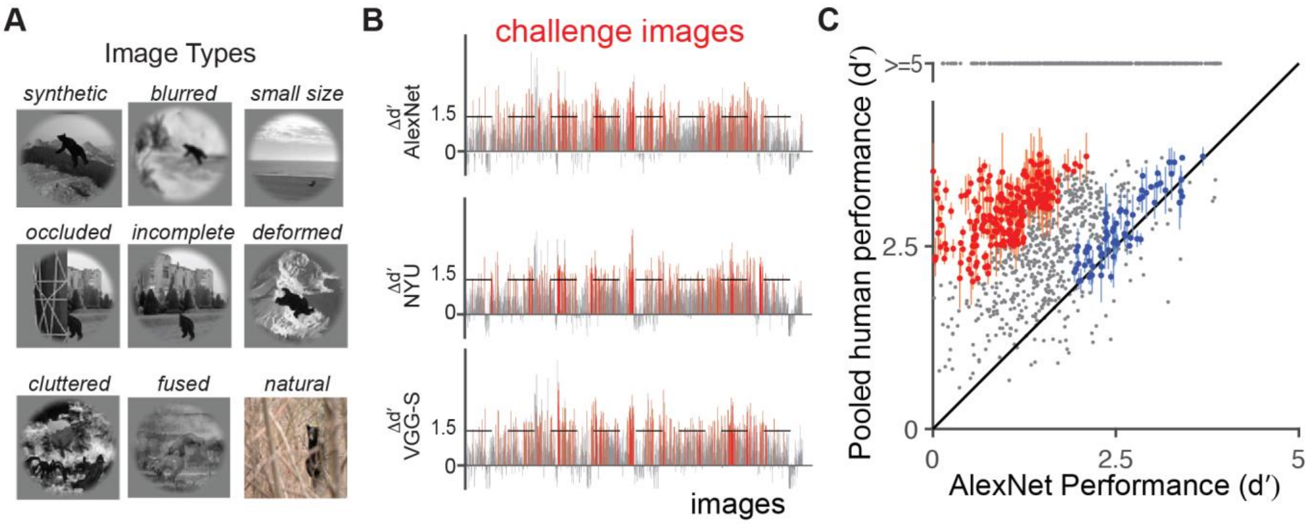
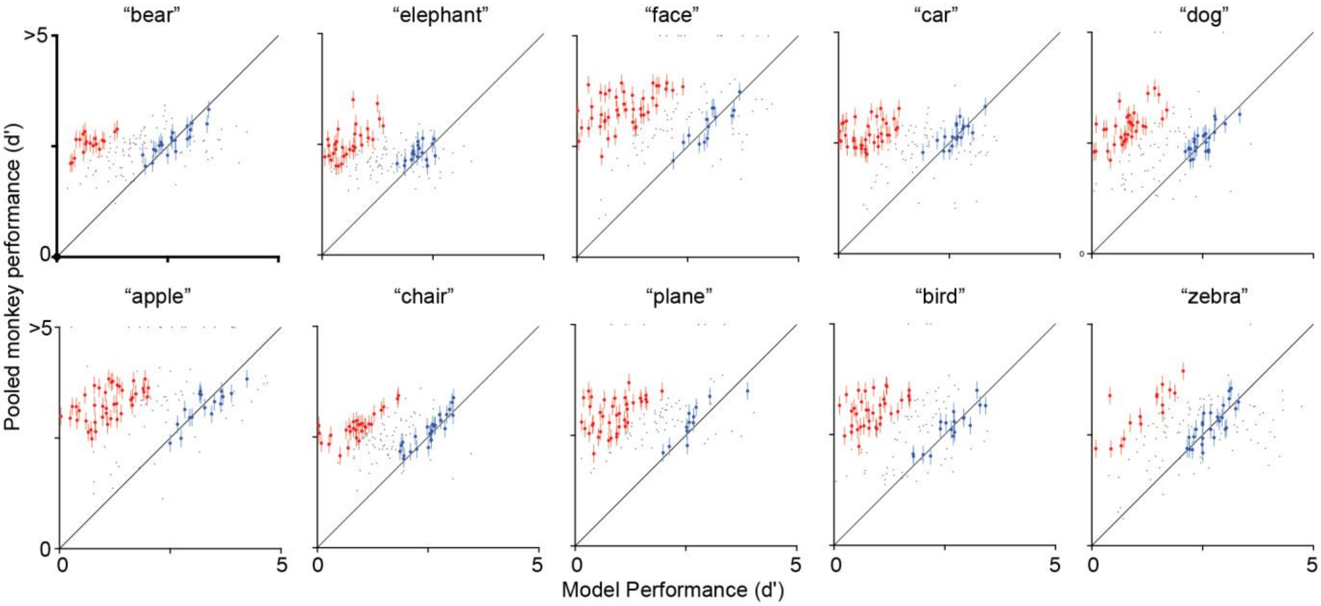


Supplementary Figure 1

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

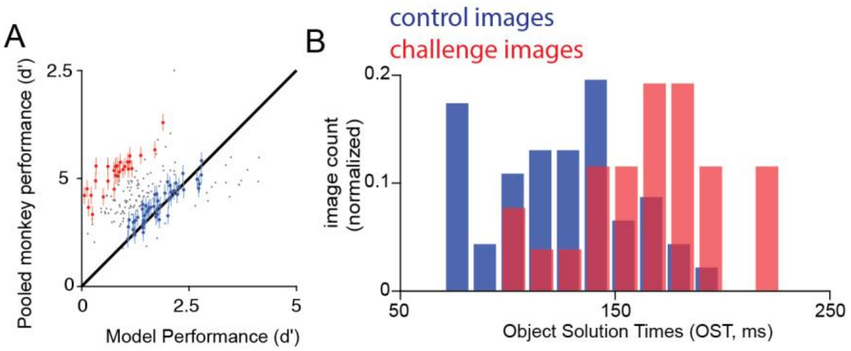
A) Examples of different image types used in the behavioral testing. Different image types included synthetic images containing an object in an uncorrelated background, images with blur, small object sizes, occlusion, incomplete objects, deformed objects, cluttered scenes, fused objects, and natural 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').

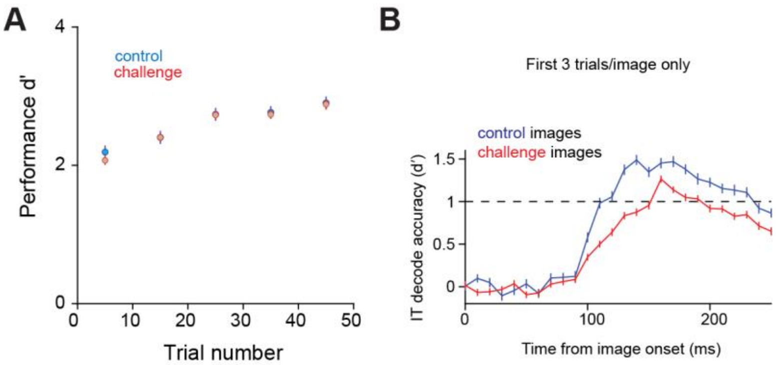
Each dot represents the behavioral task performance (I_1; refer Methods) for a single image of the corresponding object. We reliably identified *challenge* (red dots) and *control* (blue dots) images. Error bars are bootstrapped s.e.m. across 1000 resamples for 123 trials per image. n=132 images per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



Supplementary Figure 3

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

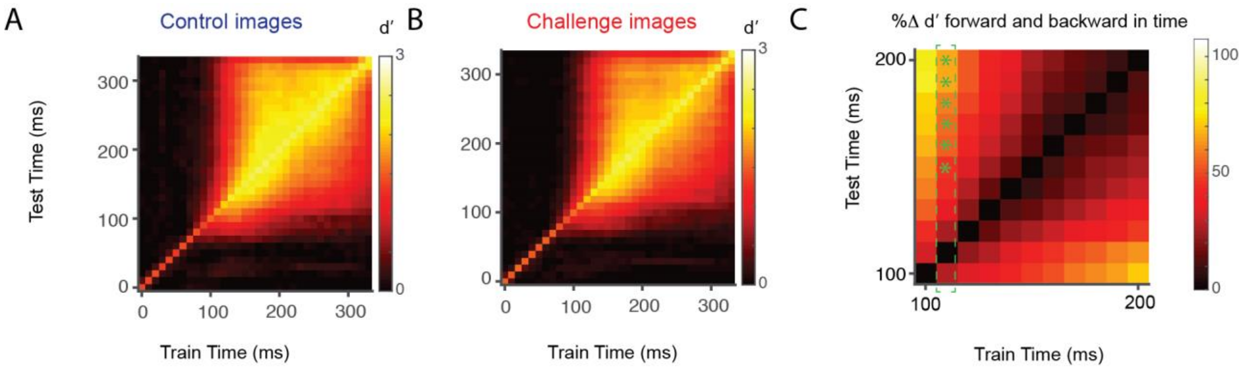
A) Comparison of AlexNet (‘fc7’) performance and pooled monkey behavior on the MS COCO images (n=200; 47 control and 38 challenge images). Errorbars show the s.e.m across 1000 resamples from 123 trials per image. B). Distribution of challenge (red) and control (blue) image OST. ∆OST was estimated at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

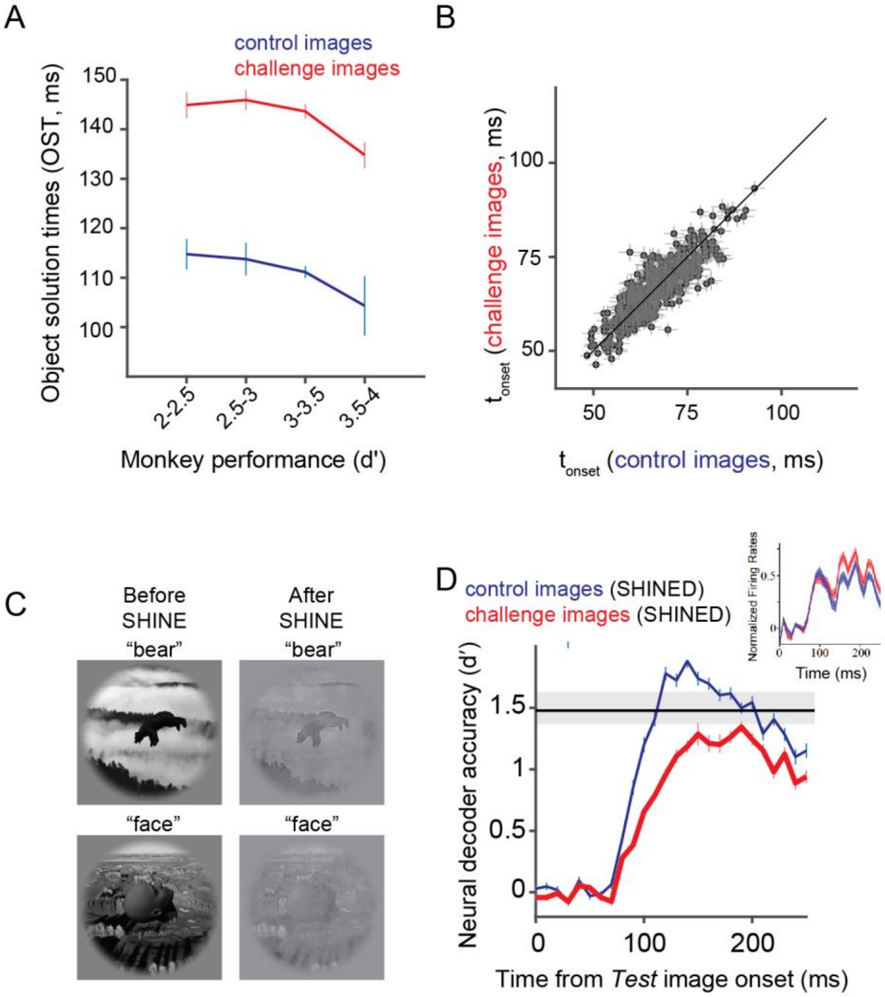
A) Change of pooled monkey behavioral performance I_1 with repeated exposure of the control (blue) and challenge (red) images. Each data point was estimated by pooling together 10 trials (around the trial numbers indicated in the x-axis). The figure shows that the control and challenge images did not 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.

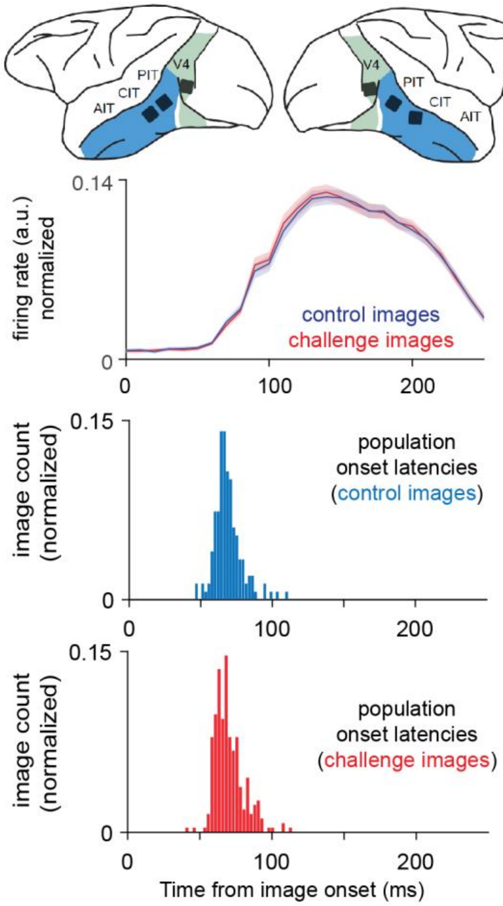
A) and B) Temporal cross training matrix for control (n=149) and challenge (n=266) images respectively, shown separately. To estimate the value at each element of the matrix, we trained a IT neural population (n=424) decoder (refer Methods) at a time ‘t1’ ms and test it at time ‘t2’ ms. C) The color 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.

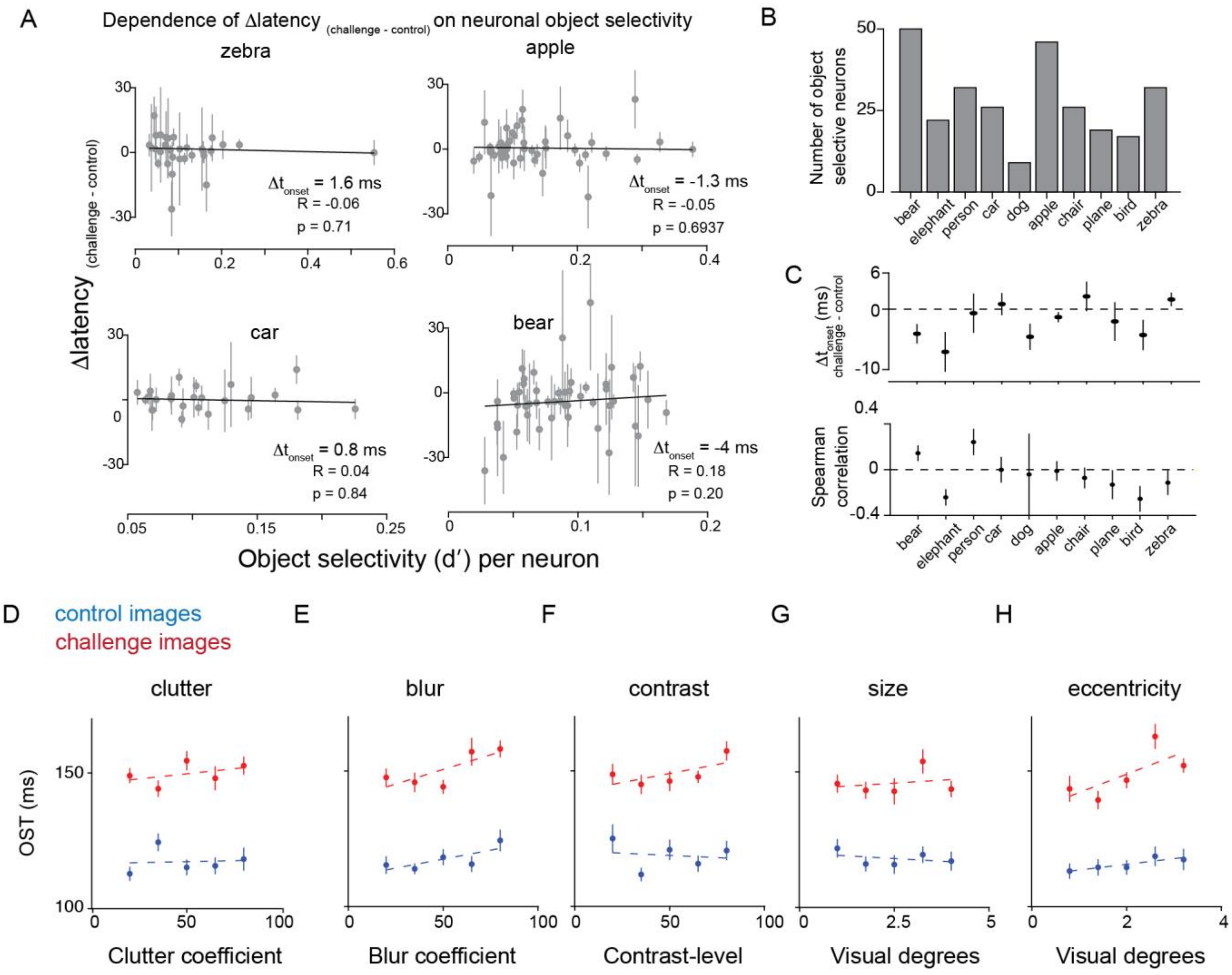
A) Dependence of OST on the pooled monkey I_1 level. The red and the blue curves show the OST values averaged across images with behavioral I_1 accuracy within the limits shown on the x-axis, for challenge (n = 67,145, 42, 12 images for each x-value) and control (n=54,44,41,10 images for each x-value) 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.

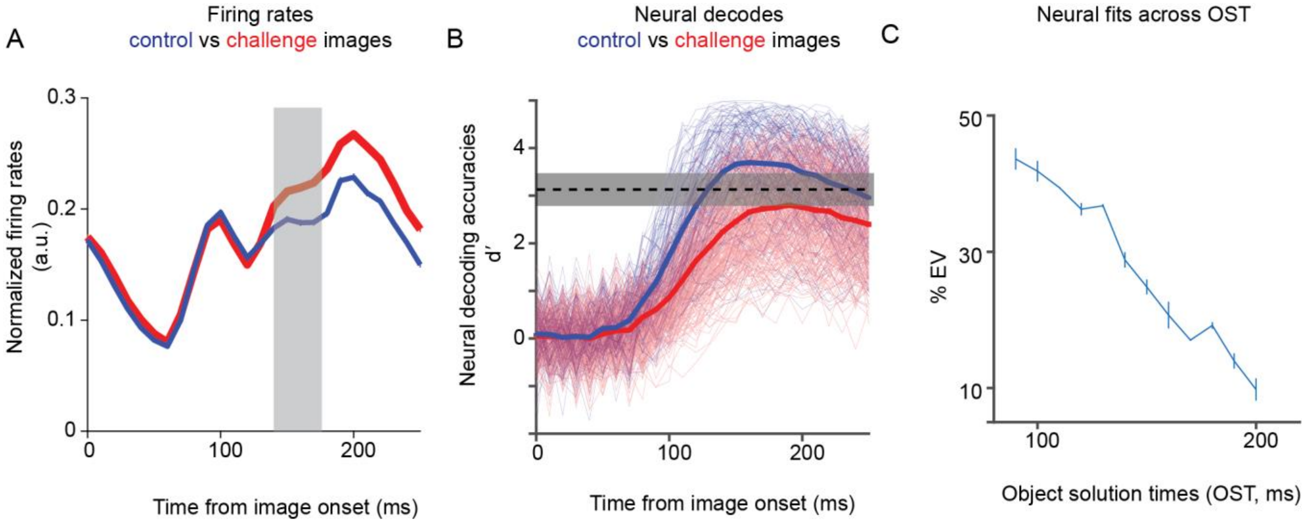
The top panel shows the placement of chronic Utah array implants in IT and V4 of two monkeys. Below it, we show the time course of normalized neural firing rates (averaged across the V4 population of 151 sites) for control (n=149 images; blue) and challenge (n=266 images; red) images. Errorclouds indicate 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.

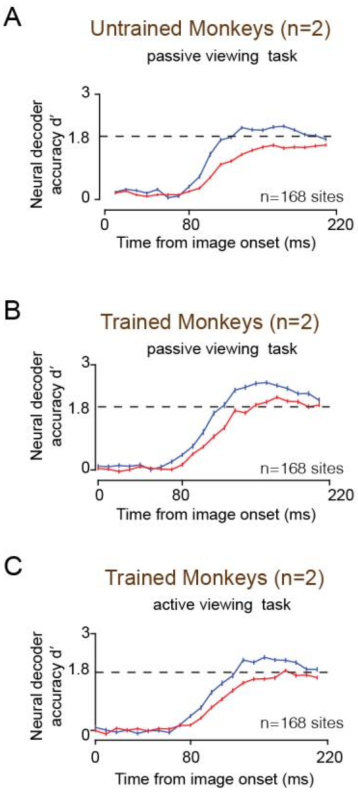
We considered the possibility that the difference in the OST between control and challenge images for each object category is primarily driven by neurons that specifically prefer that category (*object relevant neurons:* number for each category shown in B). To address this, we first asked whether the 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.

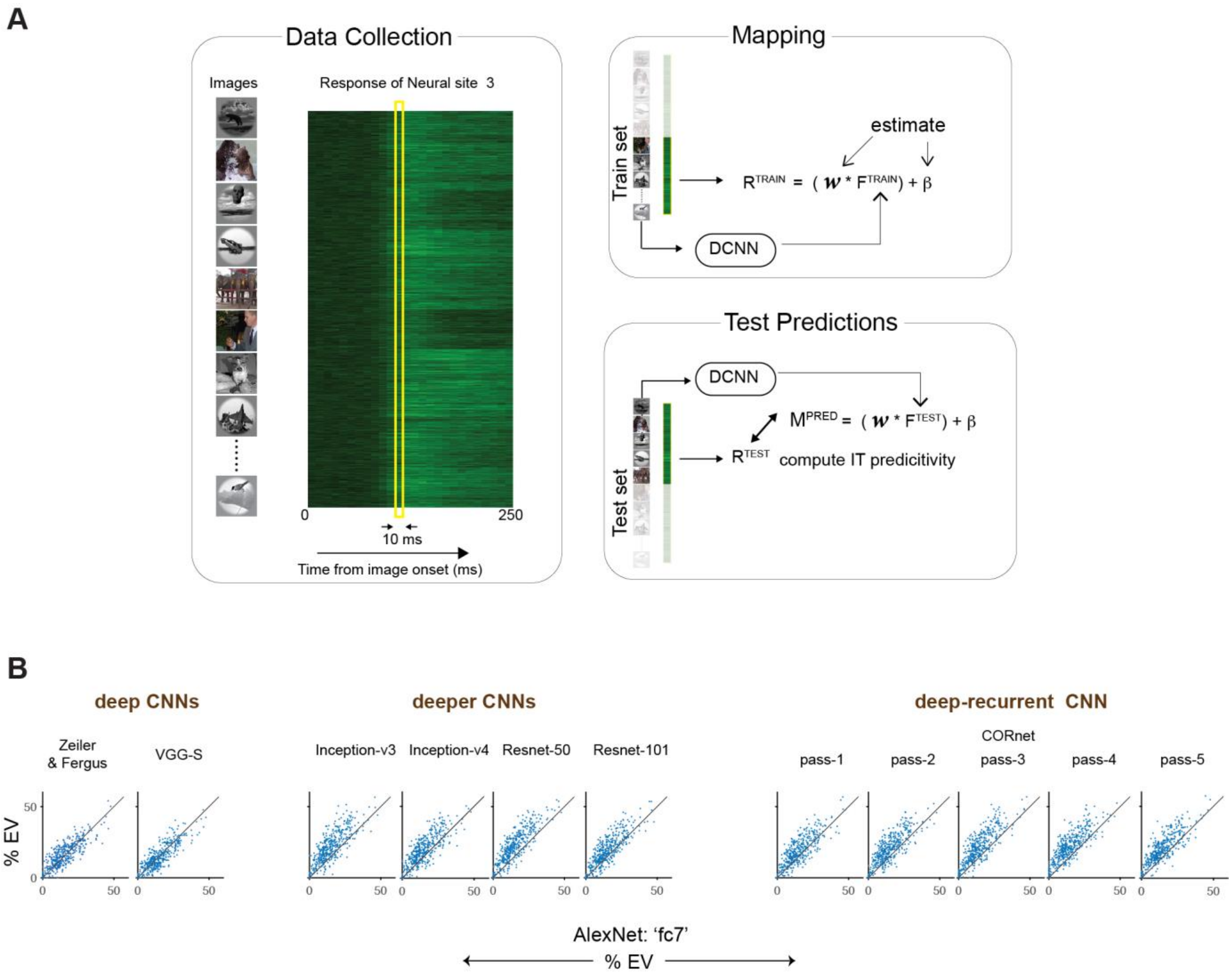
A) Comparison of normalized firing rate responses (averaged across all 424 IT sites) to the control (n=149 images; blue) and challenge images (n=266 images; red). The initial dip in the firing rate is caused by the offset responses related to the previous stimulus. The gray bar shows the time bins for 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.

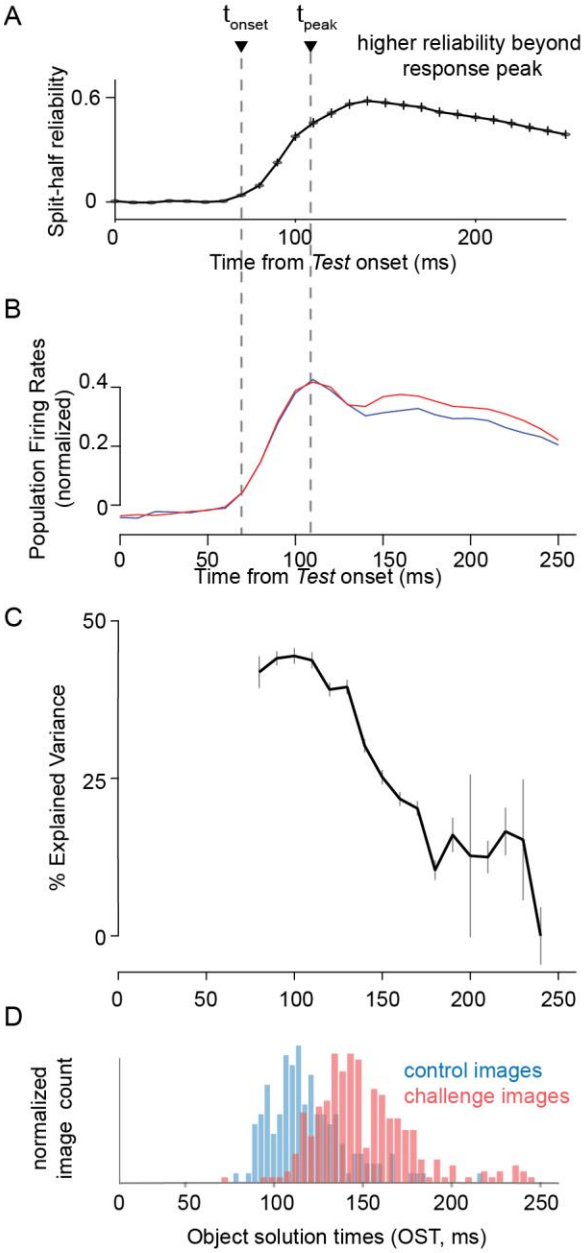
A) Results from untrained monkeys: IT population decodes over time for control (blue curve; 86 images) and challenge (red curve; 117 images) images. The threshold to estimate the decode latency, denoted by the dashed black line, was set at 1.8. The recordings were done from 168 sites (refer [6]). B) 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.

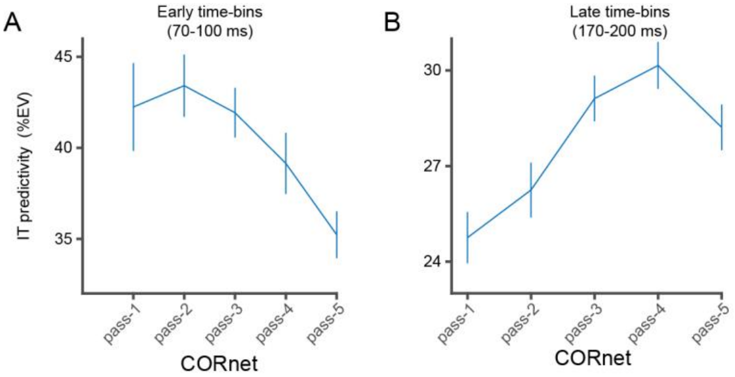
A) Schematic of the DCNN neural fitting and prediction testing procedure. This includes three main steps. Data collection: neural responses are collected for each of the 1320 images (50 repetitions), e.g. shown is that of example neural site #3, across 10 ms time-bins. Mapping: We divide the images and 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.

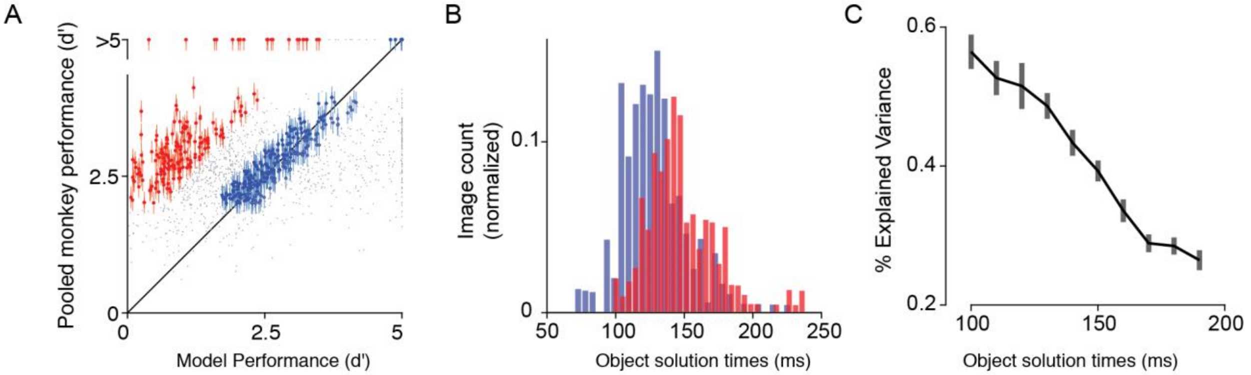
A) Reliability (or internal consistency) of neural responses as a function of time. The internal consistency was computed as a Spearman-Brown corrected correlation between two split halves (trial based) of each IT neural site's responses across all tested images. Errorbar indicates s.e.m across neurons (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.

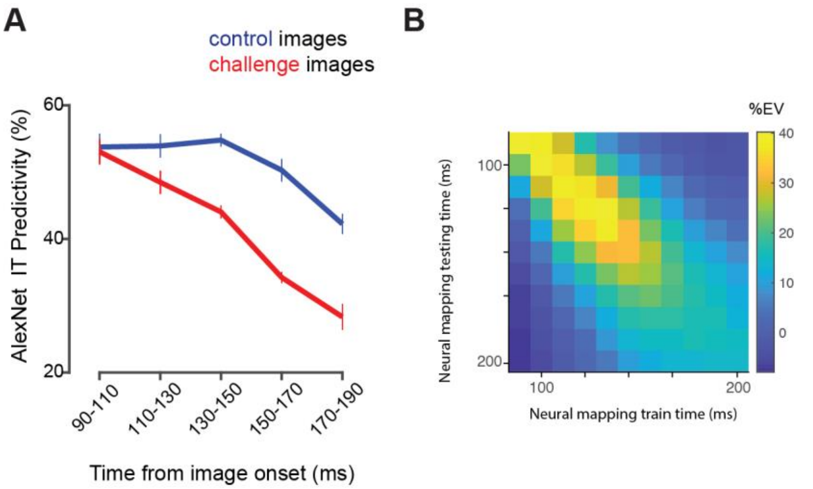
We observe that the earlier passes (pass 1 and pass 2) are better predictors of the early time bins and the prediction deteriorates for the later passes. B) IT predictivity (%EV) computed at late (170-200 ms) phases of IT responses. Here we observe that the late passes (especially pass 4) is better at 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).