# Constructing scenes from objects in human occipitotemporal cortex

Sean P. MacEvoy[1,2] & Russell A. Epstein[1]

[1]*Department of Psychology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, Massachusetts 02467, USA*

[2]*Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, Pennsylvania 19104, USA*

**Supplementary Results**

***Choice of scene predictor model***. In the main text, we examined scene predictors that were the averages of patterns evoked by their constituent objects. This choice was rooted in previous work in both humans and non-human primates, which indicated that neural responses to multi-object arrays are well-predicted by the averages of the neural responses to the objects from those arrays. It is worth noting, however, that even if the "true" relationship is the mean, any reasonable linear predictor will produce high classification accuracy. Consider, for example, two 2-dimensional response vectors with endpoints at (1,2) and (2,1), each corresponding to the response pattern evoked by a single scene category. Assume further that these are the "true" (i.e., noise-free) response vectors for these scenes and also that they are the averages of the "true" patterns evoked by their two signature objects. In this scenario, scene predictors derived from the averages of the actual responses to those objects will classify the scene vectors accurately (allowing for some error due to noise in the "actual" – as opposed to "true" – response vectors). Imagine next that we compute scene predictors as the sums (rather than the means) of the object vectors; the resulting predictors will have endpoints roughly (due to noise) at (2,4) and (4,2), respectively. Simple geometry dictates that the Euclidean distance from each of these vector endpoints to the actual vector endpoint of its corresponding scene must be shorter than the distance to the other scene – a consequence of their collinearity. Thus the sum predictors will perform similarly to the mean predictors.

To confirm this, we performed a simulation in Matlab that replicated the analyses of our imaging data. We first defined eight random 100-dimensional "object vectors" which were paired off into scene contexts, and were then averaged within those contexts to generate four "scene vectors". Data from eight independent scans were simulated by adding normally-distributed noise to eight copies of each scene and object vector. From these noisy object vectors we generated two scene predictors: one based on the average of the two pairs object vectors and one based on their sum. We then used these prediction vectors to classify scenes according to the same scheme used with our real data, including dividing data into halves and applying cocktail subtractions. After adding sufficient noise to drive classification using the average predictor to approximately 60% (matching the results with our real data), we found that the performance of the sum vector was not significantly different ($p>0.9$) across a sample of 100 simulated subjects. (In fact, the similarity between the mean and sum predictors

was independent of the level of applied noise.) These results demonstrate that the family of linear predictors is constrained to produce very similar performance.

To understand the relative accuracy of the mean and sum predictors in our real experiments outside the constraints of between-category classification, we computed the median Euclidean distance within each subject between each predictor type and the pattern evoked by the corresponding actual scene. As with classification, distances were accumulated across all possible half-and-half data splits. Across subjects, the resulting median error scores were significantly smaller for the mean predictors than sum predictors in each of the three experiments (Exp. 1, $t(13) = 8.33$, $p < 0.0001$, Exp. 2, $t(12) = 13.18$, $p < 0.0001$; Exp. 3, $t(13) = 10.56$, $p < 0.0001$), indicating that the mean model provided better predictions of scene patterns.

It is possible that a non-linear combination of object patterns could have produced superior scene classification accuracy that what we found with the mean predictor. However, the roughly equivalent response magnitudes for scenes and objects combined with the high performance of the mean predictor suggests that any deviation from linearity must be small.

***Control Analyses.*** In the main analyses, scene predictors based on object averages were generated from twice as much data as scene predictors based on single objects. To address the possibility that superior performance of the mean predictors simply reflected this fact, we recomputed scene classification accuracy after equating the number of stimulus presentations contributing to each predictors. Even after this step, classification performance in LO of the mean predictor was significantly better than the single object predictors, averaged across all subjects in Exps. 1 and 2 for which sufficient scans were available to perform this analysis ($t(20) = 2.1$, $p = 0.047$). (The classification scheme used for Exp. 3 did not permit the further subdivision of observations necessary for this analysis; see Methods.) To further test this point, we directly measured the relative accuracy of the mean and single object predictors by computing the median Euclidean distance within each subject between each predictor type and the pattern evoked by the corresponding actual scene. Across subjects, median distances for the mean predictor were significantly shorter than for the single-object predictor (Exp. 1, $t(9) = 4.3$, $p = 0.002$; Exp. 2, $t(12) = 8.52$, $p < 0.0001$), even after equating the number of response patterns contributing to each predictor. Thus, object-average predictor patterns provided much better estimates of actual scene patterns than single-object predictor patterns.

Another concern was that classification performance for both objects and scenes might have been driven by differences in response magnitude between stimuli, rather than by differences in response patterns. Indeed, we observed differences in the overall magnitude of responses evoked by different categories of objects and scenes, making this idea plausible (Supplementary Figure 1). However, when we repeated our classification procedures using a one-dimensional "response vector" consisting of the mean ROI response to each scene or object category (averaging over all voxels in the ROI), magnitude-based classification performed significantly worse than pattern-based classification. Furthermore, scene-from-object classification based on magnitude alone was not above chance in any ROI.

Nature Neuroscience doi:10.1038/nn.2903

A final concern is that classification performance could be driven by one or two stimulus categories, which would suggest that our results do not generalize across items. However, examination of classification performance broken down by category indicates that this is not the case (Supplementary Figure 2; see also Supplementary Figure 3 for object-from-object, scene-from-scene and scene-from-object similarity matrices).

***Coding of semantic versus visual attributes***. Our ability to classify object-evoked patterns from those evoked by their same-context object counterparts in Exp. 1 (but not Exps. 2 or 3) suggests that LO can code objects at least in part on the basis of semantic attributes. An alternative interpretation, however, is that results in all three experiments were driven entirely by similarities in visual features. In particular, the interpretation of the results from the Exp. 1 in terms of semantic coding is confounded by the fact that objects from some contexts may have shared visual properties. For example, both stoves and refrigerators have large, flat surfaces interrupted by sharp contours, whereas both slides and swings have many thin, spindly metal elements.

If the classification of same-context objects were driven by visual similarities, then one would expect same-context classification performance to depend on the visual similarity of the objects. For example, stoves and refrigerators should elicit more similar patterns than traffic lights and automobiles, because stoves and refrigerators are visually similar to each other while traffic lights and automobiles are visually dissimilar. To test whether this were the case, we directly examined pattern similarities between all eight object categories in Exp. 1 (Supplementary Figure 4). As expected from our classification results, Euclidean distances in activation space between objects from the same context (stove; refrigerator) were shorter than distances between objects drawn from different contexts (stove; bathtub). At the same time, however, there were no significant differences in pattern distances among the four contexts (ANOVA, $F(3,52) = 0.31$, $p = 0.82$). These results argue against the idea that relationships between same-context objects in LO reflect visual similarities and suggest that at least some of the features underlying scene and object representations in LO may be semantic. A delay in the neural manifestation of these semantic properties could account for the convergence of object patterns from the same context in Exp. 1 (which used long presentation times) and the absence of such convergence in Exps. 2 and 3 (which used short presentation times).

***Additional ROIs.*** In addition to the ROIs discussed in the main text, we assessed performance for each of our classification tasks in three other ROIs (Supplementary Figure 6). These included two scene-selective ROIs: the retrosplenial complex (RSC) and the transverse occipital sulcus (TOS). For brevity we report the data pooled over all three experiments. As in the PPA, activity patterns in both of these ROIs carried information about scene and individual object category, but classification of scenes from objects was at chance, suggesting little or no relationship between patterns evoked by scenes and patterns evoked by the objects within them.

A different picture emerged in early visual cortex (see Methods for definition), which exhibited above-chance classification of scenes from object-based predictors, when data were pooled across all three experiments. However, this effect appears to be driven

entirely by data from Exp. 1; for the 14 subjects in this experiment, classification of scenes from object averages was significantly above chance at 61% ($t(13) = 3.96$, $p = 0.002$), which was not significantly different from accuracy in LO ($t(13 = 0.27$, $p= 0.73$). In contrast, classification of scenes from object averages was at chance in both in Exp. 2 (50.9%, $t(13) = 0.39$, $p = 0.70$) and Exp. 3 (52.3%; $t(13) = 0.97$, $p = 0.35$). In both of these experiments, these accuracies were significantly less than in LO (Exp. 2, $t(13) = 2.24$, $p = 0.044$; Exp. 3, $t(13) = 1.63$, $p = 0.025$). One possibility is that above-chance performance in Exp. 1 reflected the long stimulus presentation time used, which may have given time for the averaging scheme initiated in LO to propagate down into early visual cortex.

***Hemispheric differences***. Our searchlight analysis indicated a greater density of voxels containing information about objects within scenes in left LO than in right LO. To investigate hemispheric differences at the ROI level, we extracted classification accuracies in for LO, pF, and PPA by hemisphere (Supplementary Figure 7). In LO we observed no significant left/right difference in classifying scenes from object averages in any of the three experiments (Exp. 1: $t(13) = 1.52$, $p = 0.15$; Exp. 2: $t(13) = 0.79$, $p = 0.44$; Exp. 3: $t(14) = 0.47$, $p = 0.64$). In pF we observed significantly higher accuracy classifying scenes from object averages in the left hemisphere in Exp. 3 ($t(13) = 2.70$, $p = 0.019$), but not in Exps. 1 or 3 (Exp. 1: $t(13) = 0.31$, $p = 0.38$; Exp. 2: $t(13) = 1.66$, $p = 0.13$). In PPA we observed no significant difference between hemispheres in any experiment (Exp. 1: $t(13) = 0.42$, $p = 0.68$; Exp. 2: $t(13) = 0.87$, $p = 0.40$; Exp. 3: $t(14) = 0.59$, $p = 0.56$).

***Adaptation analysis***. As a complement to our pattern classification approach, we examined our data for adaptation effects that could also indicate similarities between the neural representations of scenes and their constituent objects. We were particularly interested in the possibility that responses to scenes that followed one of their associated objects, and responses to objects that followed scenes containing them, might be reduced in magnitude compared to responses to scenes and objects that followed unrelated stimuli. Such an adaptation effect would indicate that scenes and their associated objects activate similar pools of neurons.

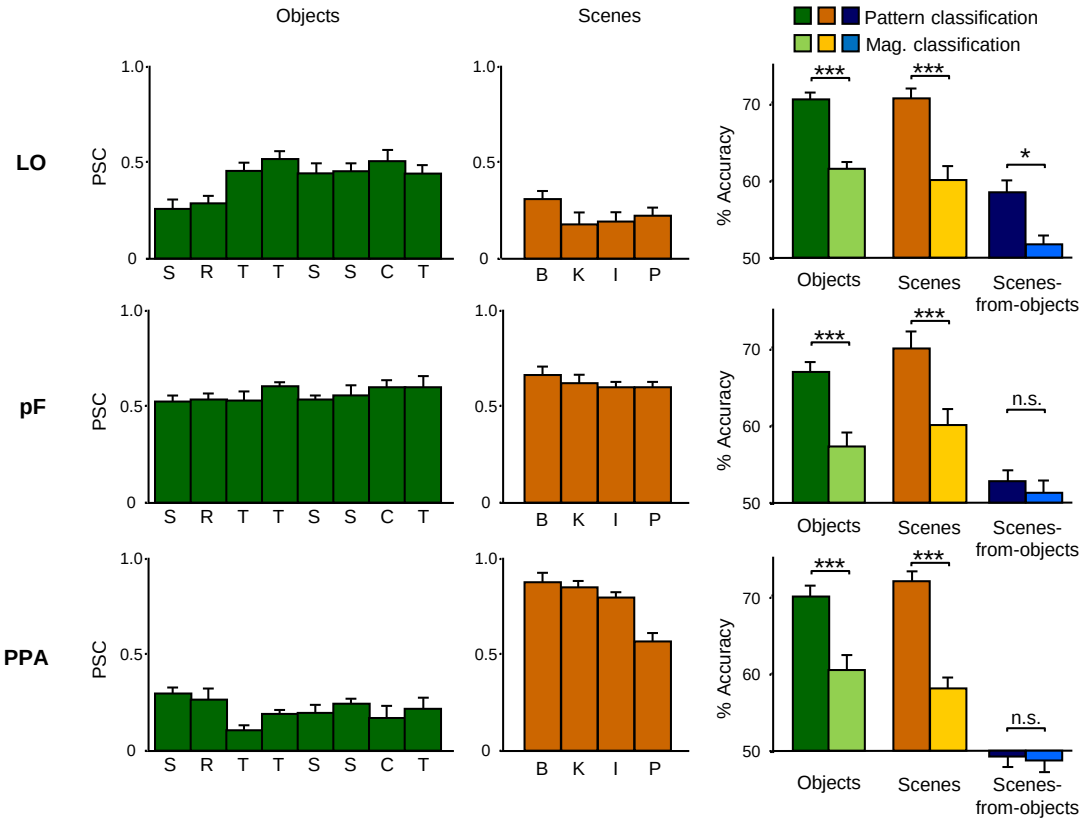
Adaptation effects were assessed by feeding spatially smoothed volumes to a separate GLM containing covariates for each of the nine possible transitions of interest between objects and scenes. For object stimuli, these conditions were: an object preceded by a scene from the same context (e.g. kitchen→refrigerator), an object preceded by a scene from a different context (e.g. playground→refrigerator), an object preceded by an object from the same category (e.g. stove→stove), an object preceded by an object from the other category in the same context (e.g. stove→refrigerator), and an object preceded by an object from a different context (e.g. swing→refrigerator). For scene stimuli, these conditions were: a scene preceded by a scene from the same category (e.g. kitchen→kitchen), a scene preceded by a scene from a different category (e.g. playground→kitchen), a scene preceded by an object from the same context (e.g. refrigerator→kitchen), and a scene preceded by an object from a different context (e.g. swing→kitchen). Note that although our primary interest was in the scene to object and object to scene transitions, the scene to scene and object to object conditions were also examined. Unlike models used for pattern analysis, functional volumes were

Nature Neuroscience doi:10.1038/nn.2903

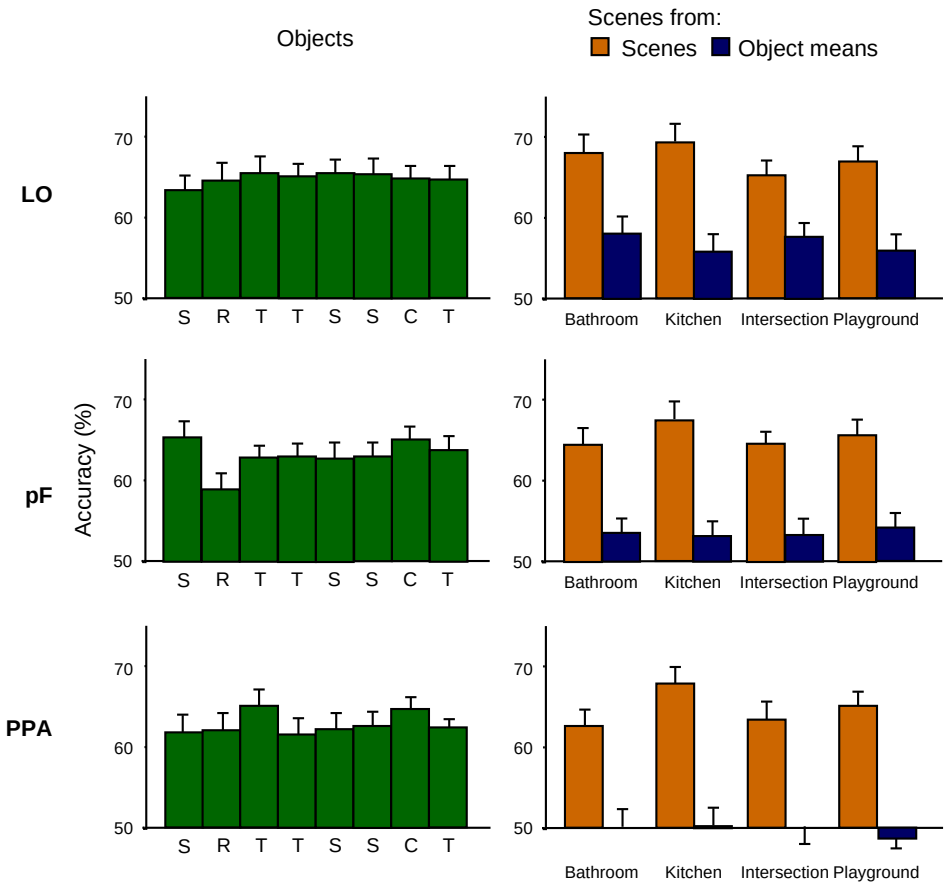
concatenated across all scans for each subject, yielding a single beta value per transition type for each subject.

We observed no significant adaptation effects either for transitions from scenes to objects from the same context or for transition from objects to scenes from the same context in the PPA, LO, or pF in any of the three experiments. To place this negative result in a proper context, though, it should be noted that none of these three regions showed significant adaptation effects even for repetition of objects or scenes from the same category (e.g., stove→stove, or playground→playground), a finding that is consistent with previous work showing an absence of same category/different exemplar adaptation for objects[1, 2] (but see[3]). Although it is possible that our design was simply not powerful enough to observe adaptation effects, an alternative hypothesis is that MVPA and adaptation reveal neural organization at different spatial scales[4]. In particular, the MVPA results might be driven by category-based clustering of neurons that are themselves more finely tuned to simpler visual feature or to individual category exemplars.

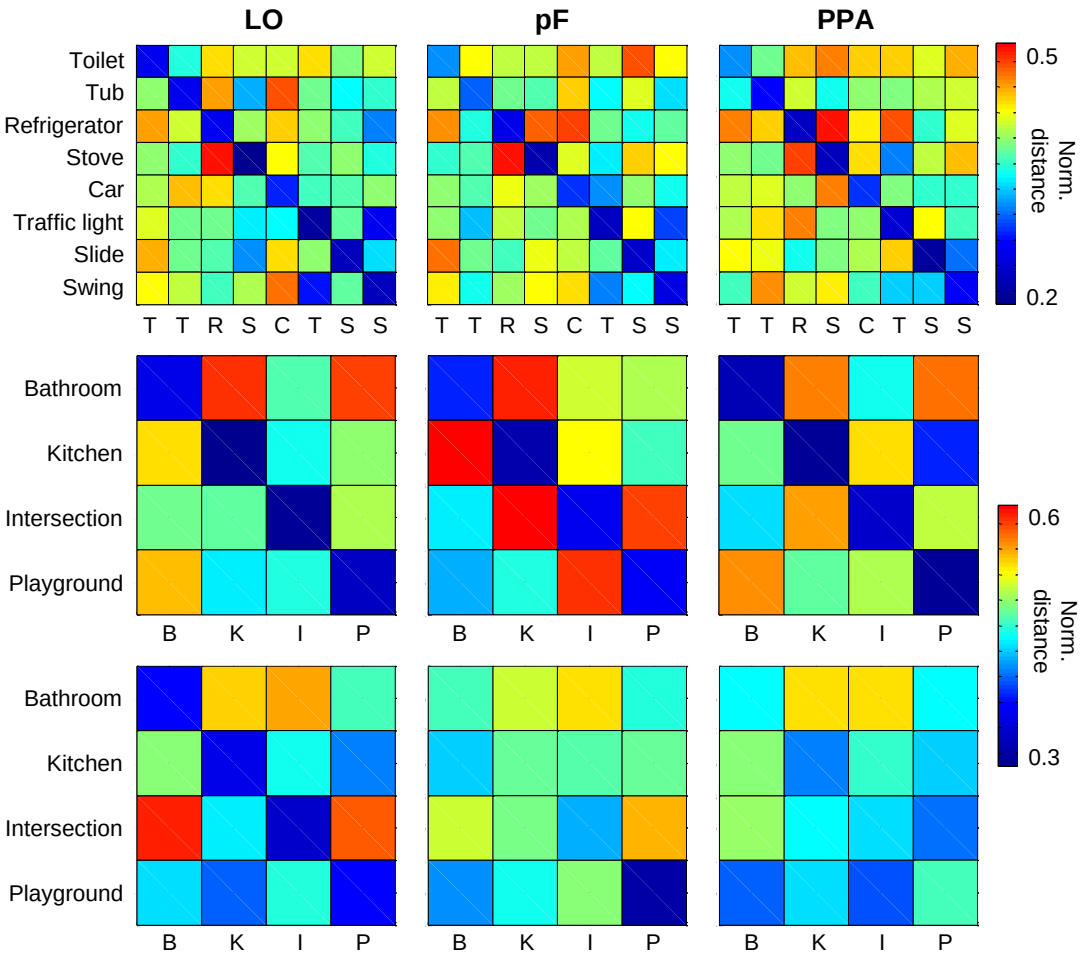
Nature Neuroscience doi:10.1038/nn.2903



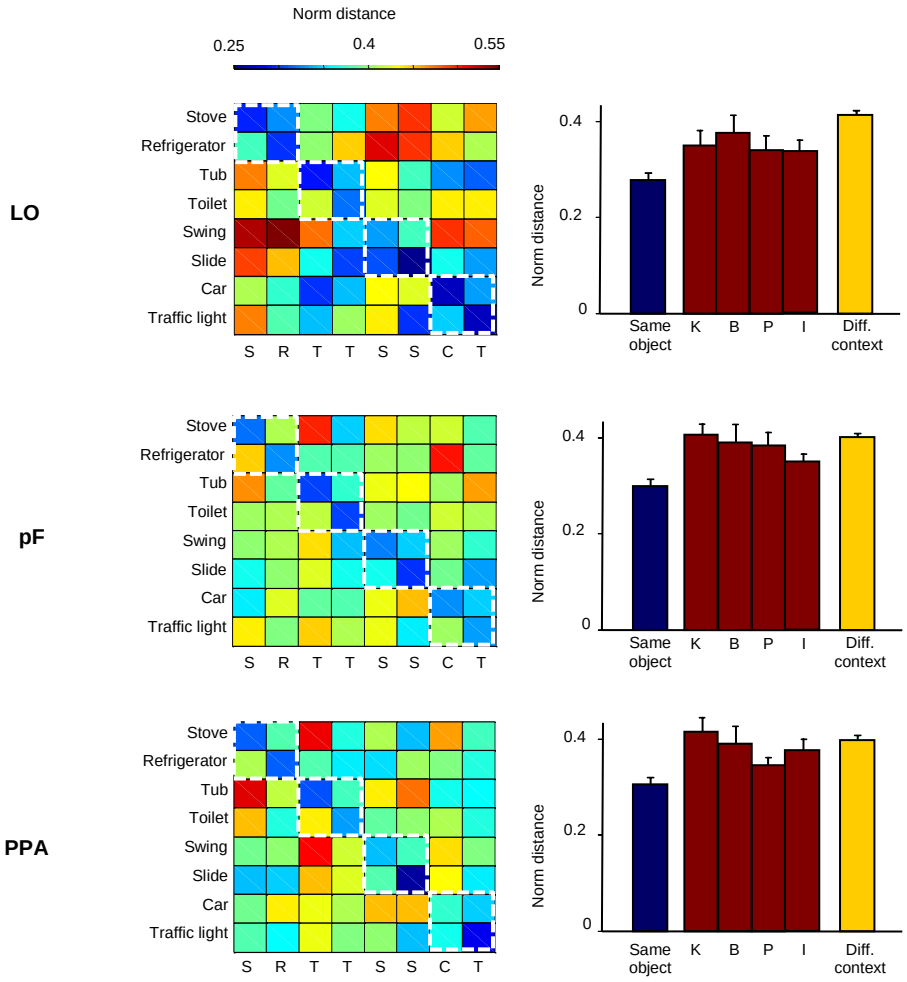

**Supplementary Figure 1**. Comparison of pattern- and magnitude-based classification averaged across all three experiments. We observed variability in the overall response magnitude evoked by objects (left column; S = strove, R = refrigerator, T = tub, T = toilet, S = slide, S = swing, C = car, T = traffic light) and scenes (middle column; B = bathroom, K = kitchen, I = intersection, P = playground). We tested whether these magnitude differences could account for our pattern classification results by recomputing classification accuracies based upon response magnitude (averaged over all voxels in the ROI) and comparing the results to accuracies based upon Euclidean distances based on multivoxel patterns (right column). Although magnitude-based classification was above chance, it was uniformly inferior to pattern-based classification for all three classification tasks: objects versus objects, scenes versus scenes, and scenes versus object mean predictors. Error bars are s.e.m. Significance levels: * p <0.05*** p < 0.001.



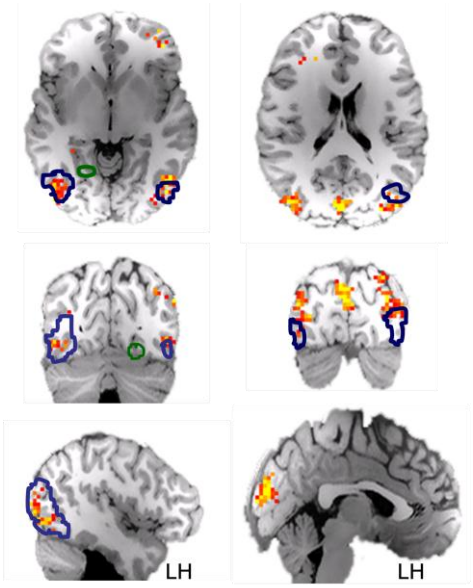

**Supplementary Figure 2**. Classification accuracy broken down by stimulus category for objects-from-objects (left column), and scenes from actual scenes and from object mean predictors (right column), averaged across all three experiments. Error bars are s.e.m.




**Supplementary Figure 3**: Pattern distance matrices for objects (top row), scenes (middle row) and scenes versus object averages (bottom row). Data were averaged across all three experiments; ROIs are labelled at top. Raw distance matrices for each subject were independently anchored at 0 by their minima and normalized to their remaining maxima, then averaged across subjects. Matrix rows denote the category in the first half of each half/half data split, while columns denote the category in the second half. In the bottom row of matrices, each entry is the average of two values: the distance between the mean predictor for the row category and the actual scene pattern for the column category, and the distance between the actual scene patterns for the row category and the mean predictor for the column category.

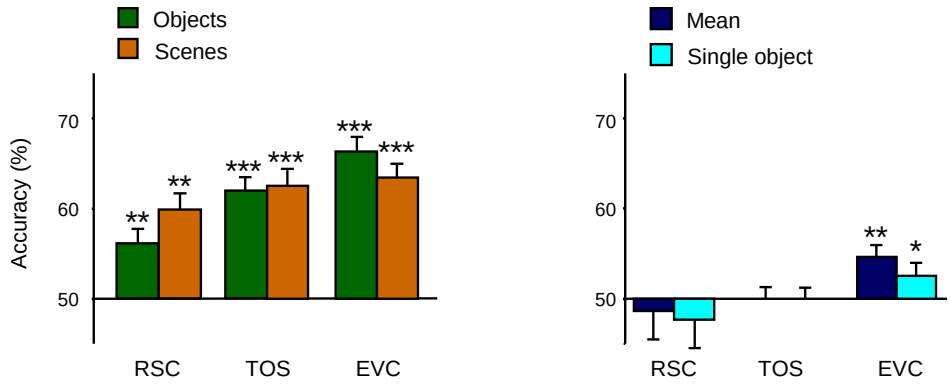



**Supplementary Figure 4**. Direct measurement of distances among object-evoked patterns in Experiment 1. Within each subject we computed the Euclidean distance between each object-evoked pattern in one half of scans and each object-evoked pattern in the remaining scans. Matrices were rescaled to range between 0 and 1 and then averaged across all half-and-half scan splits and across all subjects to produce the matrices shown in the left column. The color map for all three matrices is at top. In each ROI, pattern distances tended to be shortest (blue shades) along the main diagonal, corresponding to distances between patterns evoked by the same object category in the two data halves; LO distances also appear shorter between same-context objects (lower-left and upper right squares within dashed boxes denoting within-context distances). This is confirmed in histograms at right, which present average distances from matrices sorted according to the relationship between objects: same object category (blue bar), same-context objects (red bars) in kitchens (K), bathrooms (B), playgrounds (P), and intersections (I), and objects from different contexts (yellow bar). Consistent with our classification results, LO pattern distances between same-context objects were shorter than between objects from different contexts. There were no significant differences in within-context pattern distance among the four contexts, however. Error bars are s.e.m.

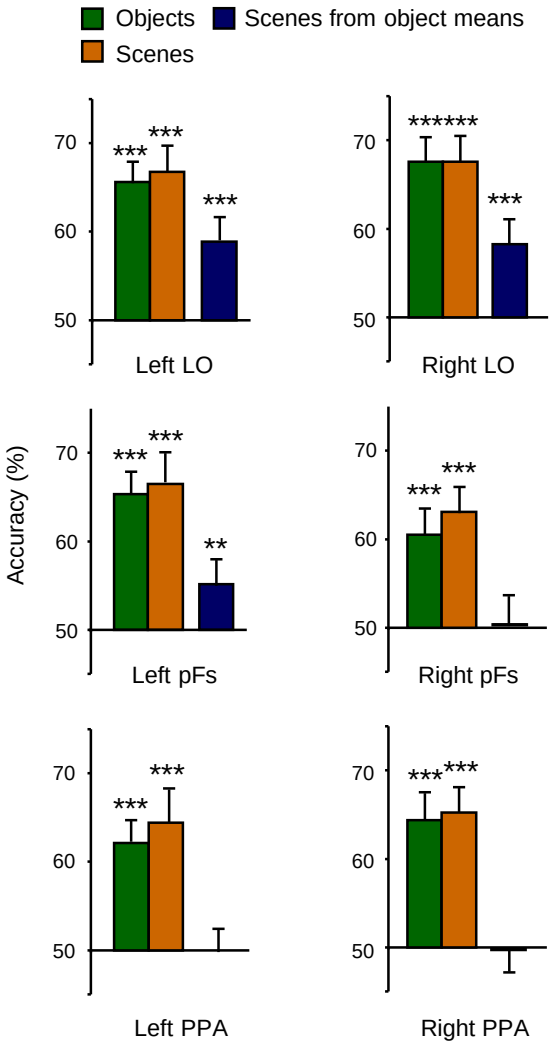



**Supplementary Figure 5**. Searchlight classification accuracy for objects from contextually-associated objects in Exp.1. Painted voxels are those at the centers of clusters that identified object-evoked patterns from the patterns of the contextually-linked object counterparts with high accuracy ($p < 0.005$, uncorrected). The left column shows cardinal slices intersecting in a region of high classification accuracy that coincided with LO. The right column shows slices intersecting at a more dorsal midline region of high classification accuracy within the cuneus. LO is outlined in blue, PPA in green; pF does not appear in these planes.

Nature Neuroscience doi:10.1038/nn.2903




**Supplementary Figure 6**: Classification accuracies in the retrosplenial complex (RSC), transverse-occipital sulcus (TOS) and early visual cortex (EVC). Data are pooled across all three experiments. Error bars are s.e.m. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The above-chance scene-from-object classification performance in early visual cortex was driven entirely by data from Exp. 1 and was not observed in the data from Exps. 2 or 3.




**Supplementary Figure 7**. Classification accuracies by hemisphere in LO, pF, and PPA, averaged across all three experiments. Although our searchlight analysis suggested that classification accuracies were higher in left LO than in right LO, there was no statistically significant effect of hemisphere on accuracy in this region. Interestingly, scene-from-object classification reached significance in left but not right pF when data were pooled across all three experiments. Error bars are s.e.m. Significance levels: ** $p < 0.01$, *** $p < 0.001$.

References

1. Grill-Spector, K., *et al.* Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* 24, 187-203 (1999).
2. Kim, J.G., Biederman, I., Lescroart, M.D. & Hayworth, K.J. Adaptation to objects in the lateral occipital complex (LOC): shape or semantics? *Vision Res.* 49, 2297-2305 (2009).
3. Koutstaal, W., *et al.* Perceptual specificity in visual object priming: functional magnetic resonance imaging evidence for a laterality difference in fusiform cortex. *Neuropsychologia* 39, 184-199 (2001).
4. Drucker, D.M. & Aguirre, G.K. Different spatial scales of shape similarity representation in lateral and ventral LOC. *Cereb. Cortex* 19, 2269-2280 (2009).

Nature Neuroscience doi:10.1038/nn.2903