nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-019-0392-5

Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior

Kohitij Kar[1,2*], Jonas Kubilius[1,3], Kailyn Schmidt[1], Elias B. Issa[1,4] and James J. DiCarlo[1,2]

Non-recurrent deep convolutional neural networks (CNNs) are currently the best at modeling core object recognition, a behavior that is supported by the densely recurrent primate ventral stream, culminating in the inferior temporal (IT) cortex. If recurrence is critical to this behavior, then primates should outperform feedforward-only deep CNNs for images that require additional recurrent processing beyond the feedforward IT response. Here we first used behavioral methods to discover hundreds of these 'challenge' images. Second, using large-scale electrophysiology, we observed that behaviorally sufficient object identity solutions emerged ~30 ms later in the IT cortex for challenge images compared with primate performance-matched 'control' images. Third, these behaviorally critical late-phase IT response patterns were poorly predicted by feedforward deep CNN activations. Notably, very-deep CNNs and shallower recurrent CNNs better predicted these late IT responses, suggesting that there is a functional equivalence between additional nonlinear transformations and recurrence. Beyond arguing that recurrent circuits are critical for rapid object identification, our results provide strong constraints for future recurrent model development.

In a single, natural fixation (~200 ms), primates can rapidly identify objects in the central visual field despite various identity-preserving image transformations, a behavior termed core object recognition[1]. Understanding the brain mechanisms that seamlessly solve this challenging computational problem has been a key goal of visual neuroscience[2,3]. Previous studies[4,5] have shown that object identities are explicitly represented in the pattern of neural activity in the primate IT cortex. Therefore, how the brain solves core object recognition boils down to building a neurally mechanistic model of the primate ventral stream that, for any image, accurately predicts the neural responses at all levels of the ventral stream, including the IT cortex.

At present, the models that best predict the individual responses of macaque IT neurons belong to the architectural family of deep CNNs (DCNNs) trained on object categorization[6–8]. These networks are also the best predictors of primate behavioral patterns across multiple core object recognition tasks[9,10]. Neural networks in this model family are almost entirely feed-forward. Specifically, unlike the ventral stream[11–14], they lack cortico-cortical, subcortical, and medium- to long-range intra-areal recurrent circuits (Fig. 1a). The short duration (~200 ms) needed to accomplish accurate object identity inferences in the ventral stream[4,15] suggests the possibility that recurrent circuit-driven computations are not critical for these inferences. In addition, it has been argued that recurrent circuits might operate at much slower time scales[16], being more relevant for processes such as regulating synaptic plasticity (learning). Therefore, a promising hypothesis is that core object recognition behavior does not require recurrent processing. The primary aim of this study was to try to falsify this hypothesis and to provide new constraints to guide future model development.

There is growing evidence to indicate that feedforward DCNNs fall short of accurately predicting primate behavior in many situations[10,17]. We therefore hypothesized that specific images for which the object identities are difficult for non-recurrent DCNNs, but are nevertheless easily solved by primates, might be critically benefiting from recurrent computations in primates. Furthermore, previous research[18] suggests that the impact of recurrent computations on the ventral stream should be most relevant at later time points in the image-driven neural responses. Therefore, we reasoned that object representations in the IT cortex for recurrence-dependent images will require an additional processing time to emerge (beyond the initial IT population response).

To discover such images, we behaviorally compared primates (humans and monkeys) and a particular non-recurrent DCNN (AlexNet 'fc7'[19]). We identified the following two groups of images: those for which object identity is easily inferred by the primate brain but not solved by DCNNs (challenge images), and those for which both primates and models easily infer object identity (control images). To test our neural hypothesis, we simultaneously measured IT population activity in response to these images using chronically implanted multielectrode arrays in two monkeys while they performed an object discrimination task.

Our results revealed that object identity decoding from IT populations for the challenge images took on average ~30-ms longer to emerge compared with the control images. Consistent with previous results, we also found that the top layers of DCNNs predicted ~50% of the image-driven neural response variance at the leading edge of the IT population response. However, this fit to the IT response was significantly worse (<20% explained variance) at later time points (150–200 ms post-stimuli onset), when the IT population solutions (linear decodes) to many of the challenge images emerged. Taken together, these results imply that recurrent computations play a behaviorally critical role during core object recognition. Notably, we also found the same neural phenomena while the monkeys

[1]McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Brain and Cognition, KU Leuven, Leuven, Belgium. [4]Present address: Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA.
*e-mail: kohitij@mit.edu

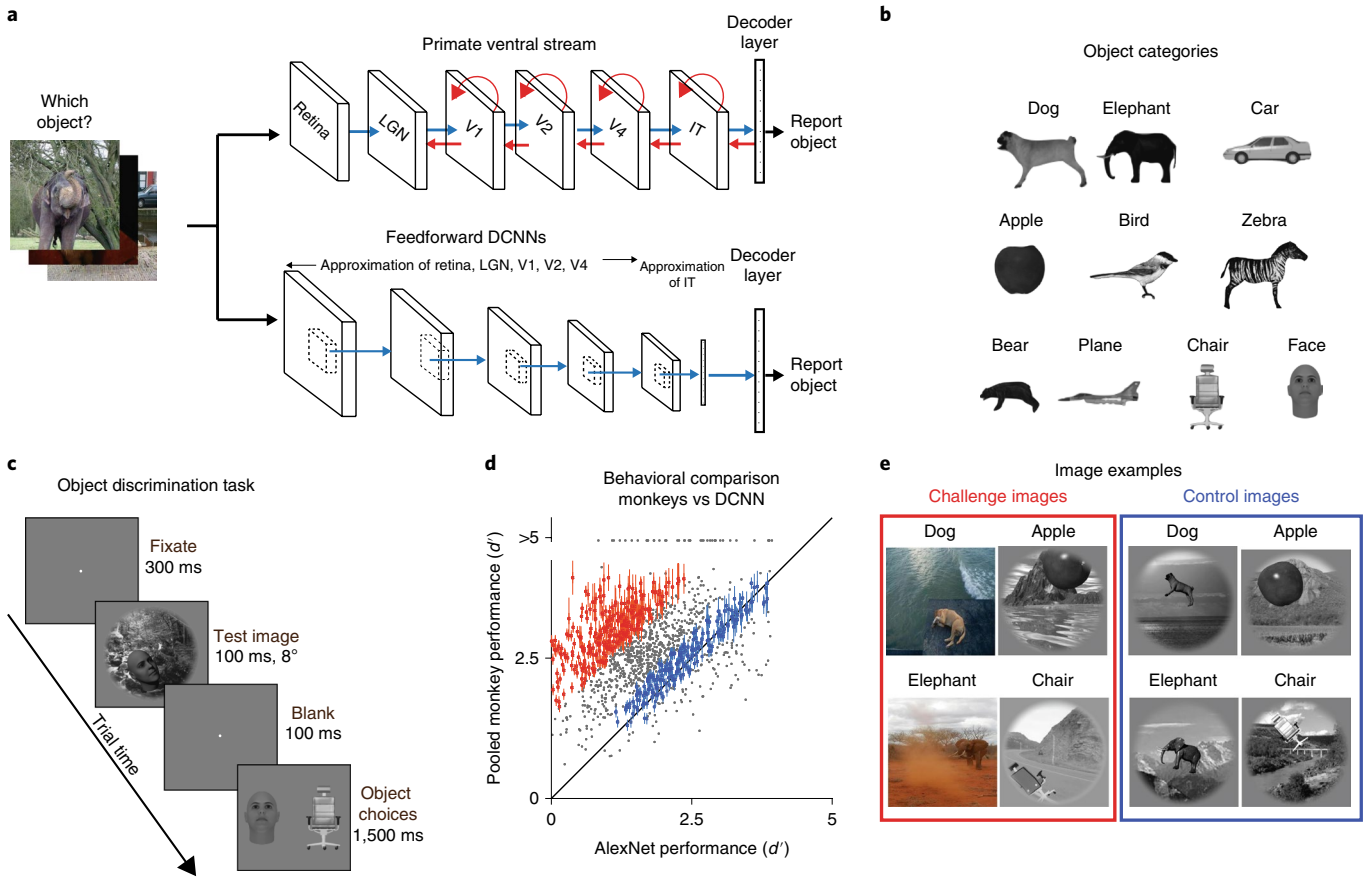


Fig. 1 | Behavioral screening and identification of control and challenge images. a Both primates (humans and macaques) and feedforward DCNNs were tasked to identify which object is present in each test image (1,320 images). Top: the stages in the primate ventral visual pathway (retina, lateral geniculate nucleus (LGN), areas V1, V2, V4, and the IT cortex), which is implicated in core object recognition. We can conceptualize each stage as rapidly transforming the representation of the image and ultimately yielding the primates' behavior (that is, producing a behavioral report of which object was present). The blue arrows indicate the known anatomical feedforward projections from one area to the other. The red arrows indicate the known lateral and top-down recurrent connections. Bottom: a schematic of a similar pathway commonly present in DCNNs. These networks contain a series of convolutional and pooling layers with nonlinear transforms at each stage, followed by fully connected layers (which approximate macaque IT neural responses) that ultimately gives rise to the models' 'behavior'. Note that the DCNNs only have feedforward (blue) connections. **b**, Illustrations of the ten different object types used in the study. **c**, Binary object discrimination task, showing the timeline of events for each trial. Subjects fixate on a circle, then the test image at 8° containing one of ten possible objects is shown for 100 ms. After a 100-ms delay, a canonical view of the target object (the same as that presented in the test image) and a distractor object (one of the other nine objects) appears, and the human or monkey indicates which object was present in the test image by clicking on or making a saccade, respectively, to one of the two choices. **d**, Comparison of monkey performance (pooled across two monkeys) and DCNN performance (AlexNet[19] fc7). Each circle represents the behavioral task performance (I_1; refer to Methods) for a single image. We reliably identified challenge images (red circles) and control images (blue circles). Error bars are bootstrapped s.e.m. **e**, Examples of four challenge and four control images.

passively viewed the images, implying that the putative recurrent mechanisms for successful core object inference in the primate are not strongly dependent on the state or task. Furthermore, we show that the observed image-by-image differences between DCNNs and primate behavior, together with precisely measured IT population dynamics for each image, better constrain the next generation of ventral stream models compared to previous qualitative approaches.

Results

As outlined above, we reasoned that if recurrent circuits are critical to core object recognition behavior, then primates should outperform current feedforward-only DCNNs for some images. The first goal of this study was to discover such challenge images. Rather than making assumptions about what types of images (for example, occluded, cluttered, or blurred) might most critically depend on feedback, we took a data-driven approach to identify such images.

Identification of DCNN challenge and control images. To compare the behavioral performance of primates (humans and macaques) and current DCNNs image-by-image, we used a binary object discrimination task (previously tested extensively[9,10]) (Fig. 1c). For each trial, monkeys used an eye movement to select one of two object choices after we briefly (100 ms) presented a test image containing one of those choice objects (see the "Primate behavioral testing" section in Methods).

We tested a total of 1,320 images (132 images per object) in which the primary visible object belonged to 1 of 10 different object categories (Fig. 1b). To make the task challenging, we included various image types (see Supplementary Fig. 1a), including synthetic objects with high view variation on cluttered natural backgrounds (similar to ref. [5]), images with occlusion, deformation, missing object-parts, and colored photographs (from the Microsoft COCO dataset[20]).

Behavioral testing of all of these images was performed in humans ($n=88$; Supplementary Fig. 1c) and in monkeys ($n=2$; Fig. 1d). We estimated the behavioral performance of the subject pool on each image, and the vector of image-wise performance is referred to as I_1 (see Methods, and refer to ref. [10]). We collected sufficient data such that the reliability of the I_1 vector was reasonably high (median split half reliability $\rho^{\sim}$, humans = 0.84 and monkeys = 0.88, where 1.0 is perfect reliability). To test the behavior of each DCNN model, we first extracted the image-evoked features from the penultimate layer, for example, the fc7 layer of AlexNet[19]. We then trained and tested (cross-validated) ten linear decoders (see Methods) to derive the binary task performances. Figure 1d shows an image-by-image behavioral comparison between the pooled monkey population and AlexNet fc7. We identified control images (blue circles in Fig. 1d) as those for which the absolute difference in primate and DCNN performance does not exceed 0.4 (d' units), and challenge images (red circles in Fig. 1d) as those for which the primate performance was at least 1.5 d' units greater than the DCNN performance. Four examples of challenge and control images are shown in Fig. 1e. The challenge images were not idiosyncratic to our choice of AlexNet (fc7) (Supplementary Fig. 1b), specific objects (Supplementary Fig. 2), or our synthetic image-generation procedure (Supplementary Fig. 3a).

Our results show that on average, both macaques and humans outperform AlexNet. We identified two groups of images: 266 challenge images and 149 control images. On visual inspection, we did not observe any specific image property that differentiated between these two groups of images. We also did not observe any difference in performance on these two image sets after the monkeys were repeatedly exposed to these images (Supplementary Fig. 4a). This result is consistent with earlier work[9] that showed that once the monkeys are trained with images of specific objects, their generalization performance to new images from the same generative space is very high and consistent with that of the training images. However, we observed that the reaction times (RTs) for both humans and macaques for challenge images were significantly higher than for the control images (monkeys: $\Delta RT = 11.9$ ms, unpaired two-sample t-test, $t(413) = 3.4$, $P < 0.0001$; humans: $\Delta RT = 25$ ms, unpaired two-sample t-test, $t(413) = 7.52$, $P < 0.0001$), suggesting that additional processing time is required for the challenge images.

Temporal evolution of image-by-image object representation in the IT cortex. Previous studies[4,21] have shown that the identity of an object in an image is often accurately conveyed in the population activity patterns of the IT cortex in the macaque. Specifically, appropriately weighted linear combinations of the activities of IT neurons can approximate how neurons in downstream brain regions could integrate this information to form a decision about the object identity. In this study, we aimed to compare these linear object decodes from the IT cortex for the challenge and control images. First, we wanted to know whether these IT object decoders were as accurate as the primates for both types of images as predicted by the leading IT decoding model[5]. This test would demonstrate whether the ventral stream successfully solves the challenge images. Second, we reasoned that if challenge image solutions required recurrent computationally driven additional processing time, then IT object decodes for the challenge images should emerge later in the IT cortex compared with the control images. To this end, we used a sliding decoding time window (10 ms) that was narrower than that of prior work[5] so that we could precisely probe the temporal dynamics of linearly decodable object category information.

To estimate the temporal evolution of the IT object decodes for each image, we used large-scale multielectrode array recordings (Fig. 2a) across the IT cortex (424 valid IT sites) in two macaques.

To determine the time at which explicit object identity representations are sufficiently formed in the IT cortex, we estimated the temporal trajectory of the IT object decode accuracy for each image. We computed the neural decoding accuracies (NDAs) per time bin (10 ms) by training and testing linear classifiers per object independently at each time bin (see Methods). Consistent with prior work[21], we observed that the linearly available information is not the same at each time bin; for example, decoders trained at early time bins (~100–130) do not generalize to late time bins (Supplementary Fig. 5). Thus, we determined the time at which the NDA measured for each image reached the level of the behavioral accuracy of each subject (pooled monkey) (see Methods; Fig. 2a, upper panel). We termed this time the object solution time (OST), and we emphasize that each image has a potentially unique solution time (OST_{image}; see examples in Fig. 2b). We also observed that the OSTs estimated by randomly subsampling half ($n=212$) the total number of sites were significantly correlated (Spearman R scores of 0.77 and 0.76 for control and challenge images, respectively, $P < 0.00001$; and ΔOST was maintained at ~30 ms) with the OSTs from the total number of sites ($n=424$).

Figure 2b shows the temporal evolution of the IT object decode (for the object bear) and the OST estimates for two control and two challenge images. Two observations are apparent in these examples. First, for both the control and the challenge images, the accuracy of the IT decodes become equal to the behavioral accuracy of the monkeys at some time point after the image onset. Second, the IT decode solutions for challenge images emerge slightly later than the solutions for the control images.

Both of these observations were also found on average in the full sets of challenge and control images. First, IT decodes achieved primate behavioral levels of accuracy on average for the challenge and control image sets (~91% of challenge and ~97% of control images). Second, and consistent with our hypothesis, we observed that IT OST_{image} values for the challenge images were on average ~30 ms later compared with the control images. Specifically, the median OST for the challenge images was 145 ± 1.4 ms (median $\pm$ s.e.) from stimulus onset, and for the control images the OST was 115 ± 1.4 ms (median $\pm$ s.e.) (Fig. 2c). The average difference (~30 ms) between the OSTs of the challenge and control images did not depend on our choice of behavioral accuracy levels (Supplementary Fig. 6a) or the type of image set (Supplementary Fig. 3b).

These results are consistent with the hypothesis that recurrent computations are critical to core object recognition (see Introduction). Thus, we next carried out a series of controls to rule out alternative explanations for these results.

Controls for initial visual drive, individual neuron-based differences, and low-level image properties. We considered the possibility that the observed OST lag for the challenge images might have been due to the IT neurons taking longer to start responding to these images; for example, if the information took longer to be transmitted by the retina. However, we observed that control and challenge images share the same population neural-onset response latencies. That is, the difference in the IT response onset latency was only 0.17 ± 0.21 ms (median $\pm$ s.e.; paired t-test, $t(423) = 0.3896$, $P = 0.69$) (Fig. 3a; Supplementary Fig. 6b), suggesting that the initial visual drive in both image sets arrive at approximately the same time in the IT cortex. We also simultaneously recorded from area V4 (upstream of the IT cortex) in the left (95 sites) and right (56 sites) hemispheres of monkey M and N, respectively, and found no significant difference in the response latencies (both onset and peak) between control and challenge images across the V4 sites (paired t-test; $t(150) = 0.2$, $P = 0.8$) (Supplementary Fig. 7). These results further support the hypothesis that the ΔOST between the challenge and the control images in the IT cortex is not driven by

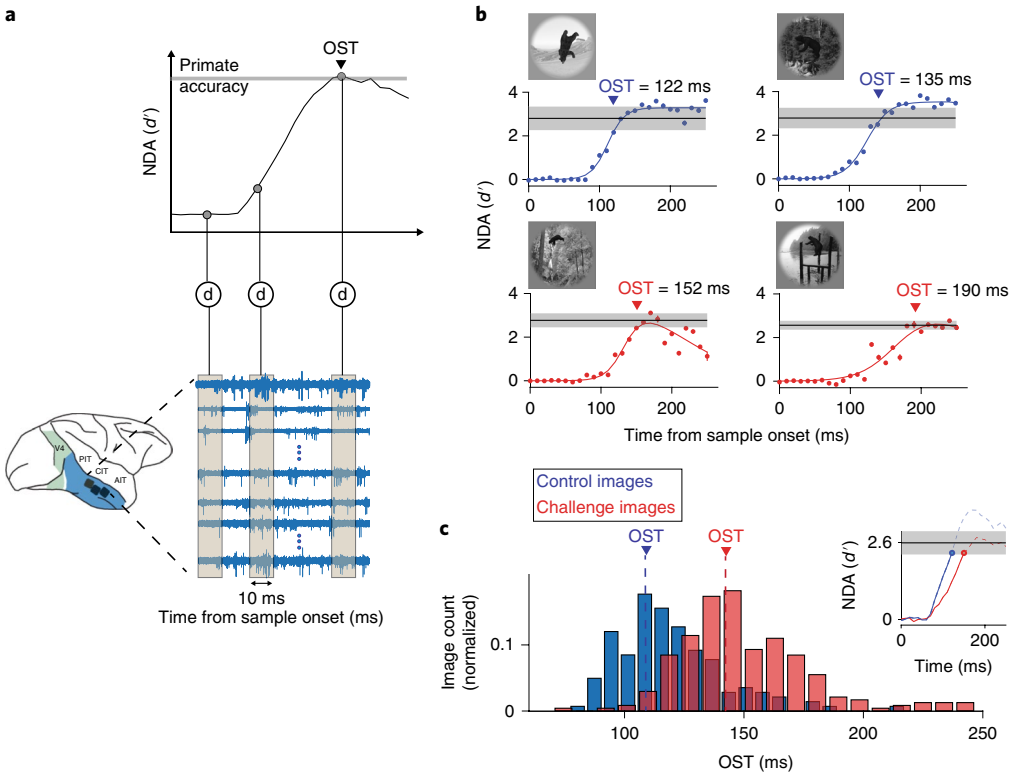


Fig. 2 | Large-scale multiunit array recordings in the macaque IT cortex. a Schematic of array placement, neural data recording and OST estimation. We recorded extracellular voltage in the IT cortex (across the posterior, central and anterior inferior temporal cortices; PIT, CIT and AIT, respectively) from two monkeys, each hemisphere implanted with two or three Utah arrays. For each image presentation (100 ms), we counted multiunit spike events (see Methods for details), per site, in non-overlapping 10-ms windows, post stimulus onset, to construct a single population activity vector per time bin. These population vectors (image-evoked neural features) were then used to train and test cross-validated linear SVM decoders (d) separately per time bin. The decoder outputs per image (over time) were then used to perform a binary match to sample task and obtain NDAs at each time bin. An example of the neural decode accuracy over time is shown in the upper panel. The time at which the neural decodes equal the primate (monkey) performance is then recorded as the OST for that specific image. **b**, Examples of IT population decodes over time, with the estimated OSTs for four images: two control (top) and two challenge images (bottom). The red and blue circles represent the estimated neural decode accuracies at each time bins. The unbroken lines are nonlinear fits of the decoder accuracies over time (see Methods). The gray lines indicate the I_1 performance of the primates (pooled monkey) for the specific images. Error bar indicates the bootstrapped s.e.m. **c**, Distribution of OSTs for control and challenge images. The median OSTs for control and challenge images are shown in the plot with broken lines. The inset on the top right shows the median evolution of IT decodes over time until the OST for control and challenge images.

image properties that evoke shorter latencies for control images at lower levels of the visual system.

When we closely examined the neural population response latencies for each image, we found that the time at which the IT population firing rates started to increase from baseline (onset latency; t_{onset}) and when the population firing rate reached its peak (t_{peak}) were on average earlier than the OST for the images (Fig. 3b,c). We also found no correlation (Pearson $r=0.009$, $P=0.8$) between the population response onset latency for each image (see Methods) and the OST for that image (Fig. 3d). For example, inspection of Fig. 3d reveals that some of the challenge images evoked faster-than-average latency responses in the IT cortex, yet have slow OSTs (~200 ms). Conversely, some of the control images evoked slower-than-average IT responses, yet have relatively fast OSTs (~110 ms). Interestingly, however, we found that firing rates (R) were significantly higher ($\Delta R = 17.3\%$; paired t-test, $t(423)=6.8848$, $P<0.0001$) for challenge images compared with control images (30 ms window centered at 150 ms post-stimuli onset) (Fig. 3a). One possible explanation for this result could be the effect of additional inputs from activated recurrent circuits into the IT neural sites at later time points (see Discussion). Regardless, these observations show that the challenge images drive IT neurons just as quickly and at least as strongly as the control images.

We considered the possibility that ΔOST between control and challenge images for each object category is primarily driven by neurons that specifically prefer that category (object-relevant neurons). To address this, we first tested whether the object-relevant neurons show a significant difference in response latency (that is, Δt_{onset} (challenge – control image) > 0) when measured for their preferred object category. Our results (Supplementary Fig. 8) showed that Δt_{onset} was not significant for any object category. In fact, a closer inspection (upper panel of Supplementary Fig. 8c) revealed that for some objects (for example, bear, elephant, and dog) Δt_{onset} was negative, indicating a trend for slightly shorter response latency for challenge images. Finally, to test the possibility that there was an overall trend for the most selective neurons to show a significant Δt_{onset}, we computed the correlation between Δt_{onset} and individual object selectivity per neuron, per object category. We observed (lower panel of Supplementary Fig. 8c) that there was no dependence of object selectivity per neuron on the response latency differences. In summary, the later mean OST for challenge images cannot be simply explained by longer response latencies of IT neurons that 'care' about the object categories.

From previous research, we know that temporal properties of IT neurons depend critically on low-level image features such as total image contrast energy[22], spatial frequency power distribution[23], and

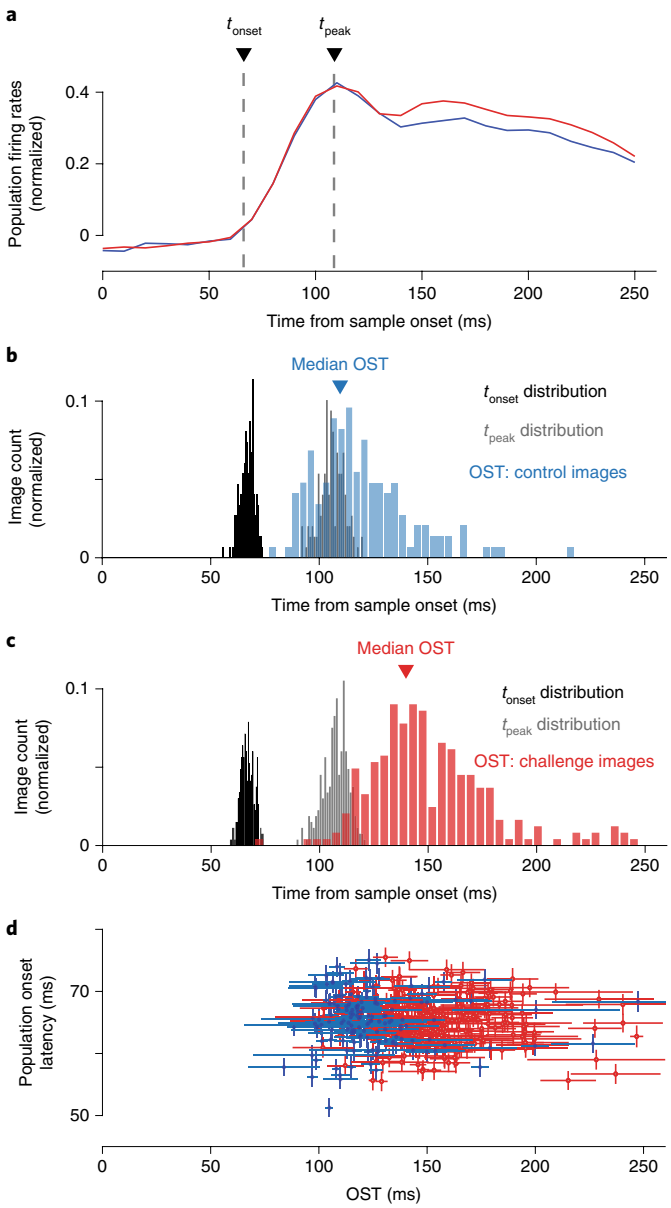


Fig. 3 | Relationship between OSTs and neural response latencies. **a** Comparison of neural responses evoked by control (blue) and challenge (red) images. We estimated two measures of population response latency: population onset latency (t_{onset}) and population peak latency (t_{peak}). **b**, Distributions of the population onset latencies (median across 424 sites), population peak response latencies (median across 424 sites) and OSTs for control images ($n=149$). **c**, Same as in **b** but for challenge images ($n=266$). **d**, Comparison of population onset latencies and OSTs for both control (blue; $n=149$ images) and challenge images (red; $n=266$ images). Vertical error bars show s.e.m. across neurons, and horizontal error bars show bootstrap (across trial repetition) standard deviations of OST estimates.

location of the visual objects[24]. So, we tested whether these low-level explanations might explain the longer OSTs of the challenge image. First, we observed that OSTs were not significantly correlated with image contrast (Spearman $\rho=-0.04$, $P=0.47$). Second, we used the SHINE (spectrum, histogram, and intensity normalization and equalization) (Supplementary Fig. 6c) technique[25] to equate low-level image properties across the control and challenge sets of images, and re-ran the recording experiment (subsampling 118 images each from the control and challenge image sets, with 44 repetitions per image). The average estimated difference in OST values between the challenge and control images after applying SHINE was still ~24 ms (Supplementary Fig. 6d). Third, we tested whether ΔOST (challenge – control) was specific to certain low or high values of various image-based properties (for example, image clutter, blur, contrast, object size, and object eccentricity; for definitions, see Methods). We observed that although certain image properties were significantly correlated with the absolute OST values, ΔOST was consistently ~30 ms at different levels of these factors (Supplementary Fig. 8d–h).

To test whether ΔOST (challenge – control) depends on neurons with higher or lower absolute latencies, we divided the neural population into the following two groups: low latencies (< 25 percentile of the neural latencies; $n=67$) and high latencies (> 75 percentile of all neural latencies; $n=67$). We found that both neural groups conveyed similar information about the two types of images. Specifically, we observed that there was no significant difference between the control and challenge image decoding accuracies estimated at the OST of each image for both the low and high latency populations (median $d'^{control}_{high-latency}=1.23$, $d'^{challenge}_{high-latency}=1.3$, $d'^{control}_{low-latency}=1.05$, $d'^{challenge}_{low-latency}=1.04$; unpaired t-test for high latency group, $t(388)=0.17$, $P=0.86$; unpaired t-test for low latency group, $t(388)=1.2$, $P=0.2$). Consistent with our main result, we also found that the low latency group of neurons and the high latency group of neurons each showed a positive lag for the decoding of the challenge images relative to the control images ($\Delta\text{Decode latency}^{low}_{th=1.0}=\sim22$ ms, $\Delta\text{Decode latency}^{high}_{th=1.0}=\sim18$ ms; note that we set a decoding threshold of 1.0 to compensate for the smaller number of neurons relative to the ~400 needed to achieve monkey behavioral d').

Object solution estimates during passive viewing. To test whether the late-emerging object solutions in the IT cortex only emerge when the animal is actively performing the task, we also recorded IT population activity during passive viewing of all the images. Monkeys fixated on a circle while images were each presented for 100 ms followed by 100 ms of no image, followed by the next image for 100 ms, and so on, until reward (typically 5 images were presented per fixation trial; see Methods).

First, similar to the active condition, we observed that the challenge images evoked a significantly higher firing rate (ΔR = 13.2%, paired t-test; $t(423)=8.27$, $P<0.0001$) at later time points (30 ms window centered at 150 ms post-stimuli onset) compared with the control images (Supplementary Fig. 9a). Second, we observed that we could successfully estimate the OSTs for 92% of the challenge images and 98% of the control images. The OSTs estimated during the active and passive conditions were also strongly correlated (Spearman $\rho=0.76$, $P<0.0001$). Similar to the active condition, challenge image solutions required an additional time of ~28 ms (on average) to achieve full solution compared with the control images (Supplementary Fig. 9b). Taken together, this suggests that the putative recurrent computations that underlie the late-emerging IT solutions are not task-dependent but are instead automatically triggered by the images. This is consistent with previous findings[26]. Similar results have also been reported in humans[27].

However, because these animals were trained on the object discrimination task, the OST difference might be due to internal processes that are only activated in trained monkeys (for example, mental task performance) or somehow due to the training history. To test this, we performed the same analyses on smaller sets of previously published data from two untrained animals[6–8]. To appropriately compare the results from the trained monkeys, we matched the set of common images (640), array implant locations, number of neural sites (168), and number of image repetitions (43). We observed a small but significant overall decrease in IT-based

decoding accuracy across all images in the untrained monkey (paired t-test; median $\Delta d' = 0.23$, $t(639) = 7.78$, $P < 0.0001$). Most importantly, however, similar to trained monkeys, we found that the IT cortex of untrained monkeys demonstrated lagged decode solutions for the challenge images relative to the control images (estimated at a decoding accuracy threshold of 1.8; $\Delta\text{Decode latency}^{\text{untrained}}_{th=1.8} = \sim 34$ ms, $\Delta\text{Decode latency}^{\text{trained}}_{th=1.8} = \sim 30$ ms) (Supplementary Fig. 10). In summary, our main experimental observation (lagged OST for challenge images) appears to be largely automatic and it does not require, and is not the result of, laboratory training.

IT predictivity across time using current feedforward deep neural network models of the ventral stream. We reasoned that if the late-emerging IT solutions are indeed dependent on recurrent computations, then perhaps the previously demonstrated ability of feedforward DCNNs to (partially) predict individual IT neurons[7] was mostly due to the similarity of the DCNN activations to the feedforward portion of the IT population response. To test this idea, we asked how well the DCNN features (which are not temporally evolving) could predict the time-evolving IT population response pattern up to and including the OST of each image. To do this, we used previously described methods (similar to ref. [8]). Specifically, we quantified the IT population goodness of fit as the median (over neurons) of the noise-corrected explained response variance score (IT predictivity) (Supplementary Fig. 11a).

First, we observed that the fc7 layer of AlexNet predicted $44.3 \pm 0.7\%$ of the explainable IT neural response variance (percentage EV) during the early response phase (90–110 ms; Fig. 4a). This result further confirms that feedforward DCNNs indeed approximate the initial (putative largely feedforward) IT population response. However, we observed that the ability of this DCNN to predict the IT population pattern significantly worsened (<20% EV) as that response pattern evolved over time (Fig. 4a). This drop in IT predictivity was not due to a low signal-to-noise ratio (SNR) of the neural responses during those time points. This is because our percentage EV measure already compensates for any changes in the SNR, and because the SNR remained relatively high in the late part of the IT responses (Supplementary Fig. 12). This gradual drop in IT predictivity of feedforward DCNNs is consistent with the hypothesis that late-phase IT population responses are modified by the action of recurrent circuits. Consistent with our hypothesis that challenge images rely more strongly on those recurrent circuits than control images, we observed that the drop in IT predictivity coincided with the solution times of the challenge images (refer to the OST distributions of challenge and control images in the upper panel of Fig. 4a).

Evaluation of deeper CNNs as models of ventral visual stream processing. It is understood in the artificial neural network community that finite-time recurrent neural networks can be constructed as very deep, feedforward-only neural networks with weight sharing across layers that are recurrently connected in the original recurrent network[28]. We reasoned that the actions of recurrent circuits in the ventral stream might be computationally equivalent to stacking further nonlinear transformations onto the initially evoked (~feedforward) IT population response pattern. To test this idea, we determined whether existing very-deep CNNs[29] (that outperform AlexNet) provide a better neural match to the IT response at its late phase. Based on the number of layers (nonlinear transformations), we divided the tested DCNN models into the following two groups: deep (eight layers; AlexNet, Zeiler and Fergus model, and VGG-S) and deeper (>20 layers, inception-v3[30], inception-v4[31], ResNet-50[32], and ResNet-101[32]) CNNs. We made three observations that corroborate our speculation.

First, we observed that the model IT layers (the layer with the highest behavioral (I_1) consistency to that of primates) of deeper CNNs predicted IT neural responses at the late phases (150–250 ms) significantly higher ($\Delta\text{Predictivity} = 5.8\%$, paired t-test; $t(423) = 14.26$, $P < 0.0001$) than 'regular-deep' models such as AlexNet (Fig. 4b; scatter plot comparisons with AlexNet shown separately in Supplementary Fig. 11b). This observation suggests that deeper CNNs might indeed be approximating 'unrolled' versions of the recurrent circuits of the ventral stream. Second, as expected from the ImageNet challenge results[33], we observed an increased performance and therefore reduced number of challenge images for deeper CNNs. Third, we found that the images that remain unsolved by these deeper CNNs (that is, challenge images for these models) showed even longer OSTs in the IT cortex than the original full set of challenge images (Fig. 4c). Assuming that a longer OST is a signature of more recurrent computations, this suggests that the newer, deeper CNNs have implicitly, but only partially, approximated—in a feedforward network—some of the computations that the ventral stream implements recurrently to solve some of the challenge images.

Evaluation of CORnet (a regular deep-recurrent CNN) as a model of the ventral visual stream. To more directly determine whether the experimental observations above might indeed be the result of recurrent computations, we directly tested a four-layered recurrent neural network model, termed CORnet[34]. The IT layer of CORnet has within-area recurrent connections (with shared weights). The model currently implements five time-steps (pass 1 to pass 5; Fig. 4b). The activity arising at the first time-step in the model IT layer is nonlinearly transformed to arrive at the output of the second time step, and so on. Indeed, we observed that CORnet had higher IT predictivity (Fig. 4c) for the late-phase of responses. In addition, pass 1 and pass 2 (corresponding to time-step 1) of the network had a significant (multiple comparison-corrected paired t-test; $t(423) = 12.78$, $P < 0.00001$) lower IT predictivity than pass 3 and pass 4 for later time-steps, whereas the opposite was true for earlier time-steps (Supplementary Fig. 13). Taken together, these results further argue for recurrent computations in the ventral stream.

Comparison of backward visual masking between challenge and control images. Based on our results so far, we hypothesized that the late IT population responses are critical for successful core object recognition behavior for many of the challenge images (~57% of challenge images have OSTs of >140 ms). To further test this idea, we performed an additional experiment. We modified the object discrimination paradigm by adding a visual mask (phase-scrambled image[35]) for 500 ms (Fig. 5a), immediately following the test image presentation. Such backward masking has previously been associated with selective disruption of recurrent inputs to an area[36], limiting the visual processing to the initial feedforward response[37]. We reasoned that such visual mask-based disruptions will produce larger behavioral deficits for challenge images compared with control images at earlier times. However, these differences should subside at longer presentation times when enough time is provided for the recurrent processes to build a sufficient object representation for both control and challenge images in the IT cortex. Therefore, we tested a range (34, 67, 100, 167, and 267 ms) of masking disruption times by randomly interleaving the sample image duration (and thus the mask onset). Our results (Fig. 5b) showed that visual masking indeed had a significantly stronger effect on the challenge images at smaller presentation durations compared with the control images. Consistent with our hypothesis, we did not observe any measurable masking differences between the two image sets at longer presentation times (~267 ms). Median $\Delta d'$ (difference between control and challenge images grouped by objects) averaged across all 10 objects were 0.5, 0.81, 0.33, 0.40, and –0.02 for 34, 67, 100, 167, and 267-ms presentation durations, respectively. The difference in performance was significant at the 0.05 significance level

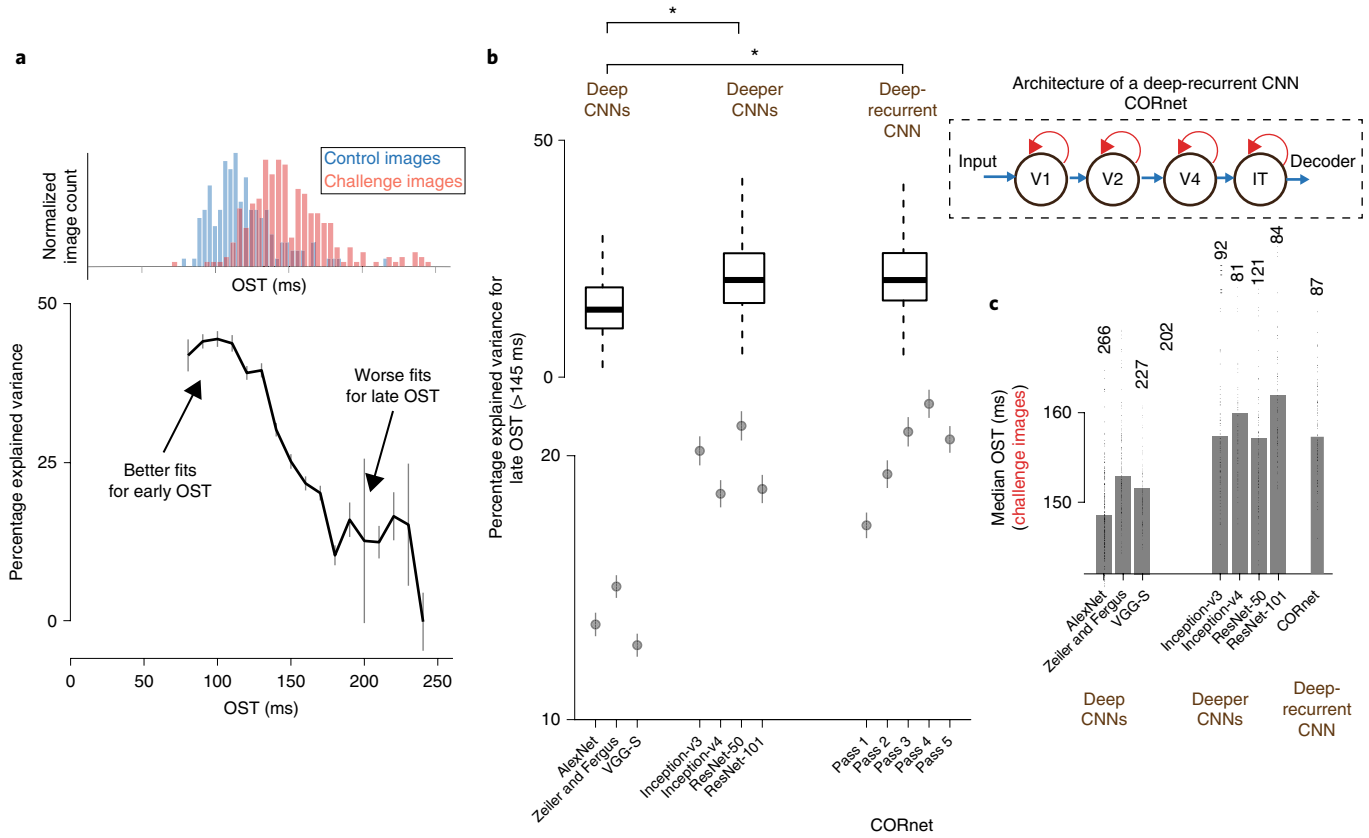


Fig. 4 | Predicting IT neural responses with DCNN features. a, IT predictivity of AlexNet's fc7 layer as a function of OST. For each time bin, we considered IT predictivity only for images that have a solution time equal to or higher than that time bin. Error bars indicate s.e.m. across neurons ($n = 424$ neurons considered for each time bin). Top: the distribution of OSTs for control ($n = 149$ images) and challenge ($n = 266$ images) images. **b**, IT predictivity computed separately for late OST images (OST > 150 ms, total of 349 images, $n = 424$ neurons) at the corresponding OSTs as a function of deep CNNs (AlexNet, Zeiler and Fergus, and VGG-S), deeper CNNs (Inception and ResNet) and deep-recurrent CNNs (CORnet). Circles indicate medians, and error bars indicate s.e.m. across neurons. Asterisks indicate a significant difference between two groups (obtained using paired t-tests). Deep (average of all three networks used) versus deeper CNNs (average of all four networks used): $t(423) = 14.26$, $P < 0.0001$. Deep (average of all three networks used) versus deep-recurrent (average of pass 3 and pass 4) CNNs: $t(423) = 15.13$, $P < 0.0001$. The inset to the right shows a schematic representation of CORnet that has recurrent connections (shown in red) at each layer (V1, V2, V4 and IT). For the boxplots, on each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers (W) extend to the most extreme data points that the algorithm considers not to be outliers. Outliers are data points that are larger than Q3 + W × (Q3 − Q1) or smaller than Q1 − W × (Q3 − Q1), where Q1 and Q3 are the 25th and 75th percentiles, respectively. Asterisk indicates a significant difference between two groups. **c**, Comparison of median OSTs for different sets of challenge images. The set of challenge images is defined with respect to each DCNN model. Thus, the exact set of images is different for each bar, the number of images is indicated on top of each bar, and the OST per image is plotted as a circle around each bar. In each case, the challenge images are defined as the set of images that remain unsolved by each model (using the fixed definitions of this study; see main text). Note that the use of deeper CNNs and the deep-recurrent CNN resulted in the discovery of challenge images that required even longer OSTs in the IT cortex than the original set challenge images (defined for AlexNet fc7).

(Bonferroni-adjusted) for all presentation durations except for 267 ms. Together with the neurophysiology results, these observations provide converging evidence that rapid, recurrent ventral stream computations are critical to the ability of the brain to infer object identity in the challenge images.

Model-driven versus image property-driven approaches to study recurrence. Previous research has suggested that recurrent computations in the ventral stream might be necessary to achieve pattern completion when exposed to occluded images[38,39], object-based attention in cluttered scenes[40,41], among others. Indeed, we observed that several image properties such as object size, presence of occlusion, and object eccentricity, as well as a combination of all these factors (Fig. 6), were significant, but very weak, predictors of our putative recurrence signal (the OST vector; see the "Estimation of the OST prediction strength" section in the Methods). In comparison, the performance gap between AlexNet and the monkeys ($\Delta d'$) was a significantly stronger predictor of OST. Therefore, our results suggest another possible image-wise predictor of ventral stream recurrence; that is, the difference in performance between feedforward DCNNs and primates, d'. This vector is probably itself dependent on a complex combination of image properties, such as those mentioned above. However, it is directly computable and our results show that it can serve as a much better model guide. In particular, we found that $\Delta d'$ was significantly predictive of the OST for each image (Spearman $\rho = 0.44$, $P < 0.001$), and, in this sense, is a much better predictor of the engagement of ventral stream recurrence than any of the individual image properties.

Discussion

The overall goal of this study was to determine whether recurrent circuits are critical to the execution of core recognition behavior in the ventral stream. We reasoned that if computations mediated by recurrent circuits are critical for some images, then one way to discover such images is by screening images that are difficult for non-recurrent DCNNs but are nevertheless easily solved by

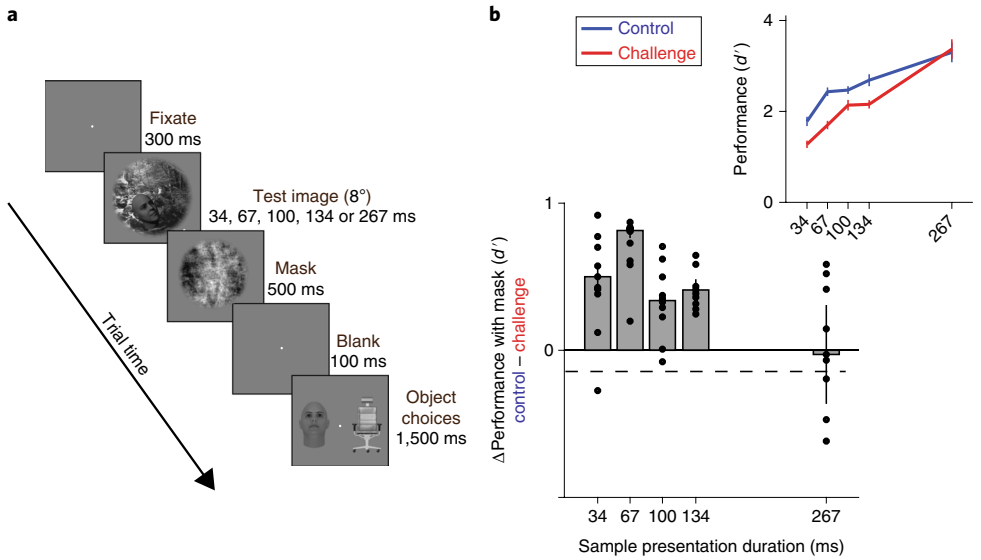


Fig. 5 | Comparison of backward visual masking between challenge and control images. a, Binary object discrimination with backward visual masking. The test image (presented for 34, 67, 100, 134 or 267 ms) was followed immediately by a visual mask (phase-scrambled image) for 500 ms, followed by a blank gray screen for 100 ms, and then the object choice screen. Monkeys reported the target object by fixating it on the choice screen. **b**, Difference in behavioral performance between control and challenge images after backward visual masking. Each bar on the plot (*y* axis) is the difference in the pooled monkey performance during the visual masking task between the control and challenge images at the respective sample image presentation durations (*x* axis). The broken black line denotes the difference in performance between the control and challenge images without backward masking at the 100-ms presentation; $n = 10$ objects considered per presentation duration. Each circle corresponds to the difference in performance per object. The upper panel inset shows the raw performance (d') for the two groups of images. Error bars denote s.e.m. across all objects.

primates. With these in hand, we aimed to look for a likely empirical signature of recurrence; that is, the requirement of additional time to complete successful processing. Large-scale neurophysiology, along with the precise estimation of the temporal evolution of the IT object identity solutions, revealed a key observation not revealed in prior work[5]. The IT solutions were lagged by on average ~30 ms for challenge images compared with the control images. In addition, we found that the late-phase IT response patterns that contained the linearly decodable object identity solutions were poorly predicted by the DCNN activations. Notably, we observed both of these findings during active task performance and passive viewing of the same images. Taken together, these results imply that automatically evoked recurrent circuits are critical for object identification behavior even at these fast timescales.

While the potential role of feedback in vision[42,43] has been previously suggested and partly explored, we believe that this is the first work to examine these questions at such large-scale and at the fast time scales of core object recognition, the first to do so using image computable models of neural processing to guide the choice of experiments (that is, the images and tasks), and the first to do so with an implemented linking model (decoder) of how the IT cortex supports recognition behavior.

Late object identity solution times in the IT complex imply that recurrent computations underlie core recognition. The most parsimonious interpretation of our results is that the late phases of the stimulus-evoked IT responses depend on recurrent computations. Our comparisons with behavior suggest that these IT dynamics are not epiphenomenal but are critical to core object recognition. But what kind (or kinds) of additional computations are taking place, and where in the brain do those recurrent circuits live? We can speculate to generate a testable set of hypotheses. Based on the number of synapses between the V1 and the IT cortex, is has been proposed[44] that the ventral stream comprises stages that are approximately 10–15 ms away from each other. Our observation of an additional processing time of ~30 ms for challenge images is therefore equivalent to at least two additional processing stages. Thus, one possible hypothesis is a cortico-cortical recurrent pathway between ventral stream cortical areas including the IT cortex and lower areas such as V4, V2, and V1 (similar to previous suggestions[45]). This possibility is consistent with observations of temporally specific effects in the response dynamics of V4 neurons[46] for images with occlusion. Alternatively, it is possible that the IT cortex is receiving important recurrent flow from downstream areas such as the prefrontal and perirhinal cortices (as previously suggested[47,48]). We also cannot rule out the possibility that all of the additional computations are due to recurrence within the IT cortex (consistent with recent models[39]) or due to subcortical circuits (for example, basal ganglia loops[49]). These hypotheses are not mutually exclusive. Given these prior work, the main contribution of our work is to take the very broad notion of feedback and pin down a narrower case that is both experimentally tractable and is guaranteed to have high behavioral relevance. The current results now motivate the need for direct perturbation studies that aim to independently suppress each of those circuit motifs to assess their relative importance. The estimated OST vector (putative recurrence signal) predicts exactly which individual images (that is, the images requiring longer solution times) will be most affected by a targeted disruption of the relevant recurrent circuits. This knowledge can be used to optimize the image sets and behavioral tasks for these next experiments.

Temporally specific failures of feedforward DCNNs imply the need to add recurrent circuits to improve those models. Prior studies[6,7] have demonstrated that feedforward DCNNs (for example, HMO[7], AlexNet[19], and VGG[42,50]) can explain ~50% of the within-animal explainable response variance in stimulus-evoked V4 and IT responses. Our results confirm that feedforward DCNNs indeed approximate ~50% of the first 30 ms (~90–120 ms) of the IT response variance, thus establishing DCNNs as a good functional approximation of the feedforward pass of the primate ventral stream. However, the ability of DCNNs to predict IT responses dropped significantly at later time-points (>150 ms post-image

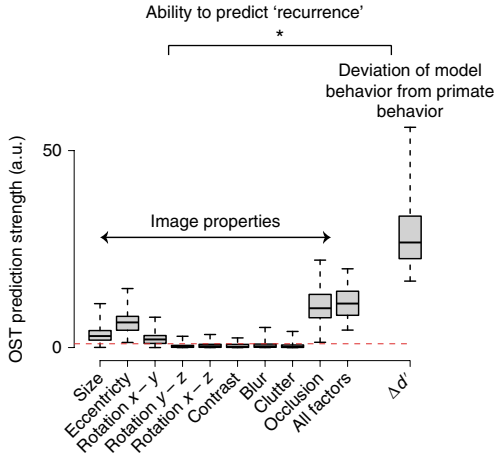


Fig. 6 | Comparison of OST prediction strength. Comparisons were made between different image properties, a combination of all estimated image properties, and the Δ*d*' vector (deviation of model behavior from pooled monkey behavior). Different image properties ($n = 64$ total: 32 high, 32 low; refer to Methods) for each image group was used. The red broken line denotes the significance threshold of the *F*-statistic. Image properties such as object size, eccentricity, presence of an occluder and a combination of these properties (referred to as All factors) significantly predict OST. However, the Δ*d*' vector provided the strongest OST predictions. Error bars denote bootstrap standard deviations over images. The asterisk denotes a significant difference between the two groups, image properties versus Δ*d*', estimated with repeated measures ANOVA ($F(1,10) > 100$, $P < 0.0001$; multiple-comparison using Turkey test showed a significant difference between Δ*d*' and all other image properties). For the boxplot, on each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points that the algorithm considers not to be outliers. Outliers are defined as in Fig. 4b.

onset; Fig. 4a). This is consistent with our inference that the late OSTs for challenge images are primarily caused by the recruitment of additional recurrent processing in the ventral stream.

Unique OSTs per image motivate the search for better decoding models. A previous study[6] showed that a simple linear decoding model, formed by linearly weighting the population activity of IT neurons (integrated from 70 ms to 170 ms post-image onset) was sufficient to predict the average performance of human subjects across 64 tested core object recognition tasks. Here, at a much finer (image-by-image) grain of testing, we observed that even for images that have statistically non-distinguishable levels of behavioral performance, the linearly decodable information in the IT population pattern varies substantially over the IT response time window used by these decoding models. Taken together, this argues that future work in this direction might successfully reject such fixed time integration-based decoding models, and thus drive the field to create better mechanistic neuronal-to-behavioral linking hypotheses.

Role of recurrent computations: deliverables from these data and insights from deeper CNNs. Prior studies have strongly associated the role of recurrent computations during visual object recognition with overcoming certain specific challenging image properties. These can be expressed as a single word or phrase such as 'occlusion'[45], high levels of 'clutter'[40], 'grouping' of behaviorally relevant image regions[43], or the need for visual 'pattern completion'[39]. While we agree that such image manipulations might recruit recurrent processes in the ventral stream, the current work argues that picking any one of these single ideas is not the most efficient approach to constrain future recurrent models of object recognition. Instead, we used the shallower models to find images for which the difference between feedforward-only DCNN and primate behavior (Δ*d*') is the largest. This difference was a better predictor of the neural phenomena of recurrence than any of the image-based properties (Fig. 6). We interpret this to mean that such image-computable models effectively embed knowledge about multiple interacting image properties that cannot be described by single words or phrases. Indeed, this knowledge better accounts for the what happens in the feedforward part of the response than those other types of explanations.

While this is a good way to focus experimental efforts, it does not yet explain the exact nature of the computational problem solved by recurrent circuits during core object recognition. Interestingly, we found that deeper CNNs such as inception-v3, v4[31], and ResNet-50,101[32], which introduce more nonlinear transformations to the image pixels compared to shallower networks such as AlexNet or VGG, are better models of the behaviorally critical late-phase of IT responses. In addition, a previous study[28] had demonstrated that a shallow recurrent neural network is equivalent to a very deep CNN (for example, ResNet) with weight sharing among the layers. Therefore, we speculate that what the computer vision community has achieved by stacking more layers into the CNNs is a partial approximation of something that is more efficiently built into the primate brain architecture in the form of recurrent circuits. That is, during core object recognition, recurrent computations act as additional nonlinear transformations of the initial feedforward IT response to produce more explicit (linearly separable) solutions. This provides a qualitative explanation for the role of recurrent computations during a variety of challenging image conditions. What is now needed are new recurrent artificial neural networks (here, we provided results from one such model, CORnet[34]) that successfully incorporated these ideas.

Constraints for future models provided by our data. Our results motivate a change in the architecture of artificial neural networks that aim to model the ventral visual stream (that is, a switch from largely feedforward DCNNs to recurrent DCNNs) However, experiments should not simply provide motivation but also validation and stronger constraints for guiding the construction of new models. Here, we first provide a behavioral vector Δ*d*′ that quantifies the performance gap between feedforward DCNNs (for example, AlexNet) and the image-by-image primate behavior I_1. Second, for each image, we have estimated the time at which object solutions are sufficiently represented in the macaque IT cortex (the $\mathrm{OST}_{\mathrm{image}}$ vector). Third, we have reliably measured neural responses to each tested image at their respective OST (potential target features for models). Next-generation dynamic ventral stream models should be constrained to produce the target features (object solutions) at these times.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-019-0392-5.

Received: 1 July 2018; Accepted: 21 March 2019;
Published online: 29 April 2019

References

1. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
2. Riesenhuber, M. & Poggio, T. Models of object recognition. *Nat. Neurosci.* **3**, 1199–1204 (2000).

3. Yamins, D. L. & DiCarlo, J. J. Eight open questions in the computational modeling of higher sensory cortex. *Curr. Opin. Neurobiol.* **37**, 114–120 (2016).
4. Hung, C. P., Kreiman, G., Poggio, T. & DiCarlo, J. J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).
5. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
6. Cadieu, C. F. et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10**, e1003963 (2014).
7. Yamins, D. L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
8. Guclu, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
9. Rajalingham, R., Schmidt, K. & DiCarlo, J. J. Comparison of object recognition behavior in human and monkey. *J. Neurosci.* **35**, 12127–12136 (2015).
10. Rajalingham, R. et al. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
11. Rockland, K. S. & Virga, A. Terminal arbors of individual "feedback" axons projecting from area V2 to V1 in the macaque monkey: a study using immunohistochemistry of anterogradely transported *Phaseolus vulgaris*-leucoagglutinin. *J. Comp. Neurol.* **285**, 54–72 (1989).
12. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
13. Rockland, K. S., Saleem, K. S. & Tanaka, K. Divergent feedback connections from areas V4 and TEO in the macaque. *Vis. Neurosci.* **11**, 579–600 (1994).
14. Rockland, K. S. & Van Hoesen, G. W. Direct temporal–occipital feedback connections to striate cortex (V1) in the macaque monkey. *Cereb. Cortex* **4**, 300–313 (1994).
15. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).
16. Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The "wake–sleep" algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).
17. Geirhos, R., et al. Comparing deep neural networks against humans: object recognition when the signal gets weaker. Preprint at *arXiv* https://arxiv.org/abs/1706.06969 (2017).
18. Lamme, V. A. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
19. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. 25th International Conference on Neural Information Processing Systems—Volume 1* (eds Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, 2012).
20. Lin, T.-Y., et al. Microsoft COCO: Common objects in context. In *Proc. 13th European Conference on Computer Vision* (eds Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) 740–755 (Springer, 2014).
21. Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K. & Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J. Neurophysiol.* **100**, 1407–1419 (2008).
22. Oram, M. W. Contrast induced changes in response latency depend on stimulus specificity. *J. Physiol. Paris* **104**, 167–175 (2010).
23. Rolls, E. T., Baylis, G. C. & Leonard, C. M. Role of low and high spatial frequencies in the face-selective responses of neurons in the cortex in the superior temporal sulcus in the monkey. *Vision Res.* **25**, 1021–1035 (1985).
24. Op De Beeck, H. & Vogels, R. Spatial sensitivity of macaque inferior temporal neurons. *J. Comp. Neurol.* **426**, 505–518 (2000).
25. Willenbockel, V. et al. Controlling low-level image properties: the SHINE toolbox. *Behav. Res. Methods* **42**, 671–684 (2010).
26. McKee, J. L., Riesenhuber, M., Miller, E. K. & Freedman, D. J. Task dependence of visual and category representations in prefrontal and inferior temporal cortices. *J. Neurosci.* **34**, 16065–16075 (2014).
27. Bugatus, L., Weiner, K. S. & Grill-Spector, K. Task alters category representations in prefrontal but not high-level visual cortex. *Neuroimage* **155**, 437–449 (2017).
28. Liao, Q. & Poggio, T. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. Preprint at *arXiv* https://arxiv.org/abs/1604.03640 (2016).
29. Schrimpf, M., et al. Brain-score: which artificial neural network for object recognition is most brain-like? Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/407007v1 (2018).
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. 29th IEEE Conference on Computer Vision and Pattern Recognition* (ed. IEEE Computer Society) 2818–2826 (IEEE Computer Society, 2016).
31. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proc. 31st AAAI Conference on Artificial Intelligence* (ed. AAAI) 4278–4284 (AAAI, 2017).
32. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. 29th IEEE Conference on Computer Vision and Pattern Recognition* (ed. IEEE Computer Society) 770–778 (IEEE Computer Society, 2016).
33. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. J. Comp. Vision* **115**, 211–252 (2015).
34. Kubilius, J., et al. CORnet: modeling the neural mechanisms of core object recognition. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/408385v1 (2018).
35. Stojanoski, B. & Cusack, R. Time to wave good-bye to phase scrambling: creating controlled scrambled images using diffeomorphic transformations. *J. Vis.* **14**, 6 (2014).
36. Fahrenfort, J. J., Scholte, H. S. & Lamme, V. A. Masking disrupts reentrant processing in human visual cortex. *J. Cogn. Neurosci.* **19**, 1488–1497 (2007).
37. Elsayed, G. F., et al. Adversarial examples that fool both human and computer vision. Preprint at *arXiv* https://arxiv.org/abs/1802.08195 (2018).
38. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 1551 (2017).
39. Tang, H. et al. Recurrent computations for visual pattern completion. *Proc. Natl Acad. Sci.* **115**, 8835–8840 (2018).
40. Walther, D., Rutishauser, U., Koch, C. & Perona, P. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Comp. Vis. Image Und.* **100**, 41–63 (2005).
41. Bichot, N. P., Heard, M. T., DeGennaro, E. M. & Desimone, R. A source for feature-based attention in the prefrontal cortex. *Neuron* **88**, 832–844 (2015).
42. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at *arXiv* https://arxiv.org/abs/1409.1556 (2014).
43. Jeurissen, D. & Self, M. W. & Roelfsema, P. R. Serial grouping of 2D-image regions with object-based attention in humans. *eLife* **5**, e14320 (2016).
44. Tovee, M. J. Neuronal processing. How fast is the speed of thought? *Curr. Biol.* **4**, 1125–1127 (1994).
45. van Kerkoerle, T. et al. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 14332–14341 (2014).
46. Fyall, A. M., El-Shamayleh, Y., Choi, H., Shea-Brown, E. & Pasupathy, A. Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. *eLife* **6**, e25784 (2017).
47. Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I. & Miyashita, Y. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* **401**, 699–703 (1999).
48. Bar, M. et al. Top-down facilitation of visual recognition. *Proc. Natl Acad. Sci. USA* **103**, 449–454 (2006).
49. Seger, C. A. How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neurosci. Biobehav. Rev.* **32**, 265–278 (2008).
50. Chatfield, K., Simonyan, K., Vedaldi, A. & Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. Preprint at *arXiv* https://arxiv.org/abs/1405.3531 (2014).

Acknowledgements

This research was primarily supported by the Office of Naval Research MURI-114407 (to J.J.D) and in part by US National Eye Institute grants R01-EY014970 (to J.J.D.) and K99-EY022671 (to E.B.I.), and the European Union's Horizon 2020 research and innovation programme under grant agreement no 705498 (to J.K.). This work was also supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. The authors thank A. Afraz for his surgical assistance.

Author contributions

K.K. and J.J.D. designed the experiments. K.K., K.S., and E.B.I. carried out the experiments. K.K. performed the data analyses. K.K. and J.K performed computational modeling. K.K. and J.J.D. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0392-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to K.K.

Journal peer review information *Nature Neuroscience* thanks Blake Richards, Pieter Roelfsema, and other anonymous reviewer(s) for their contribution to the peer review of this work.

Methods

Subjects. The nonhuman subjects in our experiments were two adult male rhesus monkeys (*Macaca mulatta*). All human studies were done in accordance with the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects. A total of 88 observers participated in the binary object discrimination task. Observers completed these 20–25-min tasks through Amazon Mechanical Turk (MTurk), an online platform in which subjects can complete experiments for a small payment.

Generation of visual stimuli. *Generation of synthetic ('naturalistic') images.* High-quality images of single objects were generated using free ray-tracing software (http://www.povray.org), similar to a previous study[6]. Each image consisted of a two-dimensional (2D) projection of a three-dimensional (3D) model (purchased from Dosch Design and TurboSquid) added to a random background. The ten objects chosen were bear, elephant, face, apple, car, dog, chair, plane, bird, and zebra (Fig. 1b). By varying six viewing parameters, we explored three types of identity while preserving object variation, position (*x* and *y*), rotation (*x*, *y*, and *z*), and size. All images were achromatic with a native resolution of 256×256 pixels (see Supplementary Fig. 1a for example images). A total of 1,120 naturalistic images (112 per object category) were used.

Generation of natural images (photographs). Images pertaining to the ten nouns were download from http://cocodataset.org. Each image was resized to 256×256×3 pixel size and presented within the central 8°. We used the same images while testing the feedforward DCNNs. A total of 200 COCO images (20 per object category) was used.

Quantification of image properties. We compared the ability of different image properties to predict the putative recurrence signal as inferred from our results. These image properties were either predefined during the image-generation process (for example, object size, object eccentricity, object rotation vectors, and presence of an object occluder) or computed after the image-generation procedure. The post image-generation properties are listed below.

Image contrast. This was defined as the variance of the luminance distribution per image (grayscale images only).

Image blur. The literature on image processing contains multiple measures of image focus based on first order differentiation or smoothing followed by differentiation. We used a previously published technique[51] to define the focus of an image.

Image clutter. This measure (Feature Congestion) of visual clutter is related to the local variability in certain key features, for example, color, contrast, and orientation[52].

Primate behavioral testing. *Humans tested via Amazon MTurk.* We measured human behavior (88 subjects) using the online Amazon MTurk platform, which enables efficient collection of large-scale psychophysical data from crowd-sourced human intelligence tasks. We did not collect information regarding the sex of the human subjects who performed the online MTurk tasks. The reliability of the online MTurk platform has previously been validated by comparing results obtained from online and in-lab psychophysical experiments[5,9]. Each trial started with a 100-ms presentation of the sample image (1 out of 1,360 images). This was followed by a blank gray screen for 100 ms followed by a choice screen with the target and distractor objects (similar to a previous study[10]). The subjects indicated their choice by touching the screen or clicking the mouse over the target object. Each subject saw an image only once. We collected the data such that there were 80 unique subject responses per image with varied distractor objects.

Monkeys tested during simultaneous electrophysiology. Active binary object discrimination task. We measured monkey behavior from two male rhesus macaques. Images were presented on a 24-inch LCD monitor (1,920×1,080 at 60 Hz) positioned 42.5 cm in front of the animal. Monkeys were head fixed. Monkeys fixated on a white circle (0.2°) for 300 ms to initiate a trial. The trial started with the presentation of a sample image (from a set of 1,360 images) for 100 ms. This was followed by a blank gray screen for 100 ms, after which the choice screen was shown containing a standard image of the target object (the correct choice) and a standard image of the distractor object. The monkey was allowed to freely view the choice objects for up to 1,500 ms, and indicated its final choice by holding fixation over the selected object for 400 ms. Trials were aborted if the gaze was not held within ±2° of the central fixation circle during any point until the choice screen was shown. Before the final behavioral testing, both monkeys were trained in their home cages on a touchscreen (for details see ref. [10]; details of the code and hardware are available at https://github.com/dicarlolab/mkturk) to perform the binary object discrimination tasks. We used a separate set of images that were synthesized using the same image-generation protocol to train the monkeys on the binary object discrimination task. Once monkeys are trained in the basic task paradigm, they readily learn each new object over full viewing and background transformations in just 1–2 days, and they easily generalize to completely new images of each learned object[9]. Once the behavioral performance stabilized during the training, we then tested the monkeys on the image set described in the manuscript along with simultaneous electrophysiology.

Passive viewing. During the passive viewing task, monkeys fixated on a white circle (0.2°) for 300 ms to initiate a trial. We then presented a sequence of 5–10 images, each one for 100 ms followed by a 100 ms gray (background) blank screen. This was followed by fluid reward and an inter-trial interval of 500 ms, followed by the next sequence. Trials were aborted if the gaze was not held within ±2° of the central fixation circle during any point.

Behavioral metrics. We used the same one-versus-all image level behavioral performance metric (I_1) to quantify the performance of the humans, monkeys, DCNNs and neural-based decoding models for the binary match sample tasks. This metric estimates the overall discriminability of each image containing a specific target object from all other objects (pooling across all nine possible distractor choices).

For example, given an image of object *i*, and all nine distractor objects ($j \neq i$) we first compute the average hit rate as follows:

$\text{Hitrate}^{i}_{\text{image}} = \frac{\sum_{j=1}^{10} Pc^{i,j\neq i}_{\text{image}}}{9}$, where *Pc* refers to the fraction of correct responses for the binary task between objects *i* and *j*. We then compute the false alarm rate for the object *i* as follows:

$$\text{Falsealarm}^{i} = 1 - \text{avg}(\text{Hitrate}^{j\neq i}_{\text{image}})$$

The unbiased behavioral performance, per image, was then computed using a sensitivity index *d'*, as follows:

$$d'_{\text{image}} = z(\text{Hitrate}^{i}_{\text{image}}) - z(\text{Falsealarm}^{i})$$

In this equation, *z* is the inverse of the cumulative Gaussian distribution. The values of *d'* were bounded between −5 and 5. Given the size of our image set, the I_1 vector contains 1,320 independent *d'* values. The estimated median false alarm rate across objects were 0.11 and 0.18 for the monkey behavior and neural decoding performance, respectively.

To compute the reliability of the estimated I_1 vector, we split the trials per image into two equal halves by resampling without substitution. The Spearman–Brown-corrected correlation of the two corresponding I_1 vectors (one from each split half) was used as the reliability score (that is, internal consistency) of our I_1 estimation.

Large-scale multielectrode recordings and simultaneous behavioral recording. *Surgical implant of chronic microelectrode arrays.* Before training, we surgically implanted each monkey with a headpost under aseptic conditions. After behavioral training, we recorded neural activity using 10×10 microelectrode arrays (Utah arrays, Blackrock Microsystems). A total of 96 electrodes were connected per array. Each electrode was 1.5-mm long and the distance between adjacent electrodes was 400 µm. Before recording, we implanted each monkey with multiple Utah arrays in the IT cortex and the V4 cortex. IT arrays were placed inferior to the superior temporal sulcus and anterior to the posterior middle temporal sulcus. In monkey M, we implanted three arrays in the right hemisphere (all three in the IT cortex) and three arrays in the left hemisphere (two in the IT cortex and one in the V4 cortex). In monkey N, we implanted three arrays in the left hemisphere (all three in the IT cortex) and three arrays in the right hemisphere (two in the IT cortex and one in the V4 cortex). In total, we recorded from 424 valid IT sites, which included 159 and 139 sites in the right hemisphere and 32 and 94 sites in the left hemisphere of monkey M (shown as inset in Fig. 2a) and monkey N, respectively. The left and right hemisphere arrays were not implanted simultaneously. We recorded for ~6–8 months from implants in one hemisphere before explanting the arrays and implanting new arrays in the opposite hemisphere. Array placements were guided by the sulcus pattern, which was visible during surgery. The electrodes were accessed through a percutaneous connector that allowed simultaneous recording from all 96 electrodes from each array. Behavioral testing was performed using standard operant conditioning (fluid reward), head stabilization, and real-time video eye tracking. All surgical and animal procedures were performed in accordance with National Institutes of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

Eye tracking. We monitored eye movements using video eye tracking (SR Research EyeLink 1000). Using operant conditioning and water reward, our two animals were trained to fixate on a central white circle (0.2°) within a square fixation window that ranged from ±2°. At the start of each behavioral session, monkeys performed an eye-tracking calibration task by making a saccade to a range of spatial targets and maintaining fixation for 500 ms. Calibration was repeated if drift was noticed over the course of the session.

Electrophysiological recording. During each recording session, band-pass filtered (0.1 Hz to 10 kHz) neural activity was recorded continuously at a sampling rate of 20 kHz using an Intan Recording Controller (Intan Technologies). The majority of the data presented here are based on multiunit activity. We detected the multiunit spikes after the raw data were collected. A multiunit spike event was defined as the threshold crossing when voltage (falling edge) deviated by more than three times the standard deviation of the raw voltage values. Of 960 implanted electrodes, five arrays (combined across the two hemispheres)×96 electrodes×two monkeys, we

focused on the 424 most visually driven, selective, and reliable neural sites. Our array placements allowed us to sample neural sites from different parts of the IT cortex along the posterior to anterior axis. However, for all the analyses, we did not consider the specific spatial location of the site, and treated each site as a random sample from a pooled IT population.

Neural recording quality metrics per site. Visual drive per neuron. We estimated the overall visual drive, d'_{visual}, for each electrode. This metric was estimated by comparing the COCO image responses of each site to a blank (gray screen) response as follows:

$$d'_{visual} = \frac{avg(R_{COCO}) - avg(R_{gray})}{\sqrt{\frac{1}{2}(\sigma^2_{R_{COCO}} + \sigma^2_{R_{gray}})}}$$

Image rank-order response reliability per neural site. To estimate the reliability of the responses per site ρ^{IRO}_{site}, we computed a Spearman–Brown-corrected, split half (trial-based) correlation between the rank order of the image responses (all images).

Selectivity per neural site. For each site, we measured selectivity as the d' for separating the best (highest response-driving) stimulus at that site from its worst (lowest response-driving) stimulus. d' was computed by comparing the response mean of the site over all trials on the best stimulus compared with the response mean of the site over all trials on the worst stimulus, and normalized by the square-root of the mean of the variances of the sites on the two stimuli as follows:

$$selectivity_i = \frac{mean(\vec{b_i}) - mean(\vec{w_i})}{\sqrt{\frac{var(\vec{b_i}) + var(\vec{w_i})}{2}}}$$

where $\vec{b_i}$ is the vector of responses of site i to its best stimulus over all trials and $\vec{w_i}$ is the vector of responses of site i to its worst stimulus. We computed this number in a cross-validated fashion, picking the best and worst stimulus on a subset of trials and then computing the selectivity measure on a separate set of trials, and averaging the selectivity value of 50 trial splits.

Inclusion criterion for neural sites. For our analyses, we only included the neural recording sites that had an overall significant visual drive (d'_{visual}), an image rank-order response reliability (ρ^{IRO}_{site}) that was greater than 0.6, and a selectivity score that was greater than 1. Given that most of our neural metrics are corrected by the estimated noise at each neural site, the criterion for selection of neural sites is not that critical. It was mostly done to reduce computation time and to eliminate noisy recordings.

Population neural response latency estimation. Onset latencies (t_{onset}) were determined as the earliest time from sample image onset when the firing rates of neurons were higher than one-tenth of the peak of its response. We averaged the latencies estimated across individual neural sites to compute the population latency.

Peak latencies (t_{peak}) were estimated as the time of maximum response (firing rate) of a neural site in response to an image. We averaged the peak latencies estimated across individual neural sites to compute the population peak latency per image.

Both of these latency measures were computed across different sets of images (control and challenge) as mentioned in the article.

Estimation of solution for object identity per image. *IT cortex.* To estimate what information downstream neurons could easily 'read' from a given IT neural population, we used a simple, biologically plausible linear decoder (that is, linear classifiers), that has been previously shown to link IT population activity and primate behavior[5]. Such decoders are simple in that they can perform binary classifications by computing weighted sums (each weight is analogous to the strength of synapse) of input features and separate the outputs based on a decision boundary (analogous to a spiking threshold of a neuron). Here, we used a support vector machine (SVM) algorithm with linear kernels. The SVM learning model generates a decoder with a decision boundary that is optimized to best separate images of the target object from images of the distractor objects. The optimization is done under a regularization constraint that limits the complexity of the boundary. We used L2 (ridge) regularization, whereby the objective function for the minimization comprises an additional term (to reduce model complexity), as follows:

$$L2(penalty) = \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2$$

where β and p are the classifier weights associated with p predictors (for example, 424 neurons). The strength of regularization, λ was optimized for each training set, and a stochastic gradient descent solver was used to estimate ten (one for each object) one-versus-all classifiers. After training each of these classifiers with a set of 100 training images per object, we generated a class score (sc) per classifier for all held-out test images given by the following:

$$sc = R\beta + \text{bias}$$

where R is the population response vector and the bias values are estimated by the SVM solver.

The training and test sets were pseudorandomly chosen multiple times until every image of our image set was part of the held-out test set. We then converted the class scores into probabilities by passing them through a softmax (normalized exponential) function, as follows:

$$P^i_{image} = \frac{e^{sc_i}}{\sum_{i=1}^{10} e^{sc_i}}$$

Our behavioral I_1 scores are all trial-averaged metrics. Therefore, to generate a comparable trial-averaged performance per image, a probability for each classifier output given any image (P^i_{image}) was generated. The decoders were therefore trained and tested with trial-averaged data.

We then computed the binary task performances by calculating the percentage correct score for each pair of possible binary task given an image. For instance, if an image was from object i, then the percentage correct score for the binary task between object i and object j, $Pr^{i,j}$ was computed as follows:

$$Pr^{i,j}_{image} = \frac{P^i_{image}}{P^i_{image} + P^j_{image}}$$

From each percentage correct score, we then estimated a neural I_1 score (per image), following the same procedures as the behavioral metric.

OST per image in the IT cortex. The OST per image, OST_{image}, was defined as the time it takes for linear IT population decodes to reach within the error margins of the pooled monkey behavioral I_1 score for that image. To estimate this time, we first computed a neural I_1 vector for nonoverlapping 10-ms time bins after the sample image onset. We then used linear interpolation to predict the value of the I_1 vector per image at any given time between 0 and 250 ms. We then used the Levenberg–Marquardt algorithm to estimate the time at which the neural I_1 vector reached the error margins of the pooled monkey behavioral I_1. Because we recorded many repetitions of each image, we were able to measure OST_{image} very accurately (standard error of ~9 ms on average, as determined via bootstrapping across repetitions).

We balanced the control and challenge image populations at each level of the performance of the monkeys. Therefore, we discarded challenge images that showed a d' of 5 or higher since there were no equivalent control images at that behavioral-accuracy level. However, we estimated the average OST for the challenge images at $d' \geq 5$ to be 150.2 ms (well within the range of other challenge image OSTs).

Binary object discrimination tasks with DCNNs. We used two different techniques to train and test the DCNN features on the binary object discrimination task.

Back-end training (transfer learning). Here, we used the same linear decoding scheme mentioned above (for the IT neurons) to estimate the object solution strengths per image for the DCNNs. Briefly, we first obtained an ImageNet pre-trained DCNN (for example, AlexNet). We then replaced the last three layers (that is, anything beyond fc7) of this network with a fully connected layer containing ten nodes (each representing one of the ten objects we used in this study). We then trained this last layer with a back-end classifier (L2-regularized linear SVM; similar to the one mentioned for the IT cortex) on a subset of images from our image set (containing both control and challenge images). These images were randomly selected from our image set and used as the training set. The remaining images were then used for testing (such that there was no overlap between the training and test images). Repeating this procedure multiple times allowed us to use all images as test images, providing us with the performance of the model for each image. The features extracted from each of the DCNN models were projected onto the first 1,000 principle components (ranked in the order of variance explained) to construct the final feature set used. This was done to maintain consistency while comparing different layers across various DCNNs (some include ~20,000 features) and to control for the total number of features used in the analyses.

Fine-tuning. Although the steps mentioned above (transfer learning) is more similar to how we think the monkey implements the learning of the task in his brain, we cannot completely rule out the possibility that the representations of the images in the IT cortex do not change after training with our image set. Prior work suggests that such IT population response changes are modest at best[53]. Therefore, we also fine-tuned (end-to-end) the ImageNet pre-trained AlexNet with images

(randomly selected from our own image set) and tested them on the remaining held-out images. This technique also involves first obtaining an ImageNet-pertained DCNN and replacing the final three layers (for example, beyond AlexNet fc7) with a fully connected layer of ten nodes. However, the key difference of this technique compared to the transfer learning technique is that the new network is now trained end-to-end with a stochastic gradient descent on separate training images from our own image set used to test the monkeys. Supplementary Fig. 14 shows that the three main findings of our article (discovery of challenge images, lagged solutions for challenge images, and lower IT predictivity for late-phase IT responses) are well replicated even with a fine-tuned ImageNet pre-trained AlexNet.

Prediction of neural response from DCNN features. We modeled each IT neural site as a linear combination of the DCNN model features (illustrated in Supplementary Fig. 11a). We first extracted the features per image from the layers of the DCNNs. The features extracted were then projected onto its first 1,000 principle components (ranked in the order of variance explained) to construct the final feature set used. For example, we used the features from AlexNet's[19] fc7 layer to generate Fig. 4a. Using a 50%/50% training/test split of the images, we then estimated the regression weights (that is, how we can linearly combine the model features to predict the responses of the neural site) using a partial least squares (MATLAB command: plsregress) regression procedure using 20 retained components. The neural responses used for training (R^{TRAIN}) and testing (R^{TEST}) the encoding models were averaged firing rates (measured at the specific sites) within the time window considered. We treated each time window (10-ms bins) independently for training and testing. The training images used for regressing the model features onto a neuron, at each time point, were sampled randomly (repeats included random subsampling) from the entire image set. For each set of regression weights (w) estimated on the training image responses (R^{TRAIN}), we generated the output of that 'synthetic neuron' for the held-out test set (M^{PRED}) as follows:

$$M^{PRED} = (w \times F^{TEST}) + \beta$$

where w and β are estimated via the PLS regression command, and F^{TEST} are the model activation features for the test image set.

The percentage of explained variance, IT predictivity (for details, refer to ref. [7]), for that neural site was then computed by normalizing the r^2 prediction value for that site by the self-consistency of the test image responses ($\rho^{R^{TEST}}$) for that site and the self-consistency of the regression model predictions ($s\rho^{M^{PRED}}$) for that site (estimated using a Spearman–Brown-corrected trial-split correlation score) as follows:

$$\text{IT predictivity} = \left(\frac{\text{corr}(R^{TEST}, M^{PRED})}{\sqrt{\rho^{R^{TEST}} \times \rho^{M^{PRED}}}} \right)^2$$

To achieve accurate cross-validation results, we had to test the prediction of the model on held-out image responses. But to make sure we had exposed the mapping procedure (mapping the model features onto individual IT neural sites) to images from the same full generative space and especially from both the control and challenge image categories, for each time step, we randomly subsampled image responses from the entire image set (measured at that specific time step). This ensured that the mapping step was exposed to exemplars from both the control and the challenge images groups. IT neural predictivity was also tested independently for control and challenge images (Supplementary Fig. 15a). We also tested the effect of time bins used for mapping on percentage EV (Supplementary Fig. 15b).

Estimation of the OST prediction strength. We compared how well different factors and $\Delta d'$ between monkey behavior and AlexNet fc7 predicted the differences in the OST estimates. Each image had an associated value for different image properties, either categorical (for example, occcluded or non-occluded) or continuous (for example, object size). We first divided the image sets into two groups, high and low, for each factor. The high group for each factor contained images with values higher than the 95th percentile of the factor distribution, and the low group contained the ones with values less than the 5th percentile of the distribution. For the categorical factor such as occlusion, the high group contained images with occlusion and the low group contained images without occlusion. Then, for each factor we performed a one-way analysis of variance (ANOVA) with OST as the dependent variable. The rationale behind this test was that if the experimenter (or experimenters) was to create image sets based on any one of these factors, how likely is it expose a large difference between the OST values? Therefore, we used the F-value of the test (y axis in Fig. 6) to quantify the OST prediction strength.

Statistics. As tests of significant difference between two variables, we used (Bonferroni-corrected) paired and unpaired t-tests and one-way ANOVA. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications[4,6,7]. All inclusion and exclusion criteria have been clearly mentioned in the corresponding Methods section and the Reporting Summary. Data distributions were assumed to be normal, but this was not formally tested. All trials during the task were randomized and drawn without replacement from the full set of images. Once the image set was exhausted, the entire randomization and sampling process was repeated. Data collection and analyses were performed blind to the conditions of the experiments.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The images used in this study and the behavioral and object solution time data will be publicly available at the time of publication from our GitHub repository (https://github.com/kohitij-kar/image_metrics).

Code availability

The code to generate the associated figures will be available upon reasonable request. The images, primate behavioral scores, estimated object solution times, and the modeling results will be hosted at http://brain-score.org[29].

References

51. Santos, A. et al. Evaluation of autofocus functions in molecular cytogenetic analysis. *J. Microsc.* **188**, 264–272 (1997).
52. Rosenholtz, R., Li, Y. & Nakano, L. Measuring visual clutter. *J. Vis.* **7**, 11–22 (2007).
53. Baker, C. I., Behrmann, M. & Olson, C. R. Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat. Neurosci.* **5**, 1210–1216 (2002).

nature research

Corresponding author(s): Kohitij Kar

Last updated by author(s): Mar 13, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected using a 512 channel Intan Recording Controller, and custom software provided by Intan Technologies, LLC. Human data was collected on an online platform: Amazon mechanical turk.
Data analysis	All behavioral and neural data were analyzed using MATLAB 2017b and 2018a. The pre-trained convolutional neural networks were downloaded using MATLAB 2018a and the matconvnet (Vedaldi et al. 2015). "MatConvNet - Convolutional Neural Networks for MATLAB", A. Vedaldi and K. Lenc, Proc. of the ACM Int. Conf. on Multimedia, 2015.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

At the time of publishing, the images used in this study and the behavioral and object solution time data will be publicly available at our github repository (https://github.com/kohitij-kar/image_metrics). The code to generate the associated figures will be available upon reasonable request. We will also host the images, primate behavioral scores, estimated object solution times, and the modeling results at http://brain-score.org

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the macaques, two different types of sample sizes were used. First, total number of monkeys: n = 2. This number was picked based on standard practices in electrophysiology literature. Second, total number of trials or repetitions for the behavioral and neural data. These numbers were computed based on split half reliability of the data collected and metric used. This has been mentioned in the corresponding locations in the manuscript. For humans, the number of subjects and the number of trials per image were determined based on the reliability (>0.8) of the data collected.
Data exclusions	Some of the neural data collected were excluded due to noisy recordings. To retain the best quality data for analysis, three specific data inclusion criteria were used as mentioned in the manuscript (and elaborated in the Methods). These were significant visual drive, high selectivity and high image rank order response reliability across trials.
Replication	All results reported were replicated by cross-validation across different train and test splits of the data. In addition, the main effects reported here, were replicated separately for each of the monkeys.
Randomization	The experimental design did not require specific randomizations. However, all our statistics and results were cross validated by randomly subsampling a part of the data collected, and re-doing the analyses on that specific part. This has been mentioned in the corresponding sections of the manuscript.
Blinding	The presentation of the images to the subjects (both humans and macaques) were randomly interleaved and therefore both the experimenters and the participants were blind to the experimental conditions / groupings. The group allocations for "challenge" and "control" image-sets were done post-hoc.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Human subjects were allowed to participate in the tasks without any specific prerequisites (e.g. age, country of origin, sex etc).
Recruitment	Humans subjects were recruited on the online platform Amazon Mechanical Turk. They individually consented to perform the tasks, by reading our instructions and voluntarily initiating the paradigm.
Ethics oversight	All human studies were done in accordance with the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Study did not involve clinical trial registration
Study protocol	Study did not involve clinical trial
Data collection	Study did not involve clinical trial
Outcomes	Study did not involve clinical trial

nature neuroscience

SUPPLEMENTARY INFORMATION

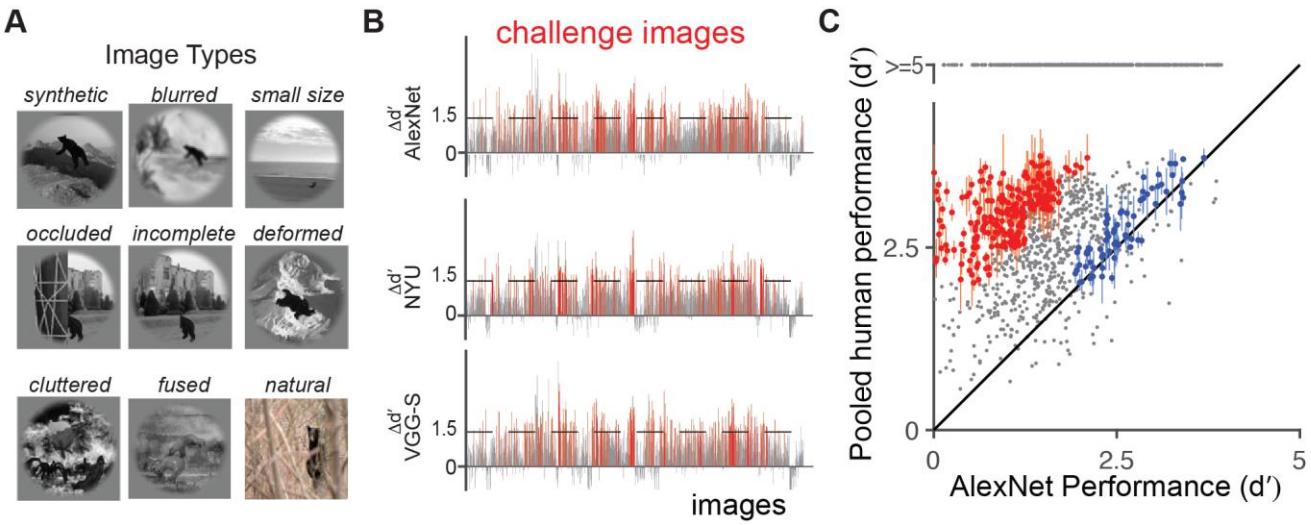
https://doi.org/10.1038/s41593-019-0392-5

In the format provided by the authors and unedited.

Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior

Kohitij Kar[1,2]*, Jonas Kubilius[1,3], Kailyn Schmidt[1], Elias B. Issa[1,4] and James J. DiCarlo[1,2]

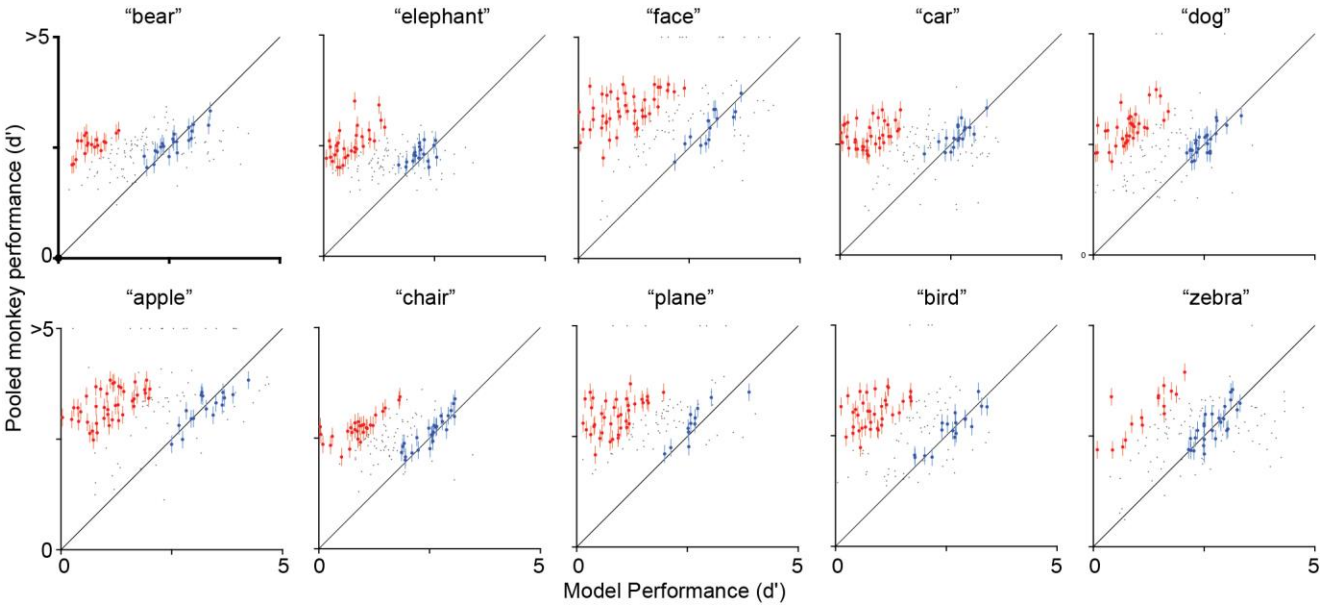
[1]McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Brain and Cognition, KU Leuven, Leuven, Belgium. [4]Present address: Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA.
*e-mail: kohitij@mit.edu



Supplementary Figure 1

Types of images used, performances across different shallower DCNNs and comparison of models with humans.

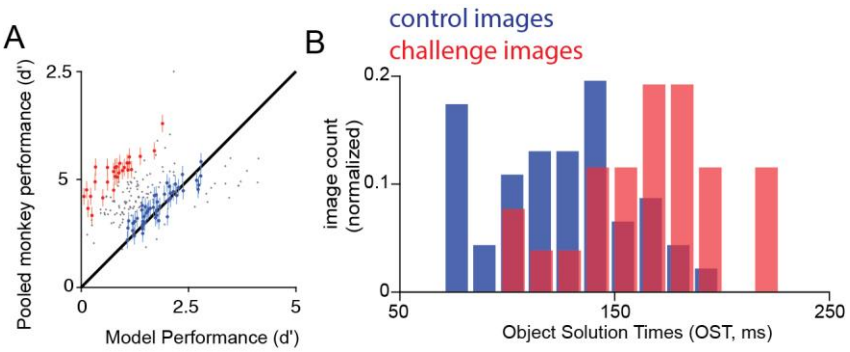
A) Examples of different image types used in the behavioral testing. Different image types included synthetic images containing an object in an uncorrelated background, images with blur, small object sizes, occlusion, incomplete objects, deformed objects, cluttered scenes, fused objects, and natural photographs. B) Comparison of pooled monkey behavioral performance and three DCNN models with similar architecture, VGG-S, NYU, and AlexNet. Each bar corresponds to an image. Red bars indicate the challenge images. The black dashed line shows the threshold difference (set at 1.5) used to determine the challenge images. C) Comparison of human performance (data pooled across 88 human subjects) and DCNN performance (AlexNet; 'fc7'). Each dot represents the behavioral task performance (I_1; refer Methods) for a single image. We reliably identified challenge (red dots; n=266 images) and control (blue dots; n=149 images) images. Error bars are bootstrapped s.e.m over 1000 resamples over n=88 trials per image.



Supplementary Figure 2

Object by object comparison of pooled monkey performance (data pooled across 2 monkeys) and DCNN performance (AlexNet; 'fc7').

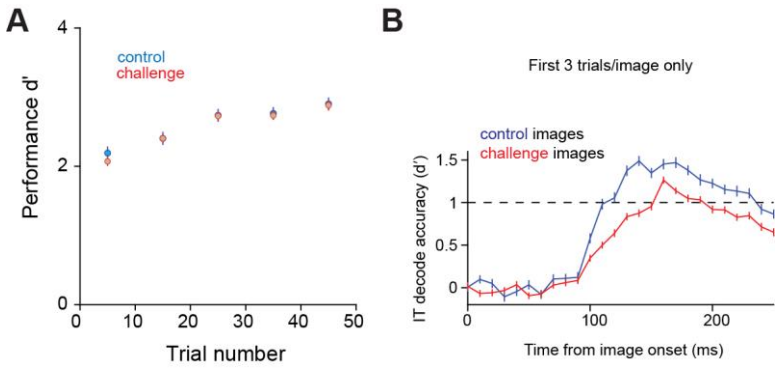
Each dot represents the behavioral task performance (I_1; refer Methods) for a single image of the corresponding object. We reliably identified *challenge* (red dots) and *control* (blue dots) images. Error bars are bootstrapped s.e.m. across 1000 resamples for 123 trials per image. n=132 images per object (corresponding to each sub-panel).



Supplementary Figure 3

Challenge image and object solution time estimation done separately for the MS COCO images.

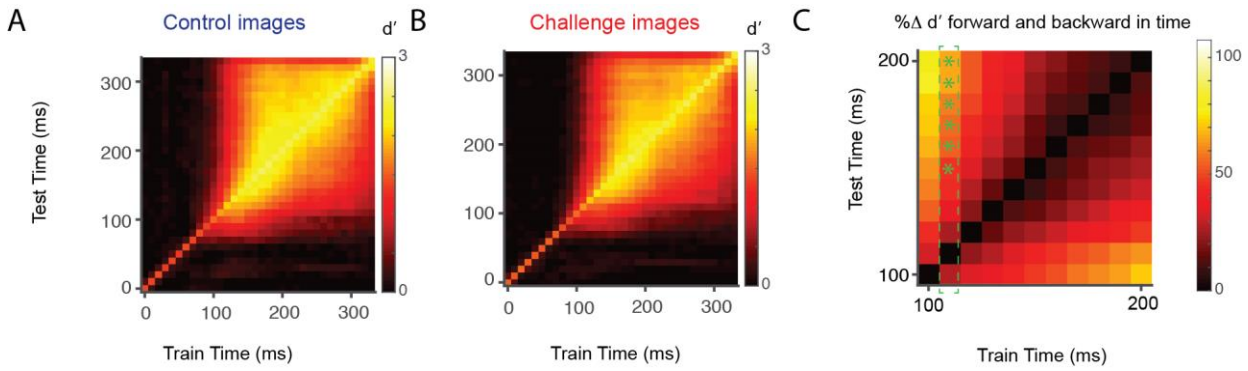
A) Comparison of AlexNet ('fc7') performance and pooled monkey behavior on the MS COCO images (n=200; 47 control and 38 challenge images). Errorbars show the s.e.m across 1000 resamples from 123 trials per image. B). Distribution of challenge (red) and control (blue) image OST. ΔOST was estimated at ~33ms.



Supplementary Figure 4

Comparison of control and challenge image performance, both behavioral and neural decoding accuracy, during repeated exposures of images and for the first three trials respectively.

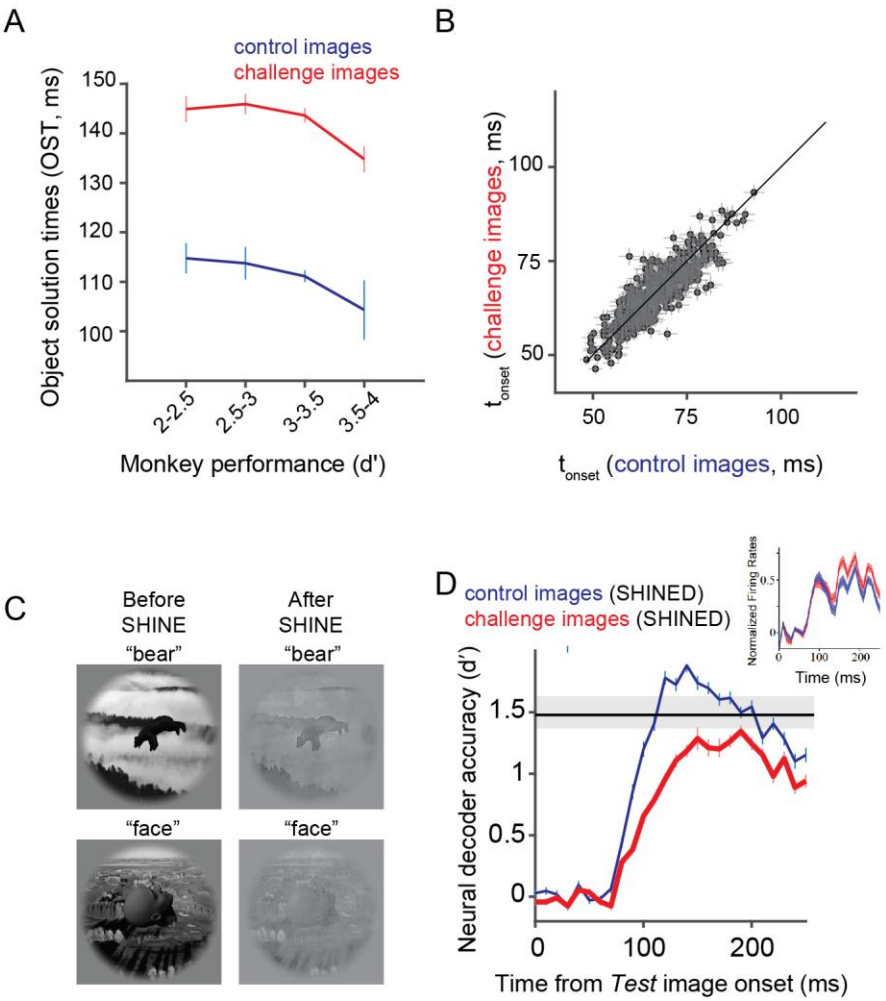
A) Change of pooled monkey behavioral performance I_1 with repeated exposure of the control (blue) and challenge (red) images. Each data point was estimated by pooling together 10 trials (around the trial numbers indicated in the x-axis). The figure shows that the control and challenge images did not show a different learning-curve across time after they were introduced during testing. Error bars are s.e.m across images. B) IT decode accuracies over time for control (blue) and challenge (red) images estimated for the first 3 trials per image only. This shows that the lagged solutions for the challenge images exist from the very early exposure periods of the images during the behavioral testing and is not a result of changes in IT responses due to repeated exposure (or some form of reinforcement learning). The dashed line at d′=1 was used as a threshold to approximate the difference in decoder latencies between these two image-sets. Errorbars are s.e.m. across images.



Supplementary Figure 5

Estimating how good the decoding accuracies are when trained and tested at different times.

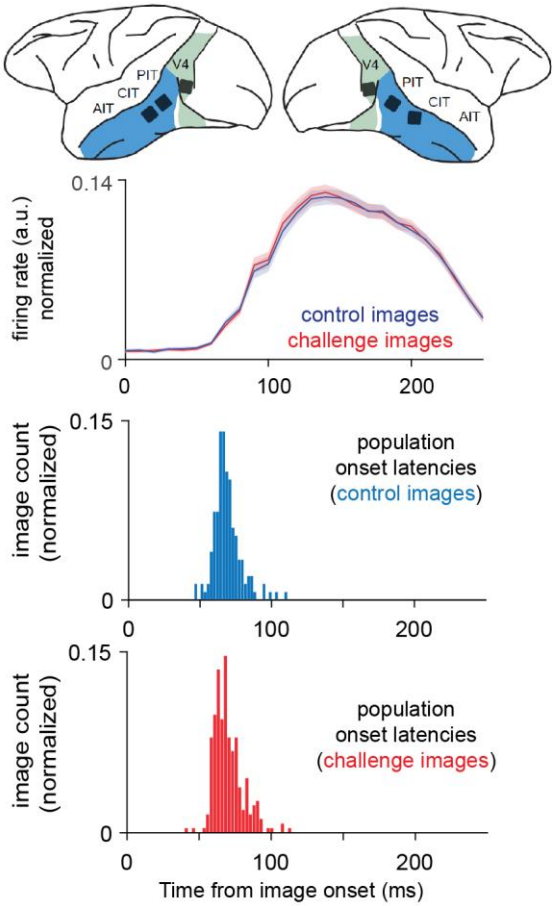
A) and B) Temporal cross training matrix for control (n=149) and challenge (n=266) images respectively, shown separately. To estimate the value at each element of the matrix, we trained a IT neural population (n=424) decoder (refer Methods) at a time 't1' ms and test it at time 't2' ms. C) The color denotes the percentage difference in performance from the diagonal (i.e. when the decoder was trained and tested at the same time point; therefore, all diagonal values are zeros). This is similar to the classification endurance (CE) metric used by Salti et al. 2015. We observed a lack of generalization across the train and test times. For instance, a closer inspection (shown in green dotted rectangle) of C) reveals that decoders trained at e.g. 110 - 120 ms (avg. OST of control images) loses greater than 50% of its decoding accuracy (shown as green *) when tested at >140 ms (avg. OST of challenge images). This suggests that object-information is coded by a dynamic population code consistent with the entry of recurrent inputs during late phases of the IT response.



Supplementary Figure 6

Controls analyses to rule out alternative hypotheses.

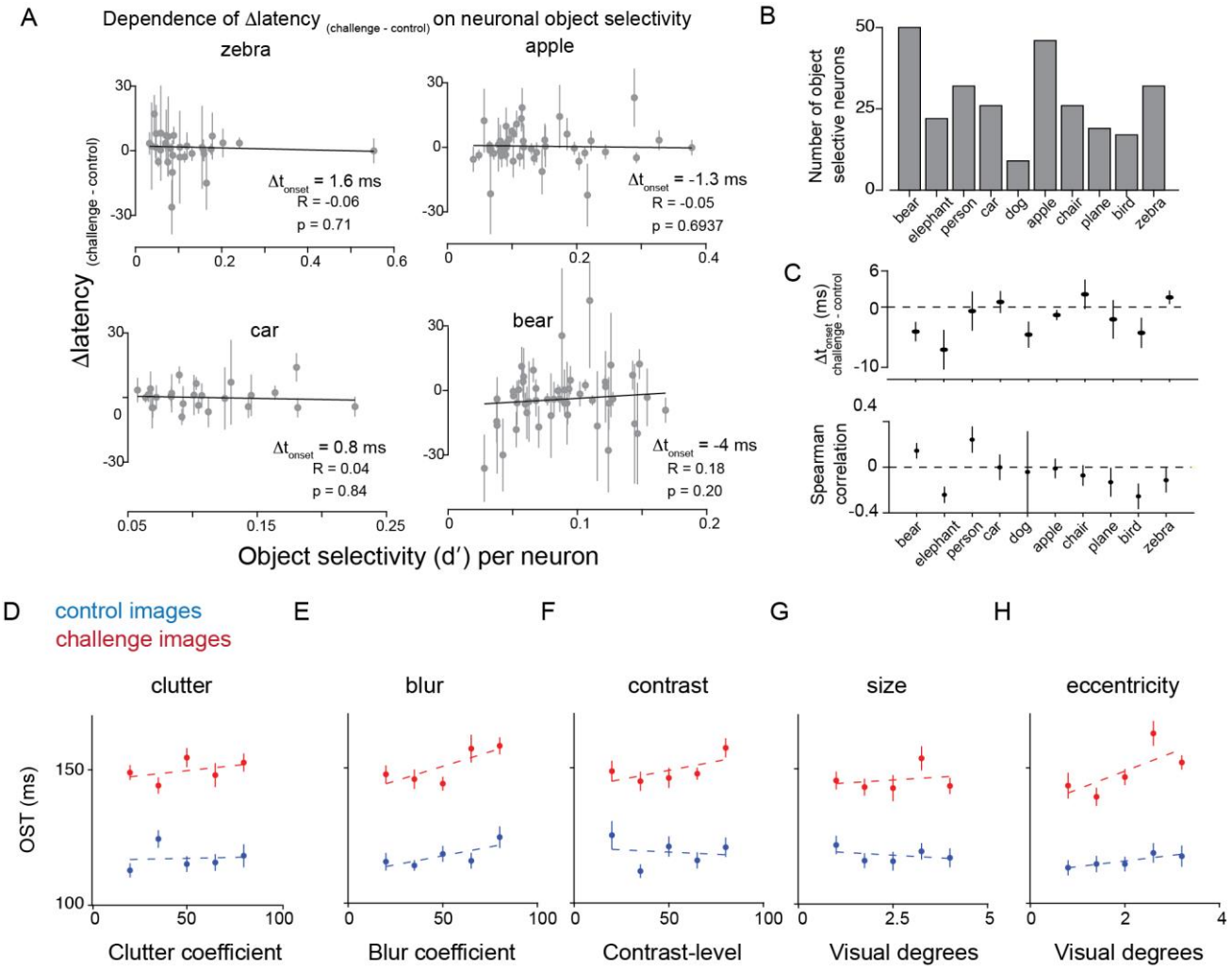
A) Dependence of OST on the pooled monkey I_1 level. The red and the blue curves show the OST values averaged across images with behavioral I_1 accuracy within the limits shown on the x-axis, for challenge (n = 67,145, 42, 12 images for each x-value) and control (n=54,44,41,10 images for each x-value) images respectively Errorbars are s.e.m across images . B) Comparison of the onset latencies (t_{onset}) per neuron(n=424 neurons), between the 266 *challenge* (y-axis) and 149 *control* (x-axis) images averaged across images of each group. Horizontal and vertical error-bars denotes s.e.m across images. C) Examples of two images, before and after the SHINE[31] (Spectrum, histogram, and intensity normalization and equalization) algorithm was implemented. D) Average IT population decodes over time after the SHINE technique was implemented, for the *control* (blue) and *challenge* (red) images. The error-bars denote s.e.m across images. The black line indicates the average behavioral I_1 for the pooled monkey population across all images. The gray shaded region indicates the standard deviation of the behavioral I_1 for the pooled monkey population across all images. The inset shows a comparison of the average normalized firing rates (across 424 neurons) over time, for both challenge (n=266 images; red) and control (n=149 images; blue) images after SHINING. Errorbars indicates s.e.m across images.



Supplementary Figure 7

Comparison of latencies in control and challenge image evoked neural responses in area V4.

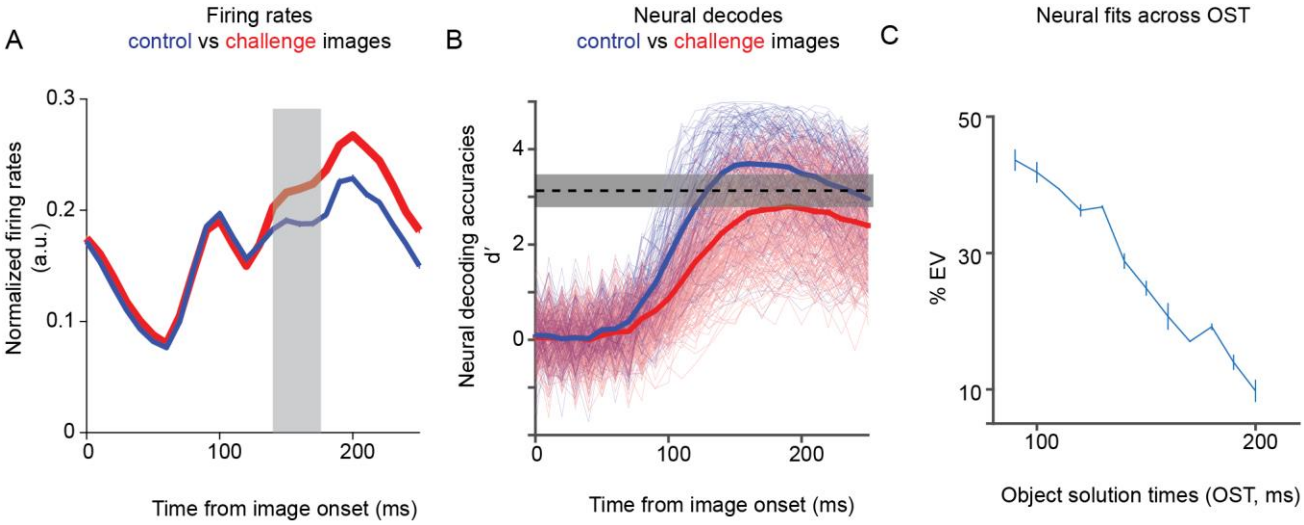
The top panel shows the placement of chronic Utah array implants in IT and V4 of two monkeys. Below it, we show the time course of normalized neural firing rates (averaged across the V4 population of 151 sites) for control (n=149 images; blue) and challenge (n=266 images; red) images. Errorclouds indicate s.e.m across neurons (n=151). The distribution of average onset latencies across the control (blue) and challenge (red) images is shown in the two bottom panels respectively. These two distributions are not significantly different.



Supplementary Figure 8

Testing the dependence of the decoding lags on category selectivity of neurons and image properties.

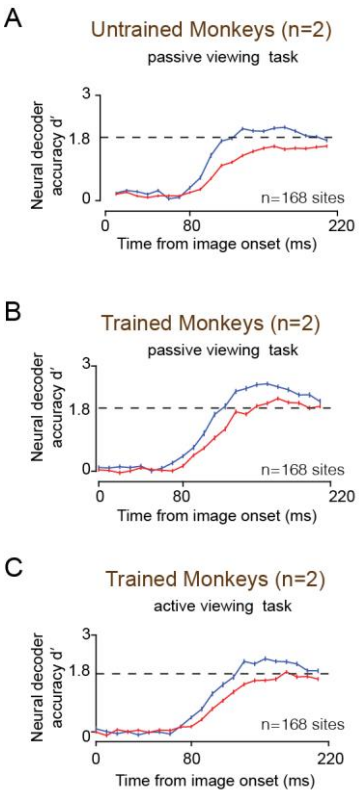
We considered the possibility that the difference in the OST between control and challenge images for each object category is primarily driven by neurons that specifically prefer that category (*object relevant neurons:* number for each category shown in B). To address this, we first asked whether the object relevant neurons show a significant difference in response latency (i.e. Δt_{onset} (challenge - control image) > 0) when measured for their preferred object category. A) shows 4 example object categories and the dependence of Δt_{onset} (Δonset latency, ms: challenge - control) on neuronal object selectivity. The Spearman correlation value, R and associated p-values are denoted as insets. The top panel of C) summarizes these examples and shows that the overall Δt_{onset} was not significantly greater than zero (unpaired t-test; p>0.5). In fact a closer inspection (top panel of C) reveals that for some objects (e.g. bear, elephant, dog) Δt_{onset} was actually negative — that is, a trend for slightly *shorter* response latency for challenge images. Finally, to test the possibility that there was an overall trend for the most selective neurons to show a significant Δt_{onset}, we computed the correlation between the Δt_{onset} and the individual object selectivity per neuron, per object category as indicated in A). Bottom panel of C) shows that there was no dependence of object selectivity per neuron on the response latency differences. In sum, the later mean OST for challenge images cannot be simply explained by longer response latencies in the IT neurons that "care" about the object categories. D-H) Dependence of object solution times on different image-based factors tested separately for control and challenge images. D-H shows the factors clutter, blur, contrast, size and eccentricity respectively. Despite some overall dependence of OST on one or more of these factors, Δ OST$_{(challenge-control)}$ is maintained ~30 ms at each tested level of these factors. The dashed lines show a linear fit of the data.



Supplementary Figure 9

Results from the passive fixation task.

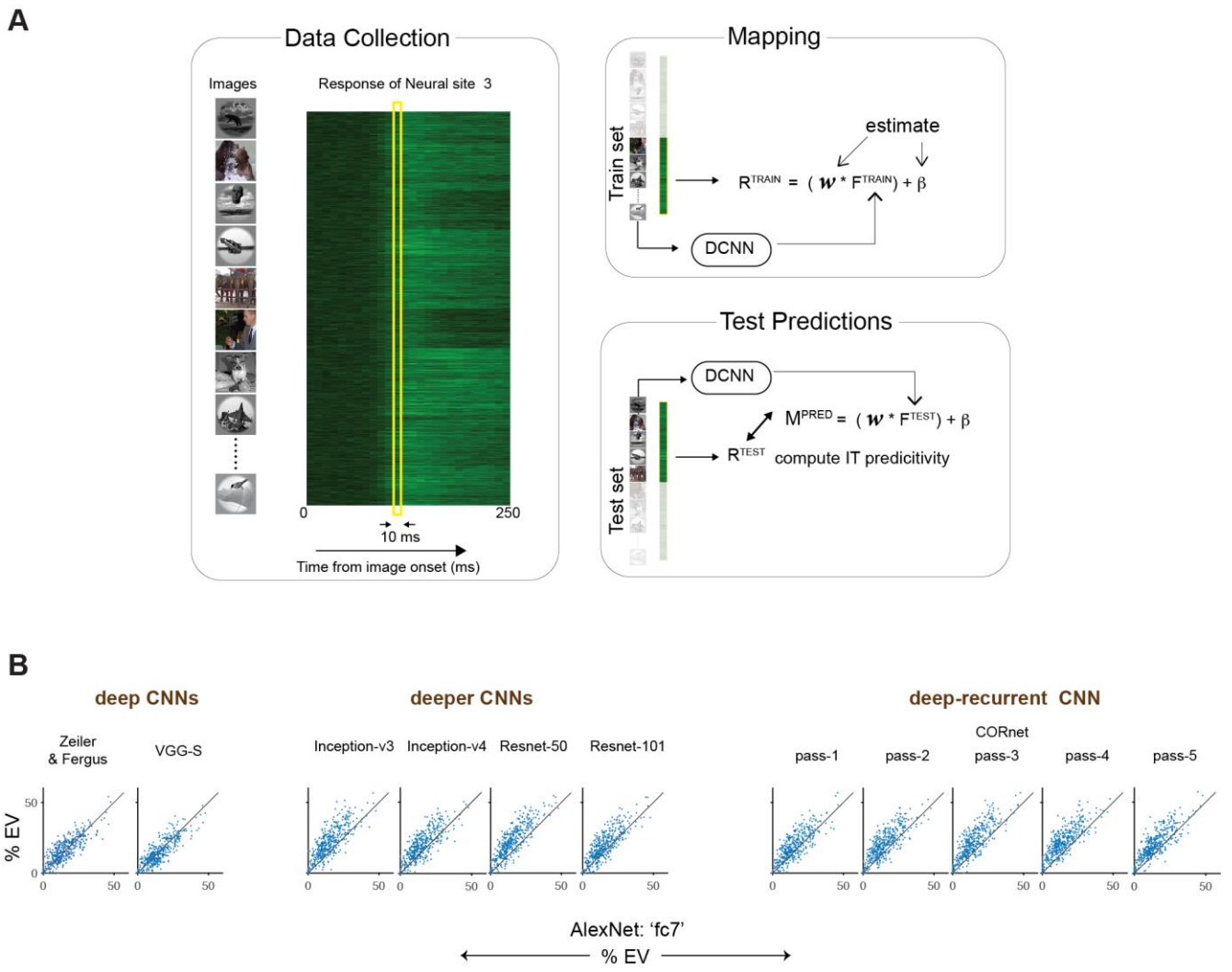
A) Comparison of normalized firing rate responses (averaged across all 424 IT sites) to the control (n=149 images; blue) and challenge images (n=266 images; red). The initial dip in the firing rate is caused by the offset responses related to the previous stimulus. The gray bar shows the time bins for comparison of challenge vs control image responses, reported in the manuscript. B) Estimates of neural decodes over time. Each thin line represents a single control (blue) or challenge (red) image. The thick blue and red line represent the average control and challenge image decodes over time respectively. The horizontal dashed line represents the average performance across control and challenge images (gray area being the standard deviation across images). This demonstrates the lagged solution times for the challenge images. C) Drop of IT predictivity over object solution time. Errorbars shows s.e.m across 424 IT sites.



Supplementary Figure 10

Comparison of neural decodes over time between trained and untrained monkey IT cortex during the passive viewing and active discrimination tasks.

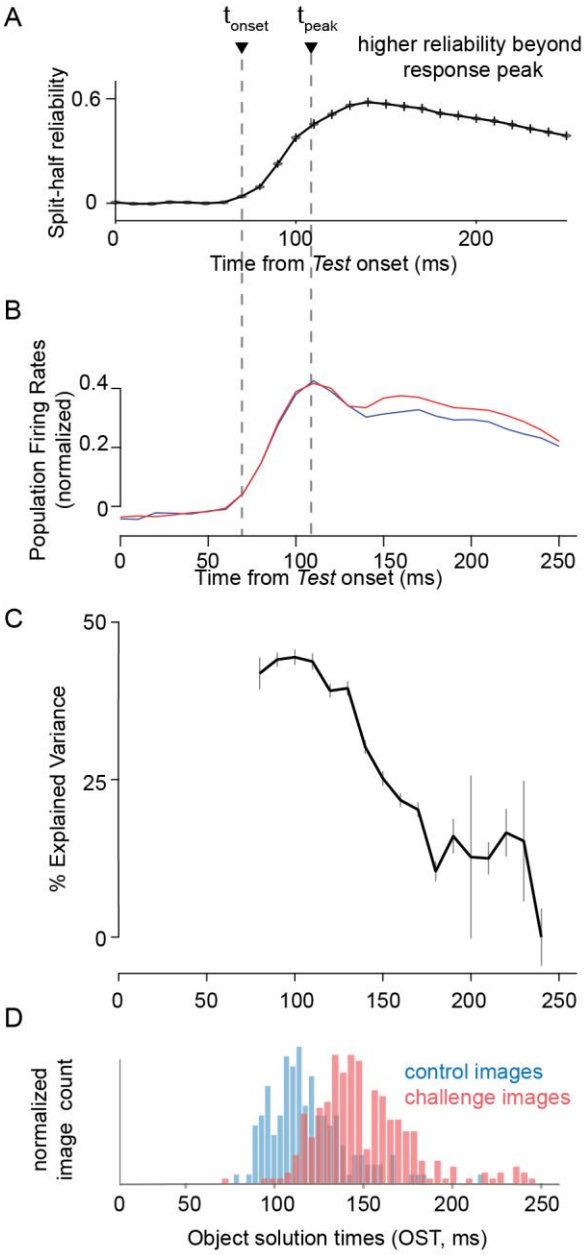
A) Results from untrained monkeys: IT population decodes over time for control (blue curve; 86 images) and challenge (red curve; 117 images) images. The threshold to estimate the decode latency, denoted by the dashed black line, was set at 1.8. The recordings were done from 168 sites (refer [6]). B) Results from trained monkeys during the passive viewing task: IT population decodes over time for control (blue curve) and challenge (red curve) images. The threshold to estimate the decode latency, denoted by the dashed black line, was set at 1.8. The recordings were subsampled randomly from 168 sites (out of 424; however, the selection was restricted to the left hemisphere and pIT and cIT arrays). C) Results from trained monkeys during active object discrimination tasks: IT population decodes over time for control (blue curve) and challenge (red curve) images. The threshold to estimate the decode latency, denoted by the dashed black line, was set at 1.8. The recordings were subsampled randomly from 168 sites (out of 424; however, the selection was restricted to the left hemisphere and pIT and cIT arrays). For A-C we plot the median accuracy for the corresponding timebin across all tested images for each time bin. All errorbars are s.e.m across images (n=117 for challenge images, n = 86 for control images).



Supplementary Figure 11

Predicting IT neural responses with DCNN features.

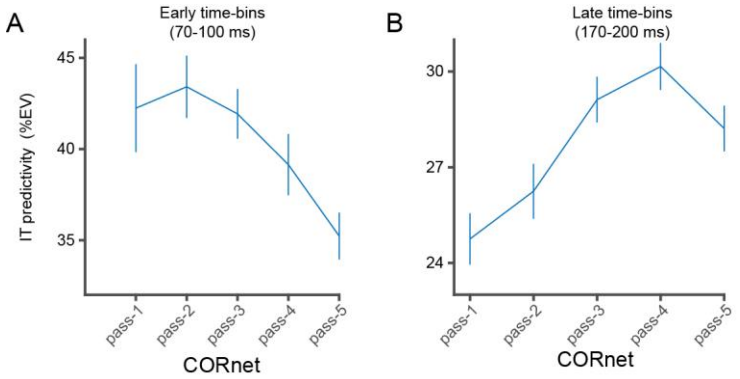
A) Schematic of the DCNN neural fitting and prediction testing procedure. This includes three main steps. Data collection: neural responses are collected for each of the 1320 images (50 repetitions), e.g. shown is that of example neural site #3, across 10 ms time-bins. Mapping: We divide the images and the corresponding neural features (R^{TRAIN}) into a 50-50 train-test split. For the train images, we compute the image evoked activations (F^{TRAIN}) of the DCNN model from a specific layer. We then use partial least square regression to estimate the set of weights (w) and biases (β) that allows us to best predict R^{TRAIN} from F^{TRAIN}. Test Predictions: Once we have the best set of weights (w) and biases (β) that linearly map the model features onto the neural responses, we generate the predictions (M^{PRED}) from this synthetic neuron for the test image evoked activations of the model F^{TEST}. We then compare these predictions with the test image evoked neural features (R^{TEST}) to compute the IT predictivity of the model. B) Scatterplots of IT (n=424 neurons) predictivity (% EV) of different deep, deeper and deep-recurrent CNNs with respect to AlexNet with images (n=319) that are solved between 150-250 ms post onset. We observe that IT predictivity of deep CNNs are not significantly different than AlexNet. However, both the deeper CNNs and late passes of CORnet (a deep-recurrent CNN) are better at IT predictivity compared to AlexNet.



Supplementary Figure 12

Comparison of internal consistency (reliability) of the IT neural responses across time with respect to other variables.

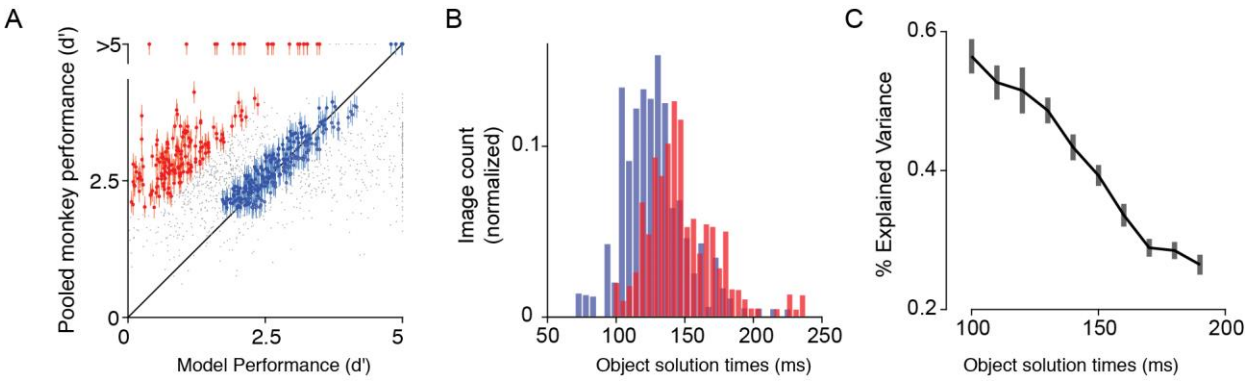
A) Reliability (or internal consistency) of neural responses as a function of time. The internal consistency was computed as a Spearman-Brown corrected correlation between two split halves (trial based) of each IT neural site's responses across all tested images. Errorbar indicates s.e.m across neurons (n=424 neurons) B) Normalized averaged population firing rate across time. Vertical dashed lines indicate onset and peak response latency., C) temporal profile of IT predictivity. D), object solution time distribution for challenges (red) and control (blue) images. Error-bar in C shows s.e.m across neural sites (n=424 sites). B), C) and D) are identical to Figure 3A, Figure 4A, and Figure 2C respectively.



Supplementary Figure 13

Evaluation of CORnet IT predictivity. A) IT predictivity (% EV) computed at early (70-100ms) response times.

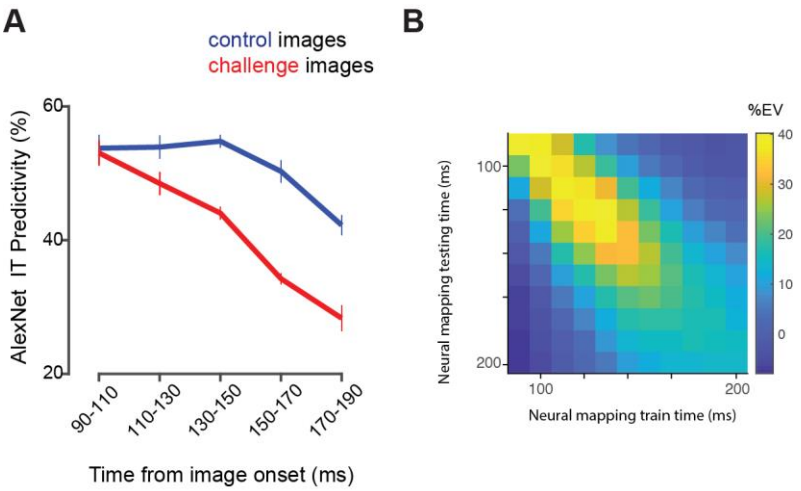
We observe that the earlier passes (pass 1 and pass 2) are better predictors of the early time bins and the prediction deteriorates for the later passes. B) IT predictivity (%EV) computed at late (170-200 ms) phases of IT responses. Here we observe that the late passes (especially pass 4) is better at predicting the IT response compared to the early passes. Error bars denote s.e.m across neurons (n=424).



Supplementary Figure 14

Evaluation of a fine-tuned AlexNet (ImageNet pre-trained).

We first downloaded a version of AlexNet (pre-trained with the imagenet classification dataset). We then cropped the network at the 'fc7' layer, and added a customized classification layer (containing 10 output nodes; corresponding to our objects) at the backend. We then trained this network end-to-end on a subset of our images (that contained a mixture of both control and challenge images). We then tested this fine-tuned network on the rest of the held-out images. This process was repeated until all images were used as (held-out) test images, achieving a full set of image-by-image cross-validated behavioral accuracies. Although the overall performance of this fine-tuned DCNN was higher than that of the pre-trained (transfer-learned) AlexNet, all of our main findings — presence of challenge images (A), lagged IT decodes (B) and lower IT predictivity (C) for those images (n=1320 images), were replicated using such a fine-tuned network. Errorbars in A are bootstrapped STD for I_1 estimates per image. Errorbars in C are s.e.m across neurons (n=424)



Supplementary Figure 15

IT neural predictivity (% EV) of AlexNet 'fc7' layer tested across time independently for the control (blue) and the challenge (red) images and IT neural predictivity of AlexNet 'fc7' layer trained and tested at different time bins (10 ms bins from 90 ms to 200 ms post image onset).

A) The data was divided into 20 ms time bins (starting from 90 ms to 190 ms). At each time bin, the image-response neural data from a subset of images (sub-sampled from the entire image-set) was used to train the mapping between 'fc7' activations and the neural response. After training, this model was tested on the responses of the control (n=149) and challenge (n=266) image present in the held-out test set. The procedure was repeated to get multiple tests for every control and challenge image. The figure shows that both control(blue) and challenge (red) image IT predictivity drops over time. However, the drop is significantly larger for the challenge images (significant interaction between image-type and time; $F(1,4) = 6.3$; $p<0.005$; post hoc Turkey test shows that IT predictivity at time bins > 130 ms are significantly different between control and challenge images). Errorbars are s.e.m across neurons (n=424). B) The diagonal of this plot (showing the strongest predictivity) corresponds to the cases where the models were trained and tested at the same time bins. Off-diagonal boxes show that IT predictivity gets worse when trained and tested at separate time bins. Of note, the strength of IT predictivity drops even along the diagonal (recapturing the phenomenon demonstrated in Figure 4A).