partial objects (**Figure 4**). However, there exist 1347 infinitely many possible bottom-up models. Hence, even though we examined state-1348 of-the-art models that are quite successful in object recognition, the failure to 1349 account for the behavioral and physiological results in the bottom-up models 1350 examined here (as well as similar failures reported in other studies, e.g. (41, 42)) 1351 should be interpreted with caution. We do not imply that it is impossible for *any* 1352 bottom-up architecture to recognize partially visible objects. In fact, it is possible to 1353 unfold a recurrent network with a finite number of time steps into a bottom-up 1354 model by creating an additional layer for each additional time step. However, there 1355 are several advantages to performing those computations with a recurrent 1356 architecture including a drastic reduction in the number of units required as well as 1357 in the number of weights that need to be trained and the fact that such unfolding is 1358 applicable only when we know *a priori* the fixed number of computational steps 1359 required, in contrast with recurrent architectures that allow an arbitrary and

1360 variable number of computations.

1361

1362 **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

1369 accounted by longer integration times at all levels of the visual hierarchy. First, all 1370 images were contrast normalized to avoid any overall intensity effects. Second, 1371 neural delays for poor visibility images were not observed in early visual areas (25). 1372 Third, the correlations between the effects of backward masking and neural delays 1373 persisted even after accounting for difficulty level (Fig. 3). Fourth, none of the state-1374 of-the-art purely bottom-up computational models were able to account for human 1375 level performance (see further elaboration of this point below). These arguments 1376 rule out slower processing throughout the entire visual system due to low intensity 1377 signals in the lower visibility conditions. However, the results presented here are 1378 still compatible with the notion that the inputs to higher-level neurons in the case of 1379 partial objects could be weaker and could require further temporal integration. This 1380 possibility is consistent with the model proposed here. Because the effects of 1381 recurrent computations are delayed with respect to the bottom-up inputs, we 1382 expect that any such slow integration would have to interact with the outputs of 1383 recurrent signals.

1384

1385 Extensions to the proposed proof-of-concept architecture

1386 A potential challenge with attractor network architectures is the pervasive 1387 presence of spurious attractor states, particularly prominent when the network is 1388 near capacity. Furthermore, the simple instantiation of a recurrent architecture 1389 presented here still performed below humans, particularly under very low visibility 1390 conditions. It is conceivable that more complex architectures that take into account 1391 the known lateral connections in every layer as well as top-down connections in 1392 visual cortex might improve performance even further. Additionally, future 1393 extensions will benefit from incorporating other cues that help in pattern 1394 completion such as relative positions (front/behind), segmentation, movement, 1395 source of illumination, and stereopsis, among others.

1396

1397 Mixed training regime

1398All the computational results shown in the main text and discussed thus far1399involve training models *exclusively* with whole objects and testing performance with

1400 images of partially visible objects. Here we discuss a "mixed training" regime where 1401 the models are trained with access to partially visible objects. As emphasized in the 1402 main text, these are weaker models since they show less extrapolation (from 1403 partially visible objects to other partially visible objects as opposed to from whole 1404 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 1405 1406 testing at other rotations). Furthermore, humans do not require this type of 1407 additional training as described in the novel object experiments reported in **Figures** 1408 **S8** and **S9**. Despite these caveats, the mixed training regime is interesting to explore 1409 because it seems natural to assume that, at least in some cases, humans may be 1410 exposed to both partially visible objects and their whole counterparts while learning 1411 about objects. We emphasize that we cannot directly compare models that are 1412 trained only with whole objects and models that are trained with both whole objects 1413 and partially visible ones.

1414 We considered two different versions of RNN models that were trained to 1415 reconstruct the feature representations of the whole objects from the feature representations of the corresponding partial objects. These models were based on a 1416 1417 mixed training regime whereby both whole objects and partial objects were used 1418 during training. The state at time t>0 was computed as the activation of the 1419 weighted sum of the previous state and the input form the previous layer: $\mathbf{h}_{t} = \operatorname{Re} \operatorname{LU}(\mathbf{W}_{h}\mathbf{h}_{t-1}, \mathbf{x}_{t})$ where $\operatorname{ReLU}(z) = max(0, z)$. The loss function was the 1420 1421 mean squared Euclidean distance between the features from the partial objects and 1422 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 1423 from fc6 constant. Thus, letting \mathbf{h}_{t}^{i} be the RNN state at the last time step for a given 1424 image *i* and $_{whole} \mathbf{h}_{t0}^{i}$ be the feed-forward feature vector of the corresponding whole 1425 1426 image, the loss function has the form

1427
$$E = \frac{1}{T_I} \sum_{i=1}^{T_I} \left[\frac{1}{T_u} \sum_{j=1}^{T_u} (h_{t_{\max}}^i[j] - {}_{whole} h_{t0}^i[j])^2 \right]$$

1428 where *i* goes over all the T_{ij} units in fc7 and *i* goes over all the T_{ij} images in the 1429 training set. The RNN was trained in a cross validated fashion (5 folds) using the 1430 same cross validation scheme as with the feed-forward models and using the 1431 RMSprop algorithm for optimization. In RNN₅, the weights of the RNN were trained 1432 with 260 objects for each fold. All of the partial objects from the psychophysics 1433 experiment for the given 260 objects, as well as one copy of the original 260 images, 1434 were used to train the RNN for the corresponding split. In the case where the input 1435 to the RNN was the original image itself, the network did not change its 1436 representation over the recurrent iterations. Given the high number of weights to 1437 be learned by the RNN as compared to the number of training examples, the RNNs 1438 overfit fairly quickly. Therefore, early stopping (10 epochs) was implemented as 1439 determined from the validation set, i.e., we used the weights at the time step where 1440 the validation error was minimal.

1441 To evaluate the extent of extrapolation across categories, we considered an 1442 additional version, RNN₁. In RNN₁, the recurring weights were trained using objects 1443 from only one category and the model was tested using objects from the remaining 1444 4 categories. In all RNN versions, once \mathbf{W}_{h} was fixed, classification performance was 1445 assessed using a linear SVM, as with the feed-forward models. Specifically, the SVM 1446 boundaries were trained using the responses from the feed-forward model to the 1447 whole objects and performance was evaluated using the representation at different 1448 time steps of recurrent computation.

1449 The RNN₅ model had 40962 recurrent weights trained on a subset of the 1450 objects from all five categories. The RNN₅ model matched or surpassed human performance (Figure S11). Considering all levels of visibility, the RNN₅ model 1451 1452 performed slightly above human levels ($p=3x10^{-4}$, Chi-squared test). While the RNN₅ 1453 model can extrapolate across objects and categorize images of partial objects that it 1454 has not seen before, it does so by exploiting features that are similar for different 1455 objects within the 5 categories in the experiment. RNN₁, a model where the 1456 recurrent weights were trained using solely objects from one of the categories and 1457 performance was evaluated using objects from the remaining 4 categories, did not 1458 perform any better than the purely feed-forward architecture (p=0.05, Chi-squared

1459 test). Upon inspection of the fc7 representation, we observed that several of the 1460 features were sparsely represented across categories. Therefore, the recurrent 1461 weights in RNN₁ only modified a fraction of all the possible features, missing many 1462 important features to distinguish the other objects. Thus, the improvement in 1463 RNN_5 is built upon a sufficiently rich dictionary of features that are shared among 1464 objects within a category. These results show that recurrent neural networks 1465 trained with subsets of the partially visible objects can achieve human level 1466 performance, extrapolating across objects, as long as they are trained with a 1467 sufficiently rich set of features.

1468 We also evaluated the possibility of training the bottom-up model (AlexNet) 1469 using the mixed training regime and the same loss function as with RNN₅ and RNN₁, 1470 i.e. the Euclidean distance between features of whole and occluded images. Using 1471 the fc7 representation of the AlexNet model trained with partially visible objects 1472 also led to a model that either matched or surpassed human level performance at 1473 most visibility levels (Figure S11). The bottom-up model in the mixed training 1474 regime showed slightly worse performance than humans at very high visibility 1475 levels, including whole objects, perhaps because of the extensive fine-tuning with 1476 partially visible objects (note performance above humans at extremely low visibility 1477 levels). Within the mixed-training regimes, the RNN₅ model slightly outperformed 1478 the bottom-up model (Figure S11).

1479 A fundamental distinction between the models presented in the text, 1480 particularly RNN_h, and the models introduced here, is that the mixed training 1481 models require training with partial objects from the same categories in which they 1482 will be evaluated. Although the specific photographs of objects used in the 1483 psychophysics experiments presented here were new to the subjects, humans have 1484 extensive experience in recognizing similar objects from partial information. It 1485 should also be noted that there is a small number of partially visible images in 1486 ImageNet, albeit not with such low visibility levels as the ones explored here, and all 1487 the models considered here were pre-trained using ImageNet. Yet, the results 1488 shown in **Figures S8-S9** demonstrate that humans can recognize objects shown 1489 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 thecorresponding whole objects.

1492

1493 **Temporal scale for recurrent computations**

1494 The models presented here, and several discussions in the literature, 1495 schematically and conceptually separate feed-forward computations from within-1496 layer recurrent computations. Physiological signals arising within \sim 150 ms after 1497 stimulus onset have been interpreted to reflect largely feed-forward processing (1, 1498 3, 5, 8, 10, 11, 43), whereas signals arising in the following 50 to 100 ms may reflect 1499 additional recurrent computations (27, 44, 45). This distinction is clearly an 1500 oversimplification: the dynamics of recurrent computations can very well take place quite rapidly and well within \sim 150 ms of stimulus onset (46). Rather than a 1501 1502 schematic initial feed-forward path followed by recurrent signals within the last 1503 layer in discrete time steps as implemented in RNN_h, cortical computations are 1504 based on continuous time and continuous interactions between feed-forward and 1505 within-layer signals (in addition to top-down signals). A biologically plausible 1506 implementation of a multi-layered spiking network including both feed-forward and 1507 recurrent connectivity was presented in ref. (46), where the authors estimated that 1508 recurrent signaling can take place within ~ 15 ms of computation per layer. Those 1509 time scales are consistent with the results shown here. Recurrent signals offer 1510 dynamic flexibility in terms of the amount of computational processing. Under noisy 1511 conditions (an injected noise term added to modify the input to each layer in (46), 1512 more occlusion in our case, and generally any internal or external source of noise), 1513 the system can dynamically use more computations to solve the visual recognition 1514 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

1521	delay of 50 to 100 ms (Fig. 1-2, S12) (25, 48), which are consistent with a
1522	continuous time implementation of recurrent processing (46).
1523	
1524	3. Supplementary Figures Legends
1525	
1526	Figure S1: Robust performance with occluded stimuli
1527	We measured categorization performance with masking (solid lines) or without
1528	masking (dashed lines) for (A) partial and (B) occluded stimuli on a set of 16
1529	exemplars belonging to 4 categories (chance = 25%, dashed lines). There was no
1530	overlap between the 14 subjects that participated in (A) and the 15 subjects that
1531	participated in (B). The effect of backward masking was consistent across both
1532	types of stimuli. The black lines indicate whole objects and the gray lines indicate
1533	the partial and occluded objects. Error bars denote SEM.
1534	
1535	Figure S2: Example half-split reliability of psychophysics data
1536	Figure 2E in the main text reports the masking index, a measure of how much
1537	recognition of each individual image is affected by backward masking. This measure
1538	is computed by averaging performance across subjects. In order to evaluate the
1539	variability in this metric, we randomly split the data into two halves and computed
1540	the masking index for each image for each half of the data. This figure shows one
1541	such split and how well one split correlates with the other split. Figure 2F shows
1542	error bars defined by computing standard deviations of the masking indices from
1543	100 such random splits.
1544	
1545	Figure S3: Bottom-up models can recognize minimally occluded images
1546	A. Extension to Figure 3A showing that bottom-up models successfully recognize
1547	objects when more information is available (Figure 3A showed visibility values up
1548	to 35% whereas this figure extends visibility up to 100%). The format and
1549	conventions are the same as those in Figure 3A . The black dotted line shows
1550	interpolated human performance between the psychophysics experimental values

1551 measured at 35% and 100% visibility levels.

- 1552 (B) Stochastic neighborhood embedding dimensionality reduction (t-SNE, Methods)
- 1553 to visualize the fc7 representation in the AlexNet model for whole objects (open
- 1554 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
- 1556 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).
- 1559
- Figure S4: All of the purely feed-forward models tested were impaired under
 low visibility conditions
- 1562 The human, AlexNet-pool5 and AlexNet-fc curves are the same ones shown in
- 1563 **Figure 3A** and are reproduced here for comparison purposes. This figure shows
- 1564 performance for several other models: VGG16-fc2, VGG19-fc2, ResNet50-flatten,
- 1565 inceptionV3-mixed10, VGG16-block5 (see text for references). In all cases, these
- 1566 models were pre-trained to optimize performance under ImageNet 2012 and there
- 1567 was no additional training (see also **Figure S5**). An expanded version of this figure
- 1568 with many other layers and models can be found on our web site:
- 1569 <u>http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html</u>
- 1570

1571 **Figure S5: Fine-tuning did not improve performance under heavy occlusion**

- 1572 The human and fc7 curves are the same ones shown in **Figure 3A** and are
- 1573 reproduced here for comparison purposes. The pre-trained AlexNet network used
- 1574 in the text was fine tuned using back-propagation with the set of *whole* images from
- 1575 the psychophysics experiment (in contrast with the pre-trained Alexnet network
- 1576 which was trained using the Imagenet 2012 data set). The fine-tuning involved all
- 1577 layers (**Methods**).
- 1578

1579 Figure S6: Correlation between RNN_h model and human performance for 1580 individual objects as a function of time

- 1581 At each time step in the recurrent neural network model (RNN_h), the scatter plots
- 1582 show the relationship between the model's performance on individual partial

- 1583 exemplar objects and human performance. Each dot is an individual exemplar
- 1584 object. In **Figure 4E** we report the average correlation coefficient across all
- 1585 categories.
- 1586

1587 **Figure S7: Adding recurrent connectivity to VGG16 also improved**

1588 **performance**

- 1589 This Figure parallels the results shown in **Figure 4B** for AlexNet, here using the
- 1590 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)
- 1592 corresponds to the original model without any recurrent connections. The
- 1593 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
- 1595 AlexNet fc7 activations. An expanded version of this figure with similar results for
- 1596 several other layers and models can be found on our web site:
- 1597 <u>http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html</u>
- 1598

1599 **Figure S8: Robust recognition of** *novel* **objects under low visibility conditions**

- 1600 **A**. Single exemplar from each of the 5 novel object categories (**Methods**).
- 1601 (B-C) Behavioral performance for the unmasked (B) and masked (C) trials. The
- 1602 experiment was identical to the one in **Figure 1** and the format of this figure follows
- 1603 that in **Figure 1F-G**. The colors denote different SOAs. Error bars=SEM. Dashed line
- 1604 = chance level (20%). Bin size=2.5%. Note the discontinuity in the x-axis to report
- 1605 performance for whole objects (100% visibility). (**D**) Average recognition
- 1606 performance as a function of the stimulus onset asynchrony (SOA) for partial objects
- 1607 (same data and conventions as **B-C**, excluding 100% visibility). Error bars=SEM.
- 1608 Performance was significantly degraded by masking (solid) compared to the
- 1609 unmasked trials (dotted) (*p*<0.0001, Chi-squared test, d.f.=4).
- 1610

1611 **Figure S9: The performance of feed-forward and recurrent computational**

1612 models for *novel* objects was similar to those for known object categories

- A. Performance of feed-forward computational models (format as in Figure 3A) fornovel objects.
- 1615 B. Performance of the recurrent neural network RNN_h (format as in Figure 4B) for1616 novel objects.
- 1617 **C**. Temporal evolution of the feature representation for RNN_h (format as in **Figure**
- 1618 **4C**). The colors and greek letters denote the five object categories (see examples in
- 1619 **Figure S8A**).
- 1620 **D**. Performance of RNN_h as a functon of recurrent time for novel objects (format as
- 1621 in **Figure 4D**).
- 1622

1623 Figure S10: Side-by-side comparison of neurophysiological signals,

1624 psychophysics and computational model

- 1625 **A.** Adaptation of Figure 6C from Tang et al 2014. This figure shows the dynamics of
- 1626 decoding object information for whole objects and (black) and partial objects (gray)
- 1627 from neurophysiological recordings as a function of time post stimulus onset (see
- 1628 Tang et al 2014 for details.
- 1629 **B**. Reproduction of **Figure 1H** (behavior).
- 1630 **C**. Reproduction of **Figure 4F** (RNN_h model).
- 1631 Above each subplot, the experiment schematic highlights that part **A** involves no
- 1632 masking and fixed SOA = 150 ms whereas parts **B** and **C** involve masking and
- 1633 variable SOAs. The inset in part ${\bm C}$ directly overlays the results of the RNN_h model in
- 1634 part **C** onto the results of the psychophysics experiment in part **B**. In order to create
- 1635 this plot, we mapped 0 time steps to 25ms, 256 time steps to 150 ms and linearly
- 1636 interpolated the time steps in between.
- 1637

1638 **Figure S11: Mixed training regimes**.

- 1639 **A.** This figure follows the format of **Fig3A**, **4B** and **S3**, **S4**, **S5**, **S7**, **S9A-B**. The black
- 1640 line shows human performance and is copied from **Fig. 3A**. The green and blue lines
- 1641 show the recurrent model (RNN₅) and bottom-up model (AlexNet fc7), respectively,
- 1642 trained in a mixed regime that included the occluded objects with visibility levels
- 1643 within the gray rectangle (the same ones used to evaluate human psychophysics

- 1644 performance). In the RNN5 model, there were ~16 million weights trained (all-to-all
- 1645 in the fc7 layer) whereas in the Alexnet fc7 model, there were \sim 60 million weights
- 1646 trained (all the weights across layers in the Alexnet model). Cross-validated test
- 1647 performance is shown here as well as in the other figures throughout the
- 1648 manuscript. As noted in the text, we emphasize that this figure involves a different
- 1649 training regime from the ones in the previous figures and therefore one cannot
- 1650 directly compare performance with the previous figures.
- 1651 **B**. This figure follows the format of **Fig. 4E**. The green and blue bars show the
- 1652 correlation between human and model for the recurrent model and bottom-up
- 1653 model, respectively, both trained using occluded objects. The gray rectangle shows
- 1654 human-human correlation, see **Fig. 4E** for details..
- 1655

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

- 1658 Expanding on **Figure 4E**, this figure shows the correlation coefficient between
- 1659 human recognition performance in the masked condition (**Figure 1B**) at a given
- 1660 SOA (y-axis) and RNN_h model performance at a given time step (x-axis). The top row
- 1661 shows the unmasked condition (**Figure 1A**). In this figure, there is no mask for the
- 1662 model (see **Figure 4F** for model performance with a mask). The computation of the
- 1663 correlation coefficient follows the same procedure illustrated in **Figure S6** and **4E**.
- 1664 The color scale for the correlation coefficient is shown on the right. As an upper
- 1665 bound and as shown in **Figure 4E**, the correlation coefficient between different
- 1666 human subjects was 0.41 for the unmasked condition. The yellow boxes highlight
- 1667 the highest correlation for a given SOA value.
- 1668
- 1669 **4.** Author contributions
- 1670 Conceptualization: HT, BL, MS, DC, GK
- 1671 Physiology experiment design: HT, GK
- 1672 Physiological data collection and analyses: HT
- 1673 Psychophysics experiment design: HT, BL, MS, CM, GK
- 1674 Psychophysics data collection: HT, BL, MS, AP, JO, WH, CM

1675	Computational models: HT, BL, MS, DC, CM, GK					
1676	Resources: DC, GK					
1677	Manus	Manuscript writing: HT, BL, MS, GK				
1678						
1679	5.	Data availability				
1680	All relevant data and code (including image databases, behavioral measurements,					
1681	physiological measurements and computational algorithms) are publicly available					
1682	through the lab's website and through the lab's GitHub page:					
1683	http://klab.tch.harvard.edu/resources/Tangetal_RecurrentComputations.html					
1684						
1685	6.	References				
1686 1687 1688 1689	1.	Kirchner H & Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. <i>Vision research</i> 46(11):1762-1776				
1690 1691	2.	Potter M & Levy E (1969) Recognition memory for a rapid sequence of pictures. <i>Journal of experimental psychology</i> 81(1):10-15.				
1692 1693	3.	Keysers C, Xiao DK, Foldiak P, & Perret DI (2001) The speed of sight. <i>Journal</i> of Cognitive Neuroscience 13(1):90-101.				
1694 1695	4.	Hung CP, Kreiman G, Poggio T, & DiCarlo JJ (2005) Fast Read-out of Object Identity from Macaque Inferior Temporal Cortex, <i>Science</i> 310:863-866				
1696 1697 1698	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. <i>Neuron</i> 62(2):281-290.				
1699 1700 1701	6.	Tovee M & Rolls E (1995) Information encoding in short firing rate epochs by single neurosn in the primate temporal visual cortex. <i>Visual Cognition</i> 2(1):35-58.				
1702 1703 1704	7.	Pinto N, Doukhan D, DiCarlo JJ, & Cox DD (2009) A high-throughput screening approach to discovering good forms of biologically inspired visual representation. <i>PLoS Comput Biol</i> 5(11):e1000579.				
1705 1706	8.	Riesenhuber M & Poggio T (1999) Hierarchical models of object recognition in cortex. <i>Nature Neuroscience</i> 2(11):1019-1025.				
1707 1708	9.	Wallis G & Rolls ET (1997) Invariant face and object recognition in the visual system. <i>PROGRESS IN NEUROBIOLOGY</i> 51(2):167-194.				
1709 1710 1711	10.	Yamins DL, <i>et al.</i> (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. <i>Proceedings of the National Academy of Sciences of the United States of America</i> 111(23):8619-8624.				
1712 1713	11.	Serre T, <i>et al.</i> (2007) A quantitative theory of immediate visual recognition. <i>Progress In Brain Research</i> 165C:33-56				
1714 1715	12.	Breitmeyer B & Ogmen H (2006) Visual Masking: Time Slices through Conscious and Unconscious Vision (Oxford University Press, New York).				

1716	13.	Bridgeman B (1980) Temporal response characteristics of cells in monkey
1717		striate cortex measured with metacontrast masking and brightness
1718		discrimination. <i>Brain Res</i> 196(2):347-364.
1719	14.	Macknik SL & Livingstone MS (1998) Neuronal correlates of visibility and
1720		invisibility in the primate visual system. <i>Nature neuroscience</i> 1(2):144-149.
1721	15.	Lamme VA. Zipser K. & Spekreijse H (2002) Masking interrupts figure-
1722		ground signals in V1. <i>I Coan Neurosci</i> 14(7):1044-1053.
1723	16.	Kovacs G. Vogels R. & Orban GA (1995) Cortical correlate of pattern
1724	201	hackward masking. Proceedings of the National Academy of Sciences
1725		92(12)·5587-5591
1726	17.	Rolls ET. Toyee ML & Panzeri S (1999) The neurophysiology of backward
1727	17.	visual masking: information analysis <i>Journal of Cognitive Neuroscience</i>
1728		11(3)·300-311
1729	18	Keysers C & Perrett DI (2002) Visual masking and RSVP reveal neural
1730	10.	competition Trends Coan Sci 6(3):120-125
1731	19	Funs IT & Di Lollo V (2000) What's new in visual masking? Trends Coan Sci
1732	17.	4(9)·345-352
1733	20	Thompson KG & Schall ID (1999) The detection of visual signals by macaque
1734	20.	frontal eve field during masking <i>Nature neuroscience</i> 2(3):283-288
1735	21	Kellman PI, Guttman S & Wickens T (2001) Geometric and neural models of
1736	2 1.	object perception. From framents to objects: Seamentation and arouning in
1737		vision eds Shinley TF & Kellman PI (Elsevier Science Publishers, Oxford, IIK)
1738	22	Murray RF Sekuler AB & Bennett PI (2001) Time course of amodal
1739	<i></i> .	completion revealed by a shape discrimination task <i>Psychon Bull Rev</i>
1740		8(4)·713-720
1741	23	Kosai Y El-Shamavleh Y Evall AM & Pasunathy A (2014) The role of visual
1742	20.	area V4 in the discrimination of partially occluded shapes <i>Journal of</i>
1743		Neuroscience 34(25):8570-8584
1744	24.	Nakavama K. He Z. & Shimoio S (1995) Visual surface representation: a
1745		critical link between lower-level and higher-level vision. Visual cognition, eds
1746		Kosslyn S & Osherson D (The MIT press Cambridge) Vol 2
1747	25	Tang H <i>et al</i> (2014) Spatiotemporal dynamics underlying object completion
1748	20.	in human ventral visual cortex <i>Neuron</i> 83:736-748
1749	26	Johnson IS & Olshausen BA (2005) The recognition of partially visible natural
1750	20.	objects in the presence and absence of their occluders. <i>Vision research</i> 45(25-
1751		26):3262-3276.
1752	27.	Lee TS (2003) Computations in the early visual cortex. <i>J Physiol Paris</i> 97(2-
1753	_/.	3)·121-139
1754	28	Gosselin F & Schyns PG (2001) Bubbles: a technique to reveal the use of
1755	20.	information in recognition tasks <i>Vision research</i> 41(17):2261-2271
1756	29	Williams P (1998) Representational organization of multiple evenplars of
1757	<i>L)</i> .	object categories
1758	30	Willenbockel V <i>et al.</i> (2010) Controlling low-level image properties: the
1759	50.	SHINF toolbox <i>Rehav Res Methods</i> 42(3).671-684
1760	31	Krizhevsky A Sutskever I & Hinton G (2012) ImageNet Classification with
1761	51.	Deen Convolutional Neural Networks in NIPS (Montreal)
1,01		200p convolutional recurat recoveries in rul 5 (nontreal).

1762	32.	Russakovsky O, et al. (2014) ImageNet Large Scale Visual Recognition
1763		Challenge. in <i>CVPR</i> (arXiv:1409.0575, 2014).
1764	33.	Yangqing J, et al. (2014) Caffe: Convolutional Architecture for Fast Feature
1765		Embedding. <i>arXiv</i> :1408.5093.
1766	34.	Simonyan K & Zisserman A (2014) Very deep convolutional networks for
1767		large-scale image recognition. <i>arXiv</i> 1409.1556.
1768	35.	He K, Zhang X, Ren S, & Sun J (2015) Deep residual learning for image
1769		recognition. <i>arXiv</i> 1512.03385.
1770	36.	Szegedy C, Vanhoucke V, Ioffe S, Shlens J, & Wojna Z (2015) Rethinking the
1771		inception architecture for computer vision. <i>arXiv</i> 1512.005673v3.
1772	37.	van der Maaten L & Hinton G (2008) Visualizing High-Dimensional Data
1773		Using t-SNE. J. Machine Learning Res. 9:2579-2605.
1774	38.	Hopfield JJ (1982) Neural networks and physical systems with emergent
1775		collective computational abilities. <i>PNAS</i> 79:2554-2558.
1776	39.	Li J, Michel A, & Porod W (1989) Analysis and synthesis of a class of neural
1777		networks: linear systems operating on a closed hypercube. <i>IEEE Transactions</i>
1778		on Circuits and Systems 36(11):1405-1422.
1779	40.	Bregman AL (1981) Asking the "what for" question in auditory perception
1780		(Erlbaum, Hillsdale, NJ) p 19.
1781	41.	Pepik B, Benenson R, Ritschel T, & Schiele B (2015) What is holding back
1782		convnets for detection? 1508.
1783	42.	Spoerer CJ, McClure P, & Kriegeskorte N (2017) Recurrent Convolutional
1784		Neural Networks: A Better Model of Biological Object Recognition. Frontiers
1785		in psychology 8:1551.
1786	43.	DiCarlo JJ & Cox DD (2007) Untangling invariant object recognition. <i>Trends</i>
1787		Cogn Sci 11(8):333-341.
1788	44.	Lamme VA & Roelfsema PR (2000) The distinct modes of vision offered by
1789		feedforward and recurrent processing. <i>Trends Neurosci</i> 23(11):571-579.
1790	45.	Gilbert CD & Li W (2013) Top-down influences on visual processing. Nat Rev
1791		Neurosci 14(5):350-363.
1792	46.	Panzeri S, Rolls ET, Battaglia F, & Lavis R (2001) Speed of feedforward and
1793		recurrent processing in multilayer networks of integrate-and-fire neurons.
1794		Network 12(4):423-440.
1795	47.	Koch C (1999) <i>Biophysics of Computation</i> (Oxford University Press, New
1796		York).
1797	48.	Fyall AM, El-Shamayleh Y, Choi H, Shea-Brown E, & Pasupathy A (2017)
1798		Dynamic representation of partially occluded objects in primate prefrontal
1799		and visual cortex. <i>eLife</i> 6.
1800		



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

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.



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





The human and fc7 curves are the same ones shown in **Figure 3A** and are reproduced here for comparison purposes. The pretrained 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 **Fig. 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

Supplementary Figure 8 β ε α В С 100 100 ф Ф 80 80 Performance Performance 60 60 150 ms -00 ms 75 ms _ 40 40 50 ms 20 20 n=23 subjects n=23 subjects 0 0 30 20 10 20 40 10 30 40 100 0 0 Percent visible Percent visible D



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).

<u>.</u>

ᡏ

100







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. 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.



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 (RNN₅) 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 (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.



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.