Distinct contributions of functional and deep neural network features to scene representation in brain and behavior

Groen, Iris I. A.[1], Greene, Michelle R.[2], Baldassano, Christopher [3], Fei-Fei, Li.[4], Beck, Diane M. [5], Baker, Chris I.[1]

[1] National Institutes for Health, Laboratory of Brain and Cognition, Bethesda, MD
[2] Bates College, Neuroscience Program, Lewiston, ME
[3] Princeton University, Princeton Neuroscience Institute, Princeton, NJ
[4] Stanford University, Stanford Vision Lab, Stanford, CA
[5] University of Illinois, Department of Psychology and Beckman Institute, Urbana-Champaign, IL

Corresponding author:

Iris I.A. Groen
10 Center Drive
Building 10, Room 4C108, MSC 1240
Besthesda, MD 20892

Phone: 301-435-8905
E-mail: iris.groen@nih.gov

Abstract

Real-world scenes are rich, heterogeneous stimuli that contain inherent correlations between many visual and semantic features, making it difficult to determine how different scene properties contribute to neural representations. Here, we assessed the unique contributions of three behaviorally relevant feature spaces by a) selecting stimuli for which inherent correlations were minimized *a priori* and b) partitioning the neural variance attributed to each individual feature space. We found that while scene categorization behavior is best explained by a functional feature space reflecting potential actions in scenes, cortical responses in scene-selective areas are best explained by mid- and high-level layers of computational deep neural network models (DNNs). While other regions of extrastriate cortex represented some functional features, our findings reveal a striking dissociation of functional versus DNN features in their contribution to scene categorization and brain responses, indicating that scene-selective cortex and DNNs represent only a subset of behaviorally relevant scene information.

Introduction

Although researchers of visual perception often use simplified, highly controlled images in order to isolate the underlying neural processes, real-life visual perception requires the continuous processing of complex visual environments to support a variety of behavioral goals, including recognition, navigation and action planning (Malcolm et al, 2016). In the human brain, the perception of complex scenes is characterized by the activation of three scene-selective regions, the Parahippocampal Place Area (PPA; Aguirre et al. 1998; Epstein and Kanwisher 1998), Occipital Place Area (OPA; Hasson et al. 2002; Dilks et al. 2013), and Medial Place Area (MPA; Silson et al. 2016), also referred to as the Retrosplenial Complex (Bar and Aminoff 2003). A growing body of fMRI literature focuses on how these regions might facilitate scene understanding by investigating what information drives neural responses in these regions when human observers view scene stimuli. Currently, a large set of candidate low- and high-level characteristics of scenes have been identified, including but not limited to: a scene's constituent objects and their co-occurrences; spatial layout; surface textures; contrast and spatial frequency, as well as scene semantics, contextual associations, and navigational affordances (see Epstein 2014; Malcolm et al. 2016; Groen et al. 2017, for recent reviews).

This list of candidate characteristics highlights two major challenges in uncovering neural representations of complex real-world scenes (Malcolm et al. 2016). First, there are many inherent correlations between different scene properties. For example, forests are characterized by the presence of spatial boundaries and numerous vertical edges, whereas beaches are typically open with a prominent horizon, resulting in correlations between semantic category, layout and spatial frequency (Oliva and Torralba 2001; Torralba and Oliva 2003). This makes it problematic to explain neural representations of scenes based on just one of these properties (Walther et al. 2009; Kravitz et al. 2011; Park et al. 2011; Rajimehr et al. 2011) without taking into account their covariation. Indeed, an explicit test of spatial frequency, subjective distance and semantic properties found that due to inherent feature correlations, all three explained the same variance in fMRI responses, with no discernible unique contribution (Lescroart et al. 2015). Second, given the large number of possible models and the limited number that can realistically be tested in a single study, how do we select which models to focus on?

In this fMRI study, we addressed both these challenges by testing three models chosen for their behavioral relevance, and *a priori* selecting stimuli that reduced the covariance between these models. Specifically, our choice of models (*feature spaces*) was informed by a behavioral study that investigated the contribution of a large range of features to scene understanding (Greene et al. 2016). Using online crowd-sourcing on a large scene database (the SUN database, Xiao et al. 2014), Greene and colleagues found that the three models that best explained human scene categorization were 1) human-assigned object labels ('object model'), 2) a deep convolutional neural network ('DNN model'), and 3) a model based on actions that can be carried out in the scene ('function model'). To isolate the contribution of each of these models to neural scene representation, we compared them against multi-voxel patterns in fMRI data collected while participants viewed these scenes, and quantified their unique contributions using variance partitioning, accounting for any residual overlap in representational structure.

To anticipate, our data reveal a striking dissociation between the feature space that maximally drives behavioral scene categorization and the space that best explain scene-selective cortex. While the function model best predicted scene categorization, there was no unique representation of scene function in scene-selective brain regions, which instead are fully described by the deep network features. Cortical responses in scene-selective cortex were captured by both mid- and high-level DNN layers, while scene functions correlated with responses in regions outside of scene-selective cortex, some of which have been associated with action observation. This dissociation suggests a distributed functional organization of real-world scene information across visual cortex that extends beyond scene-selective regions.

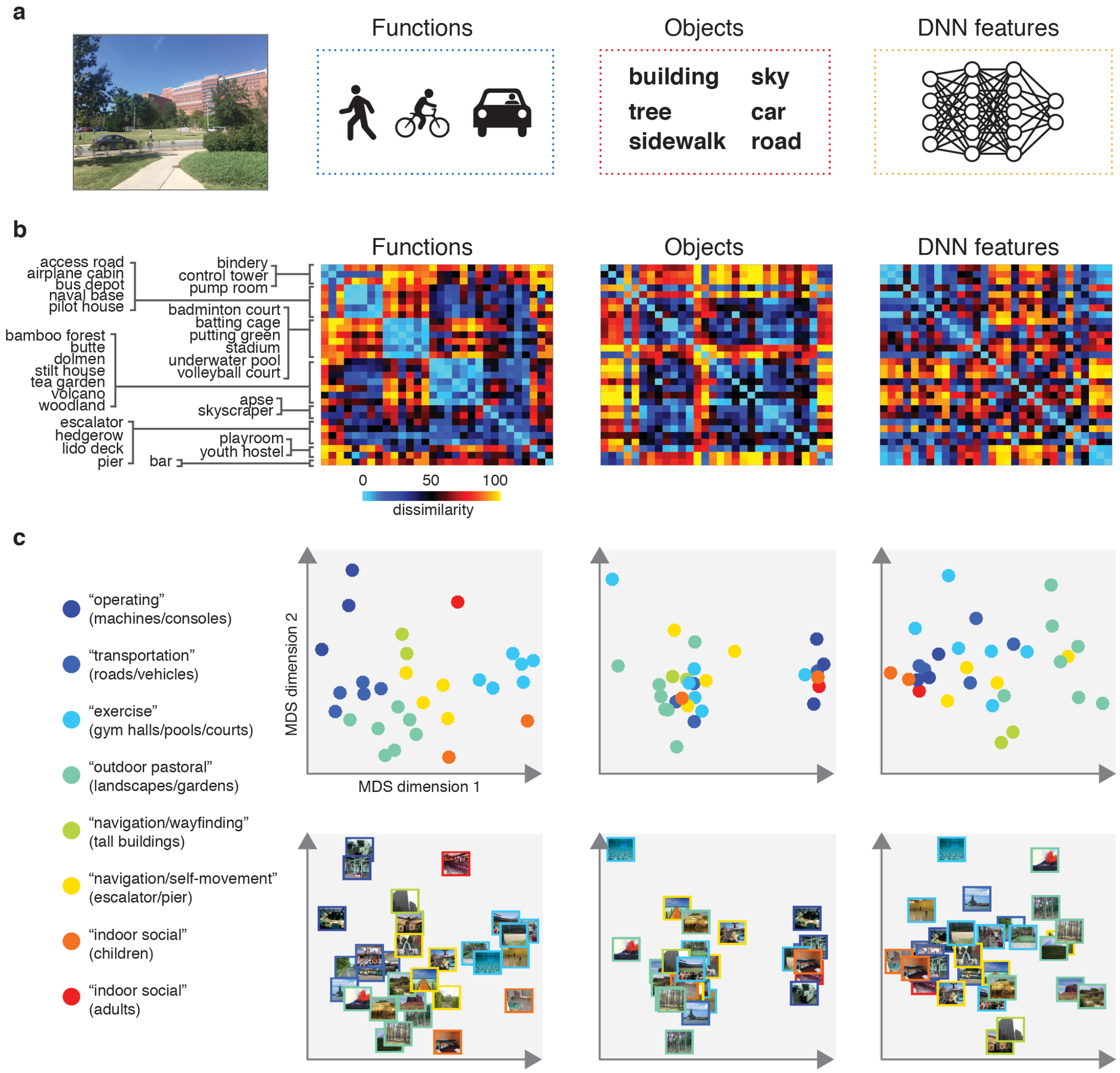

Figure 1 Models and predicted stimulus dissimilarity. **A)** Stimuli were characterized in three different ways: functions (derived using human-generated action labels), objects (derived using human-generated object labels) and DNN features (derived using layer 7 of a 1000-class trained convolutional neural network). **B)** RDMs showing predicted representational dissimilarity in terms of functions, objects and DNN features for the 30 scene categories sampled from Greene et al., (2016) for the purpose of the current study. Scenes were sampled to achieve minimal between-matrix correlations, with the constraint that the final stimulus set should be have equal portions of categories from indoor, outdoor man-made and outdoor natural scenes. The category order in the figure is determined based on a k-means clustering on the functional model RDM; clustering was performed by requesting 8 clusters, which explained 80% of the variance in the functional feature space. RDMs were rank-ordered for visualization purposes only. **C)** Multi-dimensional scaling plots of the model RDMs, color-coded based on the 8 functional clusters depicted in B). Functional model clusters indicated functions such as 'sports', and 'transportation' (note that these semantic labels were derived post-hoc after k-means clustering, and did not affect stimulus selection). Critically, representational dissimilarity based on the two other models (objects and DNN features) predicted different cluster patterns.

Results

Disentangling visual feature, object and functional information in scenes

The goal of the study was to determine the contributions of object, DNN and functional feature spaces to neural representations in scene-selective cortex. To do this, we created a stimulus set by iteratively sampling from the large set of scenes previously characterized in terms of these three types of information by Greene et al. (2016). The DNN feature space was derived using a high-level layer of an AlexNet (Alex et al. 2012; Sermanet et al. 2013) that was pre-trained using ImageNet class labels (Deng et al. 2009), while the object and function feature spaces were derived based on object and action labels assigned by human observers through Amazon Mechanical Turk (see Methods for details). On each iteration, pairwise distances between a subset of pseudo-randomly sampled categories were determined for each of these feature spaces, resulting in three representational dissimilarity matrices (RDMs) reflecting either the deep network, object or functional feature space (**Figure 1*A***) for that sample. Constraining the set to include equal numbers of indoor, urban, and natural landscape environments, our strategy was inspired by the odds algorithm of Bruss (2000), in that we rejected the first 10,000 solutions, selecting the next solution that had lower inter-feature correlations than had been observed thus far. Thus, a final selection of 30 scene categories was selected in which the three RDMs were minimally correlated (Pearson's *r*: 0.23-0.26; **Figure 1*B-C*;** see Methods).

Twenty participants viewed the selected scenes while being scanned on a high-field 7T Siemens MRI scanner using a protocol sensitive to blood oxygenation level dependent (BOLD) contrasts (see Methods). Stimuli were presented for 500 ms each while participants performed an orthogonal task on the fixation cross. To assess how each feature space contributed to scene categorization behavior for our much reduced stimulus set (30 instead of the 311 categories of Greene et al. 2016), participants performed a behavioral multi-arrangement task

(Kriegeskorte and Mur 2012) on the same stimuli, administered on a separate day after scanning. In this task, participants were presented with all stimuli in the set arranged around a large white circle on a computer screen, and were instructed to drag-and-drop these scenes within the white circle according to their similarity (see Methods and **Figure 2*A***).

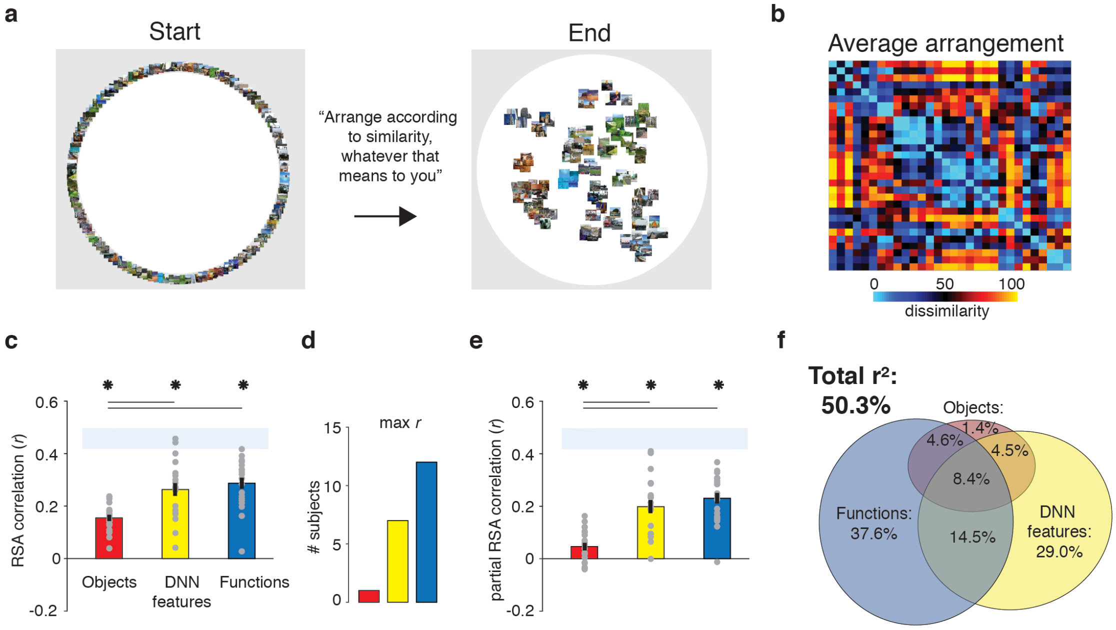

Figure 2 Behavioral multi-arrangement paradigm and results. **A)** Participants organized the scenes in inside a large white circle according to their similarity as determined by their own judgment, without receiving explicit instructions as to what information to use to determine scene similarity. **B)** RDM displaying the average dissimilarity between categories in behavioral arrangement (rank-ordered for visualization only). **C)** Average (bar) and individual participant (gray dots) correlations between the behavioral RDM and the model RDMs for objects (red), DNN features (yellow) and functions (blue) from Figure 1B. Stars (*) indicate $p < 0.05$ for model-specific one-sided signed-rank tests against zero, while horizontal bars indicate $p < 0.05$ for two-sided pairwise signed-rank tests between models; *p*-values were FDR-corrected across both types of comparisons. **D)** Count of participants with the highest correlation with either objects, DNN features or objects. **E)** Average (bar) and individual participant (gray dots) partial correlation valyes for each model RDM. Statistical significance was determined the same way as in C). **F)** Euler diagram depicting the results of a variance partitioning analysis on the behavior for objects (red circle), DNN features (yellow circle) and functions (blue circle). Unique (non-overlapping diagram portions) and shared (overlapping diagram portions) variances are expressed as percentages of the total variance explained by all models combined.

Functions uniquely predict scene categorization behavior

To determine what information contributed to scene categorization in the multi-arrangement task, we created RDMs based on each participant's final arrangement by measuring the pairwise distances between all 30 categories in the set (**Figure 2*B***), and then computed correlations of these RDMs with the three model RDMs that quantified the similarity of the scenes in terms of either functions, objects, or DNN features, respectively (see Figure 1*B*).

Replicating Greene et al., (2016), this analysis indicated that all three feature spaces were significantly correlated with scene categorization behavior, with functions having the highest correlation on average (**Figure 2*C***; objects: mean $r = 0.16$; DNN features: mean $r = 0.26$; functions: mean $r = 0.29$, Wilcoxon one-sided signed-rank test, all W(20) > 210, all $z > 3.9$, all $p < 0.0001$). The correlation with functions was higher than with objects (Wilcoxon two-sided signed-rank test, W(20) = 199, $z = 3.5$, $p = 0.0004$), but not than with DNN features (W(20) = 134, $z = 1.1$, $p = 0.28$), which also correlated higher than objects (W(20) = 194, $z = 3.3$, $p = 0.0009$). However, comparison at the level of individual participants indicated that functions outperformed both the DNN and object models for the majority of participants (highest correlation with functions: n = 12; with DNN features: n = 7; with objects: n = 1; **Figure 2*D***).

While these correlations indicate that scene dissimilarity based on the functional feature space best matched the stimulus arrangements that participants made, they do not reveal to what extent functional, DNN or object features *independently* contribute to the behavior. To assess this, we performed two additional analyses. First, we computed *partial* correlations between models and behavior whereby the correlation of each feature space with the behavior was determined while taking into account the contributions of the other two feature spaces. The results indicated that each model independently contributed to the behavioral data: significant partial correlations were obtained for the object (W(20) = 173, $z = 2.5$, $p = 0.006$), DNN (W(20) = 209, $z = 3.9$, $p < 0.0001$) and functional feature spaces (W(20) = 209, $z = 3.9$, $p < 0.0001$), with the functional model having the largest partial correlation (**Figure 2*E***). Direct comparisons yielded a similar pattern as the independent correlations, with weaker contributions of objects relative to both functional (W(20) = 201, z = 3.6, p < 0.0003) and DNN features (W(20) = 195, z = 3.4, p = 0.0008), whose partial correlations did not differ (W(20) = 135, z = 1.12, p = 0.26).

Second, we conducted a variance partitioning analysis, in which the function, DNN and object feature spaces were entered either separately or in combination as predictors in a set of multiple regression analyses aimed at explaining the categorization behavior. By comparing the explained variance based on regression on individual models versus models in combination, we computed portions of unique variance contributed by each model as well as portions of shared variance across models (see Methods for details). A full model in which all three models were included explained 50.3% of the variance in the average behavioral categorization pattern (**Figure 2*F***). Highlighting the importance of functions for scene categorization, the largest portion of this variance could be uniquely attributed to the functional feature space (unique r^2 = 37.6%), more than the unique variance explained by the DNN features (unique r^2 = 29.0%) or the object features (unique r^2 = 1.4%). This result is consistent with the findings of Greene et al., (2016), who found unique contributions of 45.2% by the function model, 7.1% by the DNN model[*], and 0.3% by objects, respectively. One interesting difference with this previous study is that the degree of shared variance across all three models in our study is notably smaller (8.4% versus 27.4%); this is presumably a result of our stimulus selection procedure that was explicitly

[*] When performing the variation partition on the behavioral categorization measured in Greene et al., (2016) but limited to the 30 scene categories that were used here, we obtained a highly similar distribution of unique variances as for the current behavioral data, namely 42.8% for the function model, 28.0% for the DNN model, and 0.003% for the objects, respectively. This suggests that the higher contribution of the DNN to the behavior relative to what is reported in Greene et al., (2016) is a result of the reduced stimulus set used here, rather than a qualitative difference in experimental results between the previous study and the current study.

aimed at minimizing correlations between the models. Importantly, a reproducibility test (see Methods)that was built-in to our design indicated that the representational space reflected in the behavior was highly generalizable, resulting in a between-set RDM correlation of r = 0.73 (95% confidence interval = [0.73-0.88], p = 0.0001), as assessed by comparison of the two different sets of scene exemplars that were evenly distributed across participants.

In sum, these results confirm an important, independent contribution of the functional feature space to scene understanding, but this time as evidenced by multi-arrangement sorting behavior (as opposed to a same/different categorization task). We also found a smaller, separate contribution of deep network features, while the unique contribution of the object feature space was negligible. Next, we examined to what extent this information is represented in brain responses to the same set of real-world scenes as measured with fMRI.

DNN features uniquely predict responses in scene-selective cortex

To determine the information that is represented in scene-selective regions PPA, OPA and MPA, we created RDMs based on the pairwise comparisons of multi-voxel activity patterns for each category in these cortical regions (**Figure 3*A***), which we subsequently correlated with the RDMs based on the object, function and DNN feature spaces. Similar to the behavioral findings, all three spaces correlated with the fMRI response patterns to scenes in PPA (objects: W(20) = 181, z = 2.8, p = 0.002; DNN: W(20) = 206, z = 3.8, p < 0.0001; functions: W(20) = 154, z = 1.8, p = 0.035, see **Figure 3*B***). Unlike our behavioral findings, however, fMRI dissimilarity in PPA correlated more strongly with the DNN model than the object (W(20) = 195, z = 2.5, p = 0.012) and function (W(20) = 198, z = 3.5, p < 0.0005) models, which did not differ (W(20) = 145 , z = 1.5, p = 0.14). In OPA, only the DNN model correlated with the fMRI response patterns (W(20) = 165, z = 2,2, p = 0.013), and this correlation was again stronger than the object model (W(20) = 172, z = 2.5, p = 0.012), but not the function model (W(20) = 134, z = 1.1, p = 0.28).

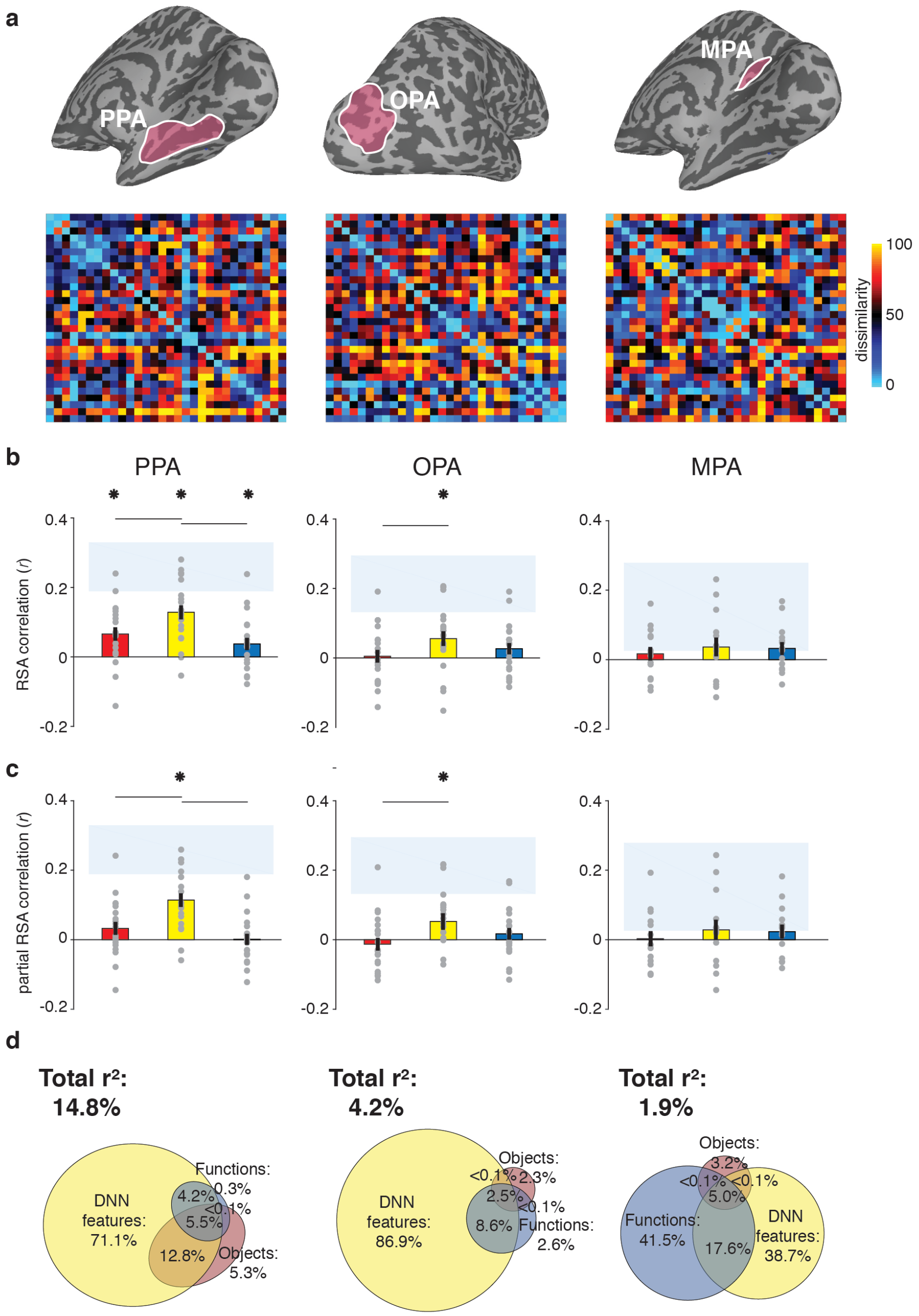

Figure 3 RDMs and model comparisons for fMRI Experiment 1 (n = 20). **A)** RDMs displaying average dissimilarity between categories in multi-voxel patterns in PPA, OPA and MPA (rank-ordered for visualization only). **B)** Average (bar) and individual (gray dots) correlations between the ROIs in A) and the model RDMs for objects (red), DNN features (yellow) and functions (blue). Stars (*) indicate $p < 0.05$ for model-specific one-sided signed-rank tests against zero, while horizontal bars indicate $p < 0.05$ for two-sided pairwise signed-rank tests between models; p-values were FDR-corrected across both types of comparisons within each ROI. **C)** Average (bar) and individual (gray dots) partial correlation coefficients for each model RDM. Statistics are the same as in B). **D)** Euler diagram depicting the variance partitioning results the average dissimilarity in each ROI for each of the three models, expressed as percentages of unique and shared variance of the variance explained by all three models together.

In MPA, none of the model correlations were significant (all W(14) < 76, all *z* < 1.4, all *p* > 0.07).

When the three models were considered in combination only the DNN model yielded a significant partial correlation (PPA: W(20) = 203, *z* = 3.6, *p* < 0.0001, OPA: W(20) = 171, *z* = 2.5, *p* = 0.007, **Figure 3*C***), further showing that DNN features best capture responses in scene-selective cortex. No significant partial correlation was found for the object model (PPA: W(20) = 148, *z* = 1.6, *p* = 0.056; OPA: W(20) = 74, *z* = 1.2, *p* = 0.88) or the function model (PPA: W(20) = 98, *z* = 0.3, *p* = 0.61, OPA: W(20) = 127, *z* = 0.8, *p* = 0.21), or for any model in MPA (all W(14) < 63, all *z* < 0.66, all *p* > 0.50). Variance partitioning of the fMRI response patterns (**Figure 3*D***) indicated that the DNN model also contributed the largest portion of unique variance: in PPA and OPA, DNN features contributed 71.1% and 68.9%, respectively, of the variance explained by all models combined, more than the unique variance explained by the object (PPA: 5.3%; OPA, 2.3%) and function (PPA: 0.3%; OPA: 2.6%) feature spaces. In MPA, a larger share of unique variance was found for the function model (41.5%) than for the DNN (38.7%) and object model (3.2%); however, overall explained variance in MPA was much lower than in the other ROIs. The direct test of reproducibility indicated that RDMs generalized across participants and stimulus sets for PPA (*r* = 0.26 [0.03-0.54], p = 0.009) and OPA (*r* = 0.23 [0.04-0.51], *p* = 0.0148), but not in MPA (*r* = 0.06 [-0.16-0.26], *p* = 0.29), suggesting that the multi-voxel patterns measured in MPA were less stable (see also the low noise ceiling in MPA in Figure 3B/C).

Taken together, the fMRI results indicate that of the three models considered, deep network features (derived using a pre-trained DNN model) best explained the coding of real-world scene information in scene-selective regions PPA and OPA, more so than object or functional information derived from semantic labels that were explicitly generated by human observers. For MPA, results were inconclusive, as none of the models adequately captured the response patterns measured for in this region, which also did not contain multi-voxel patterns that generalized across stimulus sets and participants. This result highlights a discrepancy of the brain responses with the behavioral results, which indicated a strong contribution of functions to scene representation, which was largely independent of the DNN features. To better understand if and how scene-selective cortex represented behaviorally relevant information, we next investigated how the observed behavior related to the fMRI responses.

Scene selective cortex correlation with behavior reflects DNN feature space

To assess the extent to which the patterns of response observed in scene-selective cortex predicted behavior, we correlated the RDMs in each of the ROIs with three measures of behavioral categorization: 1) the large-scale online categorization behavior measured in Greene et al., (2016), 2) the average behavior in the multi-arrangement task, and 3) each participant's own behavioral multi-arrangement data. This analysis revealed a significant correlation with behavior in all three scene-selective ROIs (**Figure 4*A***). In PPA, all three behavioral measures correlated with neural patterns of response (signed-rank test, online categorization behavior: W(20) = 168, z = 2.3, p = 0.010; average multi-arrangement behavior: W(20) = 195, z = 3.3, p = 0.0004; own arrangement behavior: W(20) = 159, z = 2.0, p = 0.023). In OPA, significant correlations were found for both of the average behavioral measures (online categorization behavior: W(20) = 181, z = 2.8, p = 0.002; average multi-arrangement behavior: W(20) = 158, z = 1.96, p = 0.025), but not for the participant's own behavior (W(20) = 106, z = 0.02, p = 0.49), possible due to higher noise in individual data. Interestingly, however, MPA showed the opposite pattern: participant's own behavior was significantly related to the observed patterns of response (W(14) = 89, z = 2.26, p = 0.011), but the average behavioral measures were not (online behavior: W(14) = 47, z = 0.4, p = 0.65; average behavior: W(14) = 74, z = 1.3, p = 0.09). Combined with the reproducibility test (see above), this suggests that the representations in MPA are possibly more idiosyncratic to individual participants or stimulus sets.

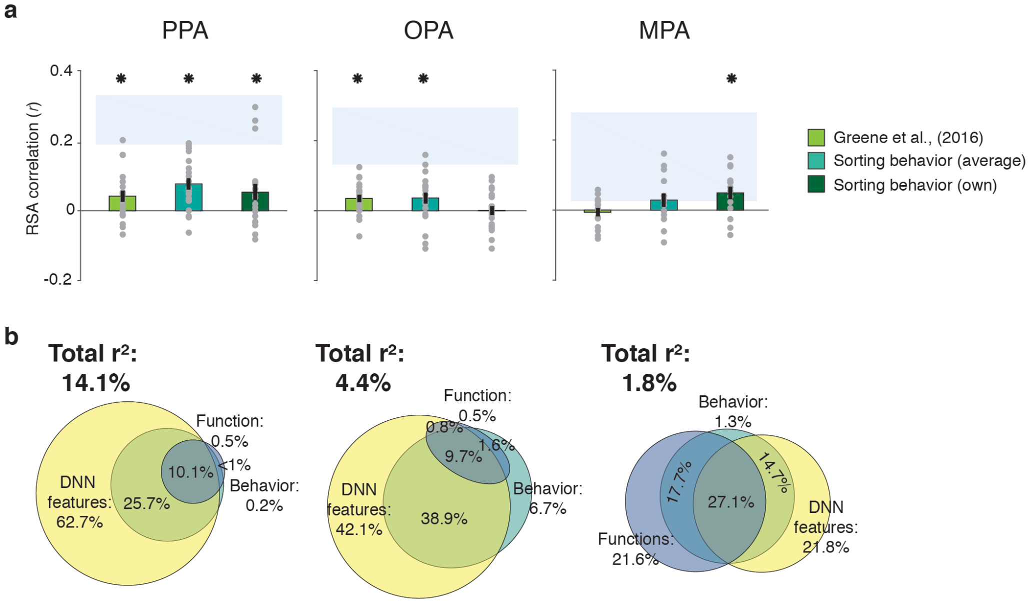

Figure 4 Correlations and variance partitioning of behavior and fMRI together. **A)** Correlations of behavioral categorization with fMRI response patterns in PPA, OPA and MPA. **B)** Euler diagram depicting the results of variance partitioning the fMRI responses in PPA, OPA and MPA for objects (red circle), DNN features (yellow circle) and average sorting behavior (green circle), indicating that the majority of behavioral variance is shared with the DNN features.

While these results support an important role for scene-selective regions in representing scene information that informs behavior, they also raise an intriguing question: what aspect of the behavior is reflected in these neural response patterns? To address this, we performed another variance partitioning analysis, now including the average multi-arrangement behavior as a predictor of the fMRI response patterns, in combination with the two models that correlated most strongly with this behavior, i.e. the DNN and function feature spaces. The purpose of this analysis was to determine how much variance in the neural responses each of the models *shared* with the behavior, and whether there was any behavioral variance in scene cortex that was not explained by our models. If the behaviorally relevant information in the fMRI responses is primarily of a functional nature, we would expect portions of the variance explained by behavior to be shared with the function feature space. Alternatively, if this variance reflects mainly DNN features (which did contribute to behavior; **Figure 2*F***), we would expect it to be shared primarily with the DNN model.

Consistent with this second hypothesis, the variance partitioning results indicated that in OPA and PPA, most of the behaviorally relevant information in the fMRI response patterns was shared with the DNN model (**Figure 4*B***). In PPA, the behavioral RDMs on average shared 25.7% variance with the DNN model, while a negligible portion was shared with the function model (less than 1%); indeed, nearly all variance shared between the function model and the behavior was also shared with the DNN model (10.1%). In OPA, a similar trend was observed, with behavior sharing 38.9% of the fMRI variance with the DNN model. In OPA, the DNN model also eclipsed nearly all variance that behavior shared with the function model (9.7% shared by behavior, functions and DNN features), leaving only 1.6% of variance shared exclusively by functions and behavior. In contrast, in MPA, behavioral variance was shared with either the DNN model or the function model to a similar degree (14.7% and 17.7%, respectively), with an additional 27.1% shared with both (note, however, again MPA's low explained variance overall).

In sum, these analyses suggest that while response patterns in PPA and OPA reflect behaviorally relevant information, this information aligns best with the DNN feature space, and does not reflect any unique contribution of functional information to behavior. While in MPA, the behaviorally relevant representations seem to partly reflect other information, the overall explained variance in MPA was again quite low, limiting interpretation of this result.

Relative model contributions to fMRI responses do not change with task manipulation

An important difference between the behavioral and the fMRI experiment was that participants had access to the entire stimulus set when performing the behavioral arrangements, which they could perform at their own pace, while they performed an orthogonal task in the fMRI scanner. Therefore, we reasoned that a possible explanation of the discrepancy between the brain and behavioral data could be a limited engagement of participants with the briefly presented scenes while in the scanner, resulting in only superficial encoding of the images in terms of basic visual features that are well captured by the DNN feature space, rather than functional or object features that might be more high-level.

To test this possible explanation, we ran Experiment 2 and collected another set of fMRI data (n = 8; four of these participants also participated in Experiment 1, allowing for comparison of tasks within individuals) using the exact same stimulation paradigm, but with a different task instruction. Specifically, instead of performing an unrelated fixation task, we instructed participants to covertly name the presented scene. Covert naming has been shown to facilitate stimulus processing within category-selective regions and to enhance semantic processing (Turennout et al. 2000; van Turennout et al. 2003). Moreover, before entering the scanner, participants were familiarized with all the individual scenes in the set, whereby they were explicitly asked to generate a name for each individual scene (see Methods). Together, these manipulations were intended to ensure that participants attended to the scenes and processed their content to a fuller extent than in Experiment 1.

Despite this task manipulation, Experiment 2 yielded similar results as Experiment 1 (**Figure 5*A***). Reflecting participant's enhanced engagement with the scenes when performing the covert naming task, overall model correlations were considerably higher than in Experiment 1, and now yielded significant correlations with the function model in both OPA and MPA (**Figure 5*B***). The direct test of reproducibility also yielded significant, and somewhat increased, correlations for PPA (r = 0.35 [0.26-0.55], p = 0.0001) and OPA (r = 0.27 [0.18-0.60], p = 0.039), but not in MPA (r = 0.10 [-0.07-0.28], p = 0.17).

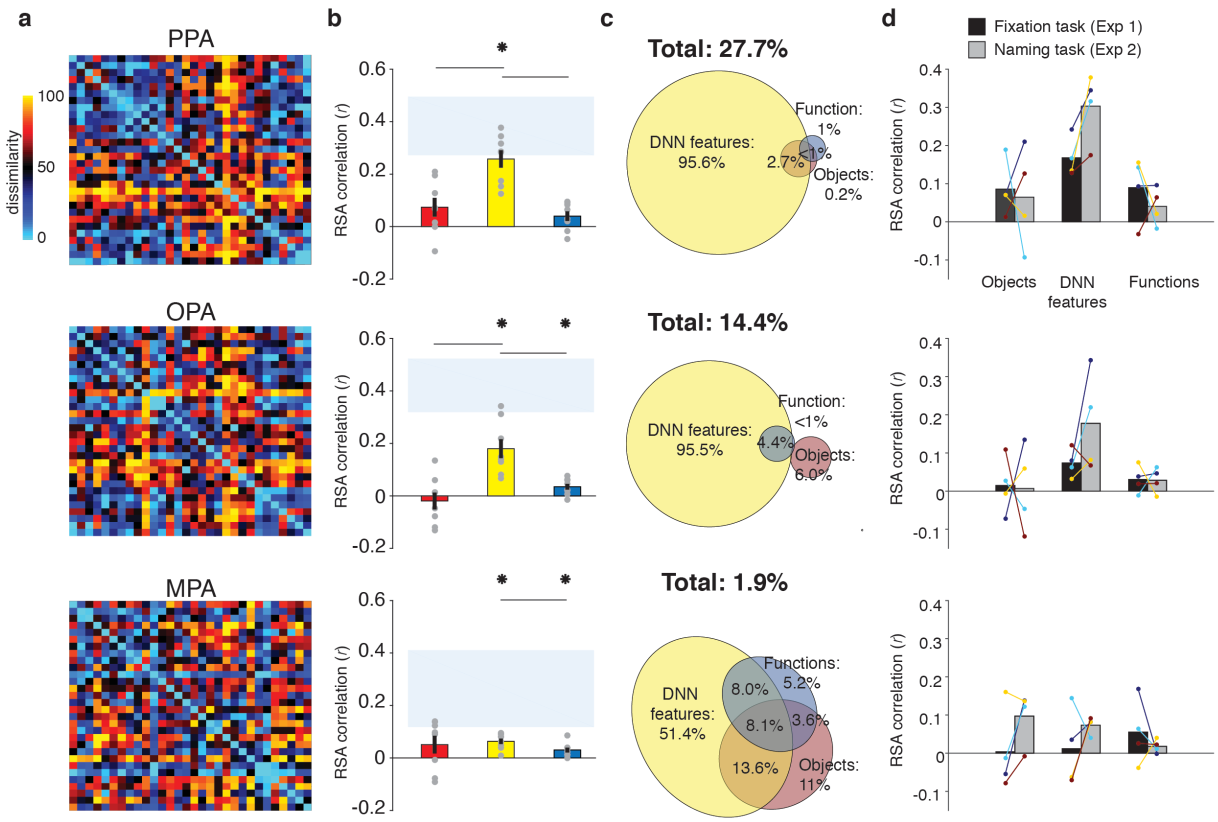

Figure 5 RDMs and model comparisons for Experiment 2 (n = 8, covert naming task). **A)** Average dissimilarity between categories in multi-voxel patterns measured in PPA, OPA and MPA (rank-ordered). **B)** Correlations between the ROIs in A) and the model RDMs. Statistics as in Figure 3. Note how in PPA, the visual feature model correlation comes close to the noise ceiling, suggesting that this is the main source of information driving neural representation in this ROI. **C)** Euler diagram depicting the variance partitioning results on the average dissimilarity in each ROI. **D)** Average (bars) and individual (dots/lines) within-participant (n = 4) comparison of fMRI-model correlations across the fixation and naming task (note that participants were presented with a different set of scenes in each task). Note how increased attention to the scenes due to the naming mainly enhances the correlation with visual features.

Importantly, in all three ROIs, the DNN model correlations were again significantly stronger than the function and object model correlations, which again contributed very little unique variance (**Figure 5*C***). Direct comparison of RDM correlations across the two tasks indicated that in PPA and OPA, the naming task resulted in increased correlations for the DNN model only (two-sided Wilcoxon ranksum test, PPA: p = 0.0048; OPA p = 0.0056), without any difference in correlations for the other models (all $p > 0.52$). In MPA, none of the model correlations differed across tasks (all $p > 0.21$). Increased correlation with the DNN model was present within the participants that participated in both experiments (n = 4; see Methods): in PPA and OPA, 4/4 and 3/4 participants showed an increased correlation, respectively, whereas no consistent patterns was observed for the other models and MPA (**Figure 5*D***).

In sum, the results of Experiment 2 indicate that the strong contribution of DNN features to scene representation in scene-selective cortex is not likely the result of limited engagement of participants with the scenes when viewed in the scanner. If anything, enhanced attention to the scenes under an explicit naming instruction resulted in even stronger representation of these features, without a clear increase in representation of functional or object feature spaces.

Visual feature coding in scene-selective cortex is not exclusive to high-level DNN layers

The DNN feature space was derived using a high-level layer (fc7) representation of an Image-Net pre-trained AlexNet computed from the large set of exemplars per scene category used in Greene et al., (2016). DNNs consist of multiple layers that capture possible transformations from pixels in the input image to a class label assigned in training. Given the strong performance of this high-level DNN feature space in explaining the fMRI responses in scene-selective cortex,it is important to determine whether this result was exclusive to higher DNN layers, and whether the task used for DNN training influences how well the features represented in individual layers explain responses in scene-selective cortex. To do so, we extracted the features for the specific scene exemplars presented in the fMRI scanners from all layers of two additional pre-trained DNNs, one trained on object labels, and another trained using scene labels (see Methods). Direct comparisons of the layer representations between these two DNNs (**Figure 6*A***) indicated that while both models have similar representations (as indicated by strong between-model RDM correlation overall; all layers $r > 0.6$), the similarity between models decreased with higher layers. This indicates that representations in higher DNN layers are more specific to the task they are trained on than the lower layers. Moreover, this suggests that higher layers of the scene-trained DNN could capture additional information that might be important for scene representation that is not captured by the object-trained DNN. To investigate this, we next evaluated the degree of correspondence between these new DNN layer representations that were specific to our scene exemplars and our original feature spaces derived from the Greene et al., (2016) database **(Figure 6*B*)**.

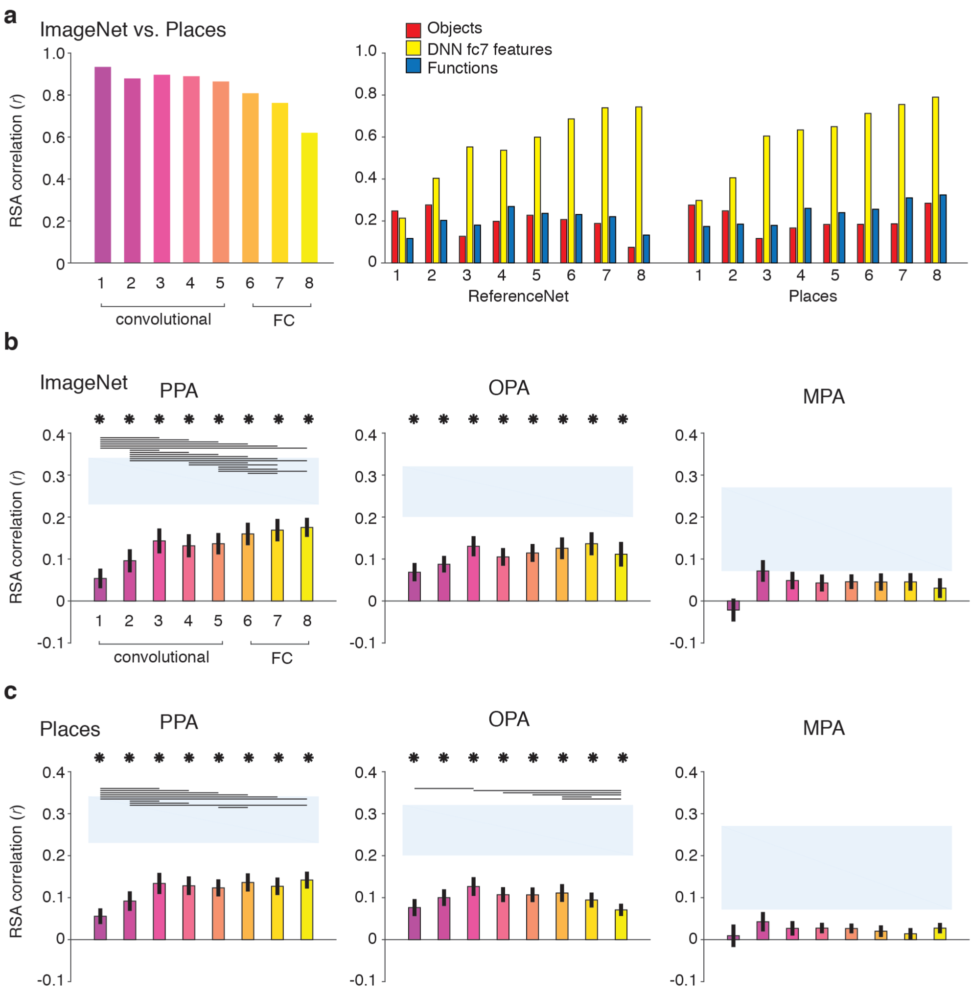

Figure 6 DNN layer and DNN training comparisons, showing layer-by-layer RDM correlations between **A)** an object- and a place-trained DNN; **B)** both DNNs and the a priori selected scene information models; **C)** the object-trained DNN and scene-selective ROIs; **D)** the scene-trained DNN and scene-selective ROIs. While the decreasing correlation between DNNs indicates stronger task-specificity of higher DNN layers, the original fc7 model correlates most strongly with high-level layers of both DNNs. Both DNNs correlate similarly with PPA and OPA, showing remarkable good performance of mid-level layers.

As expected, the original fc7 feature space model correlated most strongly with the new DNN layer representations, showing steadily increasing correlations with higher layers of both DNNs. By design, the object and functional feature spaces correlate minimally with higher layers of the object-trained DNN; however, for the scene-trained DNN, the function model correlated somewhat better with higher layers than lower layers, highlighting a potential overlap of functions with the scene labels that the scene-trained DNN was trained with, and again suggesting that the higher layers of the scene-trained DNN may capture additional scene information not represented in the object-trained DNN. We next tested whether this observation resulted in increased correlation of higher layers of the scene-trained DNN with fMRI responses in scene-selective cortex.

Layer-by-layer correlations of the object-trained (**Figure 6*C***) and the scene-trained DNN (**Figure 6*D***) with fMRI responses in PPA, OPA and MPA however did not indicate a strong evidence of a difference in DNN performance as a result of training. In PPA, both the object-trained and place-trained DNN showed increased correlation with higher DNN layers, consistent with previous work showing a hierarchical mapping of DNN layers to low vs. high-level visual cortex (Khaligh-Razavi and Kriegeskorte 2014; Güçlü and van Gerven 2015; Cichy et al. 2016). Note however that in our data, the slope of this increase is quite modest; while higher layers overall correlate better than layers 1 and 2, in both DNNs the correlation with layer 3 is not significantly different from the correlation of layers 7 and 8. In OPA, we in fact observed no evidence for increased performance with higher layers for the object-trained DNN; none of the pairwise tests survived multiple comparisons correction. For the scene-trained DNN, the OPA correlation significantly *decreased* rather than increased with higher layers, showing a peak correlation with layer 3. No significant correlations were found for any model layer with MPA.

These results suggest that despite a divergence in representation in high-level layers across differently-trained DNNs, their performance in predicting brain responses in scene-selective cortex is quite similar. While for PPA, higher layers perform significantly better than (very) low-level layers, mid-level layers already provide a relatively good correspondence with PPA activity. This result was even more pronounced for OPA where mid-level layers yielded the maximal correlations for both DNNs. Therefore, these results suggest that features represented in these scene-selective ROIs may actually be less 'high-level' than suggested by our a priori chosen, layer fc7 based DNN feature space.

Contributions of the functional feature space outside of scene-selective cortex

All our results so far indicate a dissociation between brain and behavioral assessments of the similarity structure of scenes. In the behavioral domain, functions have a large, independent contribution to scene categorization, but representational dissimilarity in scene-selective cortex is primarily driven by mid- and high-level visual features represented in a convolutional neural network, without an independent contribution of functions. Given this lack of correlation with the function model in the scene-selective cortex, we explored whether this information could be reflected in fMRI activity elsewhere in the brain by performing whole-brain searchlight analyses. Specifically, we extracted the multi-voxel patterns from spherical ROIs throughout each participant's entire volume and performed regression analyses including all three models (visual features, objects, functions) to extract the corresponding regression weights for each model. The resulting whole-brain searchlight maps were then fed into a to surface-based group analysis (see Methods) to identify clusters of positive regression weights indicating significant model contributions to brain representation of during viewing of real-world scenes.

The results of these analyses were entirely consistent with the ROI analyses: for the DNN feature space, significant searchlight clusters were found in PPA and OPA (**Figure 7*A***), but not MPA, whereas no significant clusters were found for the function model (**Figure 7*B***) in any of the scene-selective ROIs. (The object model yielded no positive clusters).

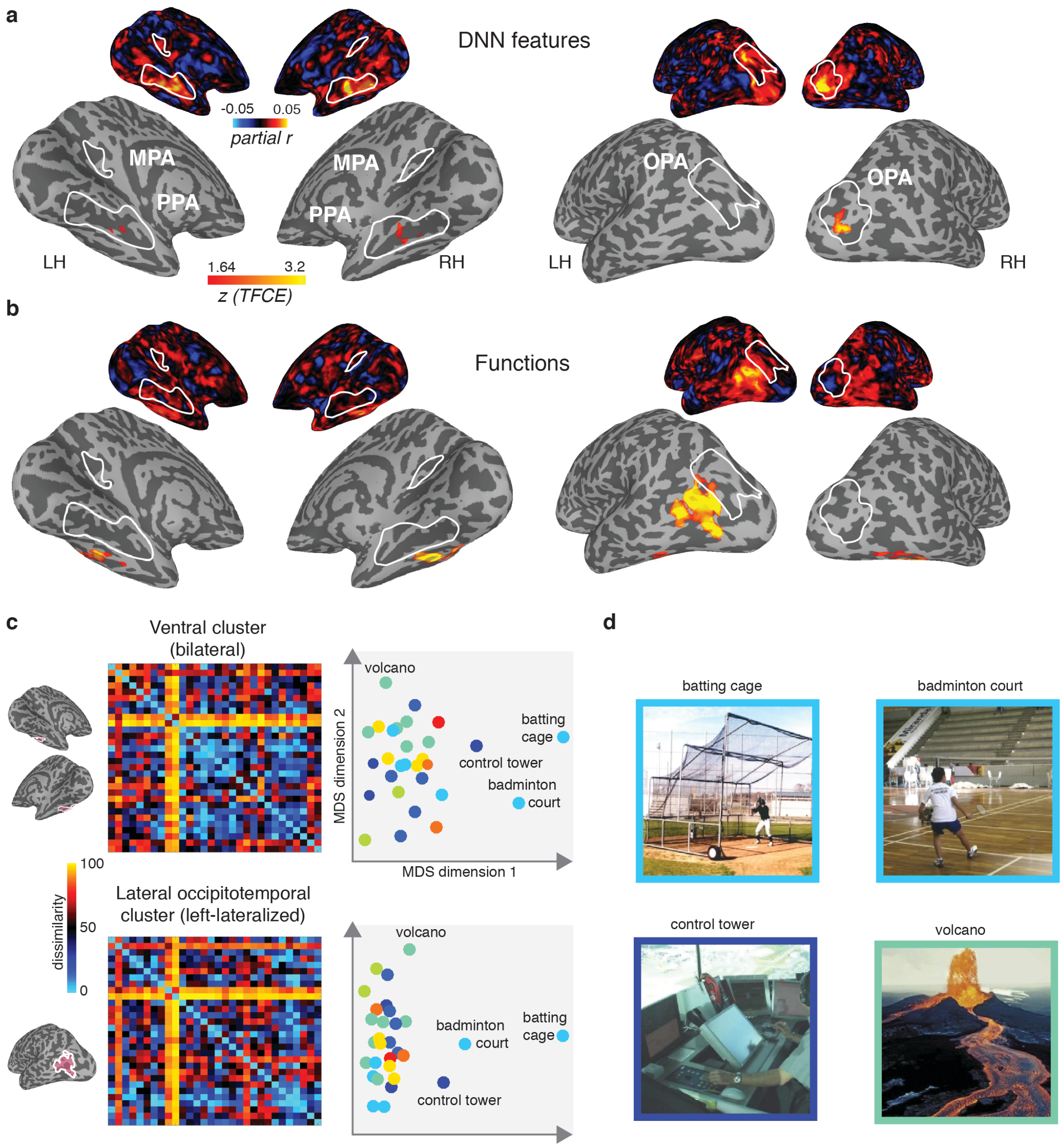

Figure 7 Searchlight results. Medial (left) and lateral (right) views of group-level searchlights for **A)** the DNN and **B)** function feature spaces, overlaid on surface reconstructions of both hemispheres of one participant. Each map was created by submitting the partial correlation maps for each model and hemisphere to one-sample tests against a mean of zero, cluster-corrected for multiple comparisons using Threshold-Free Cluster Enhancement (thresholded on z = 1.64, corresponding to one-sided $p < 0.05$). Unthresholded versions of the average partial correlation maps are inset above. Group-level ROIs PPA, OPA and MPA are highlighted in solid white lines. Consistent with the ROI analyses, the DNN feature model contributed uniquely to representation in PPA and OPA. The function model uniquely correlated with a bilateral ventral region, as well as a left-lateralized region overlapping with the middle temporal and occipital gyri. **C)** RDM and MDS plots based on the MVPA patterns in the searchlight clusters showing a significant contribution of the functional feature space in B. RDM rows are ordered as in Figure 1B and category color coding in the MDS plots is as in Figure 1C. **D)** Illustrative exemplars of the four categories that were most dissimilar from other categories within the searchlight-derived clusters depicted in B.

However, two clusters were identified for the function model outside of scene-selective cortex: a bilateral cluster on the ventral surface, lateral to PPA, overlapping with the fusiform and temporal lateral gyri, as well as a unilateral cluster on the left lateral surface, located adjacent to, but more ventral than, OPA, overlapping the posterior middle and inferior temporal gyrus.

To better understand how representational dissimilarity in these clusters related to the functional feature space, we extracted the average RDM from each searchlight cluster and inspected which scene categories were grouped together in these ROIs. Visual inspection of the RDM and MDS plots of the RDMs (**Figure 7*C***) indicates that in both the bilateral ventral and left-lateralized searchlight clusters, there is some grouping by category according to the function feature space (indicated by grouping by color in the MDS plot). However, it is also clear that the representational space in these ROIs does not exactly map onto the functional feature space in Figure 1C. Specifically, a few categories clearly 'stand out' with respect to the other categories, as indicated by a large average distance relative to the remainder of the stimulus set. Most of the scene categories that were strongly separated from the remaining categories all contained scene exemplars depicting humans that performed actions (see **Figure 7*D*)**, although it is worth noting that the fourth most distinct category, 'volcano', did not contain humans in its scene exemplars but may be characterized by implied motion. These post-hoc observations suggest that (parts of) the searchlight correlation with the functional feature space may be due to the presence of human-, body- and/or motion selective voxels in these searchlight clusters.

In sum, the searchlight analyses indicate that the maximum contributions of the DNN model were located in scene-selective cortex, while some aspects of the functional feature space may be reflected in regions outside of scene-selective cortex.

Discussion

We assessed the contribution of three feature spaces previously implicated to be important for scene understanding to neural representations of scenes in the human brain. First, we confirmed earlier reports that functions strongly contribute to scene categorization by replicating the results of Greene et al., (2016), now using a multi-arrangement task. Second, however, we found that brain responses to visual scenes in scene-selective regions were best explained by the DNN feature space, with no discernible unique contribution of the functional features. Although parts of variance in the behavioral categorization were captured by the DNN feature space - and this part of the behavior was reflected in the scene-selective cortex - there are clearly aspects of scene categorization behavior that were not reflected in the activity of these regions. Collectively, these results thus reveal a striking dissociation between the information that is most important for behavioral scene categorization and the information that best describes representational dissimilarity of fMRI responses in regions of cortex that are thought to support scene recognition. Below, we discuss two potential explanations for this dissociation.

First, one possibility is that functions are represented outside of scene-selective cortex. Our searchlight analysis indeed revealed clusters of correlations with the function model in bilateral ventral and left lateral occipito-temporal cortex. Visual inspection of these maps suggests that these clusters potentially overlap with known face- and body-selective regions such as the Fusiform Face (FFA; Kanwisher et al. 1997) and Fusiform Body (FBA; Peelen and Downing 2007) areas on ventral surface, as well as the Extrastriate Body Area (EBA; Downing 2001) on the lateral surface. This lateral cluster could possibly include motion-selective (Zeki et al. 1991; Tootell et al. 1995) and tool-selective (Martin et al. 1996) regions as well. Our results further indicated that these searchlight clusters contained distinct representations of scenes that contained *acting* bodies, and may therefore partially overlap with regions important for action observation (e.g., Hafri et al. 2017). Lateral occipital-temporal cortex in particular is thought to support action observation by containing 'representations which capture perceptual, semantic and motor knowledge of how actions change the state of the world' (Lingnau & Downing, 2015). While our searchlight results suggest a possible contribution of these non-scene-selective regions to scene understanding, more research is needed to address how the functional feature space as defined here relates to the action observation network, and to what extent the correlations with functional features can be explained by bottom-up coding of bodies and motion versus more abstract action-associated features.

The second possible explanation for the dissociation between brain and behavioral data is that the task that participants performed during fMRI did not engage the same mental processes that participants employed during the two behavioral tasks we investigated. Specifically, the behavioral tasks required participants to directly compare simultaneously presented scenes, while we employed a 'standard' fixation task in the scanner to prevent biasing our participants towards one of our feature spaces. Therefore, one possibility is that functional features only become relevant for scene categorization when participants are engaged in a *contrastive* task, i.e. explicitly comparing two scene exemplars side-by-side (as in Greene et al., 2016) or within the context of the entire stimulus set being present on the screen (as in our multi-arrangement paradigm). Thus, the fMRI results might change with an explicit contrastive task in which multiple stimuli are presented at the same time, or perhaps with a task that explicitly requires participants to consider functional aspects of the scenes. Although we investigated one possible influence of task in the scanner by using a covert naming task in Experiment 2, resulting in deeper and more conceptual processing, it did not result in a clear increase in the correlation with the behaviorally relevant function model in scene-selective cortex. However, the evidence for task effects on fMRI responses in category-selective cortex is somewhat mixed: Task differences have been reported to affect multi-voxel pattern activity in both object-selective (Harel et al. 2014) and scene-selective cortex (Lowe et al. 2016), but other studies suggest that task has a minimal influence on representation in ventral stream regions, instead being reflected in fronto-parietal networks (Erez and Duncan 2015; Bracci et al. 2017; Bugatus et al. 2017). Overall, our findings suggest that not all the information that contributes to scene categorization is reflected in scene-selective cortex activity ‘by default’, and that explicit task requirements may be necessary in order for this information to emerge in the neural activation patterns in these regions of cortex.

Importantly, the two explanations outlined above are not mutually exclusive. For example, it is possible that a task instruction to explicitly label the scenes with potential actions will activate components of both the action observation network (outside scene-selective cortex) as well as task-dependent processes within scene-selective cortex. Furthermore, given reports of potentially separate scene-selective networks for memory versus perception (Baldassano et al. 2016; Silson et al. 2016), it is likely that differences in mnemonic demands between tasks may have an important influence on scene-selective cortex activity. Indeed, memory-based navigation or place recognition tasks (Epstein et al. 2007; Marchette et al. 2014) have been shown to more strongly engage the medial parietal cortex and MPA. In contrast, our observed correlation with DNN features seems to support a primary role for PPA and OPA in bottom-up visual scene analysis, and fits well with the growing literature showing correspondences between extrastriate cortex activity and DNN features (Cadieu et al. 2014; Khaligh-Razavi and Kriegeskorte 2014; Güçlü and van Gerven 2015; Cichy et al. 2016; Horikawa and Kamitani 2017). Our analyses further showed that DNN correlations with scene-selective cortex were not exclusive to higher DNN layers, but emerged at earlier layers, independent of DNN training (i.e. object versus scene classification), suggesting that the neural representation in PPA/OPA may be driven more by basic visual features than by semantic information (Watson et al. 2017).

At a behavioral level, however, our current results suggest that when participants perform scene categorization, either explicitly (Greene et al. 2016) or within a multi-arrangement paradigm (Kriegeskorte and Mur 2012), they incorporate information that is not reflected in either the DNNs or in PPA and OPA. Our results thus highlight a significant gap between the real-world information that is captured both in scene–selective cortex and current generations of deep neural networks, and the information that drives human understanding of visual environments. Visual environments are highly multidimensional, and scene understanding encompasses many behavioral goals, including not just visual object or scene recognition, but also navigation and action planning (Malcolm et al. 2016). While visual/DNN features likely feed into multiple of these goals - for example, by signaling navigable paths in the environment (Bonner and Epstein 2017), or landmark suitability (Troiani et al. 2014) - it is probably not appropriate to think about the neural representations relevant to all these different behavioral goals as being contained within one single brain region of even a single network of brain regions. Ultimately, unraveling the neural coding of scene information will require careful manipulations of both multiple tasks and multiple scene feature spaces, as well as a potential expansion of our focus on a broader set of regions than those characterized by the presence of scene-selectivity.

Summary and conclusion

We successfully disentangled the type of information represented in scene-selective cortex: out of three behaviorally relevant feature models, only one provided a robust correlation with activity in scene-selective cortex. This model was derived from deep neural network features in state-of-the-art computer vision algorithms of object and scene recognition. Intriguingly, however, the DNN model was not the best model to explain scene categorization behavior, which was strongly driven by functional representations. This highlights both a limitation of DNNs in explaining scene understanding, as well as a potentially more distributed representation of scene information in the human brain beyond scene-selective cortex.

Methods

Participants. Twenty healthy participants (13 female, mean age 25.4 yrs, SD = 4.6) completed the first fMRI experiment and subsequent behavioral experiment. Four of these participants (3 female, mean age 24.3 yrs, SD = 4.6) additionally participated in the second fMRI experiment, as well as four new participants (2 female, mean age 25 yrs, SD = 1.6), yielding a total of eight participants. All participants had normal or corrected-to-normal vision and gave written consent. The National Institutes of Health Institutional Review Board approved the consent and protocol.

MRI acquisition. Participants were scanned on a research-dedicated Siemens 7T Magnetom scanner in the Clinical Research Center on the National Institutes of Health Campus (Bethesda, MD). Partial T2*-weighted functional image volumes were acquired using a gradient echo planar imaging (EPI) sequence with a 32-channel head coil (47 slices; 1.6 x 1.6 x 1.6 mm; 10% interslice gap; TR, 2s; TE, 27 ms; matrix size, 126 x 126; FOV, 192 mm). Oblique slices were oriented approximately parallel to the base of the temporal lobe and were positioned such that they covered the occipital, temporal, parietal cortices, and as much as possible of frontal cortex. After the functional imaging runs, standard MPRAGE (magnetization-prepared rapid-acquisition gradient echo) and corresponding GE-PD (gradient echo–proton density) images were acquired, and the MPRAGE images were then normalized by the GE-PD images for use as a high-resolution anatomical image for the following fMRI data analysis (Van de Moortele, 2009).

Stimuli & models. Experimental stimuli consisted of color photographs of real-world scenes (256 x 256 pixels) from 30 difference scene categories that were selected from a larger database described in (Greene et al. 2016). These scene categories were picked using an iterative sampling procedure that minimized the correlation between the categories across three models of scene information: functions, object labels and DNN features, with the additional constraint that the final stimulus set should be have equal portions of categories from indoor, outdoor man-made and outdoor natural scenes, which is the largest superordinate distinction present in the largest scene-database that is publicly available, the SUN database (Xiao et al. 2014). As obtaining a guaranteed minimum was impractical, we adopted a variant of the odds algorithm (Bruss 2000) as our stopping rule. Specifically, we created 10,000 sets of 30 categories and measured the correlations between functional, object, and DNN RDMs (distance metric: Spearman's *rho*), noting the minimal value from the set. We persisted in this procedure until we observed a set with lower inter-feature correlations than was observed in the initial 10,000. From each scene category, 8 exemplars were randomly selected and divided across two separate stimulus sets of 4 exemplars for each scene category. Stimulus sets were assigned randomly to individual participants (Experiment 1: stimulus set 1, n = 10; stimulus set 2, n = 10; Experiment 2, stimulus set 1, n = 5; stimulus set 2, n = 3). Participants from Experiment 2 that had also participated in Experiment 1 were presented with the other stimulus set than the one they saw in Experiment 1.

fMRI procedure. Participants were scanned while viewing the stimuli on a back-projected screen through a rear-view mirror that was mounted on the head coil. Stimuli were presented at a resolution of 800 x 600 pixels such that stimuli subtended ~10 x 10 degrees of visual angle. Individual scenes were presented in an event-related design for a duration of 500 ms, separated by a 6s interval. Throughout the experimental run, a small fixation cross (< 0.5 degrees) was presented in the center of the screen. In Experiment 1, participants performed a task on the central fixation cross that was unrelated to the scenes. Specifically, simultaneous with the presentation of each scene, either the vertical or horizontal arm of the fixation cross became slightly elongated and participants indicated which arm was longer by pressing one of two buttons indicated on a hand-held button box. Both arms changed equally often within a given run and arm changes were randomly assigned to individual scenes. In Experiment 2, the fixation cross had a constant size, and participants were instructed to covertly name the scene whilst simultaneously pressing one button on the button box. To assure that participants were able to generate a name for each scene, they were first familiarized with the stimuli. Specifically, prior to scanning, participants were presented with all scenes in the set in randomized order on a laptop in the console room. Using a self-paced procedure, each scene was presented in isolation on the screen accompanied by the question 'How would you name this scene?'. The participants were asked to type one or two words to describe the scene; as they typed, their answer appeared under the question, and they were able to correct mistakes using backspace. After typing the self-generated name, participants hit enter and the next scene would appear until all 120 scenes had been seen by the participant. This procedure took about ~10 minutes.

In both Experiment 1 and 2, participants completed 8 experimental runs of 6.4 minutes each (192 TRs per run); one participant from Experiment 1 only completed 7 runs due to time constraints. Each run started and ended with a 12s fixation period. Each run contained 2 exemplar presentations per scene category. Individual exemplars were balanced across runs such that all stimuli were presented after two consecutive runs, yielding 4 presentations per exemplar in total. Exemplars were randomized across participants such that each participant always saw the same two exemplars within an individual run; however the particular combination was determined anew for each individual participant and scene category. Stimulus order was randomized independently for each run. Stimuli were presented using PsychoPy v1.83.01 (Peirce 2007).

Functional localizers. Participants additionally completed four independent functional block-design runs (6.9 minutes, 208 TRs) that were used to localize scene-selective regions of interest (ROIs). Per block, twenty gray-scale images (300 x 300 pixels) were presented from one of eight different categories: faces, man-made and natural objects, buildings, and four different scene types (man-made open, man-made closed, natural open, natural closed; Kravitz et al., 2011) while participants performed a one-back repetition-detection task. Stimuli were presented on a gray background for 500 ms duration, separated by 300 ms gaps, for blocks of 16s duration, separated by 8s fixation periods. Categories were counterbalanced both within runs (such that each category occurred twice within a run in a mirror-balanced sequence) and across runs (such that each category was equidistantly spaced in time relative to each other category across all four runs). Two localizer runs were presented after the first four experimental runs and two after the eight experimental runs were completed but prior to the T1 acquisition. For four participants, only two localizer runs were collected due to time constraints.

Behavioral experiment. On a separate day following the MRI data acquisition, participants performed a behavioral multi-arrangement experiment. In a behavioral testing room, participants were seated in front of a desktop computer with a flat screen monitor (size?) on which all 120 stimuli that the participant had previously seen in the scanner were displayed as small thumbnails around a white circular arena. A mouse-click on an individual thumbnail displayed a larger version of that stimulus in the upper right corner. Participants were instructed to arrange the thumbnails within the white circle in such a way that the arrangement would reflect 'how similar the scenes are, whatever that means to you', by means of dragging and dropping the individual exemplar thumbnails. We purposely avoided provided specific instructions in order to not bias participants towards using either functions, objects or visual features to determine scene similarity. Participants were instructed to perform the task at their own pace; if the task took longer than 1hr, participants were encouraged to finish the experiment (almost all participants took less time, averaging a total experiment duration of ~45 mins). Stimuli were presented using the single-arrangement MATLAB code provided in (Kriegeskorte & Mur, 2012). To obtain some insight in the sorting strategies used by participants, they were asked (after completing the experiment) to take a few minutes to describe how they organized the scenes, using a blank sheet of paper and a pen, using words, bullet-points or drawings.

Behavioral data analysis. Behavioral representational dissimilarity matrices (RDMs) were constructed for each individual participant by computing the pairwise squared on-screen distances between the arranged thumbnails and averaging the obtained distances across the exemplars within each category. The relatedness of the models and the behavioral data was determined in the same manner as for the fMRI analysis, i.e. by computing both individual model correlations and unique and shared variance across models via hierarchical regression.

fMRI preprocessing. Data were analyzed using AFNI software (https://afni.nimh.nih.gov). Before statistical analysis, the functional scans were slice-time corrected and all the images for each participant were motion corrected to the first image of their first task run after removal of the first and last six TRs from each run. After motion correction, the localizer runs were smoothed with a 5mm full-width at half-maximum Gaussian kernel; the even-related data was not smoothed.

fMRI statistical analysis: localizers. Bilateral ROIs were created for each participant individually based on the localizer runs by conducting a standard general linear model implemented in

AFNI. A response model was built by convolving a standard gamma function with a 16s square wave for each condition and compared against the activation time courses using Generalized Least Squares (GLSQ) regression. Motion parameters and four polynomials accounting for slow drifts were included as regressors of no interest. To derive the response magnitude per category, t-tests were performed between the category-specific beta estimates and baseline. Scene-selective ROIs were generated by thresholding the statistical parametric maps resulting from contrasting scenes > faces at $p < 0.0001$ (uncorrected). Only contiguous clusters of voxels (>25) exceeding this threshold were then inspected to define scene-selective ROIs consistent with previously published work (Epstein 2005). For participants in which clusters could not be disambiguated, the threshold was raised until individual clusters were clearly identifiable. While PPA and OPA were identified in all participants for both Experiment 1 and 2, MPA/RSC was detected in only 14 out 20 participants in Experiment 1, and all analyses for this ROI in Experiment 1 are thus based on this subset of participants.

fMRI statistical analysis: event-related data. Each event-related run was deconvolved independently using the standard GLSQ regression model in AFNI. The regression model included a separate regressor for each of the 30 scene categories as well as motion parameters and four polynomials to account for slow drifts in the signal. The resulting beta-estimates were then used to compute representational dissimilarity matrices (RDMs; (Kriegeskorte et al. 2008) based on the multi-voxel patterns extracted from individual ROIs. Specifically, we computed pairwise cross-validated Mahalanobis distances between each of the scene 30 categories following the approach in (Walther et al. 2016). First, multi-variate noise normalization was applied by normalizing the beta-estimates by the covariance matrix of the residual time-courses between voxels within the ROI. Covariance matrices were regularized using shrinkage toward the diagonal matrix (Ledoit and Wolf 2004). Unlike univariate noise normalization, which normalizes each voxel's response by its own error term, multivariate noise normalization also takes into account the noise covariance between voxels, resulting in more reliable RDMs (Walther et al. 2016). After noise normalization, squared Euclidean distances were computed between individual runs using a leave-one-run-out procedure, resulting in cross-validated Mahalanobis distance estimates. Note that unlike correlation distance measures, cross-validated distances provide unbiased estimates of pattern dissimilarity on a ratio scale (Walther et al. 2016), thus providing a distance measure suitable for direct model comparisons.

Model comparisons: individual models. To test the relatedness of the three models of scene dissimilarity with the measured fMRI dissimilarity, the off-diagonal elements of each model RDM were correlated (Pearson's *r*) with the off-diagonal elements of the RDM of each fMRI ROI for each individual participant separately. Following (Nili et al. 2014), the significance of these correlations was determined using one-sided signed-rank tests against zero, while pairwise differences between models in terms of their correlation with fMRI dissimilarity were determined using two-sided signed-ranked tests. For each test, we report the sum of signed ranks for the number of observations W(*n*) and the corresponding p-value; for tests with n > 10 we also report the z-ratio approximation. The results were corrected for multiple comparisons (across both individual model correlations and pairwise comparisons) using FDR correction (Benjamini and Hochberg 1995) for each individual ROI separately. Noise ceilings were computed following (Walther et al. 2016): an upper bound was estimated by computing the correlation between each participant's individual RDM and the group-average RDM, while a lower bound was estimated by correlating each participant's RDM with the average RDM of the other participants (leave-one-out approach). The participant-averaged RDM was converted to rank order for visualization purposes only.

Model comparisons: partial correlations and variance partitioning. To determine the contribution of each individual model when considered in conjunction with the other models, we performed to additional types of analyses: partial correlations, in which each model was correlated (Pearsons *r*) while partialling out the other two models, as well as variation partitioning based on multiple linear regression. For the latter, the off-diagonal elements of each ROI RDM were assigned as the dependent variable, while the off-diagonal elements of the three model RDMs were entered as independent variables (predictors). To obtain unique and shared variance across the three models, 7 multiple regression analyses were run in total: one 'full' regression that included all three feature spaces as predictors; and six reduced models that included as predictors either combinations of two models in pairs (e.g., functions and objects), or including each model by itself. By comparing the explained variance (r^2) of a model used alone to the r^2 of that model in conjunction with another model, we can infer the amount of variance that is independently explained by that model, i.e. partition the variance (see also (Groen et al. 2012; Lescroart et al. 2015; Çukur et al. 2016; Greene et al. 2016) Hebart et al. (in press) for similar approaches).

Analogous to the individual model correlation analyses, partial correlations were calculated for each individual participant separately, and significance was determined using one-sided signed-rank tests across participants (FDR-corrected across all comparisons within a given ROI). To allow comparison with the results reported in (Greene et al. 2016), variance partitioning was performed on the participant-average RDMs. Similar results were found, however, when variance was partitioned for individual participant's RDMs and then averaged across participants. To visualize this information in an Euler diagram, we used the EulerAPE software (Micallef and Rodgers 2014).

Direct reproducibiilty test of representational structure in behavior and fMRI. To assess how well the obtained RDMs were reproducible in each measurement domain (behavior and fMRI), we compared the average RDMs obtained for the two separate stimulus sets. Since these two sets of stimuli were viewed by different participants (see above under 'Stimuli & models'), this comparison provides a strong test of generalizability, across both scene exemplars and across participant pools. Set-average RDMs were compared by computing inter-RDM correlations (Pearson's *r*) and 96% confidence intervals (CI) and statistically tested for reproducibility using a random permutation test based on 10.000 randomizations of the category labels.

Variance partitioning of fMRI based on models and behavior. Using the same approach as in the previous section, a second set of regression analyses was performed to determine the degree of shared variance between the behavior on the one hand, and the functions and visual features on the other hand, in terms of the fMRI response pattern dissimilarity. The Euler diagrams were derived using the group-average RDMs, taking the average result of the multi-arrangement task of these participants as the behavioral input into the analysis.

DNN comparisons To investigate the influence of DNN layer and training images on the corresponding visual features and subsequent relations with activity in scene-selective cortex, we derived two new sets of RDMs by passing our scene stimuli through two pre-trained, 8-layer AlexNet (Alex et al. 2012) architecture networks: 1) a 1000-object label ImageNet-trained (Deng et al. 2009) network implemented in Caffe (Jia et al. 2014) ('ReferenceNet') and 2) a 250-scene label Places-trained network ("Places") (Zhou et al. 2014). By extracting the node activations from each layer, we computed pairwise dissimilarity (1 - Pearson correlation) resulting in one RDM per layer and per model. These RDMs were then each correlated with the fMRI RDMs from each participant in PPA, OPA and MPA. These analyses were performed on the combined data of Experiment 1 and 2; RDMs for participants that participated in both Experiments (n = 4) were averaged prior to group-level analyses.

Searchlight analyses. To test the relatedness of functions, objects and visual features with fMRI activity recorded outside scene-selective ROIs, we conducted whole-brain searchlight analyses. RDMs were computed in the same manner as for the ROI analysis, i.e. computing cross-validated Mahalanobis distances based on multivariate noise-normalized multi-voxel patterns, but now within spherical ROIs of 3 voxel diameter (i.e. 123 voxels/searchlight). Analogous to the ROI analyses, we computed partial correlations of each feature space, correcting for the contributions of the remaining two models. These partial correlation coefficients were assigned to the center voxel of each searchlight, resulting in one whole-volume map per model. Partial correlation maps were computed for in each participant separately in their native volume space. To allow comparison at the group level, individual participant maps were first aligned to their own high-resolution anatomical scan and then to surface reconstructions of the grey and white matter boundaries created from these high-resolution scans using the Freesurfer (http://surfer.nmr.mgh.harvard.edu/) 5.3 autorecon script using SUMA (Surface Mapping with AFNI) software (https://afni.nimh.nih.gov/Suma). The surface images for each participant were then smoothed with a Gaussian 10mm FWHM filter in surface coordinate units using the SurfSmooth function with the HEAT_07 smoothing method.

Group-level significance was determined by submitting these surface maps to node-wise one-sample t-tests in conjunction with Threshold Free Cluster Enhancement (Smith and Nichols 2009) through Monte Carlo simulations using the algorithm implemented in the CoSMoMVPA toolbox (Oosterhof et al. 2016), which performs group-level comparisons using sign-based permutation testing (n = 10,000) to correct for multiple comparisons. To increase power, the data of Experiment 1 and 2 were combined; coefficient maps for participants that participated in both Experiments (n = 4) were averaged prior to proceeding to group-level analyses.

Acknowledgements

This work was supported by the Intramural Research Program (ZIAMH002909) of the National Institutes of Health – National Institute of Mental Health Clinical Study Protocol 93-M-0170, NCT00001360. IIAG was also supported by a Rubicon Fellowship from the Netherlands Organization for Scientific Research (NWO).

References

Aguirre GK, Zarahn E, D'Esposito M. 1998. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. Neuron. 21:373–383.

Alex K, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. Neural Inf Process Syst. 1097–1105.

Baldassano C, Esteva A, Fei-Fei L, Beck DM. 2016. Two distinct scene processing networks connecting vision and memory. eNeuro. 10.1523:1–14.

Bar M, Aminoff E. 2003. Cortical analysis of visual context. Neuron. 38:347–358.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. J R Stat Soc B. 57:289–300.

Bonner MF, Epstein RA. 2017. Coding of navigational affordances in the human visual system. Proc Natl Acad Sci. 201618228.

Bracci S, Daniels N, Op de Beeck H. 2017. Task Context Overrules Object- and Category-Related Representational Content in the Human Parietal Cortex. Cereb Cortex. 310–321.

Bruss FT. 2000. Sum the odds to one and stop. Ann Probab. 28:1384–1391.

Bugatus L, Weiner KS, Grill-Spector K. 2017. Task alters category representations in prefrontal but not high-level visual cortex. Neuroimage. 155:437–449.

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ. 2014. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. PLoS Comput Biol. 10.

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep. 6:1–35.

Çukur T, Huth AG, Nishimoto S, Gallant JL. 2016. Functional Subdomains within Scene-Selective Cortex: Parahippocampal Place Area, Retrosplenial Complex, and Occipital Place Area. J Neurosci. 36:10257–10273.

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conf Comput Vis Pattern Recognit. 248–255.

Dilks DD, Julian JB, Paunov AM, Kanwisher N. 2013. The occipital place area is causally and selectively involved in scene perception. J Neurosci. 33:1331–6a.

Downing PE. 2001. A Cortical Area Selective for Visual Processing of the Human Body. Science (80-). 293:2470–2473.

Epstein R. 2005. The cortical basis of visual scene processing. Vis cogn. 12:954–978.

Epstein RA. 2014. Neural systems for visual scene recognition. In: Bar M,, Kveraga K, editors. Scene Vision. Cambridge, MA: MIT Press. p. 105–134.

Epstein RA, Parker WE, Feiler AM. 2007. Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. J Neurosci. 27:6141–6149.

Epstein R, Kanwisher N. 1998. A cortical representation of the local visual environment. Nature. 392:598–601.

Erez Y, Duncan J. 2015. Discrimination of Visual Categories Based on Behavioral Relevance in Widespread Regions of Frontoparietal Cortex. J Neurosci. 35:12383–12393.

Greene MR, Baldassano C, Esteva A, Beck DM. 2016. Visual Scenes Are Categorized by Function. J Exp Psychol Gen. 145:82–94.

Groen IIA, Ghebreab S, Lamme VAF, Scholte HS. 2012. Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. PLoS Comput Biol. 8:e1002726.

Groen IIA, Silson EH, Baker CI. 2017. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. Philos Trans R Soc B. 372:1–11.

Güçlü U, van Gerven MAJ. 2015. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J Neurosci. 35:10005–10014.

Hafri A, Trueswell JC, Epstein RA. 2017. Neural Representations of Observed Actions Generalize across Static and Dynamic Visual Input. J Neurosci. 37:3056–3071.

Harel A, Kravitz DJ, Baker CI. 2014. Task context impacts visual object processing differentially across the cortex. Proc Natl Acad Sci. 962–971.

Hasson U, Levy I, Behrmann M, Hendler T, Malach R. 2002. Eccentricity bias as an organizing principle for human high-order object areas. Neuron. 34:479–490.

Horikawa T, Kamitani Y. 2017. Generic decoding of seen and imagined objects using hierarchical visual features. Nat Commun. 8:15037.

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia. MM '14. New York, NY, USA: ACM. p. 675–678.

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci. 17:4302–4311.

Khaligh-Razavi SM, Kriegeskorte N. 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol. 10.

Kravitz DJ, Peng CS, Baker CI. 2011. Real-world scene representations in high-level visual cortex: it's the spaces more than the places. J Neurosci. 31:7322–7333.

Kriegeskorte N, Mur M. 2012. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. Front Psychol. 3:1–13.

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci. 2:4.

Ledoit O, Wolf M. 2004. Honey, I Shrunk the Sample Covariance Matrix. J Portf Manag. 30:110–119.

Lescroart MD, Stansbury DE, Gallant JL. 2015. Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front Comput Neurosci. 9:135.

Lowe MX, Gallivan JP, Ferber S, Cant JS. 2016. Feature diagnosticity and task context shape activity in human scene-selective cortex. Neuroimage. 125:681–692.

Malcolm GL, Groen IIA, Baker CI. 2016. Making sense of real-world scenes. Trends Cogn Sci. 20:843–856.

Marchette SA, Vass LK, Ryan J, Epstein RA. 2014. Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. Nat Neurosci. 17:1598–1605.

Martin A, Wiggs CL, Underleider LG, Haxby J V. 1996. Neural correlates of category-specific knowledge. Nature. 379:649–652.

Micallef L, Rodgers P. 2014. euler APE: Drawing area-proportional 3-Venn diagrams using ellipses. PLoS One. 9.

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. 2014. A Toolbox for Representational Similarity Analysis. PLoS Comput Biol. 10.

Oliva A, Torralba A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J Comput Vis. 42:145–175.

Oosterhof NN, Connolly AC, Haxby J V. 2016. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab / GNU Octave. Front Neuroinform. 10:1–27.

Park S, Brady TF, Greene MR, Oliva A. 2011. Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. J Neurosci. 31:1333–1340.

Peelen M V., Downing PE. 2007. The neural basis of visual body perception. Nat Rev Neurosci. 8:636–648.

Peirce JW. 2007. PsychoPy-Psychophysics software in Python. J Neurosci Methods. 162:8–13.

Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RBH. 2011. The “parahippocampal place area” responds preferentially to high spatial frequencies in humans and monkeys. PLoS Biol. 9:e1000608.

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. 2013. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks.

Silson EH, Steel AD, Baker CI. 2016. Scene selectivity and retinotopy in medial parietal cortex. Front Hum Neurosci. 10:1–17.

Smith SM, Nichols TE. 2009. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage. 44:83–98.

Tootell RBH, Reppas JB, Kwong KK, Rosen BR, Belliveaul JW, Malach R. 1995. Functional Analysis of Human MT and Related Visual Cortical Areas Using Magnetic Resonance Imaging. J Neurosci. 15:3215–3230.

Torralba A, Oliva A. 2003. Statistics of natural image categories. Netw Comput Neural Syst. 14:391–412.

Troiani V, Stigliani A, Smith ME, Epstein RA. 2014. Multiple object properties drive scene-selective regions. Cereb Cortex. 24:883–897.

Turennout M Van, Ellmore T, Martin A. 2000. Long-lasting cortical plasticity in the object naming system. Nat Neurosci. 3:1329–1335.

van Turennout M, Bielamowicz L, Martin A. 2003. Modulation of Neural Activity during Object Naming: Effects of Time and Practice. Cereb Cortex. 13:381–391.

Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage. 137:188–200.

Walther DB, Caddigan E, Fei-Fei L, Beck DM. 2009. Natural scene categories revealed in distributed patterns of activity in the human brain. J Neurosci. 29:10573–10581.

Watson DM, Andrews TJ, Hartley T. 2017. A data driven approach to understanding the organization of high-level visual cortex. Sci Rep. 7:3596.

Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A. 2014. SUN Database: Exploring a Large Collection of Scene Categories. Int J Comput Vis.

Zeki S, Watson JDG, Lueck CJ, Friston KJ, Kennard C, Frackowiak RSJ. 1991. A direct demonstration of functional specialization in human visual cortex. J Neurosci. 11:641–649.

Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. 2014. Learning Deep Features for Scene Recognition using Places Database. Adv Neural Inf Process Syst 27. 487–495.