IT Cortex Contains a General-Purpose Visual Object Representation

Ha Hong[1,2*], Daniel Yamins[1*], Najib Majaj[1,3], and James J. DiCarlo[1] (*equal contribution)

[1]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, MIT, Cambridge MA. [2]Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge MA. [3]Present address: Center for Neural Science, New York University, New York, New York, USA.

Extensive research has shown that Inferior Temporal (IT) cortex is a key brain area underlying invariant object recognition. More recently, it has been uncovered that position is also coded in IT and human LOC. Here we show that IT neurons support robust readout of a variety of object parameters that characterize scene description in the central visual field.

We recorded neural responses in IT and V4 to set of 5760 images of photorealistic objects in a variety of categories, placed in complex realistic scenes, with significant variation in object position, size, and in-plane and out-of-plane rotation. Consistent with known results, IT achieves high invariant categorization and identification performance for these images. We also find that the IT representation of object position is highly robust, with units that encode location accurately across the full range of tested positions — even across widely varying object geometries, pose and size variation, and cluttered backgrounds that make this task very challenging for lower-level visual representations. We find similarly robust IT encodings for a variety of additional object parameters, including size, pose, perimeter, and aspect ratio, for which lower-level representations appear to have effectively no decoding power. While IT exhibits the ability to discount identity-preserving variation to solve categorization tasks, it simultaneously encodes a suite of "identity-orthogonal" dimensions, that, combined with category and object identity encodings, form a basis for a full scene description.

Moreover, while the representation of object identity and category is highly distributed across IT sites, the representations for these other properties (e.g. position) is typically more sparsely encoded, with a small proportion of highly responsive sites responsible for much of the decoding capacity. Taken together, these results suggest that IT contains a general representation of the visual environment in which key object parameters have been extracted and factored.

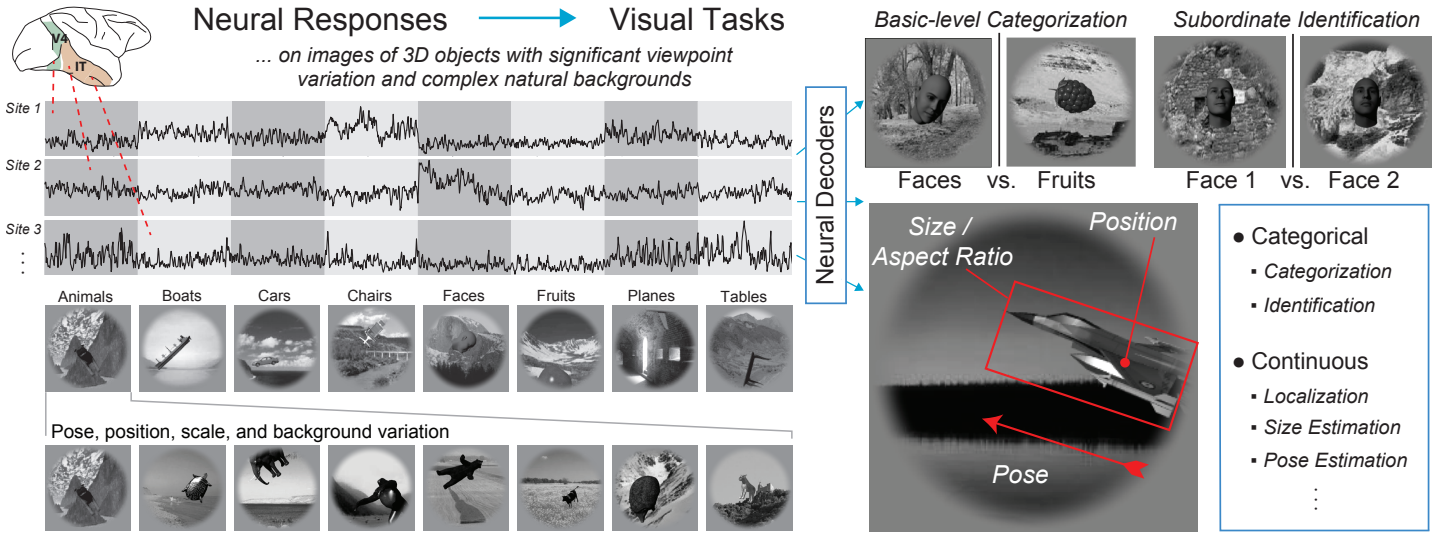


Fig. 1: We measured multi-unit neural responses to briefly presented (RSVP) images from 168 IT and 128 V4 sites in two passively fixating macaques using chronically implanted electrode arrays. We recorded 5760 images of a variety of photorealistic 3D objects in each of 8 natural categories. Objects were shown within the central 8° of the retina on complex background scenes at a broad range of position, scale, and pose views. By choosing different neural decoding rules for interpreting IT output, information relevant to a wide variety of natural tasks can be read out from the neurons. These decoding rules are simple weighted sums of neural units. The sparseness of these weightings corresponds to how distributed the task-relevant data is in the neural population. We assessed performance for a wide variety of tasks including basic-level categorization (e.g. animals vs. boats vs. cars etc.) and subordinate-level identification (e.g. face 1 vs. face 2 vs. face 3 etc.), as well as continuous-valued tasks like object position, size, pose, and aspect-ratio estimation. In the categorical cases, we used linear classifiers to find the optimal weightings; for the continuous estimation tasks, we used linear regression.

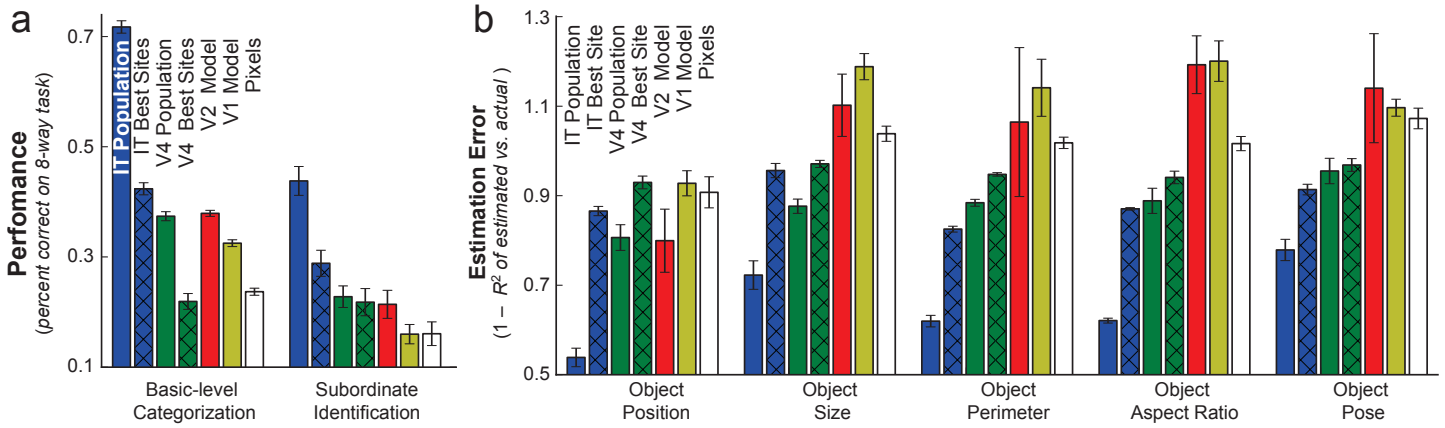
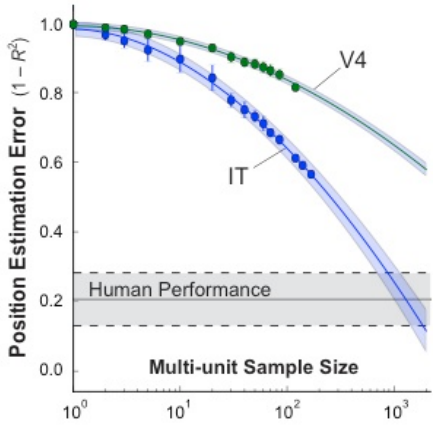
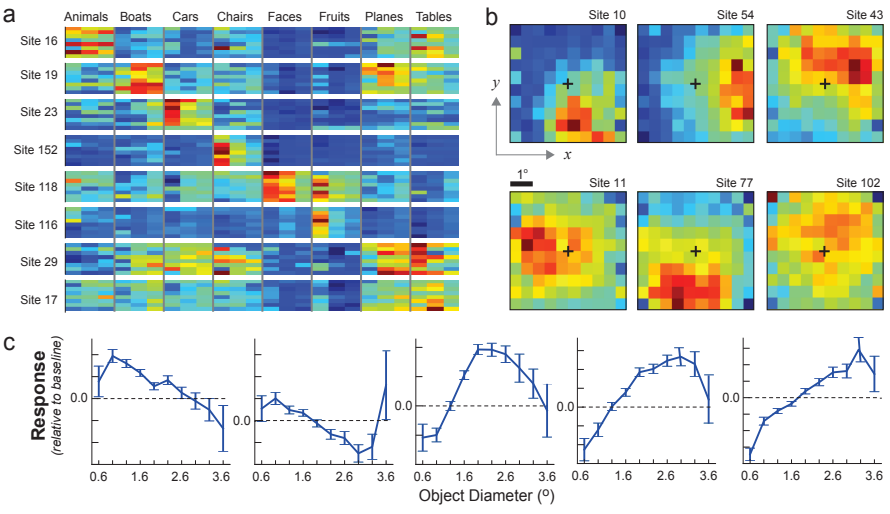
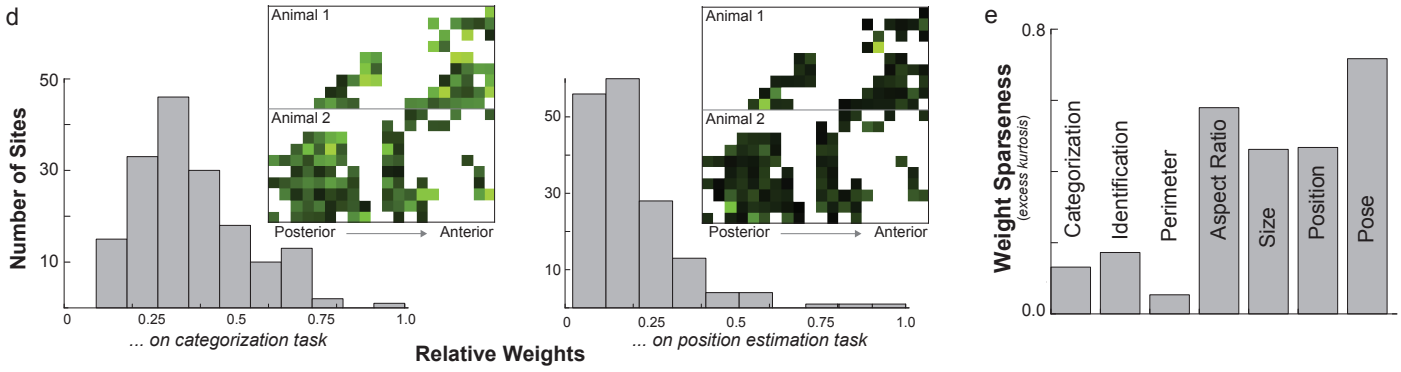


Fig. 2 (above): (***a***) Neural decoding performance on 8-way basic category and subordinate identification tasks. Results were obtained by training *L*2-regularized SVM classifiers with 75/25% train/test splits with 10-fold cross-validation. (***b***) Neural prediction accuracy for estimation tasks including object position, size, perimeter, aspect ratio, and pose. Results were obtained by *L*2-regularized linear regression with the same cross-validation protocol as for the classifiers. For both classification and estimation tasks, the population of sampled IT multi-units ($n=168$, blue bars) is able to achieve high performance, while the intermediate visual area V4 population ($n=128$, green bars) does not perform as well. Comparison to a V2-like Gabor-conjunction model [1] (red bars), a V1-like Gabor-based model (yellow bars) and image pixels (white bars) show that simpler lower-level representations fail to capture these properties for the complex images used here. **Fig. 3** (right): Extrapolating position-estimation performance as a function of neural sample size suggests that IT neurons reach human level at ~1200 multi-units, while V4 would require $\geq 10^7$. Human performance was measured in an online crowd-sourced behavioral experiment ($n=97$). **Fig. 4** (below): (***a***) Evoked responses of the most discriminative IT sites for each category (red: elevated, blue: baseline). Response average is shown for each of eight category exemplars (*y*-axis) taken over three pose, position and size parameter ranges (*x*-axis). (***b***) Spatially-binned response averages for the most position-responsive IT units. These units have poor category selectivity, but show position tuning: access to these allows effective estimation of *x*-*y* position by e.g. triangulation. It is unknown if these units are retinotopically arranged. Crosshairs denote the screen center the animals fixated at. (***c***) Tuning curves for units that robustly encode object size. Similar tuning curves are also observed for the other estimation tasks we measured, e.g. pose, aspect ratio, and perimeter. (***d***) Distribution of weights for category and position tasks. Insets show spatial layout of weightings (green: high weights, black: low weights, white: not recorded). (***e***) The encoding for estimation properties is on average much sparser than for the categorical tasks.







[1] Freeman. J and Simoncelli, E Metamers of the Ventral Stream. *Nature Neuroscience* **14**, 1195-1201 (2011).