1   **Distinct contributions of functional and deep neural network features to**

2   **representational similarity of scenes in human brain and behavior**

3

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

5   Diane M.[5], Baker, Chris I.[1]

6

7   [1] National Institutes of Health, Laboratory of Brain and Cognition, Bethesda, MD
8   [2] Bates College, Neuroscience Program, Lewiston, ME
9   [3] Princeton University, Princeton Neuroscience Institute, Princeton, NJ
10   [4] Stanford University, Stanford Vision Lab, Stanford, CA
11   [5] University of Illinois, Department of Psychology and Beckman Institute, Urbana-
12   Champaign, IL
13

14   Corresponding author:

15   Iris I.A. Groen
16   10 Center Drive
17   Building 10, Room 4C108, MSC 1240
18   Besthesda, MD 20892
19
20   Phone: 301-435-8905
21   E-mail: iris.groen@nih.gov
22

Abstract

Inherent correlations between visual and semantic features in real-world scenes make it difficult to determine how different scene properties contribute to neural representations. Here, we assessed the contributions of multiple properties to scene representation by partitioning the variance explained in human behavioral and brain measurements by three feature models whose inter-correlations were minimized *a priori* through stimulus preselection. Behavioral assessments of scene similarity reflected unique contributions from a functional feature model indicating potential actions in scenes as well as high-level visual features from a deep neural network (DNN). In contrast, similarity of cortical responses in scene-selective areas was uniquely explained by mid- and high-level DNN features only, while an object label model did not contribute uniquely to either domain. The striking dissociation between functional and DNN features in their contribution to behavioral and brain representations of scenes indicates that scene-selective cortex represents 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 functional magnetic resonance imaging (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, the presence of multiple candidate models calls for careful comparison of the contribution of each type of information to scene representation within a single study. 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? Second, 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

63 and Torralba 2001; Torralba and Oliva 2003). This makes it problematic to explain neural

64 representations of scenes based on just one of these properties (Walther et al. 2009; Kravitz et

65 al. 2011; Park et al. 2011; Rajimehr et al. 2011) without taking into account their covariation.

66 Indeed, an explicit test of spatial frequency, subjective distance and semantic properties found

67 that due to inherent feature correlations, all three properties explained the same variance in

68 fMRI responses, with no discernible unique contribution of any single property (Lescroart et al.

69 2015).

70   In the current fMRI study, we addressed the first challenge by choosing models based

71 on a prior study that investigated scene categorization behavior (Greene et al. 2016). This

72 behavioral study assessed the relative contributions of different factors that have traditionally

73 been considered important for scene understanding, including a scene's component objects

74 (e.g., Biederman 1987) and its global layout (e.g, Oliva and Torralba 2001), but also included

75 novel visual feature models based on state-of-the-art computer classification algorithms (e.g.,

76 Sermanet et al. 2013) as well as models that reflect conceptual scene properties, such as

77 superordinate categories, or the types of actions afforded by scene. Using an online same-

78 different categorization paradigm on hundreds of scene categories from the SUN database

79 (Xiao et al. 2014), a large-scale scene category distance matrix was obtained (reflecting a total

80 of 5 million trials), which was subsequently compared to predicted category distances for the

81 various candidate models. The three models that contributed most to human scene

82 categorization were 1) a model based on human-assigned labels of actions that can be carried

83 out in the scene ('functional model'), 2) a deep convolutional neural network ('DNN model') that

84 was trained to map visual features natural images to a set of a 1000 image classes from the

85 ImageNet object database (Deng et al. 2009), and 3) human-assigned object labels ('object

86 model') for all the objects in the scene. Given the superior performance of these top three

87 models in explaining scene categorization, we deemed these models most relevant to test in

88 terms of their contribution to brain representations. Specifically, we determined the relative

89   contribution of these top three models to neural scene representation by comparing them

90   against multi-voxel patterns in fMRI data collected while participants viewed a reduced set of

91   scene stimuli from Greene et al., (2016).

92   To address the second challenge, we implemented a stimulus selection procedure that

93   reduced inherent correlations between the three models of interest *a priori.* Specifically, we

94   compared predicted category distances for repeated samples of stimuli from the larger SUN

95   database, and selected a final set of stimuli for fMRI for which the predictions were minimally

96   correlated. To assess whether scene categorization behavior for this reduced stimulus set was

97   consistent with the previous behavioral findings, participants additionally performed a behavioral

98   multi-arrangement task outside the scanner. To isolate the unique contribution of each model to

99   fMRI and behavioral scene similarity, we applied a variance partitioning analysis, accounting for

100 any residual overlap in representational structure, between models.

101   To anticipate, our data reveal a striking dissociation between the feature model that best

102 describes behavioral scene similarity and the model that best explains similarity of fMRI

103 responses in scene-selective cortex. While we confirmed that behavioral scene categorization

104 was best explained a combination of unique contributions from the function model and DNN

105 features, there was no unique representation of scene functions in scene-selective brain

106 regions, which instead were best described by DNN features only. Follow-up analyses indicated

107 that scene functions correlated with responses in regions outside of scene-selective cortex,

108 some of which have been previously associated with action observation. However, a direct

109 comparison between behavioral scene similarity and fMRI responses indicated that behavioral

110 scene categorization correlated most strongly with scene-selective regions, with no discernible

111 contribution of other regions. This dissociation between the features that contribute uniquely to

112 behavioral versus fMRI scene similarity suggests that scene-selective cortex and DNN feature

113 models represent only a subset of the information relevant for scene categorization.

114

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 models 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 model was derived using a high-level layer of an AlexNet (Krizhevsky 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 models 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 models, resulting in three representational dissimilarity matrices (RDMs) reflecting either the deep network, object or functional model (**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 model 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

141   scanning. In this task, participants were presented with all stimuli in the set arranged around a

142   large white circle on a computer screen, and were instructed to drag-and-drop these scenes

143   within the white circle according to their similarity (see Methods and **Figure 2*A***).

144
145   *Function model and DNN model both contribute uniquely to scene categorization behavior*

146   To determine what information contributed to behavioral similarity judgments in the multi-

147   arrangement task, we created RDMs based on each participant's final arrangement by

148   measuring the pairwise distances between all 30 categories in the set (**Figure 2*B***), and then

149   computed correlations of these RDMs with the three model RDMs that quantified the similarity

150   of the scenes in terms of either functions, objects, or DNN features, respectively (see Figure

151   1*B*).

152   Replicating Greene et al., (2016), this analysis indicated that all three feature models

153   were significantly correlated with scene categorization behavior, with the functional feature

154   model having the highest correlation on average (**Figure 2*C***; objects: mean r = 0.16; DNN

155   features: mean r = 0.26; functions: mean r = 0.29, Wilcoxon one-sided signed-rank test, all

156   W(20) > 210, all z > 3.9, all p < 0.0001). The correlation with functions was higher than with

157   objects (Wilcoxon two-sided signed-rank test, W(20) = 199, z = 3.5, p = 0.0004), but not than

158   with DNN features (W(20) = 134, z = 1.1, p = 0.28), which also correlated higher than objects

159   (W(20) = 194, z = 3.3, p = 0.0009). However, comparison at the level of individual participants

160   indicated that functions outperformed both the DNN and object models for the majority of

161   participants (highest correlation with functions: n = 12; with DNN features: n = 7; with objects: n

162   = 1; **Figure 2*D***).

163   While these correlations indicate that scene dissimilarity based on the functional feature

164   model best matched the stimulus arrangements that participants made, they do not reveal to

165   what extent functional, DNN or object features *independently* contribute to the behavior. To

166   assess this, we performed two additional analyses. First, we computed *partial* correlations

167   between models and behavior whereby the correlation of each feature model with the behavior

168   was determined while taking into account the contributions of the other two feature models. The

169   results indicated that each model independently contributed to the behavioral data: significant

170   partial correlations were obtained for the object (W(20) = 173, z = 2.5, p = 0.006), DNN (W(20) =

171   209, z = 3.9, p < 0.0001) and functional feature models (W(20) = 209, z = 3.9, p < 0.0001), with

172   the functional model having the largest partial correlation (**Figure 2E**). Direct comparisons

173   yielded a similar pattern as the independent correlations, with weaker contributions of objects

174   relative to both functional (W(20) = 201, z = 3.6, p < 0.0003) and DNN features (W(20) = 195, z

175   = 3.4, p = 0.0008), whose partial correlations did not differ (W(20) = 135, z = 1.12, p = 0.26).

176   Second, we conducted a variance partitioning analysis, in which the function, DNN and

177   object feature models were entered either separately or in combination as predictors in a set of

178   multiple regression analyses aimed at explaining the multi-arrangement categorization behavior.

179   By comparing the explained variance based on regression on individual models versus models

180   in combination, we computed portions of unique variance contributed by each model as well as

181   portions of shared variance across models (see Methods for details). A full model in which all

182   three models were included explained 50.3% of the variance in the average multi-arrangement

183   behavior (**Figure 2F**). Highlighting the importance of functional features for scene

184   categorization, the largest portion of this variance could be uniquely attributed to the functional

185   feature model (unique r^2 = 37.6%), more than the unique variance explained by the DNN

186   features (unique r^2 = 29.0%) or the object features (unique r^2 = 1.4%). This result is consistent

187   with the findings of Greene et al., (2016), who found unique contributions of 45.2% by the

188   function model, 7.1% by the DNN model[*], and 0.3% by objects, respectively to scene

---

[*] 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

189   categorization measured using an online same-different categorization ask. One interesting

190   difference with this previous study is that the degree of shared variance between the three

191   models in our study is notably smaller (8.4% versus 27.4%); this is presumably a result of our

192   stimulus selection procedure that was explicitly aimed at minimizing correlations between the

193   models. Importantly, a reproducibility test indicated that the scene similarity reflected in the

194   multi-arrangement behavior was highly generalizable, resulting in an RDM correlation of $r = 0.73$

195   (95% confidence interval = [0.73-0.88], $p = 0.0001$), as assessed by comparison of two different

196   sets of scene exemplars that were evenly distributed across participants (see Methods).

197   In sum, these behavioral results confirm a unique, independent contribution of the

198   functional feature model to scene categorization behavior, here assessed using a multi-

199   arrangement sorting task (as opposed to a same/different categorization task). We also found a

200   unique but smaller contribution of deep network features, while the unique contribution of object

201   features was negligible. Next, we examined to what extent this information is represented in

202   brain responses to the same set of real-world scenes as measured with fMRI.

203

204   *DNN feature model uniquely predicts responses in scene-selective cortex*

205   To determine the information that is represented in scene-selective brain regions PPA, OPA and

206   MPA, we created RDMs based on the pairwise comparisons of multi-voxel activity patterns for

207   each category in these cortical regions (**Figure 3*A***), which we subsequently correlated with the

208   RDMs based on the object, function and DNN feature models. Similar to the behavioral findings,

209   all three feature models correlated with the fMRI response patterns to scenes in PPA (objects:

210   W(20) = 181, $z = 2.8$, $p = 0.002$; DNN: W(20) = 206, $z = 3.8$, $p < 0.0001$; functions: W(20) = 154,

211   $z = 1.8$, $p = 0.035$, see **Figure 3*B***). However, fMRI dissimilarity in PPA correlated more strongly

---

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.

212   with the DNN model than the object (W(20) = 195, z = 2.5, p = 0.012) and function (W(20) =

213   198, z = 3.5, p < 0.0005) feature models, which did not differ from one another (W(20) = 145 , z

214   = 1.5, p = 0.14). In OPA, only the DNN model correlated with the fMRI response patterns (W(20)

215   = 165, z = 2,2, p = 0.013), and this correlation was again stronger than for the object model

216   (W(20) = 172, z = 2.5, p = 0.012), but not the function model (W(20) = 134, z = 1.1, p = 0.28). In

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

218   When the three models were considered in combination, only the DNN model yielded a

219   significant partial correlation (PPA: W(20) = 203, z = 3.6, p < 0.0001, OPA: W(20) = 171, z =

220   2.5, p = 0.007, **Figure 3*C***), further showing that DNN features best capture responses in scene-

221   selective cortex. No significant partial correlation was found for the object model (PPA: W(20) =

222   148, z = 1.6, p = 0.056; OPA: W(20) = 74, z = 1.2, p = 0.88) or the function model (PPA: W(20)

223   = 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)

224   < 63, all z < 0.66, all p > 0.50). Variance partitioning of the fMRI response patterns (**Figure 3*D***)

225   indicated that the DNN model also contributed the largest portion of unique variance: in PPA

226   and OPA, DNN features contributed 71.1% and 68.9%, respectively, of the variance explained

227   by all models combined, more than the unique variance explained by the object (PPA: 5.3%;

228   OPA, 2.3%) and function (PPA: 0.3%; OPA: 2.6%) models. In MPA, a larger share of unique

229   variance was found for the function model (41.5%) than for the DNN (38.7%) and object model

230   (3.2%); however, overall explained variance in MPA was much lower than in the other ROIs. A

231   reproducibility test indicated that RDMs generalized across participants and stimulus sets for

232   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

233   (r = 0.06 [-0.16-0.26], p = 0.29), suggesting that the multi-voxel patterns measured in MPA were

234   less stable (see also the low noise ceiling in MPA in Figure 3B/C).

235   Taken together, the fMRI results indicate that of the three models considered, deep

236   network features (derived using a pre-trained convolutional network) best explained the coding

237   of real-world scene information in scene-selective regions PPA and OPA, more so than object

238   or functional information derived from semantic labels that were explicitly generated by human

239   observers. For MPA, results were inconclusive, as none of the models adequately captured the

240   response patterns measured in this region, which also did not contain response patterns that

241   generalized across stimulus sets and participants. This result reveals a discrepancy between

242   measurements of brain responses versus behavioral scene similarity, which indicated a large

243   contribution of functions to scene representation independent of the DNN features. To better

244   understand if and how scene-selective cortex represents behaviorally relevant information, we

245   next compared measurements of behavioral scene similarity to the fMRI responses directly.

246

247   *Scene selective cortex correlation with behavior reflects DNN feature model only*

248   To assess the extent to which fMRI response patterns in scene-selective cortex predicted

249   behavioral scene categorization, we correlated each of the scene-selective ROIs with three

250   measures of behavioral categorization: 1) the large-scale online categorization behavior

251   measured in Greene et al., (2016), 2) the average multi-arrangement behavior, and 3) each

252   participant's own multi-arrangement behavior. This analysis revealed a significant correlation

253   with behavior in all scene-selective ROIs (**Figure 4*A***). In PPA, all three measures of behavioral

254   categorization correlated with fMRI patterns of response (signed-rank test, online categorization

255   behavior: W(20) = 168, z = 2.3, p = 0.010; average multi-arrangement behavior: W(20) = 195, z

256   = 3.3, p = 0.0004; own arrangement behavior: W(20) = 159, z = 2.0, p = 0.023). In OPA,

257   significant correlations were found for both of the average behavioral measures (online

258   categorization behavior: W(20) = 181, z = 2.8, p = 0.002; average multi-arrangement behavior:

259   W(20) = 158, z = 1.96, p = 0.025), but not for the participant's own multi-arrangement behavior

260   (W(20) = 106, z = 0.02, p = 0.49), possible due to higher noise in the individual data.

261   Interestingly, however, MPA showed the opposite pattern: participant's own behavior was

262   significantly related to the observed patterns of response (W(14) = 89, z = 2.26, p = 0.011), but

263   the average behavioral measures were not (online behavior: W(14) = 47, z = 0.4, p = 0.65;

264   average behavior: W(14) = 74, *z* = 1.3, *p* = 0.09). Combined with the reproducibility test results

265   (see above), this suggests that the representations in MPA are more idiosyncratic to individual

266   participants or stimulus sets.

267   While these results support an important role for scene-selective regions in representing

268   information that informs scene categorization behavior, they also raise an intriguing question:

269   what aspect of categorization behavior is reflected in these neural response patterns? To

270   address this, we performed another variance partitioning analysis, now including the average

271   multi-arrangement behavior as a predictor of the fMRI response patterns, in combination with

272   the two models that correlated most strongly with this behavior, i.e. the DNN and function

273   feature models. The purpose of this analysis was to determine how much variance in neural

274   responses each of the models *shared* with the behavior, and whether there was any behavioral

275   variance in scene cortex that was not explained by our models. If the behaviorally relevant

276   information in the fMRI responses is primarily of a functional nature, we would expect portions of

277   the variance explained by behavior to be shared with the function features. Alternatively, if this

278   variance reflects mainly DNN features (which also contributed uniquely to the behavioral

279   categorization; **Figure 2*F***), we would expect it to be shared primarily with the DNN model.

280   Consistent with this second hypothesis, the variance partitioning results indicated that in

281   OPA and PPA, most of the behaviorally relevant information in the fMRI response patterns was

282   shared with the DNN model (**Figure 4*B***). In PPA, the behavioral RDMs on average shared

283   25.7% variance with the DNN model, while a negligible portion was shared with the function

284   model (less than 1%); indeed, nearly all variance shared between the function model and the

285   behavior was also shared with the DNN model (10.1%). In OPA, a similar trend was observed,

286   with behavior sharing 38.9% of the fMRI variance with the DNN model. In OPA, the DNN model

287   also eclipsed nearly all variance that behavior shared with the function model (9.7% shared by

288   behavior, functions and DNN features), leaving only 1.6% of variance shared exclusively by

289   functions and behavior. In contrast, in MPA, behavioral variance was shared with either the

290   DNN model or the function model to a similar degree (14.7% and 17.7%, respectively), with an

291   additional 27.1% shared with both; note, however, again MPA's low explained variance overall.

292   In sum, while fMRI response patterns in PPA and OPA reflect information that

293   contributes to scene similarity judgments, this information aligns best with the DNN feature

294   model; it does not reflect the unique contribution of functions to scene categorization behavior.

295   While in MPA, the behaviorally relevant representations may partly reflect other information, the

296   overall explained variance in MPA was again quite low, limiting interpretation of this result.

297

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

299   An important difference between the behavioral and the fMRI experiment was that participants

300   had access to the entire stimulus set when performing the behavioral multi-arrangement task,

301   which they could perform at their own pace, while they performed an task unrelated to scene

302   categorization in the fMRI scanner. Therefore, we reasoned that a possible explanation of the

303   discrepancy between our fMRI and behavioral findings could be a limited engagement of

304   participants with the briefly presented scenes while in the scanner, resulting in only superficial

305   encoding of the images in terms of visual features that are well captured by the DNN model,

306   rather than functional or object features that might be more conceptual in nature.

307   To test this possible explanation, we ran Experiment 2 and collected another set of fMRI

308   data (n = 8; four of these participants also participated in Experiment 1, allowing for comparison

309   of tasks within individuals) using the exact same visual stimulation, but with a different task

310   instruction. Specifically, instead of performing an unrelated fixation task, we instructed

311   participants to covertly name the presented scene. Covert naming has been shown to facilitate

312   stimulus processing within category-selective regions and to enhance semantic processing

313   (Turennout et al. 2000; van Turennout et al. 2003). Before entering the scanner, participants

314   were familiarized with all the individual scenes in the set, whereby they were explicitly asked to

315   generate a name for each individual scene (see Methods). Together, these manipulations were

316   intended to ensure that participants attended to the scenes and processed their content to a

317   fuller extent than in Experiment 1.

318   Despite this task manipulation, Experiment 2 yielded similar results as Experiment 1

319   (**Figure 5*A***). Reflecting participant's enhanced engagement with the scenes when performing

320   the covert naming task, overall model correlations were considerably higher than in Experiment

321   1, and now yielded significant correlations with the function model in both OPA and MPA

322   (**Figure 5*B***). The direct test of reproducibility also yielded significant, and somewhat increased,

323   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),

324   but not in MPA (r = 0.10 [-0.07-0.28], p = 0.17).

325   Importantly, in all three ROIs, the DNN model correlations were again significantly

326   stronger than the function and object model correlations, which again contributed very little

327   unique variance (**Figure 5*C***). Direct comparison of RDM correlations across the two

328   Experiments indicated that in PPA and OPA, the naming task resulted in increased correlations

329   for the DNN model only (two-sided Wilcoxon ranksum test, PPA: p = 0.0048; OPA p = 0.0056),

330   without any difference in correlations for the other models (all p > 0.52). In MPA, none of the

331   model correlations differed across tasks (all p > 0.21). Increased correlation with the DNN

332   model was present within the participants that participated in both experiments (n = 4; see

333   Methods): in PPA and OPA, 4/4 and 3/4 participants showed an increased correlation,

334   respectively, whereas no consistent patterns was observed for the other models and MPA

335   (**Figure 5*D***).

336   In sum, the results of Experiment 2 indicate that the strong contribution of DNN features

337   to fMRI responses in scene-selective cortex is not likely the result of limited engagement of

338   participants with the scenes when viewed in the scanner. If anything, enhanced attention to the

339   scenes under an explicit naming instruction resulted in even stronger representation of these

340   features, without a clear increase in contributions of the functional or object feature models.

341

342   *Contributions of the functional feature model outside of scene-selective cortex*

343   All our results so far indicate a dissociation between brain and behavioral assessments of the

344   representational similarity of scenes. In the behavioral domain, visual features in a deep

345   convolutional network uniquely contributed to behavioral scene categorization, but the function

346   model also exhibited a large unique contribution to scene categorization, regardless of whether

347   this behavior was assessed using a same-different categorization or a multi-arrangement task.

348   In contrast, fMRI responses in scene-selective cortex were primarily driven by DNN features,

349   without convincing evidence of an independent contribution of functions. Given this lack of

350   correlation with the function model in the scene-selective cortex, we explored whether this

351   information could be reflected elsewhere in the brain by performing whole-brain searchlight

352   analyses. Specifically, we extracted the multi-voxel patterns from spherical ROIs throughout

353   each participant's entire volume and performed partial correlation analyses including all three

354   models (visual features, objects, functions) to extract corresponding correlation maps for each

355   model. The resulting whole-brain searchlight maps were then fed into a to surface-based group

356   analysis (see Methods) to identify clusters of positive correlations indicating significant model

357   contributions to brain representation throughout all measured regions of cortex.

358   The results of these analyses were entirely consistent with the ROI analyses: for the

359   DNN feature model, significant searchlight clusters were found in PPA and OPA (**Figure 6*A***),

360   but not MPA, whereas no significant clusters were found for the function model in any of the

361   scene-selective ROIs. (The object model yielded no positive clusters). However, two clusters

362   were identified for the function model outside of scene-selective cortex (**Figure 6*B***): a bilateral

363   cluster on the ventral surface, lateral to PPA, overlapping with the fusiform and temporal lateral

364   gyri, as well as a unilateral cluster on the left lateral surface, located adjacent to, but more

365   ventral than, OPA, overlapping the posterior middle and inferior temporal gyrus.

366   The observed dissociation between behavioral categorization and scene-selective cortex

367   might mean that the functional features are represented outside of scene-selective cortex. If so,

368　we would expect the searchlight clusters that correlated with the function model to show a

369　correspondence with the behavioral scene categorization. To test this, we also correlated the

370　multi-arrangement behavior with multi-voxel pattern responses throughout the brain. Consistent

371　with the results reported in Figure 4, we found a significant searchlight correlation between the

372　behavioral measurements and response patterns in PPA and OPA (**Figure 7A**). Surprisingly,

373　however, behavioral categorization did not correlate with any regions outside these ROIs,

374　including the clusters that correlated with the function model.

375　　　　In order to better understand how representational dissimilarity in those clusters relates

376　to the functional feature model, we extracted the average RDM from each searchlight cluster

377　and inspected which scene categories were grouped together in these ROIs. Visual inspection

378　of the RDM and MDS plots of the RDMs (**Figure 7B**) indicates that in both the bilateral ventral

379　and left-lateralized searchlight clusters, there is some grouping by category according to the

380　function feature model (indicated by grouping by color in the MDS plot). However, it is also clear

381　that the representational space in these ROIs does not *exactly* map onto the functional feature

382　model in Figure 1C. Specifically, a few categories clearly 'stand out' with respect to the other

383　categories, as indicated by a large average distance relative to the remainder of the stimulus

384　set. Most of the scene categories that were strongly separated all contained scene exemplars

385　depicting humans that performed actions (see **Figure 7C**), although it is worth noting that the

386　fourth most distinct category, 'volcano', did not contain humans in its scene exemplars but may

387　be characterized by implied motion. These post-hoc observations suggest that (parts of) the

388　searchlight correlation with the functional feature model may be due to the presence of human-,

389　body- and/or motion selective voxels in these searchlight clusters.

390　　　　In sum, the searchlight analyses indicate that the maximum contributions of the DNN

391　model were located in scene-selective cortex. While some aspects of the functional feature

392　model may be reflected in regions outside of scene-selective cortex, these regions did not

16

393   appear to contribute to the scene categorization behavior, and may reflect selectivity for only a

394   subset of scene categories that clustered together in the functional model.

395

396   *Scene-selective cortex correlates with features in both mid- and high-level DNN layers*

397   DNNs consist of multiple layers that capture a series of transformations from pixels in the input

398   image to a class label, implementing a non-linear mapping of local convolutional filters

399   responses (layers 1-5) onto a set of fully-connected layers consisting of classification nodes

400   (layers 6-8) culminating in a vector of output 'activations' for labels assigned in the DNN training

401   phase. Visualization and quantification methods of the learned feature selectivity (e.g., Zhou et

402   al. 2014; Güçlü and van Gerven 2015; Bau et al. 2017; Wen et al. 2017) suggest that while

403   earlier layers contain local filters that resemble V1-like receptive fields, higher layers develop

404   selectivity for entire objects or object parts, perhaps resembling category-selective regions in

405   visual cortex. Our deep network feature model was derived using a single high-level layer, fully-

406   connected layer 7 ("fc7"). Moreover, this model was derived using the response patterns of a

407   DNN pretrained on ImageNet (Deng et al. 2009), an image database largely consisting of object

408   labels. Given the strong performance of the DNN feature model in explaining the fMRI

409   responses in scene-selective cortex, it is important to determine whether this result was

410   exclusive to higher DNN layers, and whether the task used for DNN training influences how well

411   the features represented in individual layers explain responses in scene-selective cortex. To do

412   so, we conducted a series of exploratory analyses to assess the contribution of other DNN

413   layers to fMRI responses, whereby we compared DNNs that were trained using either object or

414   scene labels.

415   To allow for a clean test of the influence of DNN training on features representations in

416   each layer, we derived two new sets of RDMs by passing our stimuli through 1) a novel 1000-

417   object label ImageNet-trained network implemented in Caffe (Jia et al. 2014) ('ReferenceNet')

418   and 2) a 250-scene label Places-trained network ("Places") (Zhou et al. 2014), (see Methods).

419   Direct comparisons of the layer-by-layer RDMs of these two DNNs (**Figure 8*A***) indicated that

420   while both models extracted similar features (evidenced by strong between-model correlations

421   overall; all layers *r* > 0.6). However, the similarity between models decreased with higher layers,

422   suggesting that features in higher DNN layers become tailored to the task they are trained on.

423   Moreover, this suggests that higher layers of the scene-trained DNN could potentially capture

424   different   features than the object-trained DNN. To investigate this, we next computed

425   correlations between the features in each DNN layer and the three original feature models

426   **(Figure 8*B*)**.

427   As expected, the original fc7 DNN model (which was derived using DNN responses to

428   the large set of images in the Greene et al., (2016) database, and thus not corresponding

429   directly to the reduced set of stimuli used in the current study) correlated most strongly with the

430   new DNN layer representations, showing steadily increasing correlations with higher layers of

431   both object-trained and the scene-trained DNN. By design, the object and functional feature

432   models should correlate minimally with layer 7 of the object-trained ReferenceNet DNN.

433   However, the function model correlated somewhat better with higher layers of the scene-trained

434   DNN, highlighting a potential overlap of the function model with the scene-trained DNN features,

435   again suggesting that the higher layers of the scene-trained DNN potentially capture additional

436   information that is not represented in the object-trained DNN. Therefore, we next tested whether

437   the scene-trained DNN correlated more strongly with fMRI responses in scene-selective cortex.

438   Layer-by-layer correlations of the object-trained (**Figure 8*C***) and the scene-trained DNN

439   (**Figure 8*D***) with fMRI responses in PPA, OPA and MPA however did not indicate a strong

440   evidence of a difference in DNN performance as a result of training. In PPA, both the object-

441   trained and place-trained DNN showed increased correlation with higher DNN layers, consistent

442   with previous work showing a hierarchical mapping of DNN layers to low vs. high-level visual

443   cortex (Güçlü and van Gerven 2015; Cichy et al. 2016; Wen et al. 2017). Note however that the

444   slope of this increase is quite modest; while higher layers overall correlate better than layers 1

445   and 2, in both DNNs the correlation with layer 3 is not significantly different from the correlation

446   of layers 7 and 8. In OPA, we observed no evidence for increased performance with higher

447   layers for the object-trained DNN; none of the pairwise tests survived multiple comparisons

448   correction. In fact, for the scene-trained DNN, the OPA correlation significantly *decreased* rather

449   than increased with higher layers, showing a peak correlation with layer 3. No significant

450   correlations were found for any model layer with MPA. These observations were confirmed by

451   searchlight analyses in which whole-brain correlation maps were derived for each layer of the

452   object- and scene-trained DNN (see **Figure 8-video 1 and Figure 8-video 2** for layer-by-layer

453   searchlight results in a movie format for the ReferenceNet and the Places DNN, respectively).

454   These results indicate that despite a divergence in representation in high-level layers for

455   differently-trained DNNs, their performance in predicting brain responses in scene-selective

456   cortex is quite similar. In PPA, higher layers perform significantly better than (very) low-level

457   layers, but mid-level layers already provide a relatively good correspondence with PPA activity.

458   This result was even more pronounced for OPA where mid-level layers yielded the maximal

459   correlations for both DNNs regardless of training. Therefore, these results suggest that fMRI

460   responses in scene-selective ROIs may reflect a contribution of visual features of intermediate

461   complexity rather than, or in addition to, the fc7 layer that was selected *a priori*.

462

463   **Discussion**

464

465   We assessed the contribution of three feature models previously implicated to be important for

466   scene understanding to neural representations of scenes in the human brain. First, we

467   confirmed earlier reports that functions strongly contribute to scene categorization by replicating

468   the results of Greene et al., (2016), now using a multi-arrangement task. Second, we found that

469   brain responses to visual scenes in scene-selective regions were best explained by a DNN

470   feature model, with no discernible unique contribution of the functional features. Although parts

471   of variance in the multi-arrangement behavior were captured by the DNN feature model - and

472   this part of the behavior was reflected in the scene-selective cortex - there are clearly aspects of

473   scene categorization behavior that were not reflected in the activity of these regions.

474   Collectively, these results thus reveal a striking dissociation between the information that is

475   most important for behavioral scene categorization and the information that best describes

476   representational dissimilarity of fMRI responses in regions of cortex that are thought to support

477   scene recognition. Below, we discuss two potential explanations for this dissociation.

478   First, one possibility is that functions are represented outside of scene-selective cortex.

479   Our searchlight analysis indeed revealed clusters of correlations with the function model in

480   bilateral ventral and left lateral occipito-temporal cortex. Visual inspection of these maps

481   suggests that these clusters potentially overlap with known face- and body-selective regions

482   such as the Fusiform Face (FFA; Kanwisher et al. 1997) and Fusiform Body (FBA; Peelen and

483   Downing 2007) areas on ventral surface, as well as the Extrastriate Body Area (EBA; Downing

484   2001) on the lateral surface. This lateral cluster could possibly include motion-selective (Zeki et

485   al. 1991; Tootell et al. 1995) and tool-selective (Martin et al. 1996) regions as well. Our results

486   further indicated that these searchlight clusters contained distinct representations of scenes that

487   contained *acting* bodies, and may therefore partially overlap with regions important for action

488   observation (e.g., Hafri et al. 2017). Lateral occipital-temporal cortex in particular is thought to

489   support action observation by containing 'representations which capture perceptual, semantic

490   and motor knowledge of how actions change the state of the world' (Lingnau & Downing, 2015).

491   While our searchlight results suggest a possible contribution of these non-scene-selective

492   regions to scene understanding, more research is needed to address how the functional feature

493   model as defined here relates to the action observation network, and to what extent the

494   correlations with functional features can be explained by bottom-up coding of bodies and motion

495   versus more abstract action-associated features. Importantly, the lack of a correlation between

496 these regions and the multi-arrangement behavior suggests that these regions do not fully

497 capture the representational space that is reflected in the function feature model.

498   The second possible explanation for the dissociation between brain and behavioral data

499 is that the task that participants performed during fMRI did not engage the same mental

500 processes that participants employed during the two behavioral tasks we investigated.

501 Specifically, both the multi-arrangement used here and the online same-different behavioral

502 paradigm used in (Greene et al. 2016) required participants to directly compare simultaneously

503 presented scenes, while we employed a 'standard' fixation task in the scanner to prevent

504 biasing our participants towards one of our feature models. Therefore, one possibility is that

505 functional features only become relevant for scene categorization when participants are

506 engaged in a *contrastive* task, i.e. explicitly comparing two scene exemplars side-by-side (as in

507 Greene et al., 2016) or within the context of the entire stimulus set being present on the screen

508 (as in our multi-arrangement paradigm). Thus, the fMRI results might change with an explicit

509 contrastive task in which multiple stimuli are presented at the same time, or perhaps with a task

510 that explicitly requires participants to consider functional aspects of the scenes. Although we

511 investigated one possible influence of task in the scanner by using a covert naming task in

512 Experiment 2, resulting in deeper and more conceptual processing, it did not result in a clear

513 increase in the correlation with the function model in scene-selective cortex. The evidence for

514 task effects on fMRI responses in category-selective cortex is somewhat mixed: Task

515 differences have been reported to affect multi-voxel pattern activity in both object-selective

516 (Harel et al. 2014) and scene-selective cortex (Lowe et al. 2016), but other studies suggest that

517 task has a minimal influence on representation in ventral stream regions, instead being reflected

518 in fronto-parietal networks (Erez and Duncan 2015; Bracci et al. 2017; Bugatus et al. 2017).

519 Overall, our findings suggest that not all the information that contributes to scene categorization

520 is reflected in scene-selective cortex activity 'by default', and that explicit task requirements may

521 be necessary in order for this information to emerge in the neural activation patterns in these

522 regions of cortex.

523   Importantly, the two explanations outlined above are not mutually exclusive. For

524 example, it is possible that a task instruction to explicitly label the scenes with potential actions

525 will activate components of both the action observation network (outside scene-selective cortex)

526 as well as task-dependent processes within scene-selective cortex. Furthermore, given reports

527 of potentially separate scene-selective networks for memory versus perception (Baldassano et

528 al. 2016; Silson et al. 2016), it is likely that differences in mnemonic demands between tasks

529 may have an important influence on scene-selective cortex activity. Indeed, memory-based

530 navigation or place recognition tasks (Epstein et al. 2007; Marchette et al. 2014) have been

531 shown to more strongly engage the medial parietal cortex and MPA. In contrast, our observed

532 correlation with DNN features seems to support a primary role for PPA and OPA in bottom-up

533 visual scene analysis, and fits well with the growing literature showing correspondences

534 between extrastriate cortex activity and DNN features (Cadieu et al. 2014; Khaligh-Razavi and

535 Kriegeskorte 2014; Güçlü and van Gerven 2015; Cichy et al. 2016; Horikawa and Kamitani

536 2017; Wen et al. 2017). Our analyses further showed that DNN correlations with scene-selective

537 cortex were not exclusive to higher DNN layers, but already emerged at earlier layers,

538 suggesting that the neural representation in PPA/OPA may be driven more by visual features

539 than semantic information (Watson et al. 2017).

540   One limitation of our study is that we did not exhaustively test all possible DNN models.

541 While our design - in which we explicitly aimed to minimize inherent correlations between the

542 feature models beforehand - required us to 'fix' the DNN features to be evaluated beforehand,

543 many more variants of DNN models have been developed, consisting of different architectures

544 such as VGG, GoogleNet and ResNet (Garcia-Garcia et al. 2017), as well as different training

545 regimes. Here, we explored the effect of DNN training by comparing the feature representations

546 between an object- versus a place-trained DNN, but we did not see strong differences in terms

547   of their ability to explain fMRI responses in either scene-selective cortex or other parts of the

548   brain (see whole-brain searchlights for the two DNNs in Figure 8-video 1 and Figure 8-video 2).

549   However, this does not exclude the possibility that other DNNs will map differently onto brain

550   responses, and possibly also explain more of the behavioral measures of human scene

551   categorization. For example, aDNN trained on the Atomic Visual Actions (AVA) dataset (Gu et

552   al. 2017), or the DNNs currently being developed in context of event understanding the

553   Moments in Time Dataset (Monfort et al. 2018) could potentially capture more of the variance

554   explained by the functional feature model in the scene categorization behavior. To facilitate the

555   comparison of our measurements with alternative and future models, we have made the fMRI

556   and the behavioral data accompanying this paper publicly available in Figure 1-source data 1.

557   These considerations highlight an important avenue for future research in which multiple

558   feature models (including DNNs that vary by training and architecture) and brain and behavioral

559   measurements are carefully compared. However, our current results suggest that when

560   participants perform scene categorization, either explicitly (Greene et al. 2016) or within a multi-

561   arrangement paradigm (Kriegeskorte and Mur 2012), they incorporate information that is not

562   reflected in either the DNNs or in PPA and OPA. Our results thus highlight a significant gap

563   between the real-world information that is captured both in scene–selective cortex and a set of

564   commonly used off-the-shelf DNNs relative to the information that drives human understanding

565   of visual environments. Visual environments are highly multidimensional, and scene

566   understanding encompasses many behavioral goals, including not just visual object or scene

567   recognition, but also navigation and action planning (Malcolm et al. 2016). While visual/DNN

568   features likely feed into multiple of these goals - for example, by signaling navigable paths in the

569   environment (Bonner and Epstein 2017), or landmark suitability (Troiani et al. 2014) - it is

570   probably not appropriate to think about the neural representations relevant to all these different

571   behavioral goals as being contained within one single brain region or a single neural network

572   model. Ultimately, unraveling the neural coding of scene information will require careful

573   manipulations of both multiple tasks and multiple scene feature models, as well as a potential

574   expansion of our focus on a broader set of regions than those characterized by the presence of

575   scene-selectivity.

576

577   *Summary and conclusion*

578   We successfully disentangled the type of information represented in scene-selective cortex: out

579   of three behaviorally relevant feature models, only one provided a robust correlation with activity

580   in scene-selective cortex. This model was derived from deep neural network features in a widely

581   used computer vision algorithm of object and scene recognition. Intriguingly, however, the DNN

582   model was not sufficient to explain scene categorization behavior, which was characterized by

583   an additional strong contribution of functional information. This highlights both a limitation of

584   current DNNs in explaining scene understanding, as well as a potentially more distributed

585   representation of scene information in the human brain beyond scene-selective cortex.

586

587   **Methods**

588

589   *Participants.* Twenty healthy participants (13 female, mean age 25.4 yrs, SD = 4.6) completed

590   the first fMRI experiment and subsequent behavioral experiment. Four of these participants (3

591   female, mean age 24.3 yrs, SD = 4.6) additionally participated in the second fMRI experiment,

592   as well as four new participants (2 female, mean age 25 yrs, SD = 1.6), yielding a total of eight

593   participants. Criteria for inclusion were that participants had to complete the entire experimental

594   protocol (i.e., the fMRI scan and the behavioral experiment). Beyond the participants reported,

595   three additional subjects were scanned but behavioral data was either not obtained or lost. Four

596   additional participants did not complete the scan session due to discomfort or technical

597   difficulties. All participants had normal or corrected-to-normal vision and gave written informed

598   consent as part of the study protocol (93 M-0170, NCT00001360) prior to participation in the

599   study. The study was approved by the Institutional Review Board of the National Institutes of

600   Health and was conducted according to the Declaration of Helsinki.

601

602   *MRI acquisition.* Participants were scanned on a research-dedicated Siemens 7T Magnetom

603   scanner in the Clinical Research Center on the National Institutes of Health Campus (Bethesda,

604   MD). Partial T2*-weighted functional image volumes were acquired using a gradient echo planar

605   imaging (EPI) sequence with a 32-channel head coil (47 slices; 1.6 x 1.6 x 1.6 mm; 10%

606   interslice gap; TR, 2s; TE, 27 ms; matrix size, 126 x 126; FOV, 192 mm). Oblique slices were

607   oriented approximately parallel to the base of the temporal lobe and were positioned such that

608   they covered the occipital, temporal, parietal cortices, and as much as possible of frontal cortex.

609   After the functional imaging runs, standard MPRAGE (magnetization-prepared rapid-acquisition

610   gradient echo) and corresponding GE-PD (gradient echo–proton density) images were

611   acquired, and the MPRAGE images were then normalized by the GE-PD images for use as a

612   high-resolution anatomical image for the following fMRI data analysis (Van de Moortele, 2009).

613

614   *Stimuli & models.* Experimental stimuli consisted of color photographs of real-world scenes (256

615   x 256 pixels) from 30 difference scene categories that were selected from a larger database

616   previously described in (Greene et al. 2016). These scene categories were picked using an

617   iterative sampling procedure that minimized the correlation between the categories across three

618   different models of scene information: functions, object labels and DNN features, with the

619   additional constraint that the final stimulus set should be have equal portions of categories from

620   indoor, outdoor man-made and outdoor natural scenes, which is the largest superordinate

621   distinction present in the largest scene-database that is publicly available, the SUN database

622   (Xiao et al. 2014). As obtaining a guaranteed minimum was impractical, we adopted a variant of

623   the odds algorithm (Bruss 2000) as our stopping rule. Specifically, we created 10,000 sets of 30

624   categories and measured the correlations between functional, object, and DNN RDMs (distance

625   metric: Spearman's *rho*), noting the minimal value from the set. We persisted in this procedure

626   until we observed a set with lower inter-feature correlations than was observed in the initial

627   10,000. From each scene category, 8 exemplars were randomly selected and divided across

628   two separate stimulus sets of 4 exemplars for each scene category. Stimulus sets were

629   assigned randomly to individual participants (Experiment 1: stimulus set 1, n = 10; stimulus set

630   2, n = 10; Experiment 2, stimulus set 1, n = 5; stimulus set 2, n = 3). Participants from

631   Experiment 2 that had also participated in Experiment 1 were presented with the other stimulus

632   set than the one they saw in Experiment 1.

633

634   *fMRI procedure.* Participants were scanned while viewing the stimuli on a back-projected screen

635   through a rear-view mirror that was mounted on the head coil. Stimuli were presented at a

636   resolution of 800 x 600 pixels such that stimuli subtended ~10 x 10 degrees of visual angle.

637   Individual scenes were presented in an event-related design for a duration of 500 ms, separated

638   by a 6s interval. Throughout the experimental run, a small fixation cross (< 0.5 degrees) was

639   presented in the center of the screen. In Experiment 1, participants performed a task on the

640   central fixation cross that was unrelated to the scenes. Specifically, simultaneous with the

641   presentation of each scene, either the vertical or horizontal arm of the fixation cross became

642   slightly elongated and participants indicated which arm was longer by pressing one of two

643   buttons indicated on a hand-held button box. Both arms changed equally often within a given

644   run and arm changes were randomly assigned to individual scenes. In Experiment 2, the fixation

645   cross had a constant size, and participants were instructed to covertly name the scene whilst

646   simultaneously pressing one button on the button box. To assure that participants were able to

647   generate a name for each scene, they were first familiarized with the stimuli. Specifically, prior

648   to scanning, participants were presented with all scenes in the set in randomized order on a

649   laptop in the console room. Using a self-paced procedure, each scene was presented in

650   isolation on the screen accompanied by the question 'How would you name this scene?'. The

651 participants were asked to type one or two words to describe the scene; as they typed, their

652 answer appeared under the question, and they were able to correct mistakes using backspace.

653 After typing the self-generated name, participants hit enter and the next scene would appear

654 until all 120 scenes had been seen by the participant. This procedure took about ~10 minutes.

655   In both Experiment 1 and 2, participants completed 8 experimental runs of 6.4 minutes

656 each (192 TRs per run); one participant from Experiment 1 only completed 7 runs due to time

657 constraints. Each run started and ended with a 12s fixation period. Each run contained 2

658 exemplar presentations per scene category. Individual exemplars were balanced across runs

659 such that all stimuli were presented after two consecutive runs, yielding 4 presentations per

660 exemplar in total. Exemplars were randomized across participants such that each participant

661 always saw the same two exemplars within an individual run; however the particular

662 combination was determined anew for each individual participant and scene category. Stimulus

663 order was randomized independently for each run. Stimuli were presented using PsychoPy

664 v1.83.01 (Peirce 2007).

665

666 *Functional localizers.* Participants additionally completed four independent functional block-

667 design runs (6.9 minutes, 208 TRs) that were used to localize scene-selective regions of

668 interest (ROIs). Per block, twenty gray-scale images (300 x 300 pixels) were presented from

669 one of eight different categories: faces, man-made and natural objects, buildings, and four

670 different scene types (man-made open, man-made closed, natural open, natural closed; Kravitz

671 et al., 2011) while participants performed a one-back repetition-detection task. Stimuli were

672 presented on a gray background for 500 ms duration, separated by 300 ms gaps, for blocks of

673 16s duration, separated by 8s fixation periods. Categories were counterbalanced both within

674 runs (such that each category occurred twice within a run in a mirror-balanced sequence) and

675 across runs (such that each category was equidistantly spaced in time relative to each other

676 category across all four runs). Two localizer runs were presented after the first four experimental

677 runs and two after the eight experimental runs were completed but prior to the T1 acquisition.

678 For four participants, only two localizer runs were collected due to time constraints.

679

680 *Behavioral experiment.* On a separate day following the MRI data acquisition, participants

681 performed a behavioral multi-arrangement experiment. In a behavioral testing room, participants

682 were seated in front of a desktop computer with a flat screen monitor (size?) on which all 120

683 stimuli that the participant had previously seen in the scanner were displayed as small

684 thumbnails around a white circular arena. A mouse-click on an individual thumbnail displayed a

685 larger version of that stimulus in the upper right corner. Participants were instructed to arrange

686 the thumbnails within the white circle in such a way that the arrangement would reflect 'how

687 similar the scenes are, whatever that means to you', by means of dragging and dropping the

688 individual exemplar thumbnails. We purposely avoided provided specific instructions in order to

689 not bias participants towards using either functions, objects or visual features to determine

690 scene similarity. Participants were instructed to perform the task at their own pace; if the task

691 took longer than 1hr, participants were encouraged to finish the experiment (almost all

692 participants took less time, averaging a total experiment duration of ~45 mins). Stimuli were

693 presented using the single-arrangement MATLAB code provided in (Kriegeskorte & Mur, 2012).

694 To obtain some insight in the sorting strategies used by participants, they were asked (after

695 completing the experiment) to take a few minutes to describe how they organized the scenes,

696 using a blank sheet of paper and a pen, using words, bullet-points or drawings.

697

698 *Behavioral data analysis.* Behavioral representational dissimilarity matrices (RDMs) were

699 constructed for each individual participant by computing the pairwise squared on-screen

700 distances between the arranged thumbnails and averaging the obtained distances across the

701 exemplars within each category. The relatedness of the models and the behavioral data was

702   determined in the same manner as for the fMRI analysis, i.e. by computing both individual

703   model correlations and unique and shared variance across models via hierarchical regression.

704

705   *fMRI preprocessing.* Data were analyzed using AFNI software (https://afni.nimh.nih.gov). Before

706   statistical analysis, the functional scans were slice-time corrected and all the images for each

707   participant were motion corrected to the first image of their first task run after removal of the first

708   and last six TRs from each run. After motion correction, the localizer runs were smoothed with a

709   5mm full-width at half-maximum Gaussian kernel; the even-related data was not smoothed.

710

711   *fMRI statistical analysis: localizers.* Bilateral ROIs were created for each participant individually

712   based on the localizer runs by conducting a standard general linear model implemented in

713   AFNI. A response model was built by convolving a standard gamma function with a 16s square

714   wave for each condition and compared against the activation time courses using Generalized

715   Least Squares (GLSQ) regression. Motion parameters and four polynomials accounting for slow

716   drifts were included as regressors of no interest. To derive the response magnitude per

717   category, t-tests were performed between the category-specific beta estimates and baseline.

718   Scene-selective ROIs were generated by thresholding the statistical parametric maps resulting

719   from contrasting scenes > faces at $p < 0.0001$ (uncorrected). Only contiguous clusters of voxels

720   (>25) exceeding this threshold were then inspected to define scene-selective ROIs consistent

721   with previously published work (Epstein 2005). For participants in which clusters could not be

722   disambiguated, the threshold was raised until individual clusters were clearly identifiable. While

723   PPA and OPA were identified in all participants for both Experiment 1 and 2, MPA/RSC was

724   detected in only 14 out 20 participants in Experiment 1, and all analyses for this ROI in

725   Experiment 1 are thus based on this subset of participants.

726

727   *fMRI statistical analysis: event-related data.* Each event-related run was deconvolved

728   independently using the standard GLSQ regression model in AFNI. The regression model

729   included a separate regressor for each of the 30 scene categories as well as motion parameters

730   and four polynomials to account for slow drifts in the signal. The resulting beta-estimates were

731   then used to compute representational dissimilarity matrices (RDMs; (Kriegeskorte et al. 2008)

732   based on the multi-voxel patterns extracted from individual ROIs. Specifically, we computed

733   pairwise cross-validated Mahalanobis distances between each of the scene 30 categories

734   following the approach in (Walther et al. 2016). First, multi-variate noise normalization was

735   applied by normalizing the beta-estimates by the covariance matrix of the residual time-courses

736   between voxels within the ROI. Covariance matrices were regularized using shrinkage toward

737   the diagonal matrix (Ledoit and Wolf 2004). Unlike univariate noise normalization, which

738   normalizes each voxel's response by its own error term, multivariate noise normalization also

739   takes into account the noise covariance between voxels, resulting in more reliable RDMs

740   (Walther et al. 2016). After noise normalization, squared Euclidean distances were computed

741   between individual runs using a leave-one-run-out procedure, resulting in cross-validated

742   Mahalanobis distance estimates. Note that unlike correlation distance measures, cross-

743   validated distances provide unbiased estimates of pattern dissimilarity on a ratio scale (Walther

744   et al. 2016), thus providing a distance measure suitable for direct model comparisons.

745

746   *Model comparisons: individual models.* To test the relatedness of the three models of scene

747   dissimilarity with the measured fMRI dissimilarity, the off-diagonal elements of each model RDM

748   were correlated (Pearson's *r*) with the off-diagonal elements of the RDM of each fMRI ROI for

749   each individual participant separately. Following (Nili et al. 2014), the significance of these

750   correlations was determined using one-sided signed-rank tests against zero, while pairwise

751   differences between models in terms of their correlation with fMRI dissimilarity were determined

752   using two-sided signed-ranked tests. For each test, we report the sum of signed ranks for the

753 number of observations W(n) and the corresponding p-value; for tests with n > 10 we also report

754 the z-ratio approximation. The results were corrected for multiple comparisons (across both

755 individual model correlations and pairwise comparisons) using FDR correction (Benjamini and

756 Hochberg 1995) for each individual ROI separately. Noise ceilings were computed following (Nili

757 et al. 2014): an upper bound was estimated by computing the correlation between each

758 participant's individual RDM and the group-average RDM, while a lower bound was estimated

759 by correlating each participant's RDM with the average RDM of the other participants (leave-

760 one-out approach). The participant-averaged RDM was converted to rank order for visualization

761 purposes only.

762

763 *Model comparisons: partial correlations and variance partitioning.* To determine the contribution

764 of each individual model when considered in conjunction with the other models, we performed to

765 additional types of analyses: partial correlations, in which each model was correlated (Pearsons

766 r) while partialling out the other two models, as well as variation partitioning based on multiple

767 linear regression. For the latter, the off-diagonal elements of each ROI RDM were assigned as

768 the dependent variable, while the off-diagonal elements of the three model RDMs were entered

769 as independent variables (predictors). To obtain unique and shared variance across the three

770 models, 7 multiple regression analyses were run in total: one 'full' regression that included all

771 three feature models as predictors; and six reduced models that included as predictors either

772 combinations of two models in pairs (e.g., functions and objects), or including each model by

773 itself. By comparing the explained variance (r^2) of a model used alone to the r^2 of that model in

774 conjunction with another model, we can infer the amount of variance that is independently

775 explained by that model, i.e. partition the variance (see also (Groen et al. 2012; Ramakrishnan

776 et al. 2014; Lescroart et al. 2015; Çukur et al. 2016; Greene et al. 2016; Hebart et al. 2018) for

777 similar approaches).

778   Analogous to the individual model correlation analyses, partial correlations were

779   calculated for each individual participant separately, and significance was determined using

780   one-sided signed-rank tests across participants (FDR-corrected across all comparisons within a

781   given ROI). To allow comparison with the results reported in (Greene et al. 2016), variance

782   partitioning was performed on the participant-average RDMs. Similar results were found,

783   however, when variance was partitioned for individual participant's RDMs and then averaged

784   across participants. To visualize this information in an Euler diagram, we used the EulerAPE

785   software (Micallef and Rodgers 2014).

786

787   *Direct reproducibiilty test of representational structure in behavior and fMRI.* To assess how well

788   the obtained RDMs were reproducible in each measurement domain (behavior and fMRI), we

789   compared the average RDMs obtained for the two separate stimulus sets. Since these two sets

790   of stimuli were viewed by different participants (see above under 'Stimuli & models'), this

791   comparison provides a strong test of generalizability, across both scene exemplars and across

792   participant pools. Set-average RDMs were compared by computing inter-RDM correlations

793   (Pearson's *r*) and 96% confidence intervals (CI) and statistically tested for reproducibility using a

794   random permutation test based on 10.000 randomizations of the category labels.

795

796   *Variance partitioning of fMRI based on models and behavior.* Using the same approach as in

797   the previous section, a second set of regression analyses was performed to determine the

798   degree of shared variance between the behavior on the one hand, and the functions and visual

799   features on the other hand, in terms of the fMRI response pattern dissimilarity. The Euler

800   diagrams were derived using the group-average RDMs, taking the average result of the multi-

801   arrangement task of these participants as the behavioral input into the analysis.

802

803 *DNN comparisons* The original fc7 DNN feature model was determined based on to the large

804 set of exemplars (average of 65 per scene category) used in Greene et al., (2016). To

805 investigate the influence of DNN layer and training images on the learned visual features and

806 their correspondence with activity in scene-selective cortex, we derived two new sets of RDMs

807 by passing our scene stimuli through two pre-trained, 8-layer AlexNet (Krizhevsky et al. 2012)

808 architecture networks: 1) a 1000-object label ImageNet-trained (Deng et al. 2009) network

809 implemented in Caffe (Jia et al. 2014) ('ReferenceNet') and 2) a 250-scene label Places-trained

810 network ("Places") (Zhou et al. 2014). By extracting the node activations from each layer, we

811 computed pairwise dissimilarity (1 – Pearson's *r*) resulting in one RDM per layer and per model.

812 These RDMs were then each correlated with the fMRI RDMs from each participant in PPA, OPA

813 and MPA (Pearson's *r*). These analyses were performed on the combined data of Experiment 1

814 and 2; RDMs for participants that participated in both Experiments (n = 4) were averaged prior

815 to group-level analyses.

816

817 *Searchlight analyses.* To test the relatedness of functions, objects and visual feature models

818 with fMRI activity recorded outside scene-selective ROIs, we conducted whole-brain searchlight

819 analyses. RDMs were computed in the same manner as for the ROI analysis, i.e. computing

820 cross-validated Mahalanobis distances based on multivariate noise-normalized multi-voxel

821 patterns, but now within spherical ROIs of 3 voxel diameter (i.e. 123 voxels/searchlight).

822 Analogous to the ROI analyses, we computed partial correlations of each feature model,

823 correcting for the contributions of the remaining two models. These partial correlation

824 coefficients were assigned to the center voxel of each searchlight, resulting in one whole-

825 volume map per model. Partial correlation maps were computed for in each participant

826 separately in their native volume space. To allow comparison at the group level, individual

827 participant maps were first aligned to their own high-resolution anatomical scan and then to

828 surface reconstructions of the grey and white matter boundaries created from these high-

829 resolution scans using the Freesurfer (http://surfer.nmr.mgh.harvard.edu/) 5.3 autorecon script

830 using SUMA (Surface Mapping with AFNI) software (https://afni.nimh.nih.gov/Suma). The

831 surface images for each participant were then smoothed with a Gaussian 10mm FWHM filter in

832 surface coordinate units using the SurfSmooth function with the HEAT_07 smoothing method.

833 Group-level significance was determined by submitting these surface maps to node-wise

834 one-sample t-tests in conjunction with Threshold Free Cluster Enhancement (Smith and Nichols

835 2009) through Monte Carlo simulations using the algorithm implemented in the CoSMoMVPA

836 toolbox (Oosterhof et al. 2016), which performs group-level comparisons using sign-based

837 permutation testing (n = 10,000) to correct for multiple comparisons. To increase power, the

838 data of Experiment 1 and 2 were combined; coefficient maps for participants that participated in

839 both Experiments (n = 4) were averaged prior to proceeding to group-level analyses.

840 For searchlight comparisons with scene categorization behavior and feature models

841 based on different DNN layers, we computed regular correlations (Pearson's *r)* rather than

842 partial correlations. For the behavioral searchlight, we used the average multi-arrangement

843 behavior from Experiment 1 (since the participants from Experiment 2 did not perform this task).

844 For the DNN searchlights, we used the same layer-by-layer RDMs as for the ROI analyses,

845 independently correlating those with the RDMs of each spherical ROI. Group-level significance

846 was determined in the same manner as for the *a priori* selected feature models (see above).

847 **Acknowledgements**

858   **Figure captions**

859   **Figure 1** Models and predicted stimulus dissimilarity. **A)** Stimuli were characterized in three

860   different ways: functions (derived using human-generated action labels), objects (derived using

861   human-generated object labels) and DNN features (derived using layer 7 of a 1000-class

862   trained convolutional neural network). **B)** RDMs showing predicted representational dissimilarity

863   in terms of functions, objects and DNN features for the 30 scene categories sampled from

864   Greene et al., (2016) for the purpose of the current study. Scenes were sampled to achieve

865   minimal between-matrix correlations, with the constraint that the final stimulus set should have

866   equal portions of categories from indoor, outdoor man-made and outdoor natural scenes. The

867   category order in the figure is determined based on a k-means clustering on the functional

868   model RDM; clustering was performed by requesting 8 clusters, which explained 80% of the

869   variance in the functional feature model. RDMs were rank-ordered for visualization purposes

870   only. **C)** Multi-dimensional scaling plots of the model RDMs, color-coded based on the functional

871   clusters depicted in B). Functional model clusters reflected functions such as 'sports', and

872   'transportation'; note that these semantic labels were derived post-hoc after clustering, and did

873   not affect stimulus selection. Critically, representational dissimilarity based on the two other

874   models (objects and DNN features) predicted different cluster patterns. The stimuli and model

875   RDMs, along with the behavioral and fMRI measurements, are provided in Figure 1-source data

876   1.

877 **Figure 2** Behavioral multi-arrangement paradigm and results. **A)** Participants organized the

878 scenes in inside a large white circle according to their perceived similarity as determined by

879 their own judgment, without receiving explicit instructions as to what information to use to

880 determine scene similarity. **B)** RDM displaying the average dissimilarity between categories in

881 behavioral arrangement (rank-ordered for visualization only). **C)** Average (bar) and individual

882 participant (gray dots) correlations between the behavioral RDM and the model RDMs for

883 objects (red), DNN features (yellow) and functions (blue) from Figure 1B. Stars (*) indicate $p <$

884 0.05 for model-specific one-sided signed-rank tests against zero, while horizontal bars indicate

885 $p < 0.05$ for two-sided pairwise signed-rank tests between models; p-values were FDR-

886 corrected across both types of comparisons. The light-blue shaded rectangular region reflects

887 the upper and lower bound of the noise ceiling, indicating RDM similarity between individual

888 participants and the group average (see Methods). **D)** Count of participants with the highest

889 correlation with either objects, DNN features or objects. **E)** Average (bar) and individual

890 participant (gray dots) partial correlation values for each model RDM. Statistical significance

891 was determined the same way as in C). **F)** Euler diagram depicting the results of a variance

892 partitioning analysis on the behavioral RDM for objects (red circle), DNN features (yellow circle)

893 and functions (blue circle). Unique (non-overlapping diagram portions) and shared (overlapping

894 diagram portions) variances are expressed as percentages of the total variance explained by all

895 models combined.

896 **Figure 3** RDMs and model comparisons for fMRI Experiment 1 (n = 20). **A)** RDMs displaying

897 average dissimilarity between categories in multi-voxel patterns in PPA, OPA and MPA (rank-

898 ordered for visualization only). **B)** Average (bar) and individual (gray dots) correlations between

899 the ROIs in A) and the model RDMs for objects (red), DNN features (yellow) and functions

900 (blue) (FDR-corrected). See legend of Figure 2B for explanation of the statistical indicators and

901 noise ceiling. **C)** Average (bar) and individual (gray dots) partial correlation coefficients for each

902   model RDM. Statistics are the same as in B). **D)** Euler diagram depicting the variance

903   partitioning results the average dissimilarity in each ROI for each of the three models,

904   expressed as percentages of unique and shared variance of the variance explained by all three

905   models together.

906   **Figure 4** Correlations and variance partitioning of behavioral measurements of scene

907   categorization and similarity of fMRI responses to the same scene categories. **A)** Correlations of

908   three measures of behavioral categorization (see Results section for details) with fMRI response

909   patterns in PPA, OPA and MPA. See legend of Figure 2B for explanation of the statistical

910   indicators and noise ceiling. **B)** Euler diagram depicting the results of variance partitioning the

911   fMRI responses in PPA, OPA and MPA for objects (red), DNN features (yellow) and average

912   sorting behavior (green), indicating that the majority of the variance in the fMRI signal that is

913   explained by categorization behavior is shared with the DNN features.

914   **Figure 5** RDMs and model comparisons for Experiment 2 (n = 8, covert naming task). **A)**

915   Average dissimilarity between categories in multi-voxel patterns measured in PPA, OPA and

916   MPA (rank-ordered). **B)** Correlations between the ROIs in A) and the model RDMs for objects

917   (red), DNN features (yellow) and functions (blue) (FDR-corrected). See legend of Figure 2B for

918   explanation of the statistical indicators and noise ceiling. Note how in PPA, the DNN model

919   correlation approaches the noise ceiling, suggesting that this model adequately captures the

920   information reflected in response patterns in this ROI. **C)** Euler diagram depicting the variance

921   partitioning results on the average dissimilarity in each ROI. **D)** Average (bars) and individual

922   (dots/lines) within-participant (n = 4) comparison of fMRI-model correlations across the different

923   task manipulations in Experiment 1 and 2 (participants were presented with a different set of

924   scenes in each task, see Methods). Note how increased attention to the scenes due to the

925   naming mainly enhances the correlation with DNN features.

926   **Figure 6**. Medial (left) and lateral (right) views of group-level searchlights for **A)** the DNN and **B)**

927   function feature models, overlaid on surface reconstructions of both hemispheres of one

928   participant. Each map was created by submitting the group-average partial correlation maps for

929   each model and hemisphere to one-sample tests against a mean of zero, cluster-corrected for

930   multiple comparisons using Threshold-Free Cluster Enhancement (thresholded on z = 1.64,

931   corresponding to one-sided p < 0.05). Unthresholded versions of the average partial correlation

932   maps are inset above. Group-level ROIs PPA, OPA and MPA are highlighted in solid white

933   lines. Consistent with the ROI analyses, the DNN feature model contributed uniquely to

934   representation in PPA and OPA. The function model uniquely correlated with a bilateral ventral

935   region, as well as a left-lateralized region overlapping with the middle temporal and occipital

936   gyri.

937   **Figure 7**. **A)** Group-average searchlight result for behavioral scene categorization. Maps reflect

938   correlation (Pearson's *r*) of the group-average behavior in the multi-arrangement task from the

939   participants of Experiment 1. Scene-selective ROIs are outlined in white solid lines; the

940   searchlight clusters showing a significant contribution of the functional feature model are

941   outlined in dashed white lines for reference. See Figure 6 for further explanation of the

942   searchlight display. **B)** RDM and MDS plots based on the MVPA patterns in the function model

943   searchlight clusters. RDM rows are ordered as in Figure 1B and category color coding in the

944   MDS plots is as in Figure 1C. **C)** Illustrative exemplars of the four categories that were most

945   dissimilar from other categories within the searchlight-derived clusters depicted in B.

946   **Figure 8** DNN layer and DNN training comparisons, showing layer-by-layer RDM correlations

947   between **A)** an object-trained (ReferenceNet) and a scene-trained (Places) DNN; **B)** both DNNs

948   and the *a priori* selected feature models; **C)** the object-trained DNN and scene-selective ROIs;

949   **D)** the scene-trained DNN and scene-selective ROIs (all comparisons FDR-corrected within

950   ROI: See legend of Figure 2B for explanation of the statistical indicators and noise ceiling).

951 While the decreasing correlation between DNNs indicates stronger task-specificity of higher

952 DNN layers, the original fc7 DNN feature model correlated most strongly with high-level layers

953 of both DNNs. The object-trained and the scene-trained DNN correlated similarly with PPA and

954 OPA, with both showing remarkable good performance for mid-level layers. The RDMs for each

955 individual DNN layer are provided in Source Data 1. Searchlight maps for each layer of the

956 object- and scene trained DNN are provided in Figure 8–video 1 and Figure 8-video 2,

957 respectively.

958 **Figure 8-video 1** Layer-by-layer searchlight results for the object-trained DNN (ReferenceNet).

959 The first half of the movie shows group-average correlation maps for layer 1-8, cluster-corrected

960 for multiple comparisons using Threshold-Free Cluster Enhancement (thresholded on z = 1.64,

961 corresponding to one-sided p < 0.05), overlaid on medial and lateral views of inflated surface

962 reconstructions of both hemispheres of one participant. The second half of the movie shows the

963 same data but without thresholding. Group-level ROIs PPA, OPA and MPA are highlighted in

964 solid white lines.

965 **Figure 8-video 2** Layer-by-layer searchlight results for the scene-trained DNN (Places). See

966 legend of Figure 8-video 1 for details.

967 **References**

968 Aguirre GK, Zarahn E, D'Esposito M. 1998. An area within human ventral cortex sensitive to
969   "building" stimuli: evidence and implications. Neuron. 21:373–383.
970 Baldassano C, Esteva A, Fei-Fei L, Beck DM. 2016. Two distinct scene processing networks
971   connecting vision and memory. eNeuro. 10.1523:1–14.
972 Bar M, Aminoff E. 2003. Cortical analysis of visual context. Neuron. 38:347–358.
973 Bau D, Zhou B, Khosla A, Oliva A, Torralba A. 2017. Network Dissection: Quantifying
974   Interpretability of Deep Visual Representations. In: Computer Vision and Pattern
975   Recognition (CVPR). p. 1–9.
976 Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate : A Practical and Powerful
977   Approach to Multiple Testing. J R Stat Soc B. 57:289–300.
978 Biederman I. 1987. Recognition-by-components: a theory of human image understanding.
979   Psychol Rev. 94:115–147.
980 Bonner MF, Epstein RA. 2017. Coding of navigational affordances in the human visual system.
981   Proc Natl Acad Sci. 201618228.

982   Bracci S, Daniels N, Op de Beeck H. 2017. Task Context Overrules Object- and Category-
983   Related Representational Content in the Human Parietal Cortex. Cereb Cortex. 310–321.
984   Bruss FT. 2000. Sum the odds to one and stop. Ann Probab. 28:1384–1391.
985   Bugatus L, Weiner KS, Grill-Spector K. 2017. Task alters category representations in prefrontal
986   but not high-level visual cortex. Neuroimage. 155:437–449.
987   Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ. 2014.
988   Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual
989   Object Recognition. PLoS Comput Biol. 10.
990   Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. 2016. Comparison of deep neural
991   networks to spatio-temporal cortical dynamics of human visual object recognition reveals
992   hierarchical correspondence. Sci Rep. 6:1–35.
993   Çukur T, Huth AG, Nishimoto S, Gallant JL. 2016. Functional Subdomains within Scene-
994   Selective Cortex: Parahippocampal Place Area, Retrosplenial Complex, and Occipital
995   Place Area. J Neurosci. 36:10257–10273.
996   Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. ImageNet: A large-scale hierarchical
997   image database. 2009 IEEE Conf Comput Vis Pattern Recognit. 248–255.
998   Dilks DD, Julian JB, Paunov AM, Kanwisher N. 2013. The occipital place area is causally and
999   selectively involved in scene perception. J Neurosci. 33:1331–6a.
1000   Downing PE. 2001. A Cortical Area Selective for Visual Processing of the Human Body. Science
1001   (80-). 293:2470–2473.
1002   Epstein R. 2005. The cortical basis of visual scene processing. Vis cogn. 12:954–978.
1003   Epstein RA. 2014. Neural systems for visual scene recognition. In: Bar M,, Kveraga K, editors.
1004   Scene Vision. Cambridge, MA: MIT Press. p. 105–134.
1005   Epstein RA, Parker WE, Feiler AM. 2007. Where am I now? Distinct roles for parahippocampal
1006   and retrosplenial cortices in place recognition. J Neurosci. 27:6141–6149.
1007   Epstein R, Kanwisher N. 1998. A cortical representation of the local visual environment. Nature.
1008   392:598–601.
1009   Erez Y, Duncan J. 2015. Discrimination of Visual Categories Based on Behavioral Relevance in
1010   Widespread Regions of Frontoparietal Cortex. J Neurosci. 35:12383–12393.
1011   Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. 2017. A
1012   Review on Deep Learning Techniques Applied to Semantic Segmentation. Arxiv Prepr.
1013   http://arxiv.org/abs/1704.06857.
1014   Greene MR, Baldassano C, Esteva A, Beck DM. 2016. Visual Scenes Are Categorized by
1015   Function. J Exp Psychol Gen. 145:82–94.
1016   Groen IIA, Ghebreab S, Lamme VAF, Scholte HS. 2012. Spatially pooled contrast responses
1017   predict neural and perceptual similarity of naturalistic image categories. PLoS Comput Biol.
1018   8:e1002726.
1019   Groen IIA, Silson EH, Baker CI. 2017. Contributions of low- and high-level properties to neural
1020   processing of visual scenes in the human brain. Philos Trans R Soc B. 372:1–11.
1021   Gu C, Sun C, Ross DA, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco
1022   S, Sukthankar R, Schmid C, Malik J. 2017. AVA: A Video Dataset of Spatio-temporally
1023   Localized Atomic Visual Actions. bioArchiv. http://arxiv.org/abs/1705.08421.
1024   Güçlü U, van Gerven MAJ. 2015. Deep Neural Networks Reveal a Gradient in the Complexity of
1025   Neural Representations across the Ventral Stream. J Neurosci. 35:10005–10014.
1026   Hafri A, Trueswell JC, Epstein RA. 2017. Neural Representations of Observed Actions
1027   Generalize across Static and Dynamic Visual Input. J Neurosci. 37:3056–3071.
1028   Harel A, Kravitz DJ, Baker CI. 2014. Task context impacts visual object processing differentially
1029   across the cortex. Proc Natl Acad Sci. 962–971.
1030   Hasson U, Levy I, Behrmann M, Hendler T, Malach R. 2002. Eccentricity bias as an organizing
1031   principle for human high-order object areas. Neuron. 34:479–490.
1032   Hebart MN, Bankson BB, Harel A, Baker CI, Cichy RM. 2018. The representational dynamics of

1033   task and object category processing in humans. Elife. DOI: 10.7554/7:e32816 (in press).

1034 Horikawa T, Kamitani Y. 2017. Generic decoding of seen and imagined objects using
1035   hierarchical visual features. Nat Commun. 8:15037.

1036 Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. 2014.
1037   Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22Nd
1038   ACM International Conference on Multimedia. MM '14. New York, NY, USA: ACM. p. 675–
1039   678.

1040 Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human
1041   extrastriate cortex specialized for face perception. J Neurosci. 17:4302–4311.

1042 Khaligh-Razavi SM, Kriegeskorte N. 2014. Deep Supervised, but Not Unsupervised, Models
1043   May Explain IT Cortical Representation. PLoS Comput Biol. 10.

1044 Kravitz DJ, Peng CS, Baker CI. 2011. Real-world scene representations in high-level visual
1045   cortex: it's the spaces more than the places. J Neurosci. 31:7322–7333.

1046 Kriegeskorte N, Mur M. 2012. Inverse MDS: Inferring dissimilarity structure from multiple item
1047   arrangements. Front Psychol. 3:1–13.

1048 Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis - connecting the
1049   branches of systems neuroscience. Front Syst Neurosci. 2:4.

1050 Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional
1051   neural networks. Neural Inf Process Syst. 1097–1105.

1052 Ledoit O, Wolf M. 2004. Honey, I Shrunk the Sample Covariance Matrix. J Portf Manag.
1053   30:110–119.

1054 Lescroart MD, Stansbury DE, Gallant JL. 2015. Fourier power, subjective distance, and object
1055   categories all provide plausible models of BOLD responses in scene-selective visual areas.
1056   Front Comput Neurosci. 9:135.

1057 Lowe MX, Gallivan JP, Ferber S, Cant JS. 2016. Feature diagnosticity and task context shape
1058   activity in human scene-selective cortex. Neuroimage. 125:681–692.

1059 Malcolm GL, Groen IIA, Baker CI. 2016. Making sense of real-world scenes. Trends Cogn Sci.
1060   20:843–856.

1061 Marchette SA, Vass LK, Ryan J, Epstein RA. 2014. Anchoring the neural compass: coding of
1062   local spatial reference frames in human medial parietal lobe. Nat Neurosci. 17:1598–1605.

1063 Martin A, Wiggs CL, Underleider LG, Haxby J V. 1996. Neural correlates of category-specific
1064   knowledge. Nature. 379:649–652.

1065 Micallef L, Rodgers P. 2014. euler APE: Drawing area-proportional 3-Venn diagrams using
1066   ellipses. PLoS One. 9.

1067 Monfort M, Zhou B, Bargal SA, Andonian A, Yan T, Ramakrishnan K, Brown L, Fan Q,
1068   Gutfruend D, Vondrick C, Oliva A. 2018. Moments in Time Dataset: one million videos for
1069   event understanding. Arxiv Prepr. http://arxiv.org/abs/1801.03150.

1070 Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. 2014. A Toolbox for
1071   Representational Similarity Analysis. PLoS Comput Biol. 10.

1072 Oliva A, Torralba A. 2001. Modeling the shape of the scene: A holistic representation of the
1073   spatial envelope. Int J Comput Vis. 42:145–175.

1074 Oosterhof NN, Connolly AC, Haxby J V. 2016. CoSMoMVPA: multi-modal multivariate pattern
1075   analysis of neuroimaging data in Matlab / GNU Octave. Front Neuroinform. 10:1–27.

1076 Park S, Brady TF, Greene MR, Oliva A. 2011. Disentangling scene content from spatial
1077   boundary: complementary roles for the parahippocampal place area and lateral occipital
1078   complex in representing real-world scenes. J Neurosci. 31:1333–1340.

1079 Peelen M V., Downing PE. 2007. The neural basis of visual body perception. Nat Rev Neurosci.
1080   8:636–648.

1081 Peirce JW. 2007. PsychoPy-Psychophysics software in Python. J Neurosci Methods. 162:8–13.

1082 Rajimehr R, Devaney KJ, Bilenko NY, Young JC, Tootell RBH. 2011. The "parahippocampal
1083   place area" responds preferentially to high spatial frequencies in humans and monkeys.

1084   PLoS Biol. 9:e1000608.
1085 Ramakrishnan K, Scholte HS, Groen IIA, Smeulders AWM, Ghebreab S. 2014. Visual
1086   dictionaries as intermediate features in the human brain. Front Comput Neurosci. 8:168.
1087 Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. 2013. OverFeat: Integrated
1088   Recognition, Localization and Detection using Convolutional Networks.
1089 Silson EH, Steel AD, Baker CI. 2016. Scene selectivity and retinotopy in medial parietal cortex.
1090   Front Hum Neurosci. 10:1–17.
1091 Smith SM, Nichols TE. 2009. Threshold-free cluster enhancement: Addressing problems of
1092   smoothing, threshold dependence and localisation in cluster inference. Neuroimage.
1093   44:83–98.
1094 Tootell RBH, Reppas JB, Kwong KK, Rosen BR, Belliveaul JW, Malach R. 1995. Functional
1095   Analysis of Human MT and Related Visual Cortical Areas Using Magnetic Resonance
1096   Imaging. J Neurosci. 15:3215–3230.
1097 Torralba A, Oliva A. 2003. Statistics of natural image categories. Netw Comput Neural Syst.
1098   14:391–412.
1099 Troiani V, Stigliani A, Smith ME, Epstein RA. 2014. Multiple object properties drive scene-
1100   selective regions. Cereb Cortex. 24:883–897.
1101 Turennout M Van, Ellmore T, Martin A. 2000. Long-lasting cortical plasticity in the object naming
1102   system. Nat Neurosci. 3:1329–1335.
1103 van Turennout M, Bielamowicz L, Martin A. 2003. Modulation of Neural Activity during Object
1104   Naming: Effects of Time and Practice. Cereb Cortex. 13:381–391.
1105 Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. 2016. Reliability of dissimilarity
1106   measures for multi-voxel pattern analysis. Neuroimage. 137:188–200.
1107 Walther DB, Caddigan E, Fei-Fei L, Beck DM. 2009. Natural scene categories revealed in
1108   distributed patterns of activity in the human brain. J Neurosci. 29:10573–10581.
1109 Watson DM, Andrews TJ, Hartley T. 2017. A data driven approach to understanding the
1110   organization of high-level visual cortex. Sci Rep. 7:3596.
1111 Wen H, Shi J, Zhang Y, Lu K-H, Cao J, Liu Z. 2017. Neural Encoding and Decoding with Deep
1112   Learning for Dynamic Natural Vision. Cereb Cortex. 1–25.
1113 Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A. 2014. SUN Database: Exploring a Large
1114   Collection of Scene Categories. Int J Comput Vis.
1115 Zeki S, Watson JDG, Lueck CJ, Friston KJ, Kennard C, Frackowiak RSJ. 1991. A direct
1116   demonstration of functional specialization in human visual cortex. J Neurosci. 11:641–649.
1117 Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. 2014. Learning Deep Features for Scene
1118   Recognition using Places Database. Adv Neural Inf Process Syst 27. 487–495.
1119

a

Functions

Objects

building　sky
tree　car
sidewalk　road
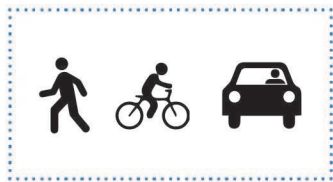
DNN features

b

access road
airplane cabin
bus depot
naval base
pilot house

bindery
control tower
pump room

bamboo forest
butte
dolmen
stilt house
tea garden
volcano
woodland

badminton court
batting cage
putting green
stadium
underwater pool
volleyball court

apse
skyscraper

escalator
hedgerow
lido deck
pier

playroom
youth hostel

bar

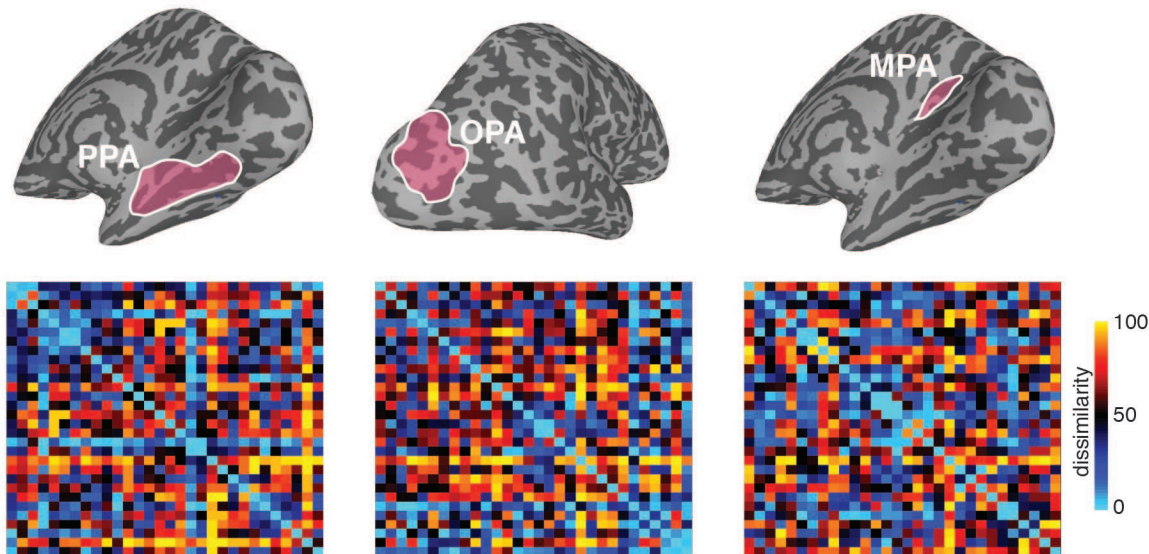Functions

Objects

DNN features

0　　50　　100
dissimilarity
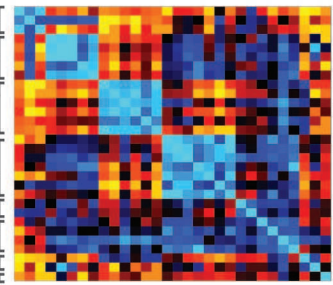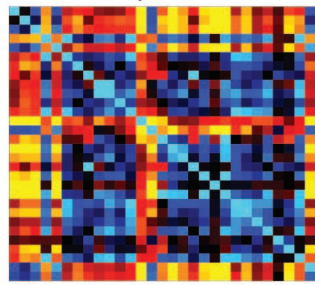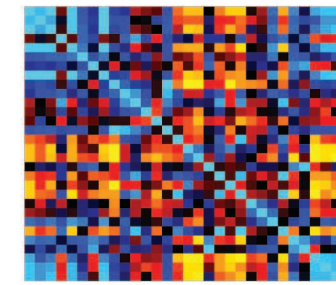
c

"operating"
(machines/consoles)

"transportation"
(roads/vehicles)

"exercise"
(gym halls/pools/courts)

"outdoor pastoral"
(landscapes/gardens)

"navigation/wayfinding"
(tall buildings)

"navigation/self-movement"
(escalator/pier)

"indoor social"
(children)

"indoor social"
(adults)

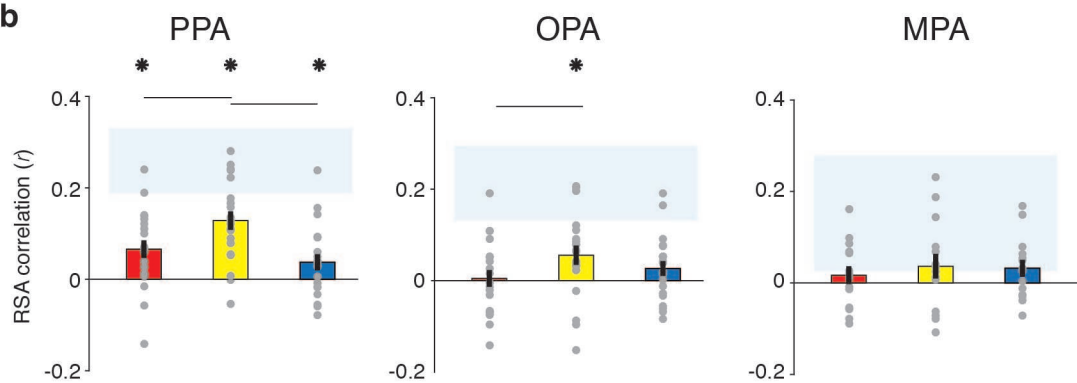MDS dimension 2

MDS dimension 1

a

Start

End

"Arrange according to similarity, whatever that means to you"

b

Average arrangement

0   50   100
dissimilarity

c

RSA correlation (*r*)

Objects   DNN features   Functions

d

subjects

max *r*

e

partial RSA correlation (*r*)

f

Total r²:
50.3%

Objects:
1.4%

4.6%   4.5%

8.4%

Functions:
37.6%

14.5%

DNN features:
29.0%

a

PPA

OPA

MPA

dissimilarity
100
50
0

b

RSA correlation (*r*)

*

*   *

*   *

c

Total: 27.7%

DNN features: 95.6%

Function: 1%

2.7%

<1%

Objects: 0.2%

Total: 14.4%

DNN features: 95.5%

Function: <1%

4.4%

Objects: 6.0%

Total: 1.9%

DNN features: 51.4%

Functions: 5.2%

8.0%

8.1%

3.6%

13.6%

Objects: 11%

d

■ Fixation task (Exp 1)
▨ Naming task (Exp 2)

RSA correlation (*r*)

Objects | DNN features | Functions

RSA correlation (*r*)

RSA correlation (*r*)

a

DNN features

partial r
-0.05 0.05

MPA
PPA
MPA
PPA

OPA
OPA

LH   RH   LH   RH

z (TFCE)
1.64   3.2

b

Functions

a

Multi-arrangement behavior

LH MPA PPA RH

−0.05   0.05
r

1.64   3.2
z (TFCE)

OPA OPA

LH   RH

b

Ventral cluster (bilateral)

dissimilarity
100

50

0

volcano

batting cage
control tower
badminton court

MDS dimension 1

MDS dimension 2

Lateral occipitotemporal cluster (left-lateralized)

volcano

badminton court
batting cage

control tower

c

batting cage

badminton court

control tower

volcano

a ReferenceNet vs. Places

b Objects / DNN features (fc7) / Functions

c ReferenceNet — PPA, OPA, MPA

d Places — PPA, OPA, MPA