Recurrent computations for visual pattern completion

Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, Gabriel Kreiman

Angaben zur Veröffentlichung / Publication details:

Tang, Hanlin, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. 2018. "Recurrent computations for visual pattern completion." *Proceedings of the National Academy of Sciences* 115 (35): 8835–40. https://doi.org/10.1073/pnas.1719397115.

Recurrent computations for visual pattern completion

Hanlin Tang[a,b,1], Martin Schrimpf[b,c,d,e,1], William Lotter[a,b,f,1], Charlotte Moerman[b], Ana Paredes[b], Josue Ortega Caro[b], Walter Hardesty[b], David Cox[f], and Gabriel Kreiman[b,2]

[a]Program in Biophysics, Harvard University, Boston, MA 02115; [b]Children's Hospital, Harvard Medical School, Boston, MA 02115; [c]Program in Software Engineering, Institut für Informatik, Universität Augsburg, 86159 Augsburg, Germany; [d]Program in Software Engineering, Institut für Informatik, Ludwig-Maximilians-Universität München, 80538 München, Germany; [e]Program in Software Engineering, Fakultät für Informatik, Technische Universität München, 85748 Garching, Germany; and [f]Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved July 20, 2018 (received for review November 10, 2017)

Making inferences from partial information constitutes a critical aspect of cognition. During visual perception, pattern completion enables recognition of poorly visible or occluded objects. We combined psychophysics, physiology, and computational models to test the hypothesis that pattern completion is implemented by recurrent computations and present three pieces of evidence that are consistent with this hypothesis. First, subjects robustly recognized objects even when they were rendered <15% visible, but recognition was largely impaired when processing was interrupted by backward masking. Second, invasive physiological responses along the human ventral cortex exhibited visually selective responses to partially visible objects that were delayed compared with whole objects, suggesting the need for additional computations. These physiological delays were correlated with the effects of backward masking. Third, state-of-the-art feed-forward computational architectures were not robust to partial visibility. However, recognition performance was recovered when the model was augmented with attractor-based recurrent connectivity. The recurrent model was able to predict which images of heavily occluded objects were easier or harder for humans to recognize, could capture the effect of introducing a backward mask on recognition behavior, and was consistent with the physiological delays along the human ventral visual stream. These results provide a strong argument of plausibility for the role of recurrent computations in making visual inferences from partial information.

visual object recognition | computational neuroscience | pattern completion | artificial intelligence | machine learning

Humans and other animals have a remarkable ability to make inferences from partial data across all cognitive domains. This inference capacity is ubiquitously illustrated during pattern completion to recognize objects that are partially visible due to noise, limited viewing angles, poor illumination, or occlusion. There has been significant progress in describing the neural machinery along the ventral visual stream responsible for recognizing whole objects (1–5). Computational models instantiating biologically plausible algorithms for pattern recognition of whole objects typically consist of a sequence of filtering and nonlinear pooling operations. The concatenation of these operations transforms pixel inputs into a feature representation amenable for linear decoding of object labels. Such feed-forward algorithms perform well in large-scale computer vision experiments for pattern recognition (6–9) and provide a first-order approximation to describe the activity of cortical neurons (e.g., ref. 10).

Spatial and temporal integration play an important role in pattern completion mechanisms (11–14). When an object is occluded, there are infinitely many possible contours that could join the object's parts together. However, the brain typically manages to integrate those parts to correctly recognize the occluded object. Multiple studies have highlighted the importance of temporal integration by demonstrating that recognizing partially visible objects takes more time than recognizing fully visible ones at the behavioral (11, 15) and physiological (12, 13) levels. We conjectured that within-layer and top-down recurrent computations are involved in implementing the spatial and temporal integrative mechanisms underlying pattern completion. Recurrent connections can link signals over space within a layer and provide specific top-down modulation from neurons with larger receptive fields (16, 17). Additionally, recurrent signals temporally lag behind their feed-forward counterparts, and therefore provide an ideal way to incorporate temporal integration mechanisms.

To examine plausible mechanisms involved in pattern completion, we combined psychophysics, neurophysiology (13), and computational modeling to evaluate recognition of partially visible objects. We show that humans robustly recognize objects even from a limited amount of information, but performance rapidly deteriorates when computations are interrupted by a noise mask. On an image-by-image basis, the behavioral effect of such backward masking correlates with an increase in latency in neurophysiological intracranial field potentials along the ventral visual stream. A family of modern feed-forward convolutional hierarchical models is not robust to occlusion. We extend previous notions of attractor dynamics by adding recurrence to such bottom-up models and providing a proof-of-concept model that captures the essence of human pattern completion behavior.

Results

Robust Recognition of Partially Visible Objects. Subjects performed a recognition task (Fig. 1 *A* and *B*) involving categorization of objects that were either partially visible (Partial in Fig. 1*C*, *Right*) or fully visible (Whole in Fig. 1*C*, *Left*). Images were followed by either a gray screen ("unmasked" in Fig. 1*A*) or a spatially overlapping noise pattern ("masked" in Fig. 1*B*). The image presentation time, referred to as stimulus onset asynchrony (SOA),

Significance

The ability to complete patterns and interpret partial information is a central property of intelligence. Deep convolutional network architectures have proved successful in labeling whole objects in images and capturing the initial 150 ms of processing along the ventral visual cortex. This study shows that human object recognition abilities remain robust when only small amounts of information are available due to heavy occlusion, but the performance of bottom-up computational models is impaired under limited visibility. The results provide combined behavioral, neurophysiological, and modeling insights showing how recurrent computations may help the brain solve the fundamental challenge of pattern completion.

Author contributions: H.T., M.S., W.L., C.M., D.C., and G.K. designed research; H.T., M.S., W.L., C.M., A.P., J.O.C., W.H., and G.K. performed research; H.T., M.S., W.L., C.M., and G.K. analyzed data; and H.T., M.S., W.L., and G.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All data and code (including image databases, behavioral measurements, physiological measurements, and computational algorithms) have been deposited on GitHub and is available at https://github.com/kreimanlab/occlusion-classification.

[1]H.T., M.S., and W.L. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: gabriel.kreiman@tch.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1719397115/-/DCSupplemental.

Published online August 13, 2018.

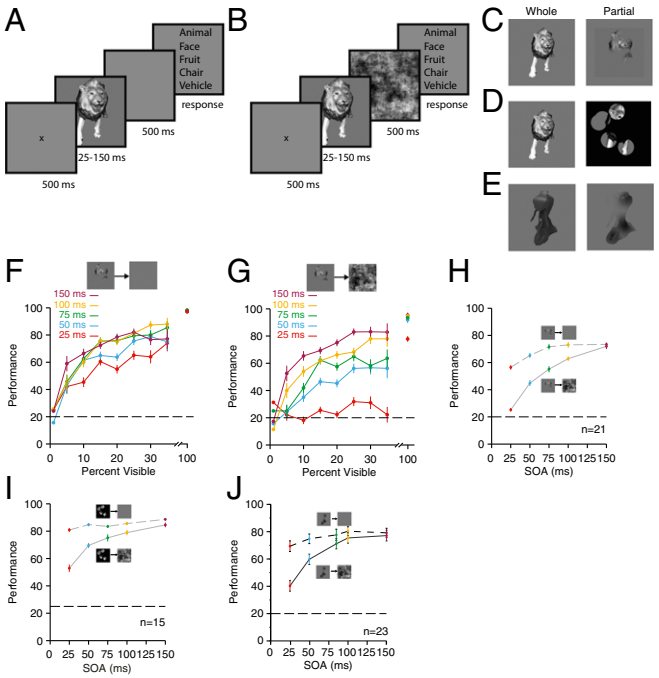

Fig. 1. Backward masking disrupts recognition of partially visible objects. (*A* and *B*) Forced-choice categorization task (n = 21 subjects). After 500 ms of fixation, stimuli were presented for variable exposure times (SOA from 25 to 150 ms), followed by a gray screen (*A*) or a noise mask (*B*) for 500 ms. Stimuli were presented unaltered (Whole; *C*, *Left* and *D*, *Left*), rendered partially visible (Partial; *C*, *Right*), or rendered occluded (*D*, *Right*) (*SI Appendix*, Fig. S1). (*E*) Experimental variation with novel objects (*SI Appendix*, Fig. S8). Behavioral performance is shown as a function of visibility for the unmasked (*F*) and masked (*G*) trials. Colors denote different SOAs. Error bars denote SEM. The horizontal dashed line indicates chance level (20%). Bin size = 2.5%. Note the discontinuity in the *x* axis to report performance at 100% visibility. (*H*) Average recognition performance as a function of SOA for partial objects (same data replotted from *F* and *G*, excluding 100% visibility). Performance was significantly degraded by masking (solid gray line) compared with the unmasked trials (dotted gray line) ($P < 0.001$, χ^2 test; df = 4). (*I*) Performance versus SOA for the occluded stimuli in *D* (note: chance = 25% here) (*SI Appendix*, Fig. S1). (*J*) Performance versus SOA for the novel object stimuli in *E*.

ranged from 25 to 150 ms in randomly ordered trials. Stimuli consisted of 325 objects belonging to five categories: animals, chairs, faces, fruits, and vehicles. The parts revealed for each object were chosen randomly. There were 40 images per object, comprising a total of 13,000 images of partial objects (*Methods*).

For whole objects and without a mask, behavioral performance was near ceiling, as expected (100% visible in Fig. 1*F*). Subjects robustly recognized partial objects across a wide range of visibility levels despite the limited information provided (Fig. 1*F*). Although poor visibility degraded performance, subjects still showed 80 ± 3% performance at 35 ± 2.5% visibility (partial versus whole objects: $P < 10^{-10}$, two-sided t test). Even for images with 10 ± 2.5% visibility, performance was well above chance levels (59 ± 2%; $P < 10^{-10}$, two-sided t test; chance = 20%). There was a small but significant improvement in performance at longer SOAs for partially visible objects (dashed lines in Fig. 1*H*; Pearson $r = 0.56$; $P < 0.001$, permutation test).

In a separate experiment, we generated images where objects appeared behind a black surface occluder (Fig. 1*D*). Consistent with previous studies (e.g., ref. 14), recognition was also robust when using heavily occluded images (Fig. 1*I*). The presence of an occluder improved recognition performance with respect to partial objects (compare *SI Appendix*, Fig. S1 *A* versus *B*; $P < 10^{-4}$, χ^2 test). We focused next on the essential aspects of pattern completion by considering the more challenging condition of partially visible objects, without help from other cues such as occluders.

While subjects had not seen any of the specific images in this experiment before, they had had extensive experience with fully visible and occluded versions of other images of animals, faces, fruits, chairs, and vehicles. We conducted a separate experiment with novel shapes (Fig. 1*E* and *SI Appendix*, Fig. S8*A*) to assess whether robustness to limited visibility (Fig. 1 *F*, *H*, and *I*) extended to unfamiliar objects. Visual categorization of such novel objects was also robust to limited visibility (Fig. 1*J* and *SI Appendix*, Fig. S8*B*).

Backward Masking Disrupts Recognition of Partially Visible Objects. Behavioral (18), physiological (19, 20), and computational studies (3, 4, 10) suggest that recognition of whole isolated objects can be described by rapid, largely feed-forward, mechanisms. Several investigators have used backward masking to force visual recognition to operate in a fast regime with minimal influences from recurrent signals (21): When an image is rapidly followed by a spatially overlapping mask, the high-contrast noise mask interrupts any additional, presumably recurrent, processing of the original image (22–24). We asked whether this fast, essentially feed-forward, recognition regime imposed by backward masking is sufficient for robust recognition of partially visible objects by randomly interleaving trials with a mask (Fig. 1*B*).

Performance for whole images was affected by the mask only for the shortest SOA values (compare Fig. 1 *F* versus *G* at 100% visibility; $P < 0.01$, two-sided t test). When partial objects were followed by a backward mask, performance was severely impaired (compare Fig. 1 *F* versus *G*). A two-way ANOVA on performance with SOA and masking as factors revealed a significant interaction ($P < 10^{-8}$). The behavioral consequences of shortening SOA were significantly stronger in the presence of backward masking (compare solid versus dashed lines in Fig. 1*H*). Additionally, backward masking disrupted performance across a wide range of visibility levels for SOAs that were ≤100 ms (Fig. 1 *G* and *H*). Similar effects of backward masking were observed when using occluded objects (Fig. 1*I*; $P < 0.001$, two-way ANOVA) as well as when using novel objects (Fig. 1*J* and *SI Appendix*, Fig. S8 *C* and *D*; $P < 0.0001$, two-way ANOVA). In sum, interrupting processing via backward masking led to a large reduction in the ability for recognition of partially visible objects, occluded images, and partially visible novel objects across a wide range of SOA values and visibility levels.

Images More Susceptible to Backward Masking Elicited Longer Neural Delays Along Human Ventral Visual Cortex. In a recent study, we recorded invasive physiological signals throughout the ventral visual stream in human patients with epilepsy while they performed an experiment similar to the one in Fig. 1*A* (13). This experiment included 25 objects presented for 150 ms without any masking, with random bubble positions in each trial. For whole objects, neural signals along the ventral visual stream showed rapid selective responses to different categories, as shown for an example electrode in the left fusiform gyrus in Fig. 2 *A* and *B*. When presenting partially visible objects, the neural signals remained visually selective (Fig. 2 *C* and *D*). The visually selective signals elicited by the partial objects were significantly delayed with respect to the responses to whole objects (compare the neural latency, defined here as the single-trial time of peak responses, in Fig. 2 *C* and *D* with the time of peak response before 200 ms in Fig. 2 *A* and *B*). Because the visible features varied from trial to trial, different renderings of the same object elicited a wide distribution in the neural latencies (Fig. 2 *C* and *D*). For example, the peak voltage occurred at 206 ms after stimulus onset in response to the first image in Fig. 2*C* and at 248 ms in response to the last image in Fig. 2*C*.

Heterogeneity across different renderings of the same object was also evident in the range of effects of backward masking at the behavioral level in the experiment in Fig. 1 *G* and *H*. We hypothesized that those images that elicited longer neural delays would also be more susceptible to backward masking. To test this hypothesis, we selected two electrodes in the neurophysiological study showing strong visually selective signals (*Methods*; one of these sites is shown in Fig. 2 *A*–*D*). We considered 650 images of partially visible objects corresponding to the 25 objects from the

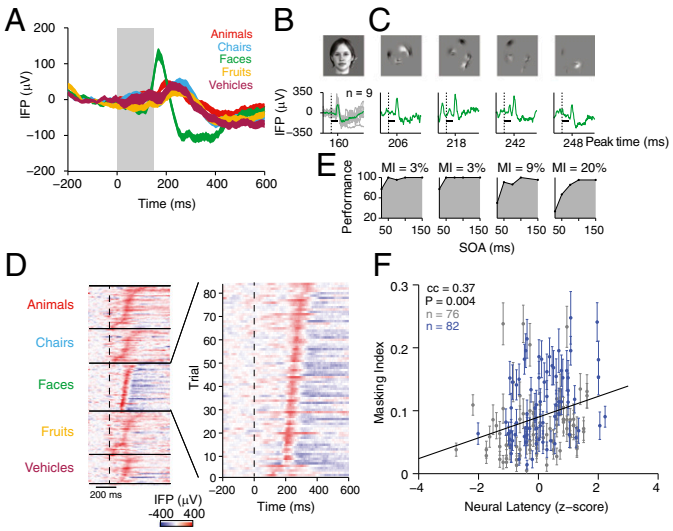

Fig. 2. Behavioral effect of masking correlated with the neural response latency on an image-by-image basis. (*A*) Intracranial field potential (IFP) responses from an electrode in the left fusiform gyrus averaged across five categories of whole objects while a subject was performing the task described in Fig. 1 (no masking, 150-ms presentation time). This electrode showed a stronger response to faces (green). The gray rectangle indicates the stimulus presentation time (150 ms). The shaded area indicates SEM (details are provided in ref. 13). (*B*) IFP responses for one of the whole objects for the electrode in *A* showing single-trial responses (gray, $n = 9$) and average response (green). The latency of the peak response is marked on the *x* axis. (*C*) Single-trial responses ($n = 1$) to four partial images of the same object in *B*. (*D*) New stimulus set for psychophysics experiments was constructed from the images in 650 trials from two electrodes in the physiology experiments. A raster of the neural responses for the example electrode in *A*, one trial per line, from partial image trials selected for psychophysics is shown. These trials elicited strong physiological responses with a wide distribution of response latencies (sorted by the neural latency). The color indicates the voltage (color scale on bottom). (*Right*, *Inset*) Zoomed-in view of the responses to the 82 trials in the preferred category. (*E*) We measured the effect of backward masking at various SOAs for each of the same partial exemplar images used in the physiology experiment ($n = 33$ subjects) and computed an MI for each image (*Methods*). The larger the MI for a given image, the stronger was the effect of masking. (*F*) Correlation between the effect of backward masking (*y* axis, MI as defined in *E*) and the neural response latency (*x* axis, as defined in *B* and *C*). Each dot is a single partial object from the preferred category for electrode 1 (blue) or 2 (gray). Error bars for the MI are based on half-split reliability (*SI Appendix*, Fig. S2), and the neural latency values are based on single trials. There was a significant correlation (Pearson $r = 0.37$; $P = 0.004$, linear regression, permutation test). cc, correlation coefficient.

neurophysiology experiment. Using the same images (i.e., the exact same features revealed for each object), we conducted a separate psychophysics experiment to evaluate the effect of backward masking on each individual image ($n = 33$ subjects). This experiment allowed us to construct a curve of behavioral performance, as a function of SOA during backward masking, for each of the selected images from the neurophysiology experiment (Fig. 2*E*). To quantify the effect of backward masking for each individual image, we defined a masking index (MI), 1 − AUC, where AUC is the normalized area under the curve in the performance versus SOA plot (gray area in Fig. 2*E*). Larger MI values correspond to larger effects of backward masking: the MI ranges from 0 (no effect of backward masking) to 0.8 (backward masking leads to chance performance). For example, in Fig. 2*C*, the first image was less affected by backward masking than the last image, particularly at short SOA values (MI values of 3% and 20%, respectively).

For those images from the preferred category for each of the two electrodes, the MI showed a weak but significant correlation with the neural response latency, even after accounting for image difficulty and recording site differences (Fig. 2*F*; Pearson $r = 0.37$; $P = 0.004$, permutation test; *Methods*). This effect was stimulus selective: The MI was not correlated with the neural response latency for images from the nonpreferred categories ($P = 0.33$, permutation test). The neural latencies are noisy measures based on single trials (*Methods* and Fig. 2*C*), the physiology and behavioral experiments were conducted in different subjects, and there was variability across subjects in the MI (Fig. 2*F* and *SI Appendix*, Fig. S2). However, despite all of these sources of noise, images that led to longer neural response latencies were associated with a stronger effect of interrupting computations via backward masking.

Standard Feed-Forward Models Are Not Robust to Occlusion. We next investigated the potential computational mechanisms responsible for the behavioral and physiological observations in Figs. 1 and 2. We began by considering state-of-the-art implementations of purely feed-forward computational models of visual recognition. These computational models are characterized by hierarchical, feed-forward processing with progressive increases in the size of receptive fields, degree of selectivity, and tolerance to object transformations (e.g., refs. 2–4). Such models have been successfully used to describe rapid recognition of whole objects at the behavioral level (e.g., ref. 4) and neuronal firing rates in area V4 and the inferior temporal cortex in macaque monkeys (e.g., ref. 10). Additionally, these deep convolutional network architectures achieve high performance in computer vision competitions evaluating object recognition capabilities (e.g., refs. 6, 7).

We evaluated the performance of these feed-forward models in recognition of partially visible objects using the same 325 objects (13,000 trials) in Fig. 1. As a representative of this family of models, we considered AlexNet (6), an eight-layer convolutional neural network trained via back-propagation on ImageNet, a large corpus of natural images (9). We used as features either activity in the last fully connected layer before readout (fc7; 4,096 units) or activity in the last retinotopic layer (pool5; 9,216 units). To measure the effect of low-level differences between categories (e.g., contrast, object area), we also considered raw pixels as baseline performance ($256 \times 256 = 65{,}536$ features).

We sought to measure the robustness of these networks to partial object visibility in the same way that tolerance to other transformations such as size and position changes is evaluated [i.e., by training a decision boundary on one condition such as specific size, viewpoint, or whole objects and testing on the other conditions such as other sizes, viewpoints, or occlusion (e.g., refs. 2, 4)]. It is not fair to compare models trained with occluded objects versus models trained exclusively with whole objects; therefore, we do not include occluded objects in the training set. Furthermore, the results in Fig. 1*J* and *SI Appendix*, Fig. S8 show that humans can perform pattern completion for novel objects without any prior training with occluded versions of those objects. We trained a support vector machine (SVM) classifier (linear kernel) on the features of whole objects and tested object categorization performance on the representation of images of partial objects. Importantly, all of the models were trained exclusively with whole objects, and performance was evaluated in images with partially visible objects. Cross-validation was performed over objects: Objects used to train the decision boundary did not appear as partial objects in the test set. The performance of raw pixels was essentially at chance level (Fig. 3*A*). In contrast, the other models performed well above chance ($P < 10^{-10}$, two-sided *t* test; also *SI Appendix*, Fig. S4). While feed-forward models performed well above chance, there was a significant gap with respect to human performance at all visibility levels below 40% ($P < 0.001$, χ^2 test; Fig. 3*A*). These results are consistent with those reported in other simulations with occluded objects and similar networks (25). The decrease in performance of feed-forward models compared with humans depends strongly on the stimuli and on the amount of information available: Bottom-up models were comparable to humans at full visibility (26) (*SI Appendix*, Fig. S3).

The decline in performance with low visibility was not specific to the set of images used in this study: AlexNet pool5 and fc7 also performed below human levels when considering novel objects (*SI Appendix*, Fig. S9*A*). The decline in performance with low visibility was not specific to using pixels or the AlexNet pool5 or fc7 layer.

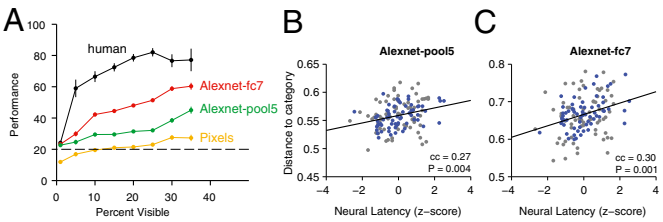

Fig. 3. Standard feed-forward models were not robust to occlusion. (*A*) Performance of feed-forward computational models (colors) compared with humans (black) (also *SI Appendix*, Figs. S4, S5, and S9*A*). We used the feature representation of a subset of whole objects to train an SVM classifier and evaluated the model's performance on the feature representation of partial objects (*Methods*). The objects used to train the classifier did not appear as partial objects in the test set. Human performance is shown here (150-ms SOA) for the same set of images. Error bars denote SEM (fivefold cross-validation). The single-trial neural latency for each image (Fig. 2*B*) was correlated with the distance of each partial object to its whole object category center for AlexNet pool5 (*B*) and AlexNet fc7 (*C*). Each dot represents a partial object, with responses recorded from either electrode 1 (blue dots) or electrode 2 (gray dots). The correlation coefficients (cc) and *P* values from the permutation test are shown for each subplot.

All of the feed-forward models that we tested led to the same conclusions, including different layers of VGG16, VGG19 (7), InceptionV3 (8), and ResNet50 (*SI Appendix*, Fig. S4). Among these models, the VGG16 architecture provided slightly better recognition performance in the low-visibility regime.

The models shown in Fig. 3*A* and *SI Appendix*, Fig. S4 were trained to optimize object classification performance in the ImageNet 2012 dataset (9) without any specific training for the set of objects used in our study, except for the SVM classifier. To assess whether fine-tuning the model's weights could alleviate the challenges with limited visibility, we fine-trained AlexNet via back-propagation using the 325 whole objects and then retested this fine-tuned model on the images with limited visibility. Fine-tuning the AlexNet architecture did not lead to improvements at low visibility (*SI Appendix*, Fig. S5). These results are consistent with a previous computational study using feed-forward models similar to the ones in the current work and evaluating a more extensive image dataset (25).

We used stochastic neighborhood embedding to project the AlexNet fc7 layer features onto two dimensions and to visualize the effects of occlusion on the model (Fig. 4*C*). The representation of whole objects (open circles) showed a clear separation among categories, but partial objects from different categories (filled circles) were more similar to each other than to their whole object counterparts. Therefore, decision boundaries trained on whole objects did not generalize to categorization of partial objects (Fig. 3*A*). Despite the success of purely feed-forward models in recognition of whole objects, these models were not robust under limited visibility.

We next sought to further understand the breakdown in the models' representations of objects under partial visibility. Removing large amounts of pixels from the objects pushed the model's representation of the partially visible object away from their whole counterparts (Fig. 4*C*). The distance between the representation of a partially visible object and the corresponding whole object category mean is indicative of the impact of partial visibility. We evaluated whether this distortion was correlated with the latencies in the neural recordings from Fig. 2. We reasoned that images of partial objects whose model representation was more distorted would lead to longer neural response latencies. We computed the Euclidean distance between the representation of each partial object and the whole object category mean. We found a modest but significant correlation at the object-by-object level between the computational distance to the category mean and the neural response latency for the pool5 (Fig. 3*B*) and fc7 (Fig. 3*C*) features. The statistical significance of these correlations was assessed by regressing the distance to category mean against the neural latency, along with the following additional predictors to account for potential confounds: (*i*) the percentage of object visibility and pixel distance to regress out any variation explained by low-level effects of occlusion and difficulty, (*ii*) the electrode number to account for the interelectrode variability in our dataset, and (*iii*) the MI (Fig. 2*E*) to control for overall recognition difficulty. The model distance to category mean in the pool5 and fc7 layers correlated with the response latency beyond what could be explained by these additional factors (pool 5: Pearson $r = 0.27$; $P = 0.004$, permutation test; fc7: Pearson $r = 0.3$; $P = 0.001$, permutation test). In sum, state-of-the-art feed-forward architectures did not robustly extrapolate from whole to partially visible objects and failed to reach human-level performance in recognition of partially visible objects. As the difference in the representation of whole and partial objects increased, the time it took for a selective neural response to evolve for the partial objects was longer.

Recurrent Neural Networks Improve Recognition of Partially Visible Objects. The behavioral, neural, and modeling results presented above suggest a need for additional computational steps beyond those present in feed-forward architectures to build a robust representation for partially visible objects. Several computational ideas, originating from models proposed by Hopfield (27), have shown that attractor networks can perform pattern completion. In the Hopfield network, units are connected in an all-to-all fashion with weights defining fixed attractor points dictated by the whole objects to be represented. Images that are pushed farther away by limited visibility would require more processing time to converge to the appropriate attractor, consistent with the behavioral and physiological observations. As a proof of principle, we augmented the feed-forward models discussed in the previous section with recurrent connections to generate a robust representation through an attractor-like mechanism (Fig. 4*A*), with one attractor for each whole object. We used the AlexNet architecture with fixed feed-forward weights from pretraining on ImageNet (as in Fig. 3) and added recurrent connections to the fc7 layer. Recurrent connectivity is ubiquitous throughout all visual neocortical areas in biological systems. The motivation to include recurrent connectivity only in the fc7 layer was to examine first a simple and possibly minimal extension to the existing architectures (*Discussion*).

We denote the activity of the fc7 layer at time t as the 4,096-dimensional feature vector $\mathbf{h}_t$. At each time step, $\mathbf{h}_t$ was determined by a combination of the activity from the previous time step $\mathbf{h}_{t-1}$ and the constant input from the previous layer $\mathbf{x}$: $\mathbf{h}_t = f(\mathbf{W}_h\mathbf{h}_{t-1}, \mathbf{x})$, where f introduces a nonlinearity (*Methods*). The input from the previous layer, fc6, was kept constant and identical to that in the feed-forward AlexNet. $\mathbf{W}_h$ is a weight matrix that governs the temporal evolution of the fc7 layer. We considered a Hopfield recurrent neural network (RNN_h) without introducing any free parameters that depended on the partial objects, where $\mathbf{W}_h$ was a symmetrical weight matrix dictated by the fc7 representation of the whole objects, using the implementation of Li et al. (28). The initial state of the network was given by the activity in the previous layer, $\mathbf{h}_0 = \mathbf{W}_{6\to7}fc6$, followed by binarization. The state of the network evolved over time according to $\mathbf{h}_t = satlins(\mathbf{W}_h\mathbf{h}_{t-1})$, where *satlins* is a saturating nonlinearity (*Methods*). We verified that the whole objects constituted an attractor point in the network by ensuring that their representation did not change over time when used as inputs to the model. We next evaluated the responses of RNN_h to all of the images containing partial objects. The model was run until convergence (i.e., until none of the feature signs changed between consecutive time steps). Based on the final time point, we evaluated the performance in recognizing partially visible objects. The RNN_h model demonstrated a significant improvement over the AlexNet fc7 layer (Fig. 4*B*; $57 \pm 0.4\%$; $P < 0.001$, χ^2 test).

The dynamic trajectory of the representation of whole and partial objects in the fc7 layer of the RNN_h model is visualized in Fig. 4*C*. Before any recurrent computations have taken place, at $t = 0$ (Fig. 4*C*, *Left*), the representations of partial objects were clustered together (closed circles in Fig. 4*C*) and separated from the clusters of whole objects in each category (open circles in Fig. 4*C*). As time progressed, the cluster of partial objects was pulled apart and moved toward their respective categories. For example, at

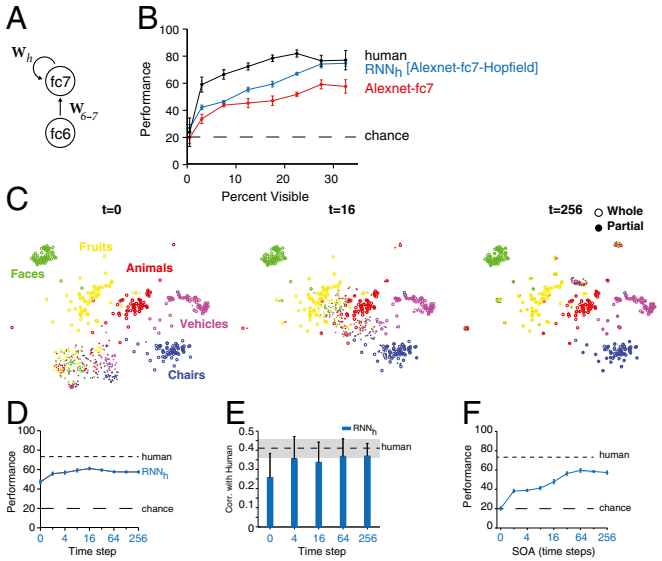

Fig. 4. Dynamic RNN showed improved performance over time and was impaired by backward masking. (*A*) Top-level representation in AlexNet (fc7) receives inputs from fc6, governed by weights $\mathbf{W}_{6\to7}$. We added a recurrent loop within the top-level representation (RNN). The weight matrix $\mathbf{W}_h$ governs the temporal evolution of the fc7 representation (*Methods*). (*B*) Performance of the RNN_h (blue) as a function of visibility. The RNN_h approached human performance (black curve) and represented a significant improvement over the original fc7 layer (red curve). The red and black curves are copied from Fig. 3*A* for comparison. Error bars denote SEM. (*C*) Temporal evolution of the feature representation for RNN_h as visualized with stochastic neighborhood embedding. Over time, the representation of partial objects approaches the correct category in the clusters of whole images. (*D*) Overall performance of the RNN_h as a function of recurrent time step compared with humans (top dashed line) and chance (bottom dashed line). Error bars denote SEM (five-way cross-validation; *Methods*). (*E*) Correlation (Corr.) in the classification of each object between the RNN_h and humans. The dashed line indicates the upper bound of human–human similarity obtained by computing how well half of the subject pool correlates with the other half. Regressions were computed separately for each category, followed by averaging the correlation coefficients across categories. Over time, the model becomes more human-like (*SI Appendix*, Fig. S6). Error bars denote SD across categories. (*F*) Effect of backward masking. The same backward mask used in the psychophysics experiments was fed to the RNN_h model at different SOA values (*x* axis). Error bars denote SEM (five-way cross-validation). Performance improved with increasing SOA values (*SI Appendix*, Fig. S10).

t = 16 (Fig. 4*C*, *Center*) and t = 256 (Fig. 4*C*, *Right*), the representation of partial chairs (closed blue circles in Fig. 4*C*) largely overlapped with the cluster of whole chairs (open blue circles in Fig. 4*C*). Concomitant with this dynamic transformation in the representation of partial objects, the overall performance of the RNN_h model improved over time (Fig. 4*D*).

In addition to the average performance reported in Fig. 4 *B* and *D*, we directly compared performance at the object-by-object level between humans and the RNN_h model (*SI Appendix*, Fig. S6). There were notable differences across categories [e.g., humans were much better than the model in detecting faces (green circles in *SI Appendix*, Fig. S6)]. For this reason, we first compared models and humans at the object-by-object level within each category and then averaged the results across categories. Over time, the RNN_h behaved more like humans at the object-by-object level (Fig. 4*E*). For each time step in the model, we computed the average correct rate on partial objects for each object, from each of the five categories, and correlated this vector with the pattern of human performance (*SI Appendix*, Fig. S6). The upper bound (dashed line in Fig. 4*E*) represents human–human similarity, defined as the correlation in the response patterns between half of the subject pool and the other half. Over time, the recurrent model–human correlation increased toward the human–human upper bound. Additionally, there was a parallel between human performance across different SOAs and the model performance across varying recurrent time steps (*SI Appendix*, Fig. S12): At fewer recurrent steps, the model showed a higher correlation with human performance at short SOAs, whereas with more recurrent steps, the model showed a higher correlation with human performance at long SOAs.

Adding a Hopfield-like recurrent architecture to AlexNet also improved performance in recognition of the novel objects (*SI Appendix*, Figs. S8*A* and S9 *B–D*). Similar conclusions were obtained when considering the VGG16 architecture and adding Hopfield-like recurrent connections to the fc1 layer (*SI Appendix*, Fig. S7).

In sum, implementing recurrent connections in an attractor-like fashion at the top of a feed-forward hierarchical model significantly improved the model's performance in pattern completion, and the additional computations were consistent with temporal delays described at the behavioral and neural levels.

Backward Masking Impaired RNN Model Performance. We reasoned that the backward mask introduced in the experiment discussed in Fig. 1 *B*, *G*, and *H* impaired performance by interrupting processing, and we set out to investigate whether we could reproduce this effect in the RNN_h model. We computed the responses of the AlexNet model to the mask and fed the fc6 features for the mask to the RNN_h model after a certain number of time steps. Switching the mask on at earlier time points was meant to mimic shorter SOAs in the psychophysical experiments. We read out performance based on the resulting fc7 activity combining the partial object and the mask at different time points (Fig. 4*F*). Presenting the mask reduced performance from 58 ± 2% (SOA = 256 time steps) to 37 ± 2% (SOA = two time steps). Although we cannot directly compare SOAs in milliseconds with time steps in the model, these results are qualitatively consistent with the behavioral effects of backward masking (Fig. 1*H*; a side-by-side comparison of the physiological, behavioral, and computational dynamics is shown in *SI Appendix*, Fig. S10).

Discussion

It is routinely necessary to recognize objects that are partially visible due to occlusion and poor illumination. The visual system is capable of making inferences even when only 10–20% of the object is visible (Fig. 1*F*), and even for novel objects (Fig. 1*J*). We investigated the mechanisms underlying such robust recognition of partially visible objects (referred to as pattern completion) at the behavioral, physiological, and computational levels. Backward masking impaired recognition of briefly presented partial images (25 ms ≤ SOA ≤ 100 ms) (Fig. 1 *G–J*). The strength of the disruptive effect of backward masking was correlated with the neural delays from invasive recordings along the ventral visual stream (13) (Fig. 2). State-of-the-art bottom-up computational architectures trained on whole objects failed to achieve robustness in recognition of partially visible objects (Fig. 3*A* and *SI Appendix*, Figs. S4 and S5). The introduction of proof-of-principle recurrent connections (Fig. 4*A*) led to significant improvement in recognition of partially visible objects at the average level (Fig. 4*B* and *SI Appendix*, Figs. S7 and S9*B*) and also in explaining which images were easier or harder for humans at the object-by-object level (Fig. 4*E* and *SI Appendix*, Figs. S6 and S12). The RNN_h model had no free parameters that depended on the partial objects: All of the weights were determined by the whole objects. The increase in performance involved recurrent computations evolving over time that were interrupted by the introduction of a mask (Fig. 4*F*).

Recognition of partially visible objects requires longer reaction times (11, 15) and delayed neural representation with respect to that of whole objects (12, 13). These delays suggest the need for additional computations to interpret partially visible images. Interrupting those additional computations by a mask impairs recognition (Fig. 1 *G–J*). Backward masking disproportionately affects recurrent computations (22–24). Accordingly, we conjectured that the disruptive effect of backward masking during pattern completion could be ascribed to the impairment of such recurrent computations. The rapid and selective signals along the ventral visual stream that enable recognition of whole objects within ~150 ms have been interpreted to reflect largely bottom-up processing

(2–4, 10, 18–20; however, ref. 29) Physiological delays of ∼50 ms during recognition of partial objects (12, 13) provide ample time for recurrent connections to exert their effects during pattern completion. These delays could involve recruitment of lateral connections (16) and/or top-down signals from higher visual areas (30).

Humans are constantly exposed to partially visible objects. While subjects had not previously seen the specific experiment images, they had had experience with occluded animals, chairs, faces, fruits, and vehicles. To evaluate whether category-specific experience with occluded objects is required for pattern completion, we conducted an experiment with completely novel objects (Fig. 1*J* and *SI Appendix*, Figs. S8 and S9). Subjects robustly categorized novel objects under low visibility even when they had never seen those heavily occluded objects or similar ones before.

There exist infinitely many possible bottom-up models. Even though we examined multiple state-of-the-art models that are successful in object recognition (AlexNet, VGG16, VGG19, InceptionV3, and ResNet50), their failure to account for the behavioral and physiological results (25, 26) (Fig. 3 and *SI Appendix*, Fig. S4) 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, a recurrent network with a finite number of time steps can be unfolded into a bottom-up model by creating an additional layer for each time step. However, there are several advantages to recurrent architectures, including a reduction in the number of units and weights. Furthermore, such unfolding of time into layers is only applicable when we know a priori the fixed number of computational steps, whereas recurrent architectures allow an arbitrary and dynamically flexible number of computations.

The RNN dynamics involve temporal evolution of the features (Fig. 4 *C*–*F*), bringing the representation of partial objects closer to that of whole objects. These computational dynamics, map onto the temporal lags observed at the behavioral and physiological levels. The RNN_h model's performance and correlation with humans saturates at around 10–20 time steps (Fig. 4 *C*–*F*); a combination of feed-forward signals and recurrent computations is consistent with the physiological responses to heavily occluded objects arising at around 200 ms (Fig. 2*D*). The RNN_h model uses discrete time steps, but a more realistic implementation should be based on spikes and continuous dynamics. A continuous time implementation of recurrent interactions shows that information can be added rapidly, particularly under noisy input conditions, and consistently with the empirically observed delays of ∼50–100 ms in Figs. 1 and 2 (29). Furthermore, these dynamics are interrupted by the presentation of a backward mask in close temporal proximity to the image (Figs. 1 *G*–*J* and 4*F* and *SI Appendix*, Fig. S10).

Multiple other cues can aid recognition of partially visible objects, including relative positions, segmentation, movement, illumination, and stereopsis. Additionally, during learning, children often encounter partially visible objects that they can continuously explore from different angles. It will be interesting to integrate these additional sources of information and to understand how they contribute to pattern completion. The convergence of behavioral, physiological, and theoretical evidence presented here provides insight into the human visual recognition neural circuitry and a biologically constrained hypothesis to understand the role of recurrent computations during pattern completion.

Methods

An expanded version is presented in *SI Appendix*.

Psychophysics. A total of 106 volunteers (62 female, aged 18–34 y) participated in the behavioral experiments. We performed an experiment with partially visible objects rendered through bubbles (Fig. 1) and three variations with occluded objects (Fig. 1 and *SI Appendix*, Fig. S1), novel objects (Fig. 1 and *SI Appendix*, Figs. S8 and S9), and stimuli matched to a previous neurophysiological experiment (13) (Fig. 2). All subjects gave informed consent and the studies were approved by the Institutional Review Board at Children's Hospital, Harvard Medical School.

Neurophysiology Experiments. The neurophysiological intracranial field potential data in Figs. 2 and 3 were taken from a study by Tang et al. (13). The neural latency for each image was defined as the time of the peak response in the intracranial field potential and was calculated in single trials (e.g., Fig. 2*C*).

Computational Models. We tested state-of-the-art feed-forward vision models, focusing on AlexNet (6) (Fig. 3; other models are shown in *SI Appendix* and *SI Appendix*, Fig. S4), with weights pretrained on ImageNet (6, 9). As a proof of principle, we proposed an RNN model by adding all-to-all recurrent connections to the top feature layer of AlexNet (Fig. 4*A*). The RNN model was defined using only information about the whole objects by setting the recurrent weights based on a Hopfield attractor network (27), as implemented in MATLAB's newhop function (28).

ACKNOWLEDGMENTS. We thank Carlos Ponce, Alan Yuille, Siamak Sorooshyari, and Guy Ben-Yosef for useful discussions and comments. This work was supported by a fellowship from the FITweltweit Programme of the German Academic Exchange Service (to M.S.), National Science Foundation Science and Technology Center Award CCF-123121 (to G.K.), and NIH Award R01EY026025 (to G.K.).

1. Logothetis NK, Sheinberg DL (1996) Visual object recognition. *Annu Rev Neurosci* 19:577–621.
2. DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci* 11:333–341.
3. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2:1019–1025.
4. Serre T, et al. (2007) A quantitative theory of immediate visual recognition. *Prog Brain Res* 165:33–56.
5. Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. *Curr Opin Neurobiol* 17:140–147.
6. Krizhevsky A, Sutskever I, Hinton G (2012) *ImageNet Classification with Deep Convolutional Neural Networks* (NIPS, Montreal).
7. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
8. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. arXiv:1512.005673v3.
9. Russakovsky O, et al. (2014) ImageNet large scale visual recognition challenge. arXiv:1409.0575.
10. Yamins DL, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci USA* 111:8619–8624.
11. Murray RF, Sekuler AB, Bennett PJ (2001) Time course of amodal completion revealed by a shape discrimination task. *Psychon Bull Rev* 8:713–720.
12. Kosai Y, El-Shamayleh Y, Fyall AM, Pasupathy A (2014) The role of visual area V4 in the discrimination of partially occluded shapes. *J Neurosci* 34:8570–8584.
13. Tang H, et al. (2014) Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* 83:736–748.
14. Johnson JS, Olshausen BA (2005) The recognition of partially visible natural objects in the presence and absence of their occluders. *Vision Res* 45:3262–3276.
15. Brown JM, Koch C (2000) Influences of occlusion, color, and luminance on the perception of fragmented pictures. *Percept Mot Skills* 90:1033–1044.
16. Gilbert CD, Li W (2013) Top-down influences on visual processing. *Nat Rev Neurosci* 14:350–363.
17. Angelucci A, Bressloff PC (2006) Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Prog Brain Res* 154:93–120.
18. Kirchner H, Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Res* 46:1762–1776.
19. Keysers C, Xiao DK, Földiák P, Perrett DI (2001) The speed of sight. *J Cogn Neurosci* 13:90–101.
20. 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* 62:281–290.
21. Serre T, Oliva A, Poggio T (2007) Feedforward theories of visual cortex account for human performance in rapid categorization. *Proc Natl Acad Sci USA* 104:6424–6429.
22. Lamme VA, Zipser K, Spekreijse H (2002) Masking interrupts figure-ground signals in V1. *J Cogn Neurosci* 14:1044–1053.
23. Kovács G, Vogels R, Orban GA (1995) Cortical correlate of pattern backward masking. *Proc Natl Acad Sci USA* 92:5587–5591.
24. Keysers C, Perrett DI (2002) Visual masking and RSVP reveal neural competition. *Trends Cogn Sci* 6:120–125.
25. Pepik B, Benenson R, Ritschel T, Schiele B (2015) What is holding back convnets for detection? arXiv:1508.02844.
26. Wyatte D, Curran T, O'Reilly R (2012) The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. *J Cogn Neurosci* 24:2248–2261.
27. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79:2554–2558.
28. Li J, Michel A, Porod W (1989) Analysis and synthesis of a class of neural networks: Linear systems operating on a closed hypercube. *IEEE Trans Circuits Syst* 36:1405–1422.
29. Panzeri S, Rolls ET, Battaglia F, Lavis R (2001) Speed of feedforward and recurrent processing in multilayer networks of integrate-and-fire neurons. *Network* 12:423–440.
30. 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:e25784.