nature neuroscience

Explicit information for category-orthogonal object properties increases along the ventral stream

Ha Hong[1–3,5], Daniel L K Yamins[1,2,5], Najib J Majaj[1,2,4] & James J DiCarlo[1,2]

Extensive research has revealed that the ventral visual stream hierarchically builds a robust representation for supporting visual object categorization tasks. We systematically explored the ability of multiple ventral visual areas to support a variety of 'category-orthogonal' object properties such as position, size and pose. For complex naturalistic stimuli, we found that the inferior temporal (IT) population encodes all measured category-orthogonal object properties, including those properties often considered to be low-level features (for example, position), more explicitly than earlier ventral stream areas. We also found that the IT population better predicts human performance patterns across properties. A hierarchical neural network model based on simple computational principles generates these same cross-area patterns of information. Taken together, our empirical results support the hypothesis that all behaviorally relevant object properties are extracted in concert up the ventral visual hierarchy, and our computational model explains how that hierarchy might be built.

Humans rapidly process visual scenes from their environment, an ability that is critical to everyday functioning. One facet of scene understanding is invariant object recognition[1], which is a challenging computational problem because different images of the same object can have vastly different low-level statistics[2]. Extensive research has uncovered the role of the ventral visual stream, a series of connected cortical areas present in humans and other primates[3], in solving this challenge. During early neural activity (~150 ms post stimulus-onset), the ventral stream functions approximately like a hierarchy of increasingly abstract processing stages[1,2,4–6]. Neurons in the earliest ventral visual processing stage, V1, can be approximated as local edge detectors[7], but cannot support decoding of object category under complex image transformations[1]. In contrast, population activity in IT cortex, the processing stage at the top of the ventral hierarchy, can directly support invariant object categorization[8–10]. These results can be summarized in this way: the amount of easily accessible information for high-variation object categorization tasks—as measured via the performance of linear classifiers seeking to decode category labels—increases along the ventral hierarchy (**Fig. 1a**). It is this pattern of relative, explicit information content between adjacent areas in the sensory cascade, rather than the absolute, implicit information in any one area, that strongly constrains the possible neural mechanisms that might be operating in the ventral hierarchy.

Visual perception involves the estimation of variety of other object-related properties besides object categorization. Many of these properties (object position, size, orientation, heading, aspect ratio, perimeter length, etc.) are often considered to be 'nuisance' variables that must be discounted to achieve invariant recognition. But humans do in fact perceive all of these category-orthogonal visual object properties in images, raising the question of what overall neural architecture underlies both tolerance to identity-preserving variable transformations needed for object categorization tasks and sensitivity to these same variables for other scene-understanding tasks. Although much research has investigated position sensitivity for simple stimuli in lower ventral visual areas such as V1 (ref. 11), relatively little work has focused on comparing such properties across ventral visual areas, especially in complex natural scenes. As a result, the patterns of information for these properties have remained largely unknown (**Fig. 1a**).

These patterns bear on a number of hypotheses about the ventral stream. One hypothesis is an intuitively appealing 'local coding' idea that directly generalizes Hubel and Wiesel's simple-to-complex dichotomy directly to higher visual areas: view-tuned units are aggregated across identity-preserving transformations at each scale to produce partially view-invariant units, which are themselves aggregated to produce invariance at a larger scale. A natural prediction from this conception is that there is a trade-off between increasing receptive field size and categorization ability on the one hand and orthogonal task performance on the other, so that explicitly available information for non-categorical properties decreases in higher ventral areas (hypothesis H1; **Fig. 1b**). This local coding mechanism is consistent with the observation of highly position and orientation-sensitive units in V1 (ref. 11), the observation of position-, size- and pose-invariant units in higher ventral areas[12,13], and the fact that higher ventral stream neurons have, on average, larger receptive fields and are less retinotopic than those in lower areas[1]. Local coding is also consistent with an multiple-streams hypothesis that identity-specific properties (for example, category membership) are represented in the ventral stream, whereas other visual variables (for example, position) are represented separately, either in the dorsal stream[14,15] or perhaps directly accessed from V1 (ref. 16).

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Present address: Center for Neural Science, New York University, New York, New York, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to J.J.D. (dicarlo@mit.edu).

Received 3 September 2015; accepted 17 January 2016; published online 22 February 2016; doi:10.1038/nn.4247

A second hypothesis is that non-categorical properties rely on intermediate visual features, analogous to 'border-ownership cells' that have been discovered in V2 (ref. 17). On this view, information for properties such as object perimeter length or aspect ratio might peak in the middle of the ventral stream (H2; **Fig. 1b**).

Experimentally, it has been observed that IT cortex maintains some sensitivity to position, pose and other properties[18–24]. Notably, this work only showed that some amount of this information was present in IT, and it did not compare across areas or reference to human levels of performance on the same tasks (**Fig. 1a**). Despite those limitations, such results have been used to argue for a third hypothesis (H3; **Fig. 1b**), in which information for low-level orthogonal properties is not lost along the ventral hierarchy, but is instead preserved because it may be behaviorally useful. A line of theoretical work has suggested factored representation schemes that retain nuisance variable information while still building category selectivity[2,18,25]. This view is suggested as one possibility in some of our own previous studies[18] and is consistent with ideas of hyperacuity[26].

A final hypothesis is that information increases for the category-orthogonal object tasks, just as it does for categorization tasks (H4; **Fig. 1b**). The ideas of coarse coding show how larger receptive fields could in theory be used to achieve greater accuracy in estimating properties such as object position[27,28]. Such ideas suggest alternatives to the multiple streams hypothesis, potentially avoiding some feature-binding problems[29] associated with that concept.

We investigated this issue systematically by recording neural responses in IT and V4 cortex and testing simulated V1 neural responses to a large set of visual stimuli containing a range of real-world objects with substantial simultaneous variation in object position, size, and pose and background scene[1,8,30]. This image set allows characterization of neural encodings for standard object categorical tasks as well as a variety of category-orthogonal object property estimation tasks. We quantified the amount of explicitly available information in each ventral stream processing stage for each task, assessing both the dependence of these measurements on the complexity of the image variation, as well as how the information is distributed across the neural populations. As a reference, we also measured human performance on each of these same category-orthogonal estimation tasks using the same images.

We found that, for all tasks in the high variation image set, including those often considered to be low level (for example, object position), the amount of explicitly available information progressively increased along the ventral stream, consistent with H4 (**Fig. 1b**). Moreover, unlike lower-area representations, we found that the decoded IT population performance pattern was consistent with measured human behavioral patterns across tasks. Task information is broadly distributed throughout the IT population, rather than concentrated in task-specific specialist units. We also found that, in lower variation image sets, the V1-V4-IT increase-of-information pattern attenuated, and in some cases reversed, suggesting that the amount of object variation, rather that the specific object-related task, is a key determinant of information patterns along the ventral hierarchy.

We also asked whether these empirically observed phenomena are readily captured by a hierarchical convolutional neural network derived from recent work modeling the ventral stream[30,31]. Even though it was not explicitly optimized for category-orthogonal task estimation, the network accurately predicts the patterns of information along the ventral hierarchy across tasks and variation levels. Taken together, our empirical results suggest that, just as with the perception of object category, the perception of category-orthogonal object properties is constructed by the ventral visual hierarchy, and our computational models provide insight into how that hierarchy is built.

RESULTS

Large-scale array electrophysiology in macaque V4 and IT

We measured macaque IT and V4 neural population responses to a stimulus set containing 5,760 images of photorealistic three-dimensional objects drawn from eight common categories (**Fig. 2a** and **Supplementary Fig. 1a**). For each image, a single foreground object was rendered at high levels of position, scale and pose variation and placed on a randomly selected cluttered natural scenes. This image set supports testing of standard object recognition tasks, including basic-level categorization (for example, faces versus cars), as well as subordinate object identification (Toyota versus BMW). Because of the high variation levels, recognition in this image set is challenging for most artificial vision systems, but is robustly solved by humans[2,30,32] and by linear IT decodes[8]. Monkeys

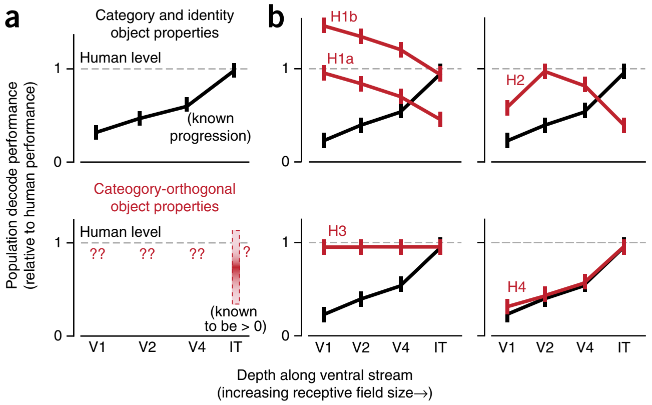

Figure 1 Illustration of possible scenarios. (**a**) Prior to this study, extensive research has shown that invariant category recognition performance increases along the ventral pathway[1] (top), whereas lower and intermediate visual areas are sensitive to various categorical-orthogonal properties (position, border continuity, etc.) in simple stimuli[11,17]. It was also known that IT contains some information for category-orthogonal properties[18–21,24]: as illustrated (bottom), performance in IT must be above floor. (**b**) However, the previous literature determined neither the relative amounts of explicitly decodable information for category-orthogonal properties between ventral cortical areas nor the ratio of neural decode performance in IT (or elsewhere) to measured behavioral performance levels. In other words, there were multiple qualitatively different hypotheses consistent with the known data as to both the red curve's shape and its height on the *y* axis. In hypothesis H1a, there is a tradeoff between increasing receptive field size and categorization ability, and performance on the orthogonal task. Early areas match human performance on these tasks, whereas later areas do not. This is probably the dominant view in the visual neuroscience community[39]. In hypothesis H1b, the same tradeoff holds as in H1a, except that the human performance is matched in IT, rather than early layers. In hypothesis H2, explicitly decodable information peaks (for at least some non-categorical properties) in intermediate visual areas, analogous with the results for V2 border-ownership cells that have been found in the context of simple visual stimuli[17]. In hypothesis H3, information is neither lost nor gained for the orthogonal variable tasks up through the ventral stream, it is simply preserved. This view is suggested as one possibility in previous studies from our group[18] and is consistent with ideas of hyperacuity[26]. Finally, in hypothesis H4, information increases for the orthogonal tasks along with the categorization tasks. Aspects of this possibility are consistent with coarse coding[27,28].

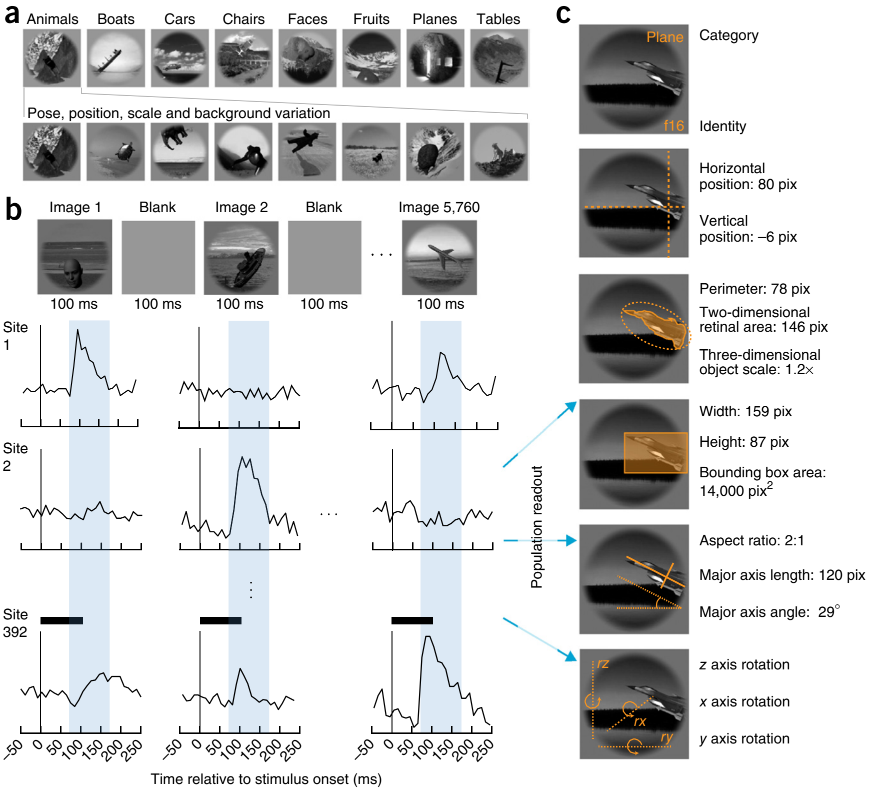

Figure 2 Large-scale electrophysiological measurement of neural responses in macaque IT and V4 cortex to visual object stimuli containing high levels of object viewpoint variation. (**a**) We recorded neural responses to 5,760 high-variation naturalistic images consisting of 64 exemplar objects in eight categories (animals, boats, cars, chairs, faces, fruits, planes and tables), placed on natural scene backgrounds, at a wide range of positions, sizes and poses. (**b**) Stimuli were presented to awake fixating animals for 100 ms in a rapid serial visual presentation (RSVP) procedure (horizontal black bars indicate stimulus-presentation period). Object centers varied within 8° of fixation center. Recordings were made using chronically implanted electrode arrays, collecting a total of 392 neuronal sites in IT (n = 266) and V4 (n = 126) visual cortex. Each stimulus was repeated between 25 and 50 times. Spike counts were binned in the time window 70–170 ms post stimulus presentation (as indicated by shaded regions) and averaged across repetitions, to produce a 5,760 × 392 neural response pattern array. (**c**) We then used linear readouts to decode a variety of types of image information from the neural responses, including categorical data such as object category and exemplar identity, as well as continuous data such as object position, retinal and three-dimensional object size, two- and three-dimensional pose angles, object perimeter, and aspect ratio.

and humans have been shown to have very similar patterns of behavioral responses on similar tasks[33].

The simultaneous variation of object viewpoint parameters in the image set also allows assessment of a battery of continuous-valued category-orthogonal object properties (**Fig. 2c**). These include object center position; object size, defined in terms of perimeter length, retinal area or three-dimensional scale; bounding-box location, area and aspect ratio; two-dimensional rotation properties including major axis length and angle; and three-dimensional pose, defined relative to category-specific canonical poses. For each property, we defined the task of estimating the value of that property, invariant to all other varied properties (including category). Using nine chronically implanted electrode arrays across three hemispheres in two macaque monkeys, we collected responses from 266 neural sites in area IT and 126 neural sites in area V4 to each image in the set[30] (**Fig. 2b** and Online Methods). We then investigated the ability of each of these neural populations, as well as a simulated V1 neural population, to support each task.

Comparing task representations across cortical areas

For many tasks, including object category, position, size and pose, we found that some individual sites in our IT sample had responses that contained reliable information for that task, despite simultaneous variation in all other variables (**Fig. 3a** and **Supplementary Fig. 2**). For categorical tasks, we defined single-site performance as the absolute value of the site's discriminability for the task on a set of held-out images (Online Methods). For estimation tasks, we defined single-site performance as the absolute value of the Pearson correlation of that site's response with the actual property value, again on a set of held-out test images. For most tasks, the best IT sites contained substantially more information than those from V4 (**Fig. 3a**).

Because information about visual properties is often distributed across multiple neural sites, we next investigated encoding at the neural population level by training linear decoders to extract the properties of interest (**Fig. 3b**). For discrete-valued categorization and subordinate identification tasks, we use linear SVM classifiers[8,9], while for estimation tasks we used L2-regularized linear regressors. Population performance levels were higher than from individual sites, as expected. The IT population (**Fig. 3b**) significantly outperformed the V4 population on all tasks, with a larger IT-V4 gap than for single sites (see **Supplementary Table 1** for statistical information). To compare these results to lower-level visual response properties, we also evaluated a Gabor-wavelet-based V1 model with local competitive normalization[32] on our stimulus set (Online Methods). In all cases, the IT sample population outperformed the V1-like model and, in most cases, the V4 population did as well. A trivial pixel control (black bars) performed least well in nearly all cases. Results were evaluated for each task using an equal number of sites from each population (n = 126). We performed additional controls to ensure that IT/V4 gap was not due to differences in recording quality, receptive field coverage, sampling sparsity, or number of decoder training examples (Online Methods and **Supplementary Figs. 2d**–**f**, **3** and **4**).

In addition to the high variation stimulus set used above, we also tested a simpler stimulus sets containing Gabor-like grating patches. In the simpler stimuli, we observed qualitatively different results from the case of the higher variation stimuli, with the V1-like population achieving higher performance than the V4 or IT populations on position and orientation estimation tasks (see below).

Neural consistency with human performance patterns

We next collected human performance data on a subset of the tasks, including categorization, position, size, pose and bounding-box estimation tasks (Online Methods). We sought to characterize, for each neural population and task, how many neural sites would be required to reach parity with human performance levels. For each neural population, we subsampled sites and trained linear decoders

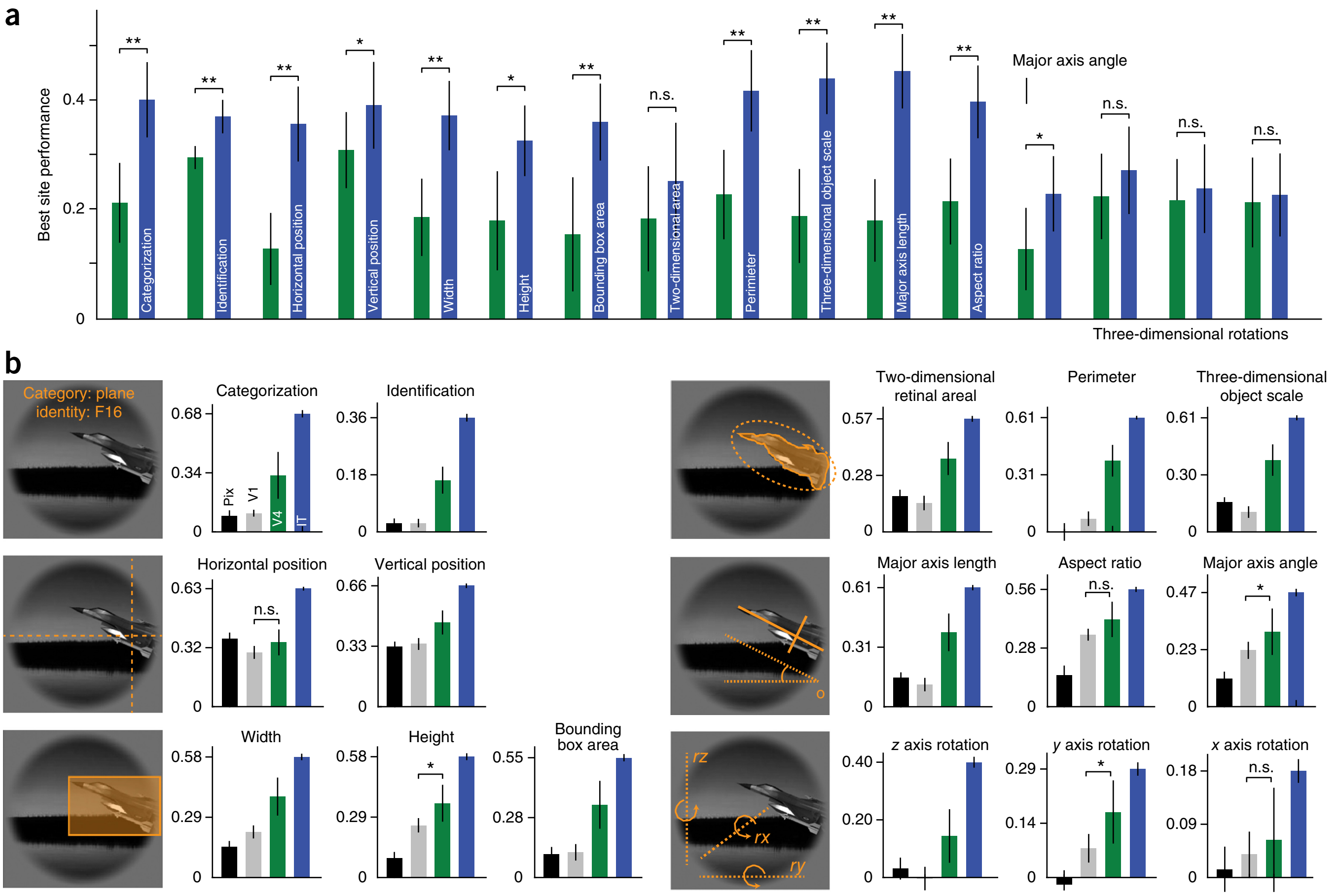

Figure 3 Comparison between ventral cortical areas of object property information encoding in high-variation stimuli. (**a**) Performance of single best sites from IT (blue bars) and V4 (green bars) on each task measured task. Best sites were chosen in a cross-validated manner, with performance being evaluated on held-out images. Chance performance is at 0. Error bars represent s.d. of the mean taken over subsets of images used to choose the best site for each task. n.s. indicates IT-V4 difference not significant, *$P < 0.05$, **$P < 0.005$. (**b**) Population decoding. For each task, we trained a linear decoder on neural output. For discrete-valued tasks, including object categorization and subordinate identification, we used support vector machine (SVM) classifiers with L2-regularization. For continuous-valued estimation tasks, we used linear regression with L2-regularization. We compared decoding performance for our recorded IT population sample (blue bars) and V4 population sample (green bars), as well as for a performance-optimized V1 Gabor wavelet model with competitive normalization (gray bars) and the trivial pixel control (black bars). For categorical properties, bar height represents balanced accuracy (0 = chance, 1 = perfect). For continuous properties, bar height represents the Pearson correlation between the predicted value and the actual ground-truth value. All values are shown on cross-validated testing images held out during classifier and regressor training. All evaluations are performed with $n = 126$ sites and a fixed number of training and testing examples. Error bars represent s.d. of the mean over cross-validation image splits and, in the case of pixel, V1 and IT data, over multiple subsamplings of 126 units from the whole population. In **b**, IT-V4 and V4-V1 separations were significant at $P < 0.005$ except where notated; n.s. indicates difference not significant, *$P < 0.05$. See **Supplementary Table 1** for statistical details.

for each sample and task using the same decoder methods described above. For each task, we produced decoding performance curves as a function of population sample size (**Fig. 4a**). For the IT and V4 neural populations, we produced curves out to the limit of the neural data, whereas we sampled increasing numbers of units up to 2,000 units for the V1 model and pixel control. We then fit each task's neural performance curve to a logarithmic functional form, to estimate performance levels at larger sample sizes. For all tasks, the estimated IT population performance curves reached human performance parity with fewer than 2,000 sites (**Table 1**), with a mean across tasks of 695 ± 142 sites (Online Methods). All tasks had similar performance-increase rates, suggesting that each additional IT site contributed a roughly similar performance benefit for each task. In contrast, V4 population performance curves were more variable over the tasks (compared with IT) and in most cases required several orders of magnitude more sites than the IT population to match human performance. The V1 model representation typically required several orders of magnitude more sites, in many cases unrealistically many more (greater than the total number of neurons in cortex). The pixel representation was not viable for any measured task.

Some of the tasks were more difficult than others for our human subjects. Pose estimation, for example, had lower raw accuracy than position estimation. Given this variability, we sought to determine whether human performance was predicted by neural population performance (**Fig. 4b**). To this end, we computed the Spearman rank correlation between the vector of human performances across tasks with equivalent vectors for each neural population (Online Methods). We found that the IT performance pattern predicted human performances substantially better than V4 or the V1-like model (**Fig. 4c**). Together with the performance parity estimation result, this suggests that IT more directly drives downstream behavior-generating neurons than lower cortical areas for all measured non-categorical and

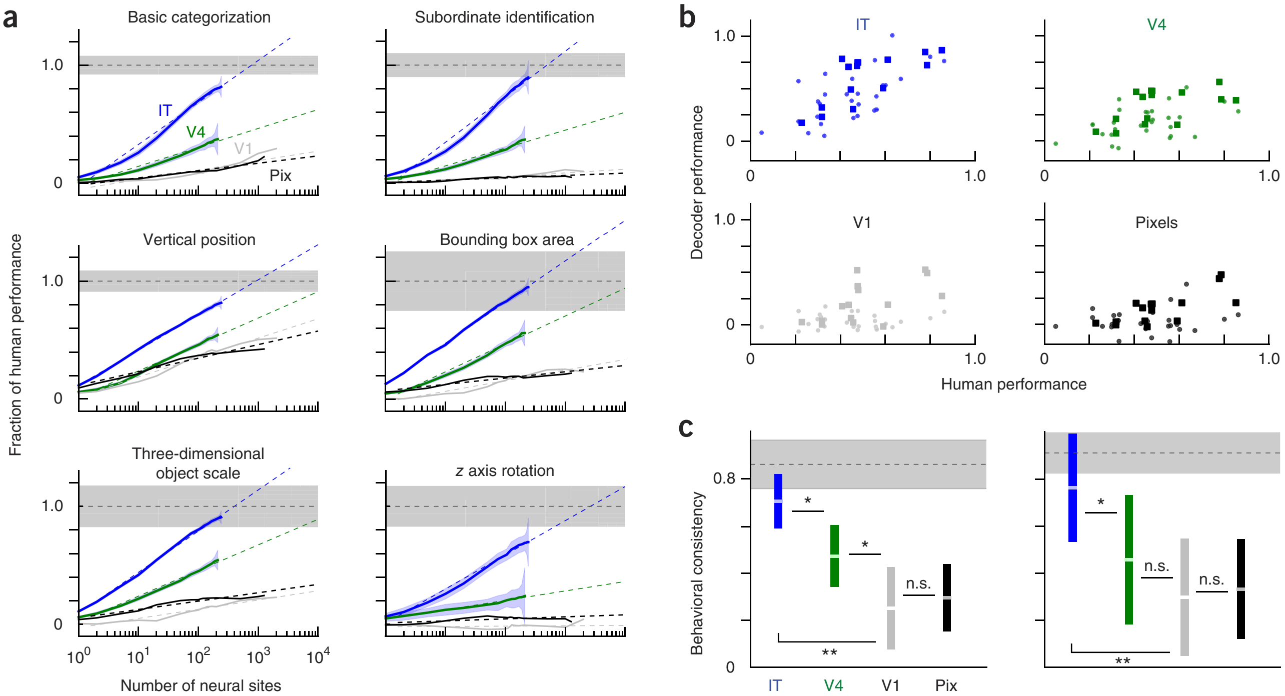

Figure 4 Comparison of neural population decoding performance to human psychophysical measurements. (**a**) Human-relative performance as a function of number of subsampled sites used to decode the property, for selected tasks. The *x* axis represents the number of sites. For each task, the *y* axis represents the performance of the decoder with the indicated number of sites, as a fraction of median human performance for that task, with a value of 1 indicating human performance parity. Decoder performance metrics and training procedures are as described in **Figure 3b**. Solid lines represent measured data and dotted lines represent log-linear extrapolations based on the measured data. Shaded areas around each solid line represent s.e. assessed by bootstrap resampling of sites and images. We evaluated our measured IT (blue lines) and V4 (green lines) neural populations out to the entire recorded populations of 266 and 126 sites, respectively, and evaluated V1 model (gray lines) and pixels (black lines) out to 2,000 units. Human performance for each indicated task was measured using large-scale web-based psychophysics (Online Methods). 1 s.d. in the human performance is indicated by gray shading flanking $y = 1$ (median performance level). (**b**) Scatters show human performance (*x* axis) versus neural performance (*y* axis) for a variety of tasks. Large squares ($n = 14$) correspond to the tasks indicated in **Table 1**. Small circles ($n = 41$) indicate values for further breakdown of the data into subordinate identification and pose estimation tasks on a per-category basis (Online Methods). (**c**) Summary of data from **b**. Bar height represents Spearman's *R* correlation between human and neural decode for the aggregated large-square tasks (left) and disaggregated small-circle tasks (right). Error bars are s.d. of the mean due to task and image variation (Online Methods). The dotted line represents the mean self-consistency of the measured human population, averaged across multiple subsets of the population sample. Horizontal gray bars represent the s.d. of the mean of human self-consistency, across population, task and image subsets. n.s. indicates difference not significant, $*P < 0.05$, $**P < 0.005$. See **Supplementary Table 1** for statistical details.

categorical tasks, and that linear decoders are a reasonable approximation of that downstream computation.

Distribution of information across IT sites

We next sought to characterize whether non-categorical properties are estimated by dedicated subpopulations of IT neurons or are instead integrated in a highly overlapping joint representation. Studies showing IT units tuned to multiple visual properties suggest that a joint representation is possible[18,20], but other studies suggesting the modularity of face, body and place-selective units[34] may point in a different direction.

To address these issues, we first considered the distribution of information across sites for each task (**Fig. 5a–c** and **Supplementary Fig. 5**). We used the weights assigned to each of the 266 IT sites by the linear decoder for that task as a proxy for the task relevance of that site, with positive weights indicating task-response correlation and negative weights indicating anticorrelation. For each task's site-weight distribution, we measured sparseness and imbalance. High sparseness would indicate only a very few sites being highly informative for the task, and low values indicate little cross-site differentiation. Imbalance measures the relative preponderance of sites correlated with the task, as compared with those that are anticorrelated.

Sparseness measurements revealed that fraction of highly weighted sites makes up between 15% and 35% of all sites, with a mean of 26.3%, and nearly half of the tasks had sparseness that was statistically indistinguishable at the $P = 0.5$ level from that of the standard normal distribution

Table 1 Estimated number of neural sites required to match median human performance

	IT	V4	V1	Pixels
Basic categorization	773 ± 185	2.2×10^6	–	–
Subordinate identification	496 ± 93	4.4×10^6	–	–
Horizontal position	1,414 ± 403	5.2×10^5	3.0×10^7	–
Vertical axis position	918 ± 309	2.5×10^4	8.7×10^6	–
Bounding box area	322 ± 90	1.7×10^4	–	–
Width	256 ± 87	9.8×10^3	3.4×10^7	–
Height	237 ± 87	3.8×10^3	9.5×10^6	–
Three-dimensional object scale	401 ± 90	3.2×10^4	–	–
Major axis length	201 ± 70	1.1×10^4	–	–
Aspect ratio	163 ± 61	951 ± 59	6.5×10^3	–
Major axis angle	774 ± 128	1.6×10^5	–	–
z axis rotation	1,932 ± 1,061	–	–	–
y axis rotation	396 ± 115	2.8×10^5	–	–
x axis rotation	1,570 ± 530	–	–	–

Error bounds are due to variation in site subsamples, and are extrapolated based on actual site subsample variation in the data (see Online Methods). Dash (–) indicates more than 10 billion sites are required.

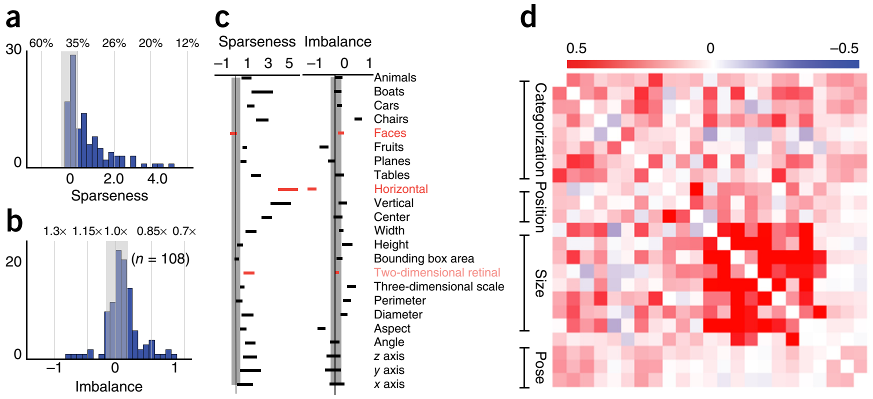

Figure 5 Distribution and overlap of IT cortex site contribution across tasks. (**a**) Histograms of values of sparseness over all tasks. Sparseness is measured via excess kurtosis (γ_2, Online Methods). Reference values show fractions of 'high-relevance' sites, as determined by three-point distribution method (Online Methods). Gray band represents 1 s.d. of distribution of sparseness values taken on size-matched samples from a Gaussian distribution. (**b**) Histograms of values of imbalance over all tasks. Imbalance is measured via skewness (γ_1, Online Methods). Reference values at the top of the imbalance panel show fractions of values above versus below means, ranging from 1.3 to 0.7. Gray band represents 1 s.d. of distribution of imbalance values taken on size-matched samples from a Gaussian distribution. (**c**) Sparseness (left) and imbalance (right) of weight distributions for selected tasks. Error bars represent s.d. over image splits on which weights were determined. Gray bands here are defined as in **a** and **b**. (**d**) Quantification of weight pattern overlap for pairs of tasks. Each colored square in the heat map is the Pearson correlation between the absolute value of the weight vectors for a pair of tasks. A high value (red color) indicates that the weight pattern for the pair of tasks is similar; a low value (blue color) indicates the opposite. White indicates a value that is not statistically significantly different from zero. The order of tasks is the same as in **c**.

of equal size (n = 266 sites). For the majority of tasks, imbalance measurements were also statistically indistinguishable at the $P = 0.5$ level from equivalently sized normal distributions. Overall, these results suggest that task information distribution is comparatively normal, with few properties having statistically especially selective units.

We then quantified information overlap between pairs of tasks. Overlap was defined as the correlation of the absolute values of the decoder weight vectors for each task pair (**Fig. 5d**). A high positive overlap for two tasks (**Fig. 5d**) indicates that downstream neurons could use common neurons for decoding the two tasks, whereas high negative correlation indicates the opposite. Across all pairs of tasks in our data set, 56.5% of pairs had positive overlap, 16.6% had negative overlap and 26.9% had overlap that was statistically indistinguishable from 0. High overlap tended to occur between groups of semantically related tasks (for example, size-related tasks). However, apparently unrelated tasks typically had more overlap than would be expected from a purely random distribution of units (Online Methods and **Supplementary Fig. 6**). An exception was the face-detection task, where the estimated overlap with other tasks was significantly less than random ($P < 0.01$). Taken together, these results provide

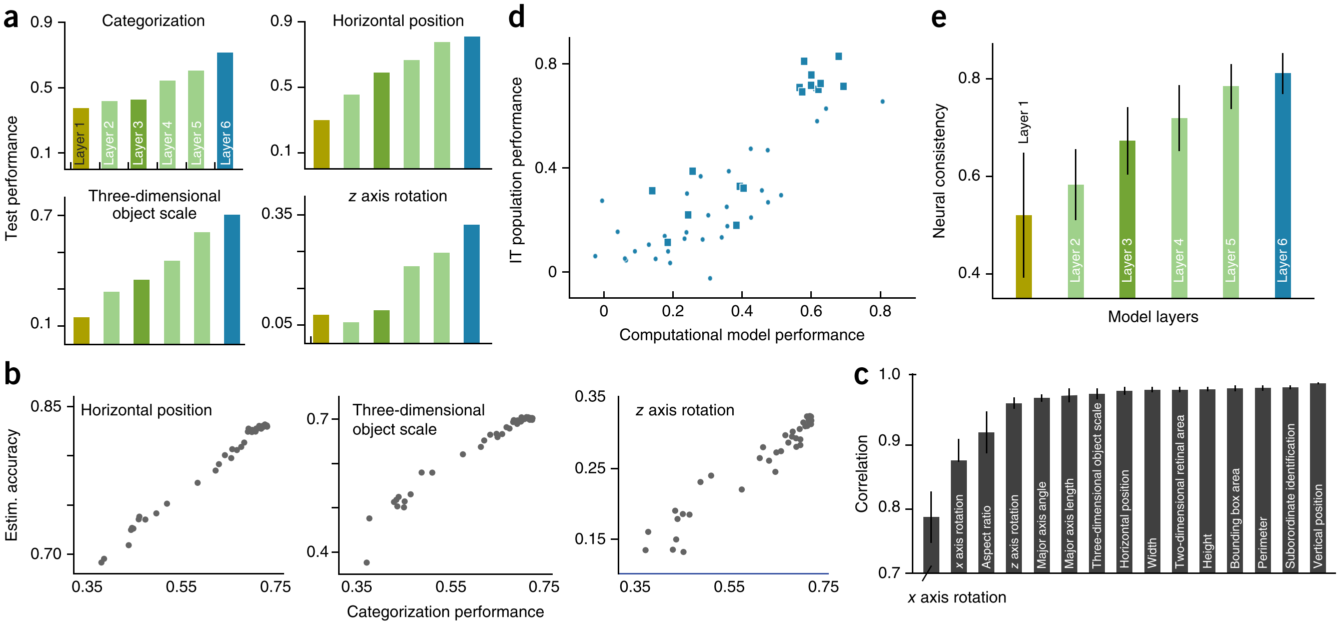

Figure 6 Computational modeling results. (**a**) Performance of fully-trained model at each hidden layer. *y* axes are as described in **Figure 3b** for corresponding tasks (**Supplementary Fig. 9**). (**b**) Scatter plots of performance of computational model's top hidden layer on training-set categorization performance versus testing-set estimation accuracy for selected non-categorical tasks. Each dot represents a state of the model during training (**Supplementary Fig. 10**). (**c**) Quantification of relationship in **b**, shown for all tested tasks aside from categorization itself (n = 15). Bar height represents Pearson correlation of accuracy on indicated task with test-set categorization performance, taken across training time steps. Error bars represent s.d. of the mean taken across both time steps as well as splits of images used for performance assessment. (**d**) Scatter plot of performance of top hidden layer of fully trained model versus performance of IT neural representation, on each task measured in **Table 1**. As in **Figure 4b**, large squares represent aggregated tasks (n = 16) and small circles represent disaggregated tasks (n = 43). Unlike **Figure 4b**, several tasks are included for which human data were not collected. (**e**) Consistency of fully trained model with neural performance pattern across layers, using the same metric described in **Figure 4c**. *y* axis and error bars are as described in **Figure 4c**. See **Supplementary Table 1** for statistical details.

additional evidence for the hypothesis that, with the possible exception of face-detection, the IT neural population jointly encodes both categorical and non-categorical visual tasks.

Computational modeling

Recent work has shown that neural responses in ventral cortex can be modeled effectively by hierarchical convolutional neural networks (HCNNs) that are optimized for performance on object categorization tasks[30,31,35]. Each HCNN layer is composed of simple, biologically plausible operations including template matching, pooling and competitive normalization (**Supplementary Fig. 7a** and ref. 30); filter is applied convolutionally, and identical filters are applied at all spatial locations. Layers are stacked hierarchically to produce complex transformations of the input images.

To determine whether HCNN models are consistent with our empirical results, we implemented one such model, containing six hidden hierarchical layers followed by one fully connected output layer. We optimized this model for category recognition performance on a subset of ImageNet, a database of natural photographs containing millions of images in thousands of every-day object categories[36] (**Supplementary Fig. 7d**). To ensure a sufficiently strong test of generalization was performed, we removed categories from the training set that overlapped with those appearing in the testing image set used in the neural and behavioral experiments discussed above.

Even though no neural data were used to learn model parameters, and the semantic content of the training was different from that of the testing images, the trained model was nonetheless highly predictive of neural responses in the test images on an image-by-image basis. Consistent with previous work[30], the model's top hidden layer was predictive of neural response patterns in IT cortex, intermediate layers were predictive of neural response patterns in V4 cortex and lower layers evidenced V1-like Gabor edge tuning (**Supplementary Fig. 7c**). These results validate the model for further investigation on non-categorical tasks.

For a series of time points during model training, we computed model activations from each layer on the test image set, which could be viewed as analogous to taking a time course of neural response measurements in a developing animal. We tested the performance of the top hidden layer of the model, with the same tasks and decoder procedures for the neural populations above. We found that performance on the categorical tasks in the test set increased throughout the course of training (**Supplementary Fig. 8**), indicating effective generalization from training to test categories.

We investigated task performance for each model layer. On the high-variation stimulus set, performance increased with each successive hidden layer, both for categorization and category-orthogonal tasks (**Fig. 6a** and **Supplementary Fig. 9**), in direct accord with our neural results (**Fig. 3b**). We found that, throughout training, performance of the model's top hidden layer improved on category-orthogonal estimation tasks (**Fig. 6b**,**c**, and **Supplementary Figs. 8** and **10**). This result may be nonintuitive, as the model's output layer, immediately downstream of the top hidden layer, was not only not explicitly supervised for estimating these category-orthogonal parameters, but in fact received supervised training to become invariant to these object parameters. Moreover, the performance pattern across tasks of the fully trained network's top hidden layer was highly consistent with the IT neural performance pattern, and consistency increased through model layers (**Fig. 6d**,**e**). Together, these results indicate that this computational model is a plausible description of the mechanism underlying our empirical results.

Dependence on amount of stimulus variation and complexity

We also recorded V4 and IT neural responses to simple grating-like patches at varying positions and orientations (**Supplementary Fig. 1b**). We then measured decoding performance for horizontal and vertical

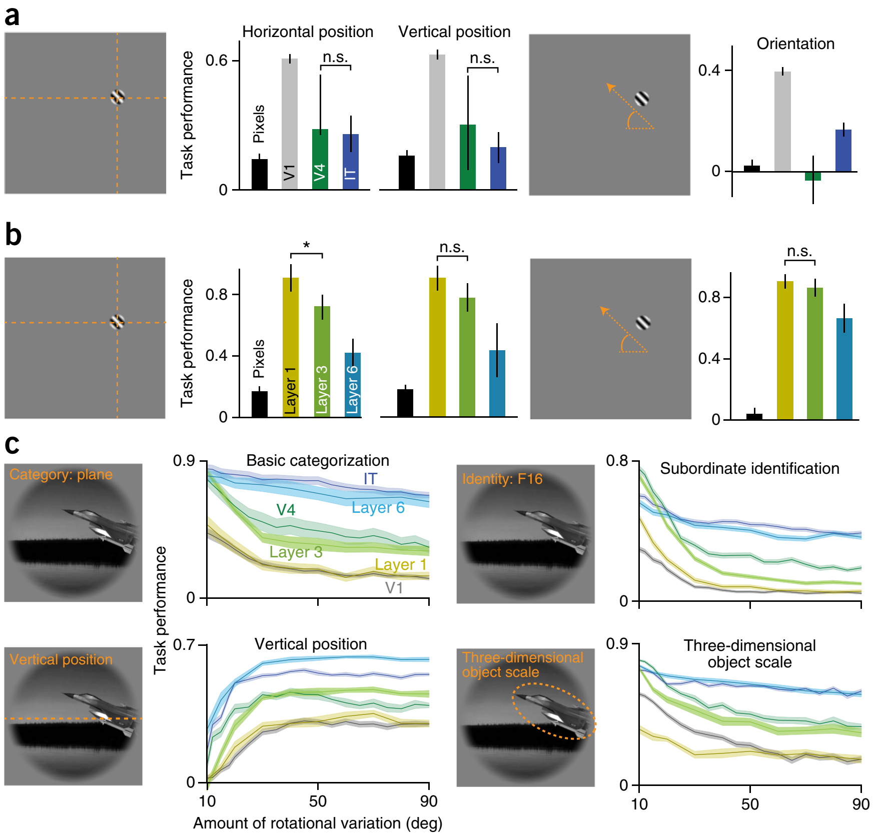

Figure 7 Dependence of linearly accessible information on the amount of variation in stimuli. (**a**) Population neural decoding results for position and orientation tasks defined on a simpler stimulus set consisting of grating patches placed on gray backgrounds. *y* axis, bar colors and error bars are as described in **Figure 3b** (**Supplementary Fig. 11**). (**b**) Performance of three selected neural network model layers (layer 1, yellow; layer 3, olive green; layer 6, cyan) for the tasks shown in **a**. (**c**) Population decoding performance as a function of amount of rotational variation in classifier training and testing data sets, for each of several representative object tasks, for measured IT neural population, V4 neural population, V1 model and for the three model layers. *x* axis represents (absolute value of) the amount of rotational variation allowed in all three rotational axes; for example, a value of 10 corresponds to rotation in *x*, *y* and *z* axes ranging from −10 to 10 degrees. *y* axis is performance evaluated using the same metrics and decoder training procedures as described in **Figure 3b**. Error bars are computed over selections of sites and units as well as image training splits (see **Supplementary Table 1** for statistical details).

position and orientation estimation tasks, again using linear classifiers (**Fig. 7a** and **Supplementary Fig. 11**). V4 and IT performance levels were significantly higher than chance, but, unlike the results for complex stimuli, the IT population was not better than the V4 population on position tasks for these simpler stimuli, and both IT and V4 populations were worse than the V1-like model (see **Supplementary Table 1** for statistical information). This clarifies our main result in relation to existing results in early visual areas[7], which contain neurons that outperform animal behavior on low-level tasks[37]: although the larger receptive fields in V4 and IT lose resolution for the pixel-level judgments needed in simplified stimuli, this information loss does not strongly interfere with decoding of similarly defined object properties in more complex image domains.

To further characterize the relationship between the cross-area information pattern and the amount of variation in stimuli, we performed analyses identical to those shown in **Figure 3b**, subsetting the image set at varying levels of rotational variation between 10° and 90° (**Fig. 7c** and **Supplementary Fig. 12**). Even at low rotational variation levels, the images contained substantial variation in object position and size, as well as background content. For each task, we found that, as the rotational variation decreased, the gap in performance between V4 and IT decreased, although the rates at which this gap closed varied between tasks. In some cases (for example, subordinate-level identification or three-dimensional object scale), the relative rank order of V4 and IT reversed at low rotational variation levels.

We evaluated the computational model using grating stimuli (**Fig. 7a**). We found that the lowest intermediate model layer (layer 1) had the highest level of performance, with a subsequence performance drop in higher layers (**Fig. 7b**), echoing the empirically observed pattern (although see the mismatch between model layer 3 and V4 data in the orientation task). We also investigated the dependence of computational model performance on the amount of rotational variation in the testing set. We found that, just as with the neural populations, the gap between the top hidden model layer and intermediate layers closed with lower amounts of rotational variation (**Fig. 7c**). The model also predicted performance characteristics on individual tasks, for example, the inversion of performance between higher and intermediate layers at low variation levels for the subordinate identification task. We also investigated the importance of high levels of variation for model correctness by training with a lower-variation image set in place of ImageNet. This alternatively trained model was much less effective at describing the observed empirical patterns of relative information (**Supplementary Figs. 13** and **14**).

DISCUSSION

We found that, for a battery of high-variation non-categorical visual tasks, there was more linearly decodable information in neural populations sampled from higher ventral stream areas than lower ones, the relative pattern of performance levels across all these tasks measured in human behavior was more consistent with that decoded from IT populations than from lower area populations, and task-related information was distributed broadly in the IT neural population, rather than factored into task-specific unit subpopulations. Unlike previous studies, we recorded population responses in two cortical areas (V4 and IT) for a large, high-variation image set, and were thus able to make empirical area comparisons for category-orthogonal tasks. Qualitatively different, but highly plausible, alternatives were consistent with the previously known data (**Fig. 1**). Our results show that only one of these scenarios (H4; **Fig. 1**) is correct for complex naturalistic stimuli.

Our results suggest that the same neural mechanisms that build tolerance to identity-preserving transforms also build explicit representation of those same transforms. Although this may sound like a contradiction, it can be interpreted in light of existing theoretical ideas about distributed, coarsely coded representations[2,18,23,25,27]. A key contribution of our experimental results is a systematic confirmation that, for complex naturalistic image domains, these theories are more consistent with the empirical data than alternatives[14,15,38].

Our study argues against mechanisms that aim to hierarchically reduce sensitivity to category-orthogonal properties with repeated 'simple/complex-cell' arrangements, trading off accuracy on orthogonal properties for increased receptive field size (H1a and H1b; **Fig. 1b**). As these mechanisms represent perhaps the dominant conception in the visual neuroscience community of how invariant object recognition is produced[39,40], in addition to the ideas implied in some of our own earlier work[18,39] (H3; **Fig. 1b**), our empirical results here are important. Previous findings suggest that the ventral stream representation strategically throws out certain stimulus information[25]. Our computational model depends crucially on the presence of pooling operations that throw out information, but our results (computational and empirical) suggest that the role of pooling is not likely to be the layer-wise discounting of object transformations. So what information is thrown out? It would be of interest to determine whether human performance patterns in simpler image domains (for example, **Fig. 7a**) are better explained by V1 than IT, especially as V1 neurons can sometimes outperform animal behavioral performance[37,41].

By exploring the dependence of the relative task performance between areas on the amount of object view variation, we found that, for some simpler and lower-variation image domains, which may sometimes by ecologically relevant, lower visual areas can have more easily accessible information than higher areas. Our data are thus not best understood as confirming that a specific type of coarse coding strategy is a complete description of the ventral stream. Our results suggest that amount of complexity in the stimulus set, rather than type of task (for example, position estimation), is a key determinant of cross-area information patterns. Future research should explore this dependence along multiple axes of variation (for example, position, size, background complexity, etc.).

These results highlight the importance of high-variation stimuli in comparing visual areas. A number of earlier studies demonstrated information for position and pose in IT[18,19,21,23,24], largely employing simpler stimuli. Had those experiments compared IT with V4 and V1, they might have found a decrease in information in higher areas (analogous to **Fig. 7a**). That the relative power of V4 and IT could be reversed for some tasks by reducing variation in one parameter (pose) while retaining substantial variation in others suggests that future studies of higher visual cortex should be careful to include sufficient variation. Future work should look to expand to more realistic image domains, with multiple foreground objects in natural visual scenes. Although it is one thing for linear decoders to report a suite of properties relevant to an object in a scene, understanding how the brain handles the full 'binding problem' posed by combinatorial property compositions is a key challenge that is beyond the scope of our current results[29].

Computationally, our main contribution is a model generated from simple principles that encompasses our main empirical findings. This model suggests why coarse encodings may have arisen to begin with: they are optimal for high-level performance goals, even when the properties they encode (for example, position) are apparently orthogonal to the optimization goal (categorization). Going beyond these ideas, however, we found that, across tasks and levels of variation, complex patterns of relative information are possible, in ways that are not predicted by any one encoding theory or 'word model'. Our results suggest that, rather than fully adopting a specific encoding strategy (coarse or local), a more general top-down goal optimization principle is at work in the ventral stream.

Feedback and/or attentional mechanisms[42,43] could account for how multiple orthogonal properties of objects can be integrated (reminiscent of the dorsoventral separation-of-roles hypothesis). However, given that our neural data were collected from passively fixating animals in a rapid serial visual presentation procedure with randomly interleaved images, reading out the earliest evoked IT responses (70–170 ms post-presentation), feedforward effects were likely dominant. Our computational model provides a neurally plausible 'existence proof' for how the experimental phenomena that we observed can be generated using largely feedforward circuitry.

Notably, our computational results indicate that learning robust category selectivity brings along performance on non-categorical tasks 'for free'. Future studies should investigate whether the converse is true, whether learning one or more non-categorical properties is enough to guarantee categorization performance or is categorization a stronger constraint driving IT neural responses. It would be interesting to identify a visual property of complex natural scenes that is supported by the IT population representation, but does not arise automatically with categorization optimization.

Our results may also be viewed as evidence that the ventral stream inverts a generative model of image space[44]. The test image set was produced by photorealistic rendering, with each image corresponding to a different choice of rendering parameters. Our results indicate that IT neural output encodes key inputs required to re-run the renderer. Such a representation could support on-line inference and long-term learning[44,45]. Although these interesting theoretical ideas have limited experimental support, our results show that some important elements are in place.

Our work is dependent on assumptions about how IT neurons are decoded by downstream units directly responsible for behavior. Linear estimators are technical tools for quantifying easily accessible task-relevant information in a population. However, because they consist only of linear weightings and at most a single threshold value, they also express a plausible rate-code model for downstream decoder neurons[1]. Future research should explore more sophisticated codes (for example, temporal decoding schemes) for the visual properties that we investigated, as well as potential columnar layout for these properties, such as those observed for shape selectivity[46].

An additional limitation is that comparisons to lower level visual areas use a V1-like model rather than actual neural recordings. However, this model is similar to state-of-the-art V1 models[7] and shows a clear and consistent pattern with data from our V4 and IT recordings. However, the model is an imperfect match to V1 (ref. 7), and it would be useful to repeat the analyses done here in V1 neural recordings. Along similar lines, although the CNN model described in **Figures 6** and **7** predicts many qualitative and quantitative features in our observed data, it is an imperfect match. Improved models will be critical in better understanding ventral stream information processing.

Another limitation of our data is that images were restricted to an 8° diameter window at the animal's center of gaze. This is large enough to allow substantial object position variability, with maximal displacement greater than the object's base diameter. However, it is not large enough to show effects in the visual periphery of the kind normally associated with parietal cortex[47,48], nor do we mean to suggest that such processing occurs exclusively in the ventral stream. Given our results and recent data showing shape and category selectivity in parietal areas[48–50], we speculate that both the dorsal and ventral stream contain representations for overlapping visual properties, categorical and otherwise, albeit at different levels of spatial resolution, with the ventral being fine scale and centrally biased and the dorsal being coarse scale with peripheral coverage. If borne out, this arrangement would naturally support behavior in which dorsal machinery directs foveation around an environmental saliency map, whereas the ventral machinery parses multiple object parameters extracted in each salient (para-)foveal snapshot, information that could then be integrated downstream across multiple foveations to produce overall scene understanding.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to K. Schmidt and C. Stawarz for key technical support, and to D. Ardila, R. Rajalingham, J. Tenenbaum, S. Gershman, C. Jennings and J. Fan for useful suggestions. The infrastructure needed for this work was supported by DARPA (Neovision2) and the NSF (IIS-0964269). The bulk of the work presented in this manuscript was supported by the US National Institutes of Health (NEI-R01 EY014970), with partial support from the Simons Center for the Global Brain and the Office of Naval Research (MURI). H.H. was supported by a fellowship from the Samsung Scholarship. We thank NVIDIA for a grant of GPU hardware and Amazon for an education grant supporting computational and psychophysical work. Additional computational infrastructure support was provided by the McGovern Institute for Brain Research (OpenMind).

AUTHOR CONTRIBUTIONS

H.H., N.J.M. and J.J.D. designed the neurophysiological experiments. H.H. and N.J.M. performed the neurophysiology experiments. D.L.K.Y., H.H. and J.J.D. designed the human psychophysical experiments. D.L.K.Y. performed the human psychophysical experiments. D.L.K.Y. and H.H. performed data analysis. D.L.K.Y. and H.H. performed computational modeling. D.L.K.Y., J.J.D., H.H. and N.J.M. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. DiCarlo, J.J., Zoccolan, D. & Rust, N.C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
2. DiCarlo, J.J. & Cox, D.D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
3. Felleman, D.J. & Van Essen, D.C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
4. Tanaka, K. Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* **19**, 109–139 (1996).
5. Logothetis, N.K. & Sheinberg, D.L. Visual object recognition. *Annu. Rev. Neurosci.* **19**, 577–621 (1996).
6. Vogels, R. & Orban, G.A. Activity of inferior temporal neurons during orientation discrimination with successively presented gratings. *J. Neurophysiol.* **71**, 1428–1451 (1994).
7. Carandini, M. *et al.* Do we know what the early visual system does? *J. Neurosci.* **25**, 10577–10597 (2005).
8. Majaj, N.J., Hong, H., Solomon, E.A. & DiCarlo, J.J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
9. Hung, C.P., Kreiman, G., Poggio, T. & DiCarlo, J.J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).
10. Rust, N.C. & Dicarlo, J.J. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
11. Movshon, J.A., Thompson, I.D. & Tolhurst, D.J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol. (Lond.)* **283**, 53–77 (1978).
12. Gochin, P.M. The representation of shape in the temporal lobe. *Behav. Brain Res.* **76**, 99–116 (1996).

13. Ito, M., Tamura, H., Fujita, I. & Tanaka, K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* **73**, 218–226 (1995).
14. Goodale, M.A. & Milner, A.D. Separate visual pathways for perception and action. *Trends Neurosci.* **15**, 20–25 (1992).
15. Ungerleider, L.G. & Haxby, J.V. 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* **4**, 157–165 (1994).
16. Bosking, W.H., Crowley, J.C. & Fitzpatrick, D. Spatial coding of position and orientation in primary visual cortex. *Nat. Neurosci.* **5**, 874–882 (2002).
17. Zhou, H., Friedman, H.S. & von der Heydt, R. Coding of border ownership in monkey visual cortex. *J. Neurosci.* **20**, 6594–6611 (2000).
18. Li, N., Cox, D.D., Zoccolan, D. & DiCarlo, J.J. What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J. Neurophysiol.* **102**, 360–376 (2009).
19. DiCarlo, J.J. & Maunsell, J.H. Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *J. Neurophysiol.* **89**, 3264–3278 (2003).
20. Logothetis, N.K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).
21. MacEvoy, S.P. & Yang, Z. Joint neuronal tuning for object form and position in the human lateral occipital complex. *Neuroimage* **63**, 1901–1908 (2012).
22. Nishio, A., Shimokawa, T., Goda, N. & Komatsu, H. Perceptual gloss parameters are encoded by population responses in the monkey inferior temporal cortex. *J. Neurosci.* **34**, 11143–11151 (2014).
23. Sayres, R. & Grill-Spector, K. Relating retinotopic and object-selective responses in human lateral occipital cortex. *J. Neurophysiol.* **100**, 249–267 (2008).
24. Sereno, A.B., Sereno, M.E. & Lehky, S.R. Recovering stimulus locations using populations of eye-position modulated neurons in dorsal and ventral visual streams of non-human primates. *Front. Integr. Neurosci.* **8**, 28 (2014).
25. Edelman, S. & Intrator, N. Towards structural systematicity in distributed, statically bound visual representations. *Cogn. Sci.* **27**, 73–109 (2003).
26. Snippe, H.P. & Koenderink, J.J. Discrimination thresholds for channel-coded systems. *Biol. Cybern.* **66**, 543–551 (1992).
27. Hinton, G., McClelland, J. & Rumelhart, D. Distributed representations. in *Parallel Distributed Processing, Vol 1* (eds. Rumelhart, D. & McClelland, J.) 77–109 (MIT Press, 1986).
28. Eurich, C.W. & Schwegler, H. Coarse coding: calculation of the resolution achieved by a population of large receptive field neurons. *Biol. Cybern.* **76**, 357–363 (1997).
29. Treisman, A. The binding problem. *Curr. Opin. Neurobiol.* **6**, 171–178 (1996).
30. Yamins, D.L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111**, 8619–8624 (2014).
31. Khaligh-Razavi, S.M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
32. Pinto, N., Cox, D.D. & DiCarlo, J.J. Why is real-world visual object recognition hard? *PLoS Comput. Biol.* **4**, e27 (2008).
33. Rajalingham, R., Schmidt, K. & DiCarlo, J.J. Comparison of object recognition behavior in human and monkey. *J. Neurosci.* **35**, 12127–12136 (2015).
34. Tsao, D.Y. & Livingstone, M.S. Mechanisms of face perception. *Annu. Rev. Neurosci.* **31**, 411–437 (2008).
35. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. in *The Handbook of Brain Theory and Neural Networks* (ed. Arbib, M.A.) 255–258 (MIT Press, 1995).
36. Deng, J. *et al.* ImageNet: a large-scale hierarchical image database. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 248–255 (2009).
37. Chen, Y., Geisler, W.S. & Seidemann, E. Optimal decoding of correlated neural population responses in the primate visual cortex. *Nat. Neurosci.* **9**, 1412–1420 (2006).
38. Mishkin, M., Ungerleider, L.G. & Macko, K.A. Object vision and spatial vision: two cortical pathways. *Trends Neurosci.* **6**, 414–417 (1983).
39. Zoccolan, D., Kouh, M., Poggio, T. & DiCarlo, J.J. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* **27**, 12292–12307 (2007).
40. Serre, T. *et al.* A quantitative theory of immediate visual recognition. *Prog. Brain Res.* **165**, 33–56 (2007).
41. Nienborg, H. & Cumming, B.G. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *J. Neurosci.* **34**, 3579–3585 (2014).
42. Chikkerur, S., Serre, T., Tan, C. & Poggio, T. What and where: a Bayesian inference theory of attention. *Vision Res.* **50**, 2233–2247 (2010).
43. Milner, P.M. A model for visual shape recognition. *Psychol. Rev.* **81**, 521–535 (1974).
44. Yildirim, I., Kulkarni, T.D., Freiwald, W.A. & Tenenbaum, J.B. Efficient analysis-by-synthesis in vision: a computational framework, behavioral tests, and modeling neuronal representations. *Proc. Annu. Conf. Cogn. Sci. Soc.* **471** (2015).
45. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
46. Tanaka, K. Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex* **13**, 90–99 (2003).
47. Brown, L.E., Halpert, B.A. & Goodale, M.A. Peripheral vision for perception and action. *Exp. Brain Res.* **165**, 97–106 (2005).
48. Sereno, A.B. & Lehky, S.R. Population coding of visual space: comparison of spatial representations in dorsal and ventral pathways. *Front. Comput. Neurosci.* **4**, 159 (2011).
49. Rishel, C.A., Huang, G. & Freedman, D.J. Independent category and spatial encoding in parietal cortex. *Neuron* **77**, 969–979 (2013).
50. Swaminathan, S.K. & Freedman, D.J. Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nat. Neurosci.* **15**, 315–320 (2012).

ONLINE METHODS

High variation stimulus set and visual task battery. Our main neural test stimulus set, which will be denoted as **Images**, consisted of 5,760 images of 64 distinct objects chosen from one of eight categories (animals, boats, cars, chairs, faces, fruits, planes, tables), with eight specific exemplars in each category (for example, BMW, Z3, Ford, etc. within the car category). The set was designed (see ref. 30) to: (i) span a range of everyday objects, (ii) support both coarse, "basic-level" category comparisons (for example, "animals" versus "cars") and finer subordinate level distinctions (for example, distinguish among specific cars)[51], and (iii) contain substantial variation in object view parameters (position, size, pose) that makes it challenging to decode any of the visual properties of objects (category, identity, position, size, pose). Objects were placed on realistic background images which were chosen randomly to prevent correlation between background content and object class identity.

As in ref. 30, the object view parameters for stimuli in **Images** were chosen randomly from uniform distributions in three levels of variation.

Low variation. All objects placed at image center (horizontal = 0, vertical =0), with a constant scale factor ($s = 1$) translating to objects occluding 40% of image on longest axis, and held at a fixed reference pose ($rx = ry = rz = 0$).

Medium variation. Object position varies within one-half multiple of total object size (|horizontal|, |vertical| ≤ 0.3), varying in scale between $s = 1 / 1.3 \sim .77$ and $s = 1.3$, and between −45 and 45 degrees of in-plane and out-of-plane rotation (≤45°).

High variation. Object position varies within one whole multiple of object size (|horizontal|, |vertical| ≤ 0.6), varying in scale between $s = 1 / 1.6 \sim 0.625$ and $s = 1.6$, and between −90 and 90 degrees of in-plane and out-of-plane rotation (≤90°).

Using this stimulus set, we defined a battery of visual tasks.

Basic-level object categorization. This is a discrete-valued eight-way object categorization task, in which the goal is to report the category of the object in the image, from the set of choices: Animals, Boat, Car, Chair, Face, Fruit, Plane, Table.

Subordinate-level object identification. These are discrete-valued eight-way object identification task, in which the goal is to report the specific identify of an object in each image from the list of eight exemplars of that object's category. There are eight such tasks, one for each category in the data set. For example, in the case of the car category, the eight-way subordinate-level object identification task is identify an image as containing one of: Beetle, Alfa Romeo, Vauxhall Astra, BMW 325, Maserati Bora, Toyota Celica, Renault Clio, or BMW z3.

Position estimation. These are a set of related continuous-valued location estimation task, in which the goal is to identify an object's center location. Tasks are to identify the location in pixels of the object center, along the horizontal axis ("Horizontal Position") and the vertical axis ("Vertical Position"), and the distance in linear pixels of the object center to any fixed point location ("Center Distance").

Bounding-box size estimation. These are a set of related continuous-valued bounding-box related tasks. The bounding box for an object is defined to be the smallest axis-aligned rectangular subset of the image that fully contains the pixels of the object. Location of each corner is measured, as is the size in linear pixels along both axes ("Width" and "Height", respectively). The area of the bounding box in square pixels is also measured ("Bounding Box Area").

Two-dimensional retinal area. This continuous-valued task measures the area in square pixels that the object takes up in the image. Each image pixel is either covered by the object, in which case the pixel is counted toward this metric, or it is not covered by the object, in which case the pixel is not counted. For example, pixels surrounded by an object but not actually covered by it (for example, the hole of a donut) do not count toward this measure.

Perimeter. This continuous-valued task measures the area in linear pixels on the boundary of the object. Pixels in the object not completely surrounded by other pixels also in the object do count toward this measure; any other pixels do not count.

Three-dimensional object scale. This continuous-valued task measures the three-dimensional scale parameter used to generate the image in the original rendering process, relative to a fixed canonical size — namely, $s = 1$ in the object parameterization discussed above. This relationship of this property to the two-dimensional retinal area depends in a complex manner on the object's geometry.

Major axis length, aspect ratio and angle. The major axis of an object is defined to be the longest line segment such that both ends of the line segment are pixels within the object. The minor axis is the shortest perpendicular line segment so that the rotated bounding box defined by the major and minor axes covers the object. The continuous-valued measure axis length is measured in linear pixels. The aspect ratio is the ratio of the lengths of minor to the major axis. The major-axis angle is the two-dimensional angle, in degrees, made by the major line with the horizontal line.

Three-dimensional rotation. These three rotations are the angles, in degrees, used by the renderer to orient the object in the original image creation process with a right-handed coordinate system, where $+y$, $+z$, $+x$ directions correspond to "right", "up", and "out of the screen" directions (**Fig. 2c**). The angles are described via standard Euler rotations using the *XYZ* order. The (0, 0, 0) rotation is defined separately for each of the 64 exemplar objects in the data set. However, the exemplar angles are fairly well-defined "semantically", meaning that they are reasonably consistent across the eight exemplars each for the eight basic object categories. Specifically, for each category the (0, 0, 0) angle is the one in which:

- Animals: animal is facing forward, with its head upright.
- Boats: boat is oriented with bow facing forward and keel point downward.
- Cars: car grille is facing forward, while tires on the bottom.
- Chairs: chair legs are facing downward, with the seat facing forward.
- Faces: looking straight the viewer, with top of the head oriented upward.
- Fruits: stem attachment at the top. Note that many of the fruits possess a rough rotational symmetry around the vertical axis.
- Planes: cockpit facing forward, with plane in upright position.
- Tables: table legs facing straight downward, with longest side along the horizontal axis.

We used the following metrics for measuring performance on these tasks, across all modalities (for example, neural data, human data, and computational model outputs). Specifically, for the discrete-valued categorization tasks, performance is measured using *balanced accuracy*. Balanced accuracy is defined for a prediction of binary task with positive and negative classes as

$$\mathbf{AccBal} = \frac{TP}{P} + \frac{TN}{N} - 1$$

where *TP* is the number of correct positive predictions, *P* is the number of positives examples in the data, *TN* is the number of correct negative predictions, and *N* is the number of negative examples in the data. Balanced accuracy for a multi-class prediction problem is the average of one-versus-all (OVA) prediction problems over the classes. For continuous-valued estimation tasks, performance is measured as the Pearson product-moment correlation between the predicted and actual values. Specifically:

$$\mathbf{Corr} = \frac{covariance(\vec{p}, \vec{a})}{\sqrt{variance(\vec{p}) \cdot variance(\vec{a})}}$$

where $\vec{p}$ is the vector of predictions for a sequence of images and $\vec{a}$ is vector of corresponding ground-truth values for that property.

We chose these metrics because they both range from −1 to 1, with 0 being chance-level prediction and 1 being perfect prediction. Slight negative values of these metrics will sometimes arise in practice because classifiers and regressors are cross-validated on held-out testing images.

We repeated our core population decoding analysis on V4 and IT neural data on a spectrum of subsets of the full high-variation image set **Images**. Specifically, we chose to subset the image set by amount of variation in the rotation parameters. We considered variation cutoff levels ϕ ranging between 10° and 90°, at 5 degree intervals. For each such ϕ, we created a subset of the original data set defined by restricting rotations of objects to at most ϕ degrees on all three axes, i.e.:

$$\mathbf{Images}_{\phi} = \{x \in \mathbf{Images}\ if\, |rot_{x,y,z}(x)| \leq \phi\}.$$

We then performed the full battery of classifier and regressor training and testing on $Images_{\phi}$ for each value of ϕ (**Fig. 7c** and **Supplementary Fig. 12**). In this work we do not make comparisons between the absolute values of the lines in 6 at different ends of the variation spectrum, and instead focus only on the *relative* values of IT, V4, and the V1-like model because there is covariation between the amount of rotational variation and variation amounts for other parameters.

Simple stimuli. We also gathered neural data on a simpler set of stimuli (**Supplementary Fig. 1b**), consisting of small grating patches placed on gray backgrounds. We will denote this set of images **Gratings**. The grating objects were shown at different positions in a 5-by-5 location grid. At each location, gratings were shown at each of 4 orientations, including 0°, 45°, 90°, and 135°, for a total of 100 images). The overall intensity of the images are all identical.

Array electrophysiology. Neural data were collected in the visual cortex of two awake behaving rhesus macaques (*Macaca mulatta*, 7 and 9 kg, both male) using parallel multi-electrode array electrophysiology recording systems (BlackRock Microsystems, Cerebus System). All procedures were done in accordance with NIH guidelines and approved by the MIT Committee on Animal Care guidelines. Nine 96-electrode arrays (three arrays in each hemisphere, with a total of three hemispheres, two left, one right, across two monkeys) were surgically implanted in anatomically-determined V4, posterior IT, central IT and anterior IT regions[3]. Of these, 392 neural sites (266 in IT and 126 in V4) were selected as being visually driven with a separate imageset. Passively fixating animals were presented with testing images in pseudo-random order with image duration comparable to those in natural primate fixations[52]. Images were presented one at a time on an LCD screen (Samsung SyncMaster 2233RZ at 120 Hz) for 100 ms, occupying a central 8° visual angle radius on top of a gray background, followed by a 100 ms gray "blank" period with no image shown. Eye movements were monitored by video tracking (SR Research, EyeLink II), and animals were given a juice reward each time fixation was maintained for 6 successive image presentations. Presentations in which eye movement jitter exceeded ±2° from screen center were discarded. In each experimental block, responses were recorded once for each image, resulting in 25–50 repeat recordings of the each testing image.

For each image repetition and electrode, scalar firing rates were obtained from spike trains by averaging spike counts in the period 70–170 ms post-stimulus presentation, a measure of neural response that has recently been shown to match behavioral performance characteristics very closely[8]. Background firing rate, defined as the mean within-block spike count for blank images, was subtracted from the raw response. Additionally, the signal was normalized such that its per-block variance is 1. Final neuron output responses were obtained for each image and site by averaging over image repetitions. Recordings took place daily over a period of several weeks, during which time neuronal selectivity patterns at each recording site were typically stable. Based on firing rates and spike-sorting analysis, we estimate that each individual electrode multi-unit site in this study picks up potentials from 1–3 single neural units.

Sorting of single units. To determine whether results would likely differ for direct single-unit recordings, we sorted single units from the multi-unit IT data by using affinity propagation[53] together with the method described in (ref. 54). Based on these analyses, we estimate that each of our multi-unit sites contains spikes from between 1 and 5 single units. In our IT sample, we obtained 154 well-isolated single units; in our V4 sample, we obtained 191 well-isolated single units. Throughout, we repeated analyses both for our raw multi-unit site data, as well as for these isolated single-unit populations. We did not see significant differences from the multi-unit analyses in the relative performance levels between V4 and IT. Absolute performances from the single units was generally lower, since the sorted single units were less reliable on average than our multi-units, but measured on a per-spike basis were generally equal to or slightly higher than for the multi-units. Moreover, we have supplemented with serially sampled, single-electrode recording[9,10], and have found that neuronal populations from arrays have very similar patterns of image encoding as assembled single-electrode unit populations.

Receptive field analysis. Using the simple grating-like stimuli, we were able to compare receptive field locations and sizes in our V4 and IT populations. We found that for both populations, receptive fields were concentrated near the center of gaze. In the case of V4 population, these fields covered the approximately central 4° relative to the center of case; in our IT population, the fields covered roughly central 8°. To investigate the effect of receptive field coverage on our results, we performed versions of each of our analyses restricting to images in the central 4 degrees of the field of view, but did not see substantial differences.

Neural performance assessment. We assessed the performance of neural sites and populations on each of the tasks in our task battery. For discrete-valued tasks, performance was assessed by training SVM classifiers (using a linear kernel) on neural output[55]. Linear SVM classifiers are a standard tool for analyzing the performance capacity of a featural representation of stimulus data on discrete classification problems[8,9,55]. For neuronal sites, the output features are defined as the vector of scalar firing rates for each unit, as is typical in neural decoding studies[8,9,56]. For any fixed population of output features (from either a model or neural population), a linear classifier determines a linear weighting of the units, followed by a discrete threshold, which best predicts classification labels on a sample set of training images. Category or identity predictions are then made for stimuli held out from the weight training set, and accuracy is assessed on these held-out images. For continuous-valued estimation tasks, performance was assessed by training support vector regression regressors with linear kernels on the output features[55]. A linear regressor determines a linear weighting of the units that best predicts the target property on a set of training images. Predictions for that property are then made for a set of held-out images, and accuracy is assessed using the Pearson correlation measure discussed above.

For both discrete classifiers and continuous regressors, to reduce the noise in estimating accuracy values, results are averaged over a number of independent cross-validation splittings of the data into training and testing portions. In the data shown in **Figures 3** and **4**, results show cross-validated test performance averaged over 50 splits in which each training split contained a randomly selected 80% of the data, and the corresponding testing split contained the remaining 20% of the data. While absolute values of performances depend on the size of training split, the results discussed in this paper do not. In all cases, classifiers and regressors were trained using an l_2 regularization penalty on the weights, and the penalty weight C was chosen separately for each split with cross-validation by sub-splitting the training data[55].

For each of the 8-way classification tasks, including the basic-categorization task and the eight subordinate identification tasks, classifiers were trained using an 8-way one-versus-all (OVA) methodology[55]. For most tasks, training and testing was done across images of all eight basic categories taken together. However, for the pose estimation tasks, the training was done within each of the eight basic category of images separately, in analogy with the human psychophysics experiments. Similarly, for eight within-category subordinate identification tasks, training and testing were performed on images from each corresponding category.

In addition to the population analyses reported in **Figure 3b**, we performed separate analyses for posterior IT cortex (PIT, n = 184 sites) and central IT cortex (CIT, n = 125 sites), though we did not have sufficiently many anterior (AIT) cortex units to perform a separate analysis there. Though we found several individual tasks with statistically-significant differences between PIT and CIT, taken as a whole with the appropriate multiple-comparison (Bonferroni) correction applied, we cannot conclude from our limited data any statistically significant differences between PIT and CIT, either for absolute performances levels (as in **Fig. 3**) or cross-task behavioral consistency (as in **Fig. 4**), similar to our previous observation on classification tasks in ref. 8.

Control models. Throughout, we use two basic models as controls against to which to compare neural population recordings.

- A V1-like model[32], that we use to provide an approximate comparison point for lower levels in the ventral visual stream. This model is based on a grid of 96 Gabor wavelets composed of filters at 16 spatial frequencies and 6 evenly-spaced spatial orientations, proceeded and followed by a local competitive normalization operation. This model is similar to those used to provide state-of-the-art predictions of neural responses in V1 (ref. 7).
- The trivial Pixel control, in which 256 × 256 square images were flattened into a 65,536-dimensional "feature" representation. The pixel features provided a control against the most basic types of low-level image confounds.

Image level controls. There is a potential that our V4 records might have been generally less reliable or of lower quality than our IT recordings, since sites in V4, with their smaller receptive fields, might be more sensitive to various factors such as (for example,) animal eye movements. To ensure that our results comparing V4 to IT were not influenced by generally lower recording quality in V4, we estimated a number of image-level controls (**Supplementary Fig. 2d–f**):

1. We measured per-site cross-trial reliability of each site (**Supplementary Fig. 2d**). To measure reliability, for each site we compared that site's responses across images on one trial (producing a vector of responses) to the same site's responses on a different trial (producing a second vector of responses). We quantified reliability as the Pearson correlation between these two vectors, averaged across all pairs of trials. (For site and each image, we had between 25 and 50 trials.) We did not observe a statistically significant difference between reliability in our V4 sites (median = 0.73 ± 0.05) as compared to our IT sites (media = 0.76 ± 0.06).
2. We measured selectivity for each site (**Supplementary Fig. 2e**). For each site, we measured selectivity as the d-prime for separating that site's best (most highly response-driving) stimulus from its worst (least highly response-driving) stimulus. D-prime 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:

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

   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 20 trial splits. We did not observe a statistically significant difference between selectivities in our V4 population (median = 1.88 ± 0.14) and IT populations (median = 1.80 ± 0.21).
3. We measured the population level separability for image pairs (**Supplementary Fig. 2f**). For each of 5000 randomly selected image pairs in our main test stimulus set (**Supplementary Fig. 1a**), we trained a classifier on our neural population to separate the first image in the pair from the second. The classifiers were trained and tested in a cross-validated way over trials (for example, classifier weights determined from one set of trials and then evaluated on another set of trials, with results averaged over 50 train/test trial splits). We measured performance as the d-prime of the 2×2 test result confusion matrices, i.e.:

$$d' = Z(TP) - Z(FP)$$

   where Z is the normal z-score function, TP is the true positive rate, and FP is the false positive rate. **Supplementary Figure 2f** show histograms of the cross-validated d-prime values, over the 5000 pairs of images, with classifiers trained on the V4 population (left panel) and the IT population (right panel). We observed a barely statistically significant difference at the $p = 0.05$ level between the median d-prime value for the V4 population (median = 3.84 ± .08) and the IT population (median = 3.66 ± .10).

These measures show that, using image-level comparison metrics, the data from our V4 site recordings were not significantly less reliable, selective, or able to separate image pairs than that from our IT population.

Extrapolation analyses. We produced performance curves by subsampling our neural populations to various sizes between 1 site and all available sites. Performance scaling appeared in all cases approximately log-linear, for example,

$$Performance(n) \sim k \cdot log(n)$$

where n is the number of neural sites, and k is a constant. We then extrapolated performance to larger n values, fitting k fit to the observed data points using a least-squares error metric[57]. For the V4 and IT neural population data, we fit to all available data (out to 126 and 266 sites, respectively), while for the V1 and pixel controls we produced random samples of the features out to size 2000. In all cases used averages of performance over 100 samples, except when fewer than 100 unique samples of a given size were available.

Counting neural sites. In understanding the counts of the number of neural sites used in various analyses in this work, it is important to recognize the use of repetition-averaged multi-unit site responses. This underestimates the number of single neurons needed to support each task in real time. To translate to single-trial single-neuron counts, it is necessary to multiply the counts reported here by factors correcting for the effects of noise reduction over multiple trials as well as the number of single sites in a multiunit. This has been done carefully in ref. 8, yielding a factor of approximately 120. Thus our median number of 695 repetition-average multi-unit sites translates to approximately 83,000 IT neurons.

Performance in simple stimuli. We estimated population decoding performance for three tasks defined on the simple grating stimuli (as shown in **Supplementary Fig. 1b**), including Horizontal Position estimation, Vertical Position estimation, and orientation estimation. We used two types of classifiers to perform these analyses, including linear SVM classifiers, using the same protocol as with the analysis in **Figure 3b**; as well as nonlinear Radial Basis Function (RBF) SVM classifiers[55], using Gaussian kernels. We found that patterns of performances were similar for both linear and nonlinear classifiers (**Supplementary Fig. 11**).

Human psychophysical experiments. Data on human object recognition judgment abilities shown in **Figure 4** were obtained using Amazon's Mechanical Turk crowdsourcing platform, an online task marketplace where subjects can complete short work assignments for a small payment. All data were collected under approval by the MIT Committee on the Use of Humans as Experimental Subjects.

We measured human performance for a subset of the tasks on which we decoded neural performance (see below for detailed list). We recruited MTurk subject pools separately for each task with a subject count of $n = 80$, though there ended up being a small amount of overlap between the subject pools for the various tasks (there were fewer than 5 overlapping subjects for any pair of tasks). For each participant and each task, task sessions consisted of a training phase containing 10 trials (except as indicated below) and a testing phase containing 100 trials. On each trial, a sample image was shown, followed by a 500ms pause, and then a response screen was shown. The nature of the response screen depended on the task type (see below for details). For each of the sessions, we measured 20 of the testing images 2 times, to enable calculation of within-subject reliability.

During the training trials, sample images were shown for an extended period of time and in which correct answers were indicated both via annotation on the original sample image and in the response screen. During the 100 testing trials, sample images were shown for 100ms, followed by a 500ms pause, and then a response screen was shown. The accuracy values reported in the figures and text were generated from the testing trials only. A small bonus was paid to subject based on their average estimation accuracy at the end of the session, and subjects were told at the beginning of each session that their bonus would depend on correctness.

The tasks we measured included:

- Basic categorization task. This was an eight-way alternate forced choice (8-AFC) task. The response screen for this task consistent of 8 response images, one for each of the eight basic categories in our image set. Subjects were required to click with their mouse on the image representing the category they thought they saw in the sample image. Average within-subject reliability for this task was 0.97.
- Subordinate identification tasks. This consisted of eight separate 8-AFC tasks, one for each category. These tasks were not intermixed, for example, sessions involving subordinate car identification were not intermixed with subordinate boat identification. The response screen for each the eight category tasks consisted of 8 response images, one for each specific object identity within that category. Average within-subject reliability for this task was 0.92. For analyses in this paper that treat subordinate categorization as a single task, that is, performance values were averaged across each of the 8 individual tasks to produce a composite value.
- Position estimation. Response screens consisted of a blank canvas the same size as the sample image, and subjects were required to click at the location where they estimated the centroid of the object in the sample image was located.

Horizontal position and vertical position estimates were computed from the indicated centroid. Average within-subject reliability for the horizontal position estimate was 0.91 and for the vertical position estimate was 0.94.
- Axis-aligned bounding box estimation. Response screens consisted of a blank canvas the same size as the sample image, and subjects were required to click on the locations where they thought the top-left and bottom-right of the axis aligned bounding box had been for the object in the sample image. Width, height and bounding-box area were computed from the indicated bounding box. Average within-subject reliability of width was 0.96, for height was 0.92, and for bounding-box area was 0.84.
- Rotated bounding box estimation. Response screens consisted of a blank canvas the same size as the sample image. Subjects were first required to click on two points indicating one side of the rotated bounding box, and then on a third point indicating the extent of the rotated bounding-box in the orthogonal direction. Major axis length, major axis angle, and aspect ratio where computed from the subject's rotated bounding box estimate. Average within-subject reliability for major axis length was 0.85, for major axis angle was 0.79, and for aspect ratio was 0.91.
- Object three-dimensional scale. Response screens consisted of a new image of the object in the sample image, but shown from a single fixed canonical angle (chosen on a per-category basis as described above). On each testing phase trial, the size of the response image was randomized by uniformly drawing from the full size range in the data set. Subjects were given a slider and were required to resize the image so that the object was at the same three-dimensional size as they perceived it to be in the sample image. Once subjects felt they had correctly resized the object they pressed a "submit" button. Average within-subject reliability for object scale estimate was 0.87.
- Object three-dimensional rotation. Response screens consisted of a three-dimensional graphical "pointer" indicating defined "top" and "front" orientations. Subjects were required to rotate the pointer into alignment with the top and front orientations that they perceived in the sample image. Once subjects felt they had correctly posed the pointer, the clicked a "submit" button. Training was provided on a per-category basis to teach subjects our definition of the canonical (0, 0, 0) angle for each category, and 32 training examples were provided (containing training images for 4 exemplars each for each of 8 categories). Average within-subject reliability for *z*-axis rotation was 0.76; for *x*-axis rotation was 0.69; and for *y*-axis rotation was 0.71.

We did not measure two-dimensional retinal area and perimeter estimation tasks in our human subjects. Across all tasks, while subject consistency was comparatively high, there were a range of levels of performance. Some tasks (for example, position estimation) were reliably easier than other tasks (for example, three-dimensional pose estimation), though all tasks were both significantly above chance and significantly below ceiling. Relatively low performance on three-dimensional pose estimation tasks is likely to be explained by the fact that objects rotated on all the axes simultaneously (while changing in size and position and background as well).

We sought to determine whether the relative difficulty of tasks for humans across our range of tasks corresponded to the relative difficulty predicted by the neural populations (**Fig. 4b**). We first constructed a vector of performances:

$$\vec{v}_{human} = (p_{human,t1}, p_{human,t2}, \ldots, p_{human,tn})$$

where $p_{human,ti}$ was the mean performance of the human subject pool on task *i*. We next constructed a vector of performances:

$$\vec{v}_{neural} = (p_{neural,t1}, p_{neural,t2}, \ldots, p_{neural,tn})$$

where $p_{neural,ti}$ was the mean performance of the trained decoder on a given neural population on task *i*. We then measured the *consistency* of the neural and human population as the Spearman rank correlation between these two vectors:

$$consistency(neural, human) = Spearman(\vec{v}_{human}, \vec{v}_{neural})$$

We estimated the human-to-human consistency in performance pattern by bootstrapping methods with $n = 1000$ bootstrap replicas, using 68% confidence intervals[58] to determine the uncertainty in this value (as shown in **Fig. 4c**, left and right panels, horizontal gray bands). The bootstrapping was done over variation caused by subsampling in the set of tasks as well as the set of images used to compute performance for each task.

Partitioning tasks for various analyses. Through this work, we have attempted to keep the set of task used in each analysis as close to identical as possible. However, there are a number of exceptions to this that we note here:

1. In the analyses in **Figure 3**, we show 16 separate tasks. These are all the tasks for which we measured neural data, and on which we computed performance for computational models. As described in a previous subsection, these tasks include two categorical (that is, classification) tasks and 14 non-categorical tasks. These tasks are basic-level categorization (for example, "Animals versus Boats versus Cars etc."), as well as subordinate identification for each of the 8 categories.
2. The analyses comparing neural data to human data in **Figure 4b** and **Supplementary Figure 4** show only 14 of the tasks. This is because, as noted above, we did not collect human data for two of the non-categorical tasks (two-dimensional retinal area and perimeter length).
3. In performing the analyses comparing neural data to human data in **Figure 4b,c** tasks were split up in two ways:

    - The first method, corresponding to the large squares in **Figure 4b** and the left panel in **Figure 4c**, is to treat each of the 14 tasks as single individual data points.
    - However, as discussed above, four of these tasks actually themselves consisted of averages of 8 per-category individual tasks, including subordinate identification, and the three three-dimensional pose tasks (that is, *x*-axis, *y*-axis, and *z*-axis rotation estimation). We wanted to be sure that our results on neural/human consistency did not depend on the fact of making these aggregations over category. Thus, we also produced versions of these analysis in which each of the 8 per-category data points were considered separately, for these four tasks. This corresponded to the small circles in **Figure 4b** and the right panel in **Figure 4c**. This analysis comprised 41 separate task points (9 tasks that were not treated on a per-category basis, plus 8 tasks for each of 4 tasks there were treated on a per-category basis).

4. The analyses in **Figure 6d** are analogous to **Figure 4b**, but are different in that this figure compares model performances to neural performances. Because we had measured all 16 tasks for the neural data and the models (as opposed to the subset of 14 measured for the humans), we were able to include all these tasks in this scatter plot – making for a total of 16 large squares and 43 small circles.
5. As discussed below in the section on weight pattern analysis, in the analyses in **Figure 5**, we computed metrics about the weight distributions of classifiers and regressors. To make each such distribution comparable, we were required to compare the individual one-dimensional weight patterns from each task. However, as discussed above, the classifiers for each of the 8-way tasks, including basic categorization and the eight subordinate classification tasks, were actually comprised of 8 separate one-versus-all binary classifiers that were combined using the maximum-margin methodology. For this reason, in the analysis in **Figure 5**, there were 107 separate linear decoders, including:

    - 8 (for the basic-categorization task,
    - plus 64, including 8 for each of the 8 subordinate identification tasks,
    - plus 11 for each of the non-categorical regression tasks whose training was not done a separate per-category basis (for example, everything except the three-dimensional rotations),
    - plus 24, including 8 for each of the three three-dimensional rotation tasks.

6. In the analyses in **Figure 6d**, 15 tasks are present (as opposed to 16) because we are showing the correlation between categorization performance (on the training data set) and the remaining tasks.

Weight pattern analysis. Having determined that the IT population is able to sustain behaviorally plausible linear coding for a variety of tasks, our next goal was to understand the distribution of information for each of the tasks amongst the various sites. To formalize the concept of "relevance of a task at a given site", we used the classifier/regressor weights trained in the population analyses described above (see below for a discussion of alternative metrics). In mathematical terms:

$$\textit{relevance of site i for task } T = w_{Ti}$$

where $\vec{w}_T = (w_{T1}, w_{T2}, \ldots, w_{Tn})$ is the vector of weights of a l_2-regularized linear estimator for task T on site i, and n is the number of neural sites. In the case of the continuous regression tasks, the weights are simply the regression coefficients, whereas in the case of the discrete categorization tasks, the weights are classifier coefficients, before the final threshold value. The absolute value of the classifier weight, $|w_{Ti}|$, is a proxy for the amount of information contributed by site i for task T. If $|w_{Ti}|$ is large compared to the weights w_{Tj} for other sites j, site i is taken to be more relevant for the task; $w_T i >> 0$ corresponds to strong correlation between the site's output at the task, while $w_{Ti} << 0$ corresponds to strong anticorrelation.

Let D_T be the distribution of weights for task T (**Supplementary Fig. 5a**). In this work, we assume that the weights in $\vec{w}_T$ are IID samples from D_T. We consider the distributions for 107 separate binary tasks, including:

- The 8 one-versus-all basic-level categorization tasks (for example, Animals versus all, Boats versus all, etc.).
- 8 one-versus-all subordinate categorization tasks for each of 8 categories, for a total of 64 binary tasks.
- 11 size, position, bounding box, and pose estimation tasks, as described above.
- 24 subordinate three-dimensional pose estimation tasks, eight each for the three pose axes, as described above.

In **Figure 5b**,**d**, we only show results for the non-subordinate tasks, for visual clarity. However, **Figure 5c** shows distributions of the γ_1 and γ_2 statistics (see below) for all 107 decoders.

We had two basic analysis goals with these distributions: (a) what do the individual task distributions of information look like for each task? and (b) how do they overlap between tasks?

Individual Task Information Distribution. In mathematical terms, our first goal was to characterize the shape of D_T for each task T. To do this, we used two statistical properties of the distributions: skewness and kurtosis.

The γ_1 skewness of the weight vector is a measure the imbalance or asymmetry of the distribution of the weights about the mean weight. Positive skewness means that the positive tail of the weight distribution is longer than the negative tail, for example, the majority of the weight distribution is below the mean. In the context of this work, high skewness for the weight distribution associated with a given task would indicate that the population was biased toward having sites that are anticorrelated with the task, while high negative skewness would indicate the opposite. Formally, skewness is a statistical third-moment measure defined as:

$$\gamma_1(\vec{w}_T) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{w_{Ti}-\mu}{\sigma}\right)^3$$

where

$$\mu = \frac{1}{n}\sum_{i=1}^{n} w_{Ti}$$

is the average weight and

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(w_{Ti}-\mu)^2}$$

is the s.d. of the weights.

We measured the sparseness of weight distributions via excess kurtosis, γ_2. Excess kurtosis measures how spread out the weights are, relative to a normal distribution. Positive excess kurtosis means that the distribution is more peaked than a Gaussian distribution with the same mean and s.d. High kurtosis values indicating that only a very few sites are highly informative for the task, and low values indicating little differentiation between sites. Formally, excess kurtosis is defined as

$$\gamma_2(\vec{w}_T) = \frac{\frac{1}{n}\sum_{i=1}^{n}(w_{Ti}-\mu)^4}{\sigma^4} - 3$$

To ensure that we accurately took into account the effects of noise and sparse sampling of image space, the skewness and sparseness shown are computed by averaging the skewness and sparseness computed separately for the weights of 50 classifiers/regressors, each trained on a different split containing 50% of the training data. We also resampled sites with replacement, to ensure we were properly accounting for uncertainty due to site sampling. Error bars shown in **Figure 5b** are s.d. computed over both site samples and image splits.

To help interpret the meaning of these skewness and sparseness values, we compared them to two types of controls:

1. Gaussian control. With a statistically large enough sample, Gaussian distributions have 0 skew and 0 excess kurtosis. However, finite samples of a Gaussian distribution will not have 0 skewness or kurtosis. We matched the size of the empirical distribution of IT sites (n = 266) and drew 1000 samples of size from a standard Gaussian, and computed the skewness and kurtosis for each sample. The gray bars in **Figure 5b** show the s.d. spread of these values.
2. Three-point distribution control. The other end of the statistical spectrum from the Gaussian control are three-point distributions, distributions that have support on three distinct points, $x_- < x_0 < x_+$. For each task T, we approximated the empirical distribution D_T with a three-point distribution by solving for x_-, x_0, x_+ as well as probabilities $0 < p_-, p_0, p_+ = 1 - p_- - p_0$, such that x_0 is the empirical mean of D_T and the three-point distribution had the same mean, s.d., skew and kurtosis as D_T. Conceptually, the interpretation of these approximations are to divide the population of sites for each task T into three subpopulations: the x_--sites that are the "highly-anticorrelated" with the task T, the x_+-sites that are highly correlated with the task, and the x_0-sites that are not highly relevant to the task. The reference values shown in the skewness histogram (**Fig. 5b**) are, by definition,

$$\textit{reference skewness} = \frac{p_+ + 0.5 \cdot p_0}{p_- + 0.5 \cdot p_0}$$

measuring the ratio of above-mean to below-mean sites. The reference numbers shown in the sparsity histogram (**Fig. 5a**) are, by definition

$$\textit{proportion of high-relevance sites} = p_- + p_+$$

As shown in **Figure 5b**,**c** and discussed in the text, we found that the distributions of weights are:

- On average, comparatively symmetric, in which most tasks are statistically indistinguishable in their skewness from size-matched Gaussian control, and the proportion of above-mean to below-mean sites range from 0.7 to 1.3.
- On average, slightly more sparse than normally distributed, in which the proportion of high-relevance sites (as defined above) ranges from 15% to 35% of the total, with a median of 26.3%. The normal distribution has 32.5% high-relevance sites, and a significant proportion of tasks are not statistically distinguishable in their sparsity from that of the size-matched Gaussian control.

Taken together, these results suggest a picture of information distribution across sites that is comparatively well distributed, as opposed to each task being supported by a small number highly-dedicated sites.

Task-pair information overlap. Having characterized the per-task distributions, we sought to characterize the *overlap* of weights for each task pair, seeking to understand how the sites that are likely to be useful for any one task are related to

doi:10.1038/nn.4247

those that are relevant for each other task. We defined the overlap between tasks i and j as the Pearson correlation between the absolute values of the weight vectors for the two tasks (see below for discussion of alternative metrics). Formally, the overlap matrix (see **Fig. 5d**) has i, j -th element defined as

$$M_{i,j} = corr(|\vec{w}_{T_i}|, |\vec{w}_{T_j}|) = \frac{cov(|\vec{w}_{T_i}|, |\vec{w}_{T_h}|)}{\sqrt{var(|\vec{w}_{T_i}|) \cdot var(|\vec{w}_{T_h}|)}}$$

where T_i and T_j are the i-th and j-th tasks, respectively. This value ranges between 1 (perfectly correlated, meaning complete overlap) and –1 (perfectly anticorrelated, meaning totally non-overlapping). In practice, given that the number of tasks is comparable to the number of sites in our sample, and that (as seen in the previous section), each task utilizes between 15% and 35% of all sites, the minimum possible average overlap will be significantly larger than –1.

In **Figure 5d**, we show the average of the correlations of 1,000 random draws of weights of classifiers/regressors over a set of 50 splits containing 50% of the training data. That is, each matrix element is the average of 1000 correlations $corr(w_{T_i}^{sk}, w_{T_j}^{sl})$ where $w_{T_i}^{sk}$ is the weight vector for the i-th task, trained on the k-th (of 50) splits, and where s_k an s_l where chosen randomly for each of the 1000 repeats.

We were particularly interested in quantifying the overlap between category-detection tasks and non-categorical tasks. To provide reference points against which to compare our results, we considered two controls:

1. Random overlap model. Weights are randomly assigned to each task subject to the constraint of matching per-task and per-site marginal weight distributions, but in which task pair overlap is unconstrained.
2. Minimum overlap model. Weight assignments are constrained as in the random overlap model but additional constrained to result in as little overlap as possible.
3. In both cases, we used gradient-based optimization methods to solve for weights $\eta_{Ti}, 0 \leq i < n$ for each task T, such that
   - $\sum_{T=0}^{N} \eta_{Ti}^2 = \sum_{T=0}^{N} w_{Ti}^2$ for each unit i, where N is the number of tasks.
   - $mean(\vec{\eta}_T) = mean(w\eta_T)$ for all tasks T
   - $variance(\vec{\eta}_T) = variance(w\eta_T)$ for all tasks T
   - $skewness(\vec{\eta}_T) = skewness(w\eta_T)$ for all tasks T
   - $kurtosis(\vec{\eta}_T) = kurtosis(w\eta_T)$ for all tasks T.

Using the l-bfgs algorithm[55], we minimized the square difference objective function summed over the above 5 terms. In the case of the minimum overlap model also simultaneously minimized $\sum_{T_i<T_j} corr(|\vec{\eta}_{T_i}|, |\vec{\eta}_{T_j}|)$. For both the random and minimum overlap models, we ran the optimization over 1,000 random initializations of the η values.

In summary, and shown in **Supplementary Figure 6** we found that:

- Overlap is generally positive.
- The average overlap of (non-face) categorical tasks with each other is higher than would be predicted by the random overlap model, except for the case of faces.
- The average overlap of the face-detection task with other categorical tasks is lower than would be predicted by random overlap, but higher than would be predicted by the minimal overlap model.
- The average overlap of (non-face) categorical tasks with non-categorical tasks is lower but not statistically different from the prediction of the random model.
- The average overlap of face detection with non-categorical tasks is not statistically distinguishable from that predicted by the minimal overlap model.

Taken together, these results suggest that, holding faces aside, the IT neural population jointly encodes both categorical and non-categorical visual tasks using an integrated representation in which many units participate in tasks. However, our this observations are consistent with well-established observations of segregated face-specific sites[34,59], and provides a positive control that the overlap-measurement methodology used here is able resolve module-like structure when it exists.

Normalized decoder weights. In **Supplementary Figure 5a**, we show "normalized" decoder weights, meaning that the weights of the decoders have been divided by the total sum across sites of the absolute values of the weights. We've done this so that the visual comparison between the weights between decoders for several different tasks can be made on the same scale.

Statistical methods. In several of the figures of this paper, we use statistical tests. These include:

- In **Figures 3**, **4**, and **6**, as well as **Supplementary Figures 3**, **4**, **6b**, **7b**,**c**, **8**, **9**, and **13**, we use confidence intervals based on bootstrapping to estimate error bars. In each figure caption, it is indicated which source(s) of variance were included in computing these bootstraps. At first, we performed 1000 replicas for each bootstrap, but upon observing the highly normal distribution of the data, we reduced the number of replicas subsequently to 100 (or in some cases 500), to aid with computational efficiency.
- In **Figures 3**, **4c**, **7**, **Supplementary Figures 2d**–**f** and **11**, we used two-way population t-tests to determine statistical significance of the differences between IT-V4 and V4-V1 populations. In all cases we use Welch's version of the test since the variation in each of the populations were typically not equal. In **Figure 7c**, we specifically tested two hypotheses:

  1. The IT and V4 population performance gap is smaller at low amounts of rotational variation (left-hand ends of axes) that high amounts of variation (right-hand end). We also made similar comparisons for the model Layer 3 and Layer 6 performances. This hypothesis was shown to be statistically significant, both for neuronal data and models, at 0.005 levels for 13 out of 16 tasks and at 0.05 for the remaining 3.
  2. There are tasks for which the V4 population performance at low variation levels is greater than the IT population performance level. We found several tasks for which this hypothesis is true (for example, subordinate identification) at a confidence level of $p < 0.001$. However, because testing this hypothesis involves multiple comparisons — one for each task — we must use a correction to achieve a meaningful statistical result. Using $m = 16$ in the Bonferroni correction, we find that the hypothesis that there is at least one such task is significant at the $P = 0.05$ level.

- In **Figure 4c**, we used a 1-way ANOVA to determine that the human consistencies of the IT, V4, V1-like and pixel were different, finding that the populations were different at a p-value of less than 10^{-5}, with F-value of 164.52. We then used standard t-tests to determine the statistical significance of the differences between each population in the human average (note that the value with respect to which these differences were computed was not 0, but rather the dotted lines in the relevant panels of **Fig. 4c**).

No statistical methods were used to pre-determine the size of our neural sample, but our sample sizes are similar to those reported in previous publications[8–10]. Whenever parametric studies were deployed, we assumed that distributions were normal, but this was not formally tested. As this study did not test the effect of a treatment condition, no blinding techniques were deemed applicable, and no such techniques were employed. Information on randomization in selection of units is discussed above in section entitled "Array Electrophysiology".

Computational modeling. Computational modeling was done using convolutional neural networks, as in previous work[30]. HCNNs are multi-layer neural networks[35]. HCNNs approximate the general retinotopic organization of the ventral stream via spatial convolution, with computations in any one region of the visual field identical to those elsewhere. Each convolutional layer is composed of simple and neurally plausible basic operations, including linear filtering, thresholding, pooling and normalization. These layers are stacked hierarchically to construct deep neural networks.

Basic definitions. Formally, an image-like array is a three-dimensional dimensional floating-point array whose shape is (*s*, *s*, *nc*), where *s* is the *image size* and *nc* is the number of channels in the image. Let's begin by defining three basic operations on image-like arrays:

- **Filter**: this is a convolutional filterbank operation[35], which applies the same filter block equally to every point in an image-like array. It's parameters include:
- The number of filters *nf*. This is a positive integer.
- The size of the filter kernel *fs*, in pixels. This is an odd integer.
- The stride of the convolution, s_f. This is a positive integer.
- The specific filter values, denoted *F*, a floating-point matrix of shape (*nc*, *fs*, *fs*, *nf*), where *nc* is the number of channels in the input.
- A bias vector *b*, of length *nf*.

For any image-like array *X* of shape (s, s, *nc*) the output of $\mathbf{Filter}_F$ on *X* is the image-like array *Y* of shape $(s / s_f, s / s_f, nf)$ where

$$Y(i,j,k) = b[k] + \frac{1}{fs^2} \sum F[:,:,:,k] \otimes N_{fs}(X, s_f \cdot i, s_f \cdot j)$$

where $\otimes$ is pointwise array multiplication, $i, j \in [1,\ldots,s / s_f]$, $k \in [1,\ldots,nf]$, and $N_{fs}(X,i,j)$ denotes the square neighborhood of diameter *fs* at location *i*, *j* in *X*. The convolution is done with "same" mode, meaning that at the edges the image is padding with 0s to produce an output of the same shape as the input

- **Thres** is a rectified linear clipping operation. Its parameters are:
- The value of the upper clipping threshold, t^{max}, which can be any floating value.
- The value of the lower clipping threshold, t^{min}, which can be any floating value less than $t_i^{\max}$.

By definition,

$$\mathbf{Thres}(X) = max(min(X, t^{max}), t^{min})$$

- **Pool** is a local pooling operation that aggregates values of the input, within each channel. Its parameters are as follows:
- The size of the pooling kernel, *ps*. This is an odd integer.
- The pooling order *po*. This is 1, an even integer, or ∞.
- The pooling stride s_p. This is a positive integer.

By definition, for any image-like array *X* of shape (*s*, *s*, *nc*), the output of **Pool** on *X* is the image-like array *Y* of shape $(s / s_p, s / s_p, nc)$ where

$$Y(i,j,k) = \left(\frac{1}{ps^2} \left(\sum N_{ps}(X^{po}, s_p \cdot i, s_p \cdot j)[:,:,k] \right) \right)^{1/po}$$

where $i, j \in [1,\ldots,s / s_f]$, $k \in [1,\ldots,nc]$, and $N_{ps}(X,i,j)$ is the square neighborhood of diameter *ps* in *X* at location *i*, *j*. Notice that when $po = 1$, this is simple local averaging, and when $po = \infty$, this is max-pooling.

A *convolutional layer* is a composition of these three basic operations; that is, a function of the form

$$F_{(\theta_P, \theta_T, \theta_F)} = \mathbf{Pool}_{\theta_P} \circ \mathbf{Thres}_{\theta_T} \circ \mathbf{Filter}_{\theta_F}$$

where (θ_p, θ_T, θ_F) are choice of parameters for the three basic operations. A *hierarchical convolutional neural network* (HCNN) is a composition of convolutional layers, for example,

$$\mathcal{F} = F_L \circ F_{L-1} \circ \ldots \circ F_1$$

The only two restriction that are required for composition to make sense are: (1) that the number of channels in layer *i* is equal to the number of filters in layer $i - 1$, that is, $nc_i = nf_{i-1}$ and (2) that the spatial size s_i of the image-like arrays is 1 or greater at every stage. If the spatial size becomes 1, then only thresholding or filtering operations with filter size 1 can be applied from then onwards. When this occurs, we say that the network is "fully connected" at that layer (and from then on).

In our case, the input image-like arrays are RGB images, so that the number of channels in in the first layer is 3, one for each color channel. (When applied to grayscale images we simply copy the grayscale values into the three channels).

Network selection. We divide the parameters that specify the layers of an HCNN into two classes, selected in two phases:

Screening. In which all the parameters *except* the filterblock and bias values where chosen. These parameters, which we refer to as the "architectural parameters", include the number of network layers, and at each layer, the number of tilers, the sizes of the filter and pooling kernels, and the pooling order.

Training. In which, once the non-filter parameters are fixed, the filter-values and bias vectors for each layer are determined via error backpropagation.

Details of error backpropagation. For any given setting of architectural parameters, we used a standard neural network backpropagation algorithm[60] to set filter filters for the parameters. The training set that we used was the 2013 ImageNet Challenge set[36], which contains approximately 1.3 million images in 1000 natural categories. We filtered out any categories that were animals, boats, cars, chairs, fruits, planes or tables from this set (some of these categories do not appear anywhere in the ImageNet challenge set to begin with), retaining 799 categories containing a total of approximately 1 million images. Actual training was performed on a subset of approximately 950,000 images, while the remaining images were used as a validation subset to monitor performance during training. **Supplementary Figure 7d** shows the percent-incorrect error rate during training, both for the actual trained subset and for the held-out validation set. Performance on the training subset was computed once every 256 images, and averaged on a running bases of 50 256-image batches (black line in panel d); performance on the validation subset was computed at the end of each 50-batch set, with no running average taken (gray line in panel d). Because the validation performance was computed at the end of a set of batches over which the training performance was averaged, the validation error rate is typically slightly lower (better performance) than its corresponding training time point as plotted in **Supplementary Fig. 7d**). Because we did not observe significant overfitting (which would have been indicated by the training curve in **Supplementary Fig. 7d** rising significantly above the validation curve), we stopped the training process when performance on the training subset appeared to stop decreasing. The significant error-rate drop at approximately 3 × 10^7 images seen in **Supplementary Figure 7d** is due to a lowering of the learning rates at each model layer by a factor of 10.

Details of screening. We used high-throughput screening techniques[30,61] to select the architectural parameters. In this process, we randomly selected 50 draws of the number of layers and within-layer architecture parameters from a parameter space (see below), ran error backpropagation on the network with those parameters for 5 epochs of ImageNet, and then recorded the final training error. We then used Tree Parzen Estimation in the Hyperopt parameter optimization framework[61] to further select 150 additional architectural parameters, and again, ran backpropagation on these networks. After having run 200 networks, we selected the best such network and subjected it to further error backpropagation for 40 epochs. This optimal model had 6 layers. At every epoch of ImageNet training, we saved out checkpoints containing the filter and bias parameters.

The parameter space that we tested was defined by the following bounds:

- Number of layers ranged in [4, 5, 6].
- Filter sizes ranged in [3, 5, 7, 9].
- Pooling kernel sizes ranged in [3, 5, 7, 9].
- Pooling order ranged in [1, 2, 3, 4, 5, ∞].
- Upper clipping thresholds ranged in [1, ∞] and lower clipping thresholds ranged in [1, –∞].

The remainder of the parameters were set to the following fixed values: number of filters at layer 1 was 96, at layer 2 was 256, at layer 3 was 512, and then at 256 for subsequent layers; strides at layer 1 was 1, at layer 2 was 2, at layer 3 was 2, and at 1 in subsequent layers.

Evaluation on the testing set. The model that achieved the best performance on the training set was selected for evaluation on the testing image set discussed earlier in the section on "Stimulus Set and Visual Task Battery" (that is, the images on which we measured neural data and human performance). For each of the 40 checkpoints saved during model training (see above), and each layer of the network, we extracted features for all the testing images. This lead to six timeseries of length 40, each point of which is a (5760, nf_i) matrix, where nf_i is the number of features at layer i. We then computed performance on each tasks on which we had earlier computed neural performance, for each layer and time point. That is, we build linear decoders for each of the testing tasks on top of the features from each layer — effectively equivalent to training a new fully-connected layer (with no nonlinearity) on top of the fixed nonlinear features up to each layer. For each model layer and each time point, we also computer the layer's ability to fit V4 and IT neural data, using procedures identical to those in ref. 30.

We also evaluated the computational model on the spectrum of rotation-limited stimulus sets $Images_\phi$, again using the same procedures as on the whole set. See **Figure 7c** showing the results for three model layers (Layer 1, Layer 3 and Layer 6), for four selected tasks. In addition, we evaluated the computational model on the simple grating stimuli (**Fig. 7b**). We found similar patterns to the neural data for these simpler tasks (**Fig. 7a**), with lower layers having more linearly-accessible information than higher layers.

Alternative computational model with lower-variation training. Both our empirical and computational results suggest that the amount of variation in the stimulus set, rather than the specific task, is a key determinant of the pattern of information through the levels of the ventral stream. However, our results only address the visual system in a "fixed" adult state, being presented with stimulus sets containing various levels of variation. A key question is whether high variation levels are themselves necessary for the proper development of the ventral stream, or whether the empirically observed pattern of information across areas would emerge from any hierarchical neural processing system. While we cannot conclusively answer this question with our existing data, we investigated this question computationally by training an alternative model using the same network architecture as our original model, but replacing the original high-variation (photographic) training set with a data set containing less object view parameter variation. Specifically, we created a synthetic training data set containing images of 1,105 three-dimensional objects in 77 categories, again containing no overlap with the categories of the high-variation test image set. Objects in this alternative training set varied in position, size, and in-plane pose, but did not vary in out-of-plane pose angles, and were presented on uniform gray backgrounds. We then trained a model on this data set to predict the category of the object (a 77-way categorization task; see **Supplementary Fig. 13a**). This trained model was then evaluated on the high-variation testing image set discussed in **Figures 3** and **6**. Although this model did achieve a significant level of generalization of categorization performance to the testing image set (**Supplementary Fig. 13b**), performance on non-categorization tasks was not strongly correlated with categorization performance in most model layers (**Supplementary Fig. 13b**,**c**). Moreover, performance did not typically monotonically increase through model layers (**Supplementary Fig. 14**), and the layers in which peak performance was achieved for one task did not always coincide with the peak-performance layer for other tasks.

These results suggest that the patterns of relative information seen in empirical neural data (for example, increasing information for all tasks) may depend critically on the fact that high levels of object view variation are present during development. They also serve as a kind of control for our computational modeling effort more generally: it is not the case that any deep convolutional network trained to solve an arbitrary object categorization task will trivially exhibit the features of the ventral stream (for example, more information at each succeeding layer for each task) that are reproduced in our original high-variation-trained computational model.

A **Supplementary Methods Checklist** is available.

51. Rosch, E., Mervis, C.B., Gray, W.D. & Johnson, D.M. Basic objects in natural categories. *Cognit. Psychol.* **8**, 382–439 (1976).
52. DiCarlo, J.J. & Maunsell, J.H.R. Inferotemporal representations underlying object recognition in the free viewing monkey. *Soc. Neurosci. Abstr.* **498.2** (2000).
53. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
54. Quiroga, R.Q., Nadasdy, Z. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput.* **16**, 1661–1687 (2004).
55. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. Rust, N.C., Mante, V., Simoncelli, E.P. & Movshon, J.A. How MT cells analyze the motion of visual patterns. *Nat. Neurosci.* **9**, 1421–1431 (2006).
57. Jones, E. *et al.* SciPy: open source scientific tools for Python (2001–) http://www.scipy.org/ (15 July 2015).
58. Efron, B. & Tibshirani, R.J. *An Introduction to the Bootstrap* (CRC Press, 1994).
59. Kanwisher, N., McDermott, J. & Chun, M.M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
60. Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst. 25* 1106–1114 (2012).
61. Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. *Proc. Int. Conf. Mach. Learn.* 115–123 (2013).

doi:10.1038/nn.4247