HHS Public Access

Author manuscript

Nat Neurosci. Author manuscript; available in PMC 2022 January 24.

Published in final edited form as:

Nat Neurosci. 2019 June ; 22(6): 974–983. doi: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,+], **James J. DiCarlo**[1,2]

[1]McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA

[2]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA

[3]Brain and Cognition, KU Leuven, Leuven, Belgium

Abstract

Non-recurrent deep convolutional neural networks (DCNNs) currently best model core object recognition; a behavior 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 DCNNs 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 ~30ms later in IT for "challenge" images compared to primate performance-matched "control" images. Third, these behaviorally-critical late-phase IT response patterns were poorly predicted by feedforward DCNN activations. Interestingly, very-deep CNNs and shallower recurrent CNNs better predicted these late IT responses, suggesting 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.

Introduction

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

[*]Correspondence should be addressed to Kohitij Kar, McGovern Institute for Brain Research, Massachusetts Institute of Technology, 77 Massachusetts Institute of Technology, 46-6161, Cambridge, MA 02139. kohitij@mit.edu.

[+]Current address: Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, NY

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 analysis. K.K., and J.K performed computational modeling. K.K. and J.J.D. wrote the manuscript.

Competing Financial Interests

The authors declare no competing interests.

Data and code 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/kohitijkar/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[29].

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 inferior temporal (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 IT.

At present, the models that best predict the individual responses of macaque IT neurons belong to the architectural family of deep convolutional neural networks (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, sub-cortical, and medium to long-range intra-areal recurrent circuits (Figure 1A). The short time 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 like 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 provide new constraints to guide future model development.

There is growing evidence 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 in the ventral stream should be most relevant at later time points in the image driven neural responses. Therefore, we reasoned that object representations in IT 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 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 decodes from IT populations for the *challenge* images took ~30ms (average) longer to emerge compared to *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) where the IT population solutions (linear

decodes) to many of the *challenge* images emerged. Taken together, these results imply a behaviorally-critical role of recurrent computations during core object recognition. Notably, we also found the same neural phenomena while the monkeys passively viewed the images, implying that the putative recurrent mechanisms for successful core object inference in the primate are not strongly state or task dependent. Furthermore, we show that the observed image-by-image differences between DCNN and primate behavior along with precisely measured IT population dynamics for each image better constrain the next generation of ventral stream models over 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 (occluded, cluttered, blurred, etc.) might most critically depend on feedback, we instead 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, Figure 1C, [9, 10]). 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 Primate Behavioral Testing in Methods)

We tested a total of 1320 images (132 images per object), in which the primary visible object belonged to one of 10 different object categories (Figure 1B). To make the task challenging, we included various image types (see Figure S1A): synthetic objects with high view variation on cluttered natural backgrounds (similar to [5]), images with occlusion, deformation, missing object-parts, and colored photographs (MS COCO dataset [20]).

Behavioral testing of all of these images was done in humans (n=88; Figure S1C) and in monkeys (n=2; Figure 1D). We estimated the behavioral performance of the subject pool on each image, and that vector of image-wise performance is referred to as I_1 (see Methods; refer [10]). We collected sufficient data such that the reliability of the I_1 vector was reasonably high (median split half reliability $\tilde{\rho}$, 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, e.g. 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 dots; Figure 1C) as those where the absolute difference in primate and DCNN performance does not exceed 0.4 (d' units), and *challenge* images (red dots; Figure 1D) as those where 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 Figure 1E. The *challenge* images

were not idiosyncratic to our choice of AlexNet ('fc7') (Figure S1B), specific objects (Figure S2) or our synthetic image generation procedure (Figure S3A).

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 as the monkeys were repeatedly exposed to these images (Figure S4A). This is consistent with earlier work[9], that showed — 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 (both humans and macaques) for *challenge* images were significantly higher than for the *control* images (monkeys: ΔRT = 11.9 ms; unpaired two-sample t-test, t(413) = 3.4; p <0.0001; humans: ΔRT = 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 IT

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 IT for the *challenge* and *control* images. First, we wanted to know if these IT object decoders were as accurate as the primates for both types of images as predicted by the leading IT decoding model [5]. That would demonstrate whether the ventral stream successfully solves the *challenge* images. Second, we reasoned that, if challenge image solutions required recurrent computation driven additional processing time, then IT object decodes for *challenge* images should emerge later in IT compared to *control* images. Thus, we here used a sliding decoding time window (10 ms) that was narrower than 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 multi-electrode array recordings (Figure 2A) across IT cortex (424 valid IT sites) in two macaques.

To determine the time at which explicit object identity representations are sufficiently formed in IT, we estimated the temporal trajectory of the IT object decode accuracy for each image. We computed the neural decoding accuracies (NDA) 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 (Figure S5). Thus, we determined the time at which the NDA measured for each image reached the level of the subject's (pooled monkey) behavioral accuracy (see Methods; Figure 2A, top 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 Figure. 2B). We also observed that the OSTs estimated by randomly subsampling half (n=212) the total number of sites were significantly correlated (Spearman R was 0.77 and 0.76 for control and challenge images respectively; p<0.00001; and *AOST* was maintained ~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 IT decodes achieve the behavioral accuracy of the monkey. 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 object solution times (OST_{image}) for the *challenge* images were, on average, ~30 ms later compared to the *control* images. Specifically, the median OST for the *challenge* images was 145 ± 1.4 ms (median ± SE) from stimulus onset and for the *control* images was 115 ± 1.4 ms (median ± SE) (Figure 2C). The average difference (~30 ms) between the OSTs of *challenge* and *control* images did not depend on our choice of behavioral accuracy levels (Figure S6A) or image-set type (Figure S3B).

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, e.g., 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 — the difference in IT response onset latency was only 0.17 ms (median; ± 0.21 ms, SE; paired t-test; t(423) = 0.3896, p = 0.69; see Figure 3A, Figure S6B), suggesting that the initial visual drive in both image-sets arrive at approximately the same time in IT. We also simultaneously recorded from area V4 (upstream of IT) 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 (Figure S7; paired t-test; t(150)=0.2; p=0.8). These results further support the hypothesis that the *ΔOST* between the challenge and the control images in IT is not driven by 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 (Figure 3B and 3C). 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 (see Figure 3D). For example, inspection of Figure 3D reveals that some of the *challenge* images evoke faster-than-average latency responses in IT, yet have slow OSTs (~200 ms). Conversely, some of the *control* images evoke slower-than-average IT responses, yet have relatively fast OSTs (~110 ms). Interestingly however, we found that firing rates (R) were significantly higher (%ΔR = 17.3%, paired t-test; t(423) = 6.8848, p <0.0001) for *challenge* images compared to *control* images (30 ms window centered at 150 ms post stimuli onset; see Figure 3A). One possible explanation of this 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 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. Our results (Figure S8) show that Δt_{onset} was not significant for any object category. In fact, a closer inspection (top panel of Figure S8C) reveals that for some objects (e.g. bear, elephant, dog) Δt_{onset} was negative — 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 (bottom panel: Figure S8C) 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 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 like total image contrast energy [22], spatial frequency power distribution [23], and location of the visual objects [24]. So, we asked if these low -evel explanations might explain the longer *challenge* image *OST*s. First, we observed that OSTs were not significantly correlated with image contrast (Spearman ρ=−0.04; p=0.47). Second, we used the SHINE (spectrum, histogram, and intensity normalization and equalization; Figure S6C) technique [25] to equate low-level image properties across the *control* and *challenge* image-sets, and re-ran the recording experiment (subsampling 118 images each from the *control* and *challenge* image-sets; number of repetitions per image = 44). The average estimated difference in *OST* values between "SHINED" *challenge* and *control* images was still ~24 ms (Figure S6D). Third, we tested whether ΔOST (challenge - control), was specific to certain low or high values of various image based properties (image clutter, blur, contrast, object size and object eccentricity; for definition — 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 (Figure S8D-H).

To test whether ΔOST (challenge - control) depends on neurons with higher or lower absolute latencies, we divided the neural population into 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 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 decoding of challenge images relative to control images ($\Delta DecodeLatency^{low}_{th=1.0} = \sim 22\,\text{ms}$, $\Delta DecodeLatency^{high}_{th=1.0} = \sim 18\,\text{ms}$; note that we here 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 IT 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 a dot, 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 *challenge* images evoked a significant 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 to the *control* images (Figure S9A). Second, we observed that we could successfully estimate the object solution times for 92% of *challenge* and 98% of *control* images. The object solution times 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 to the *control* images (Figure S9B). Taken together, this suggest 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 of McKee et al. [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 (e.g. mental task performance?) or somehow due to the training history. To test this, we performed the same analyses on smaller sets of data from two untrained animals (previously reported in [6,7,8]). To appropriately compare with 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; $\Delta DecodeLatency_{th=1.8}^{untrained} = \sim 34$ ms; $\Delta DecodeLatency_{th=1.8}^{trained} = \sim 30$ ms; see Figure S10). In sum, 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 [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; Figure S11A).

First, we observed that the 'fc7' layer of AlexNet predicted 44.3 ± 0.7% of the explainable IT neural response variance (%EV) during the early response phase (90-110 ms; Figure 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 (Figure 4A). This drop in IT predictivity was not due to low signal to noise ratio of the neural responses during those time points because our %EV measure already compensates for any changes in SNR, and also because SNR remains relatively high in the late part of the IT responses (Figure S12). 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 top panel histograms for OST distributions of *challenge* and *control* images in Figure 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 non-linear transformations onto the initially evoked (~feedforward) IT population response pattern. To test this idea, we asked if 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 (non-linear transformations),

we divided the tested DCNN models into two groups, deep (8 layers; AlexNet, Zeiler and Fergus model, VGG-S) and deeper (>20 layers, inception-v3[30], inception-v4[31], ResNet-50[32], 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 predict IT neural responses at the late phases (150-250 ms) significantly higher (ΔPredictivity = 5.8%, paired t-test; t(423) = 14.26, p <0.0001) than "regular-deep" models like AlexNet (Figure 4B; scatter plot comparisons with AlexNet shown separately in Figure S11B). This observation suggests that deeper CNNs might indeed be approximating "unrolled" versions of the ventral stream's recurrent circuits. 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 (i.e. *challenge* images for these models) showed even longer *OST*s in IT cortex than the original full set of *challenge* images (Figure 4C). Assuming that 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 ask if the experimental observations above might indeed be the result of recurrent computations, we directly tested a 4-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 (pass1- pass5 ; Figure 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 (Figure 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 compared to pass-3 and 4 for later time-steps, whereas the opposite was true for earlier time-steps (Figure S13). 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 OST>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 (Figure 5A), immediately following the test image presentation. Such backward masking has been previously 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 to *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 IT. 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 (Figure 5B) show that visual masking indeed had a significantly stronger effect on the *challenge* images at smaller presentation durations compared to 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 statistically significant at the .05 significance level (Bonferronni adjusted) for all presentation durations except 267 ms. Together with the neurophysiology results, these observations provide converging evidence that rapid, recurrent ventral stream computations are critical to the brain's ability 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], etc. Indeed, we observe that several image properties like object size, presence of occlusion, and object eccentricity, as well as a combination of all these factors (Figure 6) are significant, but very weak predictors of our putative recurrence signal (the OST vector; see Methods: Estimation of the OST prediction strength). In comparison, the performance gap between AlexNet and the monkeys ($\Delta d'$) is a significantly stronger predictor of OST. Therefore, our results suggest another possible image-wise predictor of ventral stream recurrence — the difference in performance between feed-forward DCNNs and primates, $\Delta d'$. This vector is likely 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 find that $\Delta d'$ is significantly predictive of the OST for each image (Spearman = 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 ask if recurrent circuits are critical to the ventral stream's execution of core recognition behavior. 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 primates. With these in hand, we aimed to look for a likely empirical signature of recurrence — 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 ~30ms (average) for *challenge* compared to the *control* images. In addition, we also 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 [49,50] has been previously suggested and partly explored, we believe that this is the first work to examine these questions at such large-scale, 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 (i.e. the images and tasks), and the first to do so with an implemented linking model (decoder) of how IT supports recognition behavior.

Late object identity solution times in IT imply 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(s) 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 V1 and IT, Tovee [42] proposed that the ventral stream comprises of 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 IT and lower areas like V4, V2 and V1 (similar to suggestions of [43]). This possibility is consistent with observations of temporally-specific effects in the response dynamics of V4 neurons [44] for images with occlusion. Alternatively, it is possible that IT is receiving important recurrent flow from downstream areas like the prefrontal and perirhinal cortices (as suggested by [45, 46]). We also cannot rule out the possibility that all of the additional computations are due to recurrence within IT (consistent with recent models [39]), or due to subcortical circuits (e.g. basal ganglia loops, [47]). 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 present 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 (i.e. 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 (e.g. HMO [7], AlexNet[19] and VGG [48, 49]) 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-onset; Figure 4A). This is consistent with our inference that the late object solution times for *challenge* images are primarily caused by the recruitment of additional recurrent processing in the ventral stream.

Unique object solution times per image motivate the search for better decoding models

Majaj et al. [6] showed that a simple linear decoding model, formed by linearly weighting the population activity of IT neurons (integrated from 70-170 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 observe that, even for images that have statistically non-distinguishable levels of behavioral performance, the linearly-decodable information in the IT population pattern varies quite 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” [43], high levels of “clutter” [40], “grouping” of behaviorally relevant image regions[50] or the need for visual “pattern completion”[39]. While we agree that such image manipulations might recruit recurrent processes in the ventral stream, the present 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 use the shallower models to find images for which the difference between feedforward-only DCNN and primate behavior ($\Delta d'$) is the largest. This difference is a better predictor of the neural phenomena of recurrence than any of the image-based properties (Figure. 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, but that 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 like inception-v3, v4 [31], ResNet-50,101 [32], that introduce more nonlinear transformations to the image pixels, compared to shallower networks like 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 (RNN) is equivalent to a very deep CNN (e.g. 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 non-linear 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 incorporate 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 (i.e. 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, $\Delta d'$ that quantifies the performance gap between feedforward DCNNs (e.g. 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 (the OST_{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 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's Mechanical Turk, an online platform in which subjects can complete experiments for a small payment.

Visual stimuli: generation

Generation of synthetic ("naturalistic") images—High-quality images of single objects were generated using free ray-tracing software (http://www.povray.org), similar to Majaj et al. [6]. Each image consisted of a 2D projection of a 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** (Figure 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 Figure S1A for example images). A total of 1120 naturalistic images (112 per object category) were used.

Generation of natural images (photographs)—Images pertaining to the 10 nouns, were download from http://cocodataset.org. Each image was resized to 256 x 256 x 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 have compared the ability of different image properties to predict the putative recurrence signal, inferred from our results. These image properties were either pre-defined during the image generation process (e.g. object size, object eccentricity, and the object rotation vectors, 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 image processing literature contains multiple measures of image focus based on first order differentiation or smoothing followed by differentiation. We have used a technique from Santos et al. [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, e.g., color, contrast, and orientation [52].

Primate behavioral testing

Humans tested on amazon mechanical turk—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” (HITs). We did not collect information regarding the sex of the human subjects who performed the task online on the Amazon Mechanical Turk platform. The reliability of the online MTurk platform has 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 (one out of 1360 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 [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 (1920 × 1080 at 60 Hz) positioned 42.5 cm in front of the animal. Monkeys were head fixed. Monkeys fixated a white square dot (0.2°) for 300 ms to initiate a trial. The trial started with the presentation of a sample image (from a set of 1360 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 view freely the choice objects for up to 1500 ms and indicated its final choice by holding fixation over the selected object for 400 ms. Trials were aborted if gaze was not held within ±2° of the central fixation dot during any point until the choice

screen was shown. Prior to the final behavioral testing, both monkeys were trained in their home-cages on a touchscreen (for details see[10]; details of the code and hardware 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 one or two 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 a white square dot (0.2°) for 300 ms to initiate a trial. We then presented a sequence of 5 to 10 images, each ON 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 gaze was not held within ±2° of the central fixation dot during any point.

Behavioral Metrics

We have used the same one-vs-all image level behavioral performance metric (I_1) to quantify the performance of the humans, monkeys, deep HCNNs 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 9 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,

$$HitRate^{i}_{image} = \frac{\sum_{j=1}^{10} Pc^{i, j \neq i}_{image}}{9},$$

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

$$FalseAlarm^{i} = 1 - avg(HitRate^{j \neq i}_{image})$$

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

$$d'_{image} = z(HitRate^{i}_{image}) - z(FalseAlarm^{i}),$$

where 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 1320 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 (i.e. internal consistency) of our I_1 estimation.

Large scale multielectrode recordings and simultaneous behavioral recording

Surgical implant of chronic micro-electrode arrays—Before training, we surgically implanted each monkey with a head post under aseptic conditions. After behavioral training, we recorded neural activity using 10×10 micro-electrode 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 multiple Utah arrays in the IT and 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 3 arrays in right hemisphere (all 3 in IT) and 3 arrays in the left hemisphere (2 in IT and 1 in V4). In monkey N, we implanted 3 arrays in the left hemisphere (all 3 in IT) and 3 arrays in the right hemisphere (2 in IT and 1 in V4). 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 Figure 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 2 subjects were trained to fixate a central white square (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 Intan Recording Controller (Intan Technologies, LLC). The majority of the data presented here were based on multiunit activity. We detected the multiunit spikes after the raw data was 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 IT, 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 (d'_{visual}): We estimated the overall visual drive for each electrode. This metric was estimated by comparing the COCO image responses of each site to a blank (gray screen) response.

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

Image rank-order response *reliability per neural site* (ρ^{IRO}_{site}): To estimate the reliability of the responses per 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 that site's best (highest response-driving) stimulus 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 as compared to 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:

$$selectivity_i = \frac{mean(\vec{b}_i) - mean(\vec{w}_i)}{\sqrt{\frac{var(\vec{b}_i) + var(\vec{w}_i)}{2}}}$$

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 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 (i.e., 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 neuron's spiking threshold). Here we have 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, where the objective function for the minimization comprises of an additional term (to reduce model complexity),

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

where β and p are the classifier weights associated with 'p' predictors (e.g. 424 neurons). The strength of regularization, λ was optimized for each train-set and a stochastic gradient descent solver was used to estimate 10 (one for each object) one-vs-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,

$$sc = R\beta + bias,$$

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

The train and test sets were pseudo-randomly chosen multiple times until we 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.

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

Our behavioral I_1 scores are all trial-averaged metrics. Therefore, in order 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 are therefore trained and tested with trial-averaged data.

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

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

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

Object solution time per image in IT (OST_{image})—Object solution time 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. In order to estimate this time, we first computed a neural I_1 vector for nonoverlapping 10 ms time bins post 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 ~9ms on average, as determined via bootstrapping across repetitions).

We balanced the control and challenge image populations at each level of the monkeys' performance. 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'>=5 to be 150.2 ms (well within the range of other challenge image OSTs). Deep Convolutional Neural Networks (DCNN)

Binary object discrimination tasks with DCNNs

We have used two different techniques to train and test the DCNN features on the binary object discrimination task.

1. **Back-end training (transfer learning):** Here we have 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 (e.g AlexNet). We then replaced the last three layers (i.e. anything beyond 'fc7') of this network with a fully connected layer containing 10 nodes (each representing one of the 10 objects we have 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 IT) on a subset of images from our image-set (containing both control and challenge images). These images

were selected randomly from our imageset and used as the train-set. The remaining images were then used for the testing (such that there is no overlap between the train 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 1000 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 ~20000 features) and control for the total number of features used in the analyses.

2. **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 IT 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-tune (end-to-end) the ImageNet pre-trained AlexNet with images (randomly selected from our own image-set) and test on the remaining held out images. This technique also involves first obtaining an imagenet pertained DCNN, and replacing the final 3 layers (e.g. beyond AlexNet 'fc7') with a fully connected layer of 10 nodes. However, the key difference of this technique with the transfer learning technique is that the new network is now trained end-to-end with stochastic gradient decent on separate training images from our own image-set used to test the monkeys. Figure S14 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 Figure S11A). We first extracted the features per image, from the DCNNs' layers. The features extracted were then projected onto its first 1000 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 Figure 4A. Using a 50%/50% train/test split of the images, we then estimated the regression weights (i.e how we can linearly combine the model features to predict the neural site's responses) 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

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

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

The percentage of explained variance, *IT predictivity* (for details refer [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 ($\rho^{M^{PRED}}$) for that site (estimated by a Spearman Brown corrected trial-split correlation score).

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

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 have exposed the mapping procedure (mapping the model features on to 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 sub-sampled 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 (Figure S15A). We also tested the effect of timebins used for mapping on %EV (Figure S15B).

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 object solution time (OST) estimates. Each image has an associated value for different image properties, either categorical e.g. occcluded/non-occluded or continuous e.g. object size etc. 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 95th percentile of the factor distribution, and the *low* group contained the ones with values less than 5th percentile of the distribution. For the categorical factor like 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 ANOVA with object solution time as the dependent variable. The rationale behind this test was if the experimenter(s) were 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 Figure 6) to quantify the OST prediction strength.

Statistics

As test of significant difference between two variables, we have used (Bonferroni corrected) paired and unpaired t-tests, one-way ANOVA. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications (refer [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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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

References

1. DiCarlo JJ, Zoccolan D & Rust NC How does the brain solve visual object recognition? Neuron 73, 415–434 (2012). [PubMed: 22325196]
2. Riesenhuber M & Poggio T Models of object recognition. Nat Neurosci 3 Suppl, 1199–1204 (2000). [PubMed: 11127838]
3. Yamins DL & DiCarlo JJ Eight open questions in the computational modeling of higher sensory cortex. Curr Opin Neurobiol 37, 114–120 (2016). [PubMed: 26921828]
4. Hung CP, Kreiman G, Poggio T & DiCarlo JJ Fast readout of object identity from macaque inferior temporal cortex. Science 310, 863–866 (2005). [PubMed: 16272124]
5. Majaj NJ, Hong H, Solomon EA & DiCarlo JJ Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. J Neurosci 35, 13402–13418 (2015). [PubMed: 26424887]
6. Cadieu CF, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10, e1003963 (2014). [PubMed: 25521294]
7. Yamins DL, et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111, 8619–8624 (2014). [PubMed: 24812127]
8. Guclu U & van Gerven MA Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J Neurosci 35, 10005–10014 (2015). [PubMed: 26157000]
9. Rajalingham R, Schmidt K & DiCarlo JJ Comparison of Object Recognition Behavior in Human and Monkey. J Neurosci 35, 12127–12136 (2015). [PubMed: 26338324]
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. bioRxiv, 240614 (2018).
11. Rockland KS & 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). [PubMed: 2754047]
12. Felleman DJ & Van Essen DC Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1, 1–47 (1991). [PubMed: 1822724]
13. Rockland KS, Saleem KS & Tanaka K Divergent feedback connections from areas V4 and TEO in the macaque. Vis Neurosci 11, 579–600 (1994). [PubMed: 8038130]
14. Rockland KS & Van Hoesen GW Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey. Cereb Cortex 4, 300–313 (1994). [PubMed: 8075534]
15. Thorpe S, Fize D & Marlot C Speed of processing in the human visual system. Nature 381, 520–522 (1996). [PubMed: 8632824]

16. Hinton GE, Dayan P, Frey BJ & Neal RM The "wake-sleep" algorithm for unsupervised neural networks. Science 268, 1158–1161 (1995). [PubMed: 7761831]
17. Geirhos R, et al. Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969 (2017).
18. Lamme VA & Roelfsema PR The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci 23, 571–579 (2000). [PubMed: 11074267]
19. Krizhevsky A, Sutskever I & Hinton GE ImageNet classification with deep convolutional neural networks. in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 1097–1105 (Curran Associates Inc., Lake Tahoe, Nevada, 2012).
20. Lin T-Y, et al. Microsoft coco: Common objects in context. in European conference on computer vision 740–755 (Springer, 2014).
21. Meyers EM, Freedman DJ, Kreiman G, Miller EK & Poggio T Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neurophysiol 100, 1407–1419 (2008). [PubMed: 18562555]
22. Oram MW Contrast induced changes in response latency depend on stimulus specificity. J Physiol Paris 104, 167–175 (2010). [PubMed: 19944159]
23. Rolls ET, Baylis GC & Leonard CM 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). [PubMed: 4071982]
24. Op De Beeck H & Vogels R Spatial sensitivity of macaque inferior temporal neurons. J Comp Neurol 426, 505–518 (2000). [PubMed: 11027395]
25. Willenbockel V, et al. Controlling low-level image properties: the SHINE toolbox. Behav Res Methods 42, 671–684 (2010). [PubMed: 20805589]
26. McKee JL, Riesenhuber M, Miller EK & Freedman DJ Task dependence of visual and category representations in prefrontal and inferior temporal cortices. J Neurosci 34, 16065–16075 (2014). [PubMed: 25429147]
27. Bugatus L, Weiner KS & Grill-Spector K Task alters category representations in prefrontal but not high-level visual cortex. Neuroimage 155, 437–449 (2017). [PubMed: 28389381]
28. Liao Q & Poggio T Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv preprint arXiv:1604.03640 (2016).
29. Schrimpf M, et al. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv, 407007 (2018).
30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J & Wojna Z Rethinking the inception architecture for computer vision. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2818–2826 (2016).
31. Szegedy C, Ioffe S, Vanhoucke V & Alemi AA Inception-v4, inception-resnet and the impact of residual connections on learning. in AAAI 12 (2017).
32. He K, Zhang X, Ren S & Sun J Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition 770–778 (2016).
33. Russakovsky O, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252 (2015).
34. Kubilius J, et al. CORnet: Modeling the Neural Mechanisms of Core Object Recognition. bioRxiv, 408385 (2018).
35. Stojanoski B & Cusack R Time to wave good-bye to phase scrambling: creating controlled scrambled images using diffeomorphic transformations. J Vis 14 (2014).
36. Fahrenfort JJ, Scholte HS & Lamme VA Masking disrupts reentrant processing in human visual cortex. J Cogn Neurosci 19, 1488–1497 (2007). [PubMed: 17714010]
37. Elsayed GF, et al. Adversarial examples that fool both human and computer vision. arXiv preprint arXiv:1802.08195 (2018).
38. Spoerer CJ, McClure P & Kriegeskorte N Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. Front Psychol 8, 1551 (2017). [PubMed: 28955272]
39. Tang H, et al. Recurrent computations for visual pattern completion. arXiv preprint arXiv:1706.02240 (2017).

40. Walther D, Rutishauser U, Koch C & Perona P Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. Computer Vision and Image Understanding 100, 41–63 (2005).
41. Bichot NP, Heard MT, DeGennaro EM & Desimone R A Source for Feature-Based Attention in the Prefrontal Cortex. Neuron 88, 832–844 (2015). [PubMed: 26526392]
42. Tovee MJ Neuronal processing. How fast is the speed of thought? Curr Biol 4, 1125–1127 (1994). [PubMed: 7704578]
43. van Kerkoerle T, et al. Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. Proc Natl Acad Sci U S A 111, 14332–14341 (2014). [PubMed: 25205811]
44. Fyall AM, El-Shamayleh Y, Choi H, Shea-Brown E & Pasupathy A Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. Elife 6 (2017).
45. 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). [PubMed: 10537108]
46. Bar M, et al. Top-down facilitation of visual recognition. Proc Natl Acad Sci U S A 103, 449–454 (2006). [PubMed: 16407167]
47. Seger CA 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). [PubMed: 17919725]
48. Chatfield K, Simonyan K, Vedaldi A & Zisserman A Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014).
49. Simonyan K & Zisserman A Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
50. Jeurissen D, Self MW & Roelfsema PR Serial grouping of 2D-image regions with object-based attention in humans. Elife 5 (2016).

Methods-only references

51. Santos A, et al. Evaluation of autofocus functions in molecular cytogenetic analysis. J Microsc 188, 264–272 (1997). [PubMed: 9450330]
52. Rosenholtz R, Li Y & Nakano L Measuring visual clutter. J Vis 7, 17 11–22 (2007).
53. Baker CI, Behrmann M & Olson CR Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. Nat Neurosci 5, 1210–1216 (2002). [PubMed: 12379864]

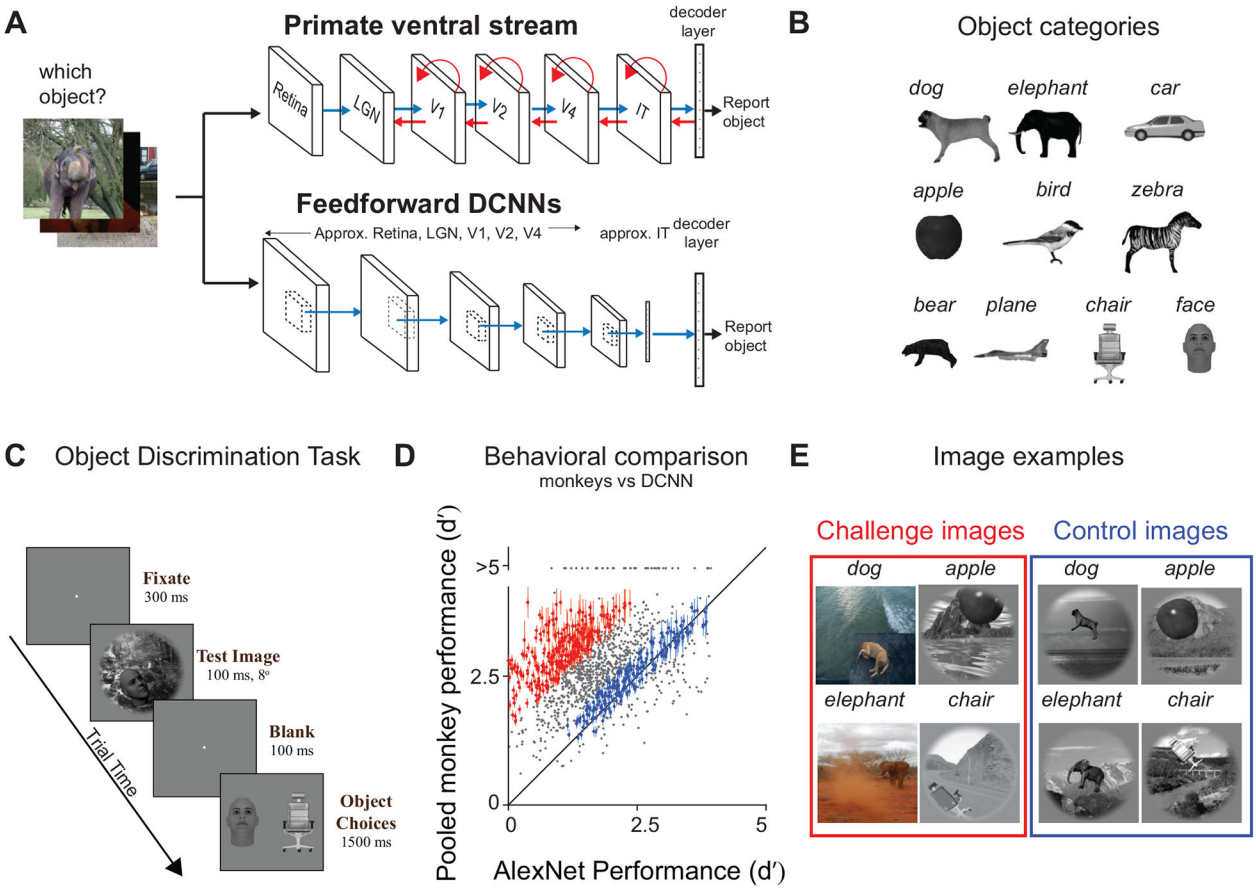

Figure 1.

Behavioral screening and identification of control and challenge images. A) We task both primates (humans and macaques; top row) and feedforward DCNNs (bottom row) to identify which object is present in each Test image (1320 images). The top row shows the stages in the ventral visual pathway in primates (retina, LGN: lateral geniculate nucleus, areas V1, V2, V4, and IT), which is implicated in core object recognition. We can conceptualize each stage as rapidly transforming the representation of the image ultimately yielding to the primates' behavior (i.e. 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. The bottom row demonstrate a schematic of a similar pathway commonly present in the DCNNs. These networks contain a series of convolutional and pooling layers with nonlinear transforms at each stage, followed by fully connected layers (which approximates macaque IT neural responses) that ultimately gives rise to the models' "behavior." Note that the DCNNs only have feedforward (blue) connections. B) Object categories. We used ten different object types; bear, elephant, face, plane, dog, car, apple, chair, bird and zebra. C) Binary object discrimination task. Here we show the timeline of events on each trial. Subjects fixate a dot. The test image 8° containing one of ten possible objects was shown for 100 ms. After a 100 ms delay, a canonical view of the target object (the same noun as that present in the test image) and a distractor object (from the other 9 objects) appeared, and the human or monkey indicated which object was present in the test image by clicking on or making a saccade to

one of the two choices respectively. D) Comparison of monkey performance (pooled across 2 monkeys) and DCNN performance (AlexNet; 'fc7'[19]). Each dot represents the behavioral task performance (I_1; refer Methods) for a single image. We reliably identified *challenge* images (red dots) and *control* images (blue dots). Error bars are bootstrapped s.e.m. E) Examples of four *challenge* and four *control* images.

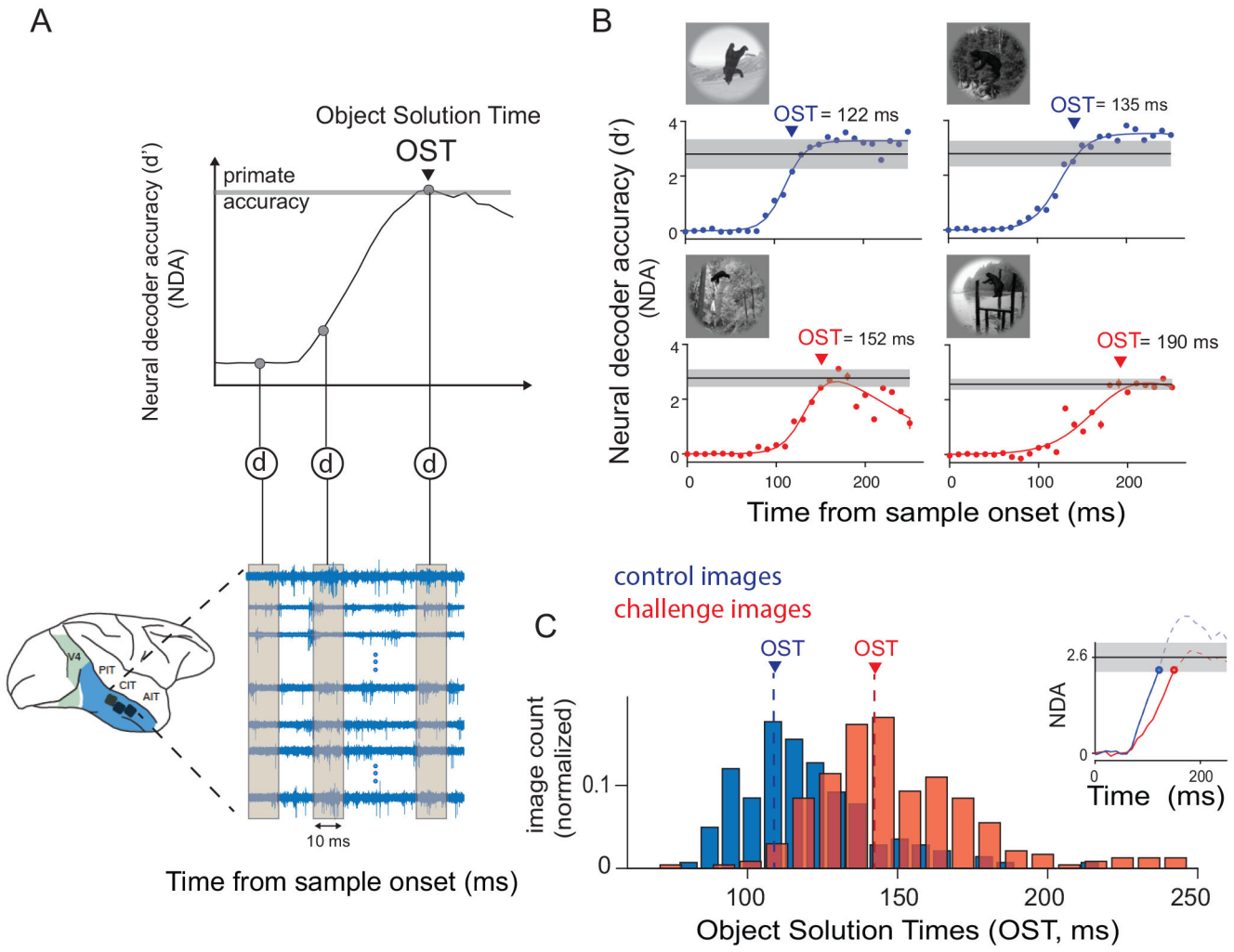

Figure 2.

Large scale multiunit array recordings in the macaque inferior temporal cortex. A) Schematic of array placement, neural data recording and object solution time estimation. We recorded extracellular voltage in IT from two monkeys, each hemisphere implanted with 2 or 3 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 support vector machine 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 neural decode accuracies (NDA) at each time bin. An example of the neural decode accuracy over time is shown in the top panel. The time at which the neural decodes equal the primate (monkey) performance, is then recorded as the object solution time (OST) for that specific image. B) Examples of IT population decodes over time, with the estimated object solution times for four images; two *control* (top panel: blue curves) and two *challenge* images (bottom panel: red curves). The red and blue dots are the estimated neural decode accuracies at each time bins. The solid 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 bootstrapped s.e.m. C) Distribution of object solution times for both *control* (blue) and *challenge* (red) images. The median OST for *control* (blue) and *challenge* (red) images are shown in the plot with dashed lines. The inset in the top shows the median evolution of IT decodes over time until the OST for control (blue) and challenge (red) images.

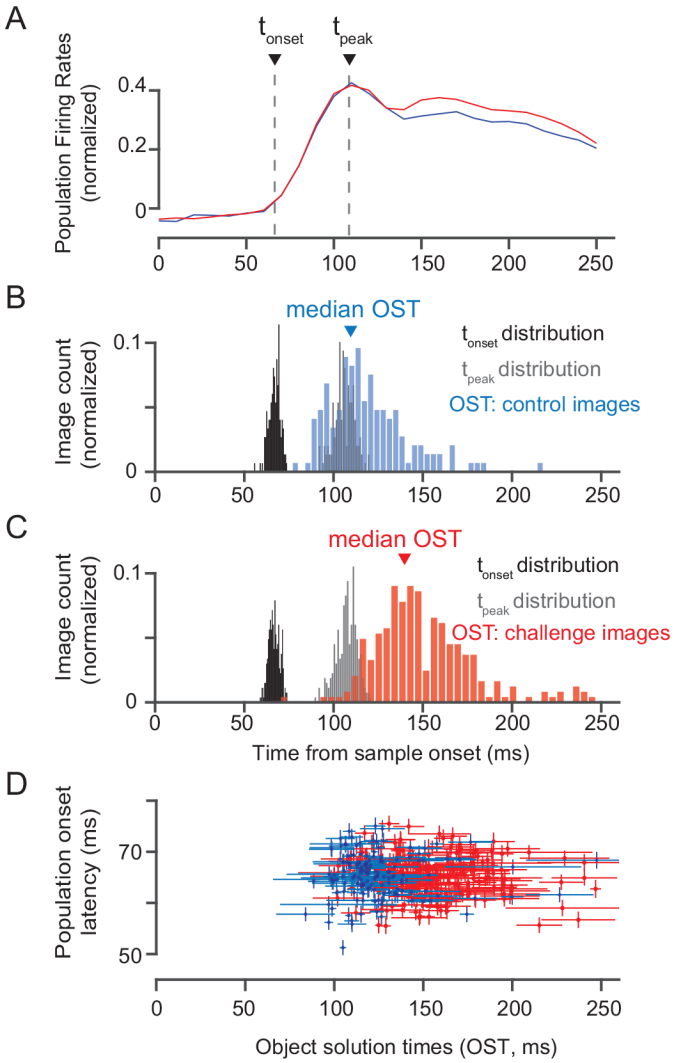

Figure 3.

Relationship between object solution times 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 object solution times for *control* images (n=149). C) Same as in B) but for *challenge* images (n = 266). D) Comparison of population onset latencies and object solution times 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 deviation of *OST* estimates.

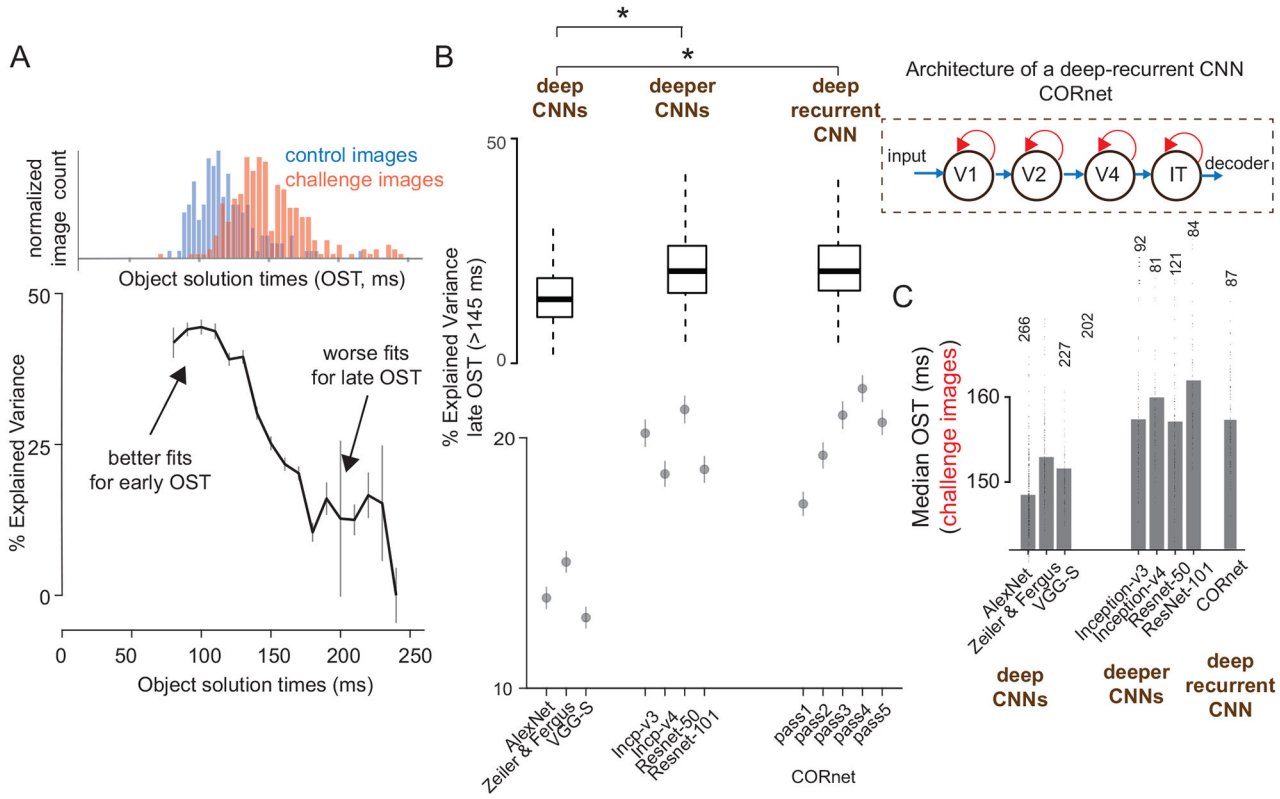

Figure 4.

Predicting IT neural responses with DCNN features. A) IT predictivity of AlexNet's 'fc7' layer as a function of object solution time (ms). For each time bin, we consider IT predictivity only for images that have a solution time equal to or higher than that time bin. Error bars indicate the standard error of mean across neurons (n = 424 neurons considered for each time bin). Top panel shows the distribution of object solution times for *control* (blue; n=149 images) and *challenge* (red; 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 object solution times, as function of deep (AlexNet, Zeiler and Fergus, VGG-S), deeper (Inception, ResNet) CNNs and deep-recurrent CNNs (CORnet). Dots indicate the median and errorbars indicate s.e.m across neurons. * indicates a statistically significant difference between two groups. We used paired t-tests; deep (avg. of all 3 networks used) vs deeper CNNs (avg. of all 4 networks used): $t(423) = 14.26$, $p < 0.0001$, deep (avg. of all 3 networks used) vs deep-recurrent (average of pass 3 and pass 4) CNN: paired t-test; $t(423) = 15.13$, $p < 0.0001$ 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, the whiskers extend to the most extreme data-points the algorithm considers to be not outliers. Outliers are datapoints 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. C) Comparison of median OST 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, and the number of images is indicated on top

of the bars as well as the OST per image is plotted as dots 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 text). Note that the use of deeper CNNs and the deep-recurrent CNN, resulted in the discovery of *challenge* images that required even longer *OST*s in IT cortex than the original set *challenge* images (defined for AlexNet 'fc7'). * indicates a statistically significant difference between two groups.

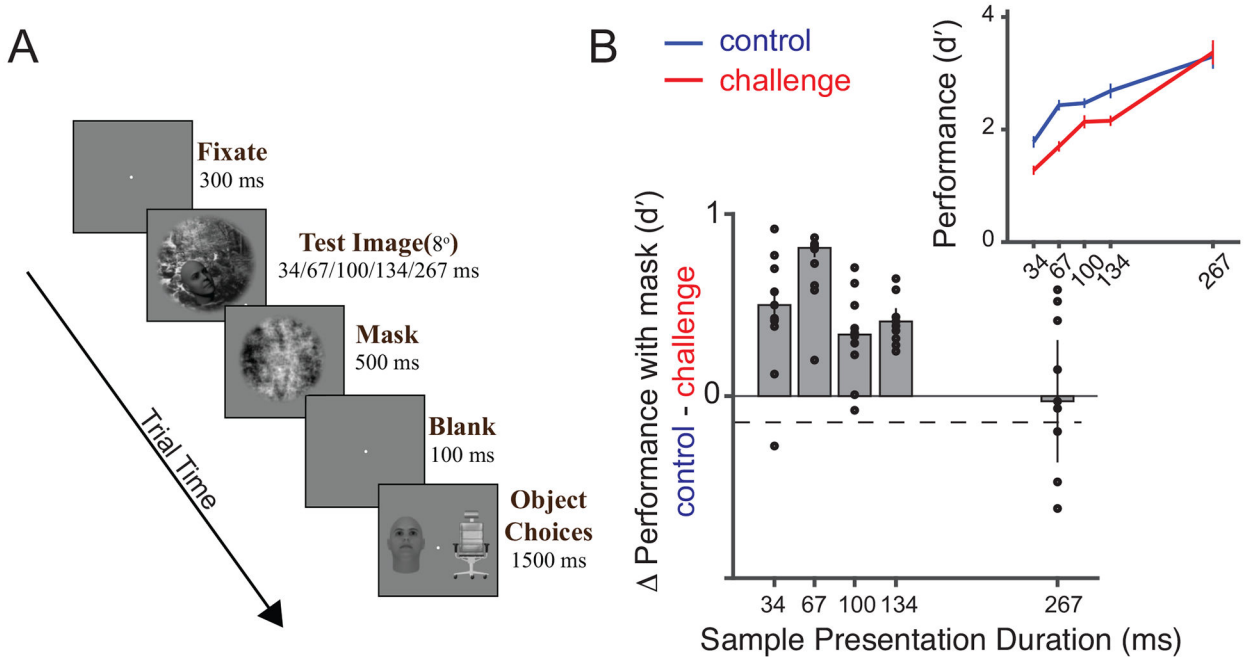

Figure 5.

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* image 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 *control* and *challenge* images at the respective sample image presentation durations (x-axis). The dashed black line denotes the difference in performance between the control and challenge images without backward masking at 100 ms presentation. n = 10 objects considered per presentation duration. Each dot corresponds to the difference in performance per object. The top panel inset shows the raw performance (d') for the two groups of images (blue: *control* images, red: *challenge* images). Error bars denote the standard error of mean across all objects.

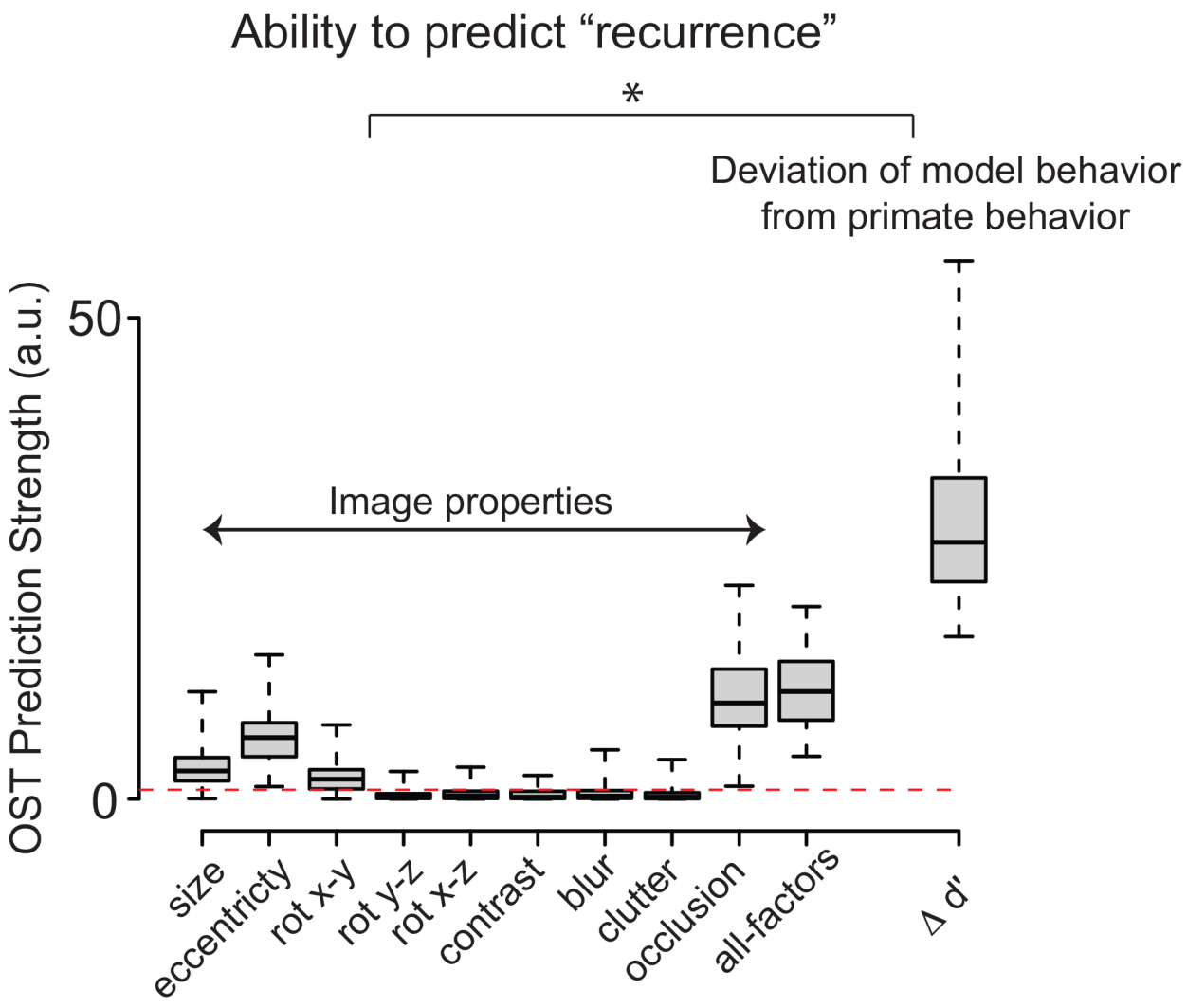

Figure 6.

Comparison of ΔST prediction strength between different image properties, a combination of all estimated image properties, and the $\Delta d'$ vector (deviation of model behavior from pooled monkey behavior). n = 64 images (32 high, 32 low; refer Methods) for each image group was used. The red dashed line denotes the significance threshold of the F-statistic. Image properties like object size, eccentricity, presence of an occluder, as well as a combination of these properties (referred to as “all-factors”) significantly predict *OST*. However, the $\Delta d'$ vector provides the strongest OST predictions. Error bars denote the bootstrap standard deviation over images. * denotes a significant difference between the two groups — image properties vs $\Delta d'$, estimated with repeated measures ANOVA ($F(1,10)>100$; $p<0.0001$; multiple-comparison using Turkey test showed a significant difference between $\Delta d'$ 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, the whiskers extend to the most extreme data-points the algorithm considers to be not

outliers. Outliers are datapoints 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.