

This is a repository copy of *A neural mechanism for contextualizing fragmented inputs during naturalistic vision*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/152312/>

Version: Accepted Version

---

**Article:**

Kaiser, Daniel [orcid.org/0000-0002-9007-3160](https://orcid.org/0000-0002-9007-3160), Turini, Jacopo and Cichy, Radoslaw M (2019) A neural mechanism for contextualizing fragmented inputs during naturalistic vision. eLife. ISSN 2050-084X

<https://doi.org/10.7554/eLife.48182>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

1 **A neural mechanism for contextualizing fragmented inputs during naturalistic vision**

2

3 Daniel Kaiser<sup>1,2,\*</sup>, Jacopo Turini<sup>2,3</sup>, Radoslaw M. Cichy<sup>2,4,5</sup>

4

5 *<sup>1</sup>Department of Psychology, University of York, York, UK*

6 *<sup>2</sup>Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany*

7 *<sup>3</sup>Institute of Psychology, Goethe-Universität Frankfurt, Frankfurt am Main, Germany*

8 *<sup>4</sup>Berlin School of Mind and Brain, Humboldt-Universität Berlin, Berlin, Germany*

9 *<sup>5</sup>Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany*

10

11 *\*Correspondence to:*

12 Dr. Daniel Kaiser

13 Department of Psychology

14 University of York

15 Heslington, York

16 YO10 5DD, UK

17 [danielkaiser.net@gmail.com](mailto:danielkaiser.net@gmail.com)

18

19 **ABSTRACT**

20 With every glimpse of our eyes, we sample only a small and incomplete fragment of the  
21 visual world, which needs to be contextualized and integrated into a coherent scene  
22 representation. Here we show that the visual system achieves this contextualization by  
23 exploiting spatial schemata, that is our knowledge about the composition of natural  
24 scenes. We measured fMRI and EEG responses to incomplete scene fragments and used  
25 representational similarity analysis to reconstruct their cortical representations in space  
26 and time. We observed a sorting of representations according to the fragments' place  
27 within the scene schema, which occurred during perceptual analysis in the occipital place  
28 area and within the first 200ms of vision. This schema-based coding operates flexibly  
29 across visual features (as measured by a deep neural network model) and different types  
30 of environments (indoor and outdoor scenes). This flexibility highlights the mechanism's  
31 ability to efficiently organize incoming information under dynamic real-world conditions.

32

33 **IMPACT STATEMENT**

34 In scene-selective occipital cortex and within 200ms of processing, visual inputs are sorted  
35 according to their typical spatial position within a scene.

## 37 INTRODUCTION

38 During natural vision, the brain continuously receives incomplete fragments of information  
39 that need to be integrated into meaningful scene representations. Here, we propose that  
40 this integration is achieved through contextualization: the brain uses prior knowledge about  
41 where information typically appears in a scene to meaningfully sort incoming information.

42 A format in which such prior knowledge about the world is represented in the brain  
43 is provided by schemata. First introduced to philosophy to explain how prior knowledge  
44 enables perception of the world (Kant, 1781), schemata were later adapted by psychology  
45 (Barlett, 1932; Piaget, 1926) and computer science (Minsky, 1975) as a means to  
46 formalize mechanisms enabling natural and artificial intelligence, respectively.

47 In the narrower context of natural vision, scene schemata represent knowledge  
48 about the typical composition of real-world environments (Mandler, 1984). Scene  
49 schemata for example entail knowledge about the distribution of objects across scenes,  
50 where objects appear in particular locations across the scene and in particular locations  
51 with respect to other objects (Kaiser et al., 2019a; Torralba et al., 2006; Vö et al., 2019;  
52 Wolfe et al., 2011).

53 The beneficial role of such scene schemata was first investigated in empirical  
54 studies of human memory performance, where memory performance is boosted when  
55 scenes are configured in accordance with the schema (Brewer and Treyns, 1981;  
56 Mandler and Johnson, 1976; Mandler and Parker, 1976).

57 Recently however, it has become clear that scene schemata not only organize  
58 memory contents, but also the contents of perception. For example, knowledge about the  
59 structure of the world can be used to generate predictions about a scene's content (Bar,  
60 2009; Henderson, 2017), or to efficiently organize the concurrent representation of multiple  
61 scene elements (Kaiser et al., 2019a; Kaiser et al., 2019b). This position is reinforced by

62 behavioral studies demonstrating a beneficial role of schema-congruent naturalistic stimuli  
63 across a variety of perceptual tasks, such as visual detection (Biederman et al., 1982;  
64 Davenport and Potter, 2004; Stein et al., 2015) and visual search (Kaiser et al., 2014;  
65 Torralba et al., 2006; Vö et al., 2019).

66 Here, we put forward a novel function of scene schemata in visual processing: they  
67 support the contextualization of fragmented sensory inputs. If sensory inputs are indeed  
68 processed in relation to the schema context, scene fragments stemming from similar  
69 typical positions within the scene should be processed similarly and fragments stemming  
70 from different positions should be processed differently. Therefore, the neural  
71 representations of scene fragments should be sorted according to their typical place within  
72 the scene.

73 We tested two hypotheses about this sorting process. First, we hypothesized that  
74 this sorting occurs during perceptual scene analysis, which can be spatiotemporally  
75 pinpointed to scene-selective cortex (Baldassano et al., 2016; Epstein, 2014) and the first  
76 250ms of processing (Cichy et al., 2017; Harel et al., 2016). Second, given that schema-  
77 related effects in behavioral studies (Mandler and Parker, 1976) are more robustly  
78 observed along the vertical dimension, where the scene structure is more rigid (i.e., the  
79 sky is almost always above the ground), we hypothesized that the cortical sorting of  
80 information should primarily occur along the vertical dimension.

81 To test these hypotheses, we used a novel visual paradigm in which participants  
82 were exposed to fragmented visual inputs, and recorded fMRI and EEG data to resolve  
83 brain activity in space and time.

84

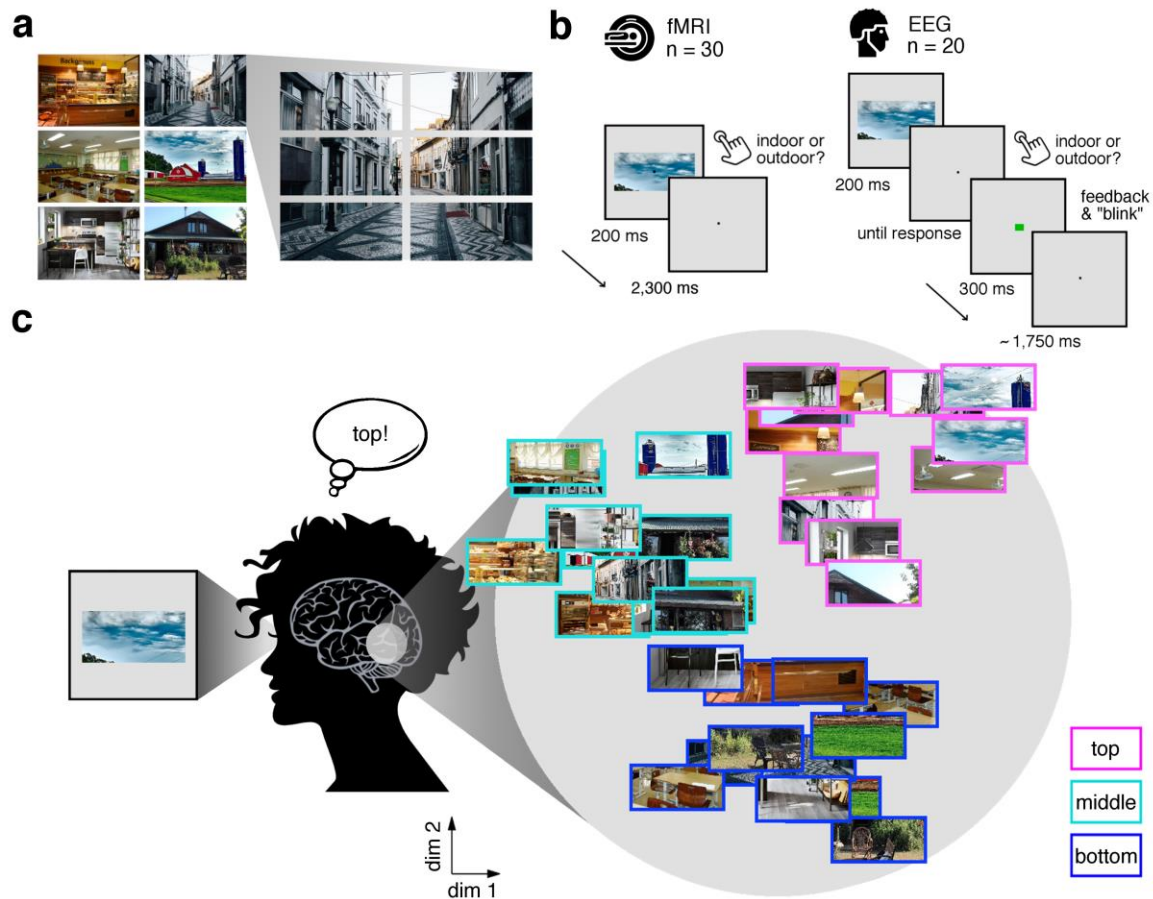
85

86 **RESULTS**

87 In our study, we experimentally mimicked the fragmented nature of naturalistic visual  
88 inputs by dissecting scene images into position-specific fragments. Six natural scene  
89 images (Fig. 1a) were each split into six equally-sized fragments (3 vertical  $\times$  2 horizontal),  
90 resulting in 36 conditions (6 scenes  $\times$  6 fragments). In separate fMRI (n=30) and EEG  
91 (n=20) experiments, participants viewed these fragments at central fixation while  
92 performing an indoor/outdoor categorization task to ensure engagement with the stimulus  
93 (Fig. 1b). Critically, this design allowed us to investigate whether the brain sorts the  
94 fragments with respect to their place in the schema in the absence of explicit location  
95 differences (Fig 1c).

96 To quantify the sorting of fragments during cortical processing we used  
97 spatiotemporally resolved representational similarity analysis (Cichy et al., 2014;  
98 Kriegeskorte et al., 2008). We first extracted representational dissimilarity matrices  
99 (RDMs) from the fMRI and EEG data, which indexed pairwise dissimilarities of the  
100 fragments' neural representations (for details on RDM construction see Figure 2 – Figure  
101 Supplement 1). In the fMRI (Fig. 2a), we extracted spatially-resolved neural RDMs from  
102 scene-selective occipital place area (OPA) and parahippocampal place area (PPA), and  
103 from early visual cortex (V1) (for temporal response profiles in these regions see Figure 2  
104 – Figure Supplement 2). In the EEG (Fig. 2b), we extracted time-resolved neural RDMs  
105 from -200ms to 800ms relative to stimulus onset from posterior EEG electrodes (for other  
106 electrode groups see Figure 2 – Figure Supplements 3-5).

107



108

109 **Fig. 1: Experimental design and rationale of schema-based information sorting.** a,  
 110 The stimulus set consisted of six natural scenes (three indoor, three outdoor). Each scene  
 111 was split into six rectangular fragments. b, During the fMRI and EEG recordings,  
 112 participants performed an indoor/outdoor categorization task on individual fragments.  
 113 Notably, all fragments were presented at central fixation, removing explicit location  
 114 information. c, We hypothesized that the visual system sorts sensory input by spatial  
 115 schemata, resulting in a cortical organization that is explained by the fragments' within-  
 116 scene location, predominantly in the vertical dimension: Fragments stemming from the  
 117 same part of the scene should be represented similarly. Here we illustrate the  
 118 hypothesized sorting in a two-dimensional space. A similar organization was observed in  
 119 multi-dimensional scaling solutions for the fragments' neural similarities (see Figure 1 –  
 120 Figure Supplement 1 and Video 1). In subsequent analyses, the spatiotemporal  
 121 emergence of the schema-based cortical organization was precisely quantified using  
 122 representational similarity analysis (Fig. 2).

123

124 We then quantified schema effects using separate model RDMs for horizontal and  
 125 vertical locations (Fig. 2c). These location RDMs reflected whether pairs of fragments

126 shared the same location or not. We additionally constructed a category model RDM,  
127 which reflected whether pairs of fragments stemmed from the same scene or not.

128 Critically, if cortical information is indeed sorted with respect to scene schemata, we  
129 should observe a neural clustering of fragments that stem from the same within-scene  
130 location – in this case, the location RDM should predict a significant proportion of the  
131 representational organization in visual cortex.

132 To test this, we modeled neural RDMS as a function of the model RDMS using  
133 general linear models, separately for the fMRI and EEG data. The resulting beta weights  
134 indicated to which degree location and category information accounted for cortical  
135 responses in the three ROIs and across time.

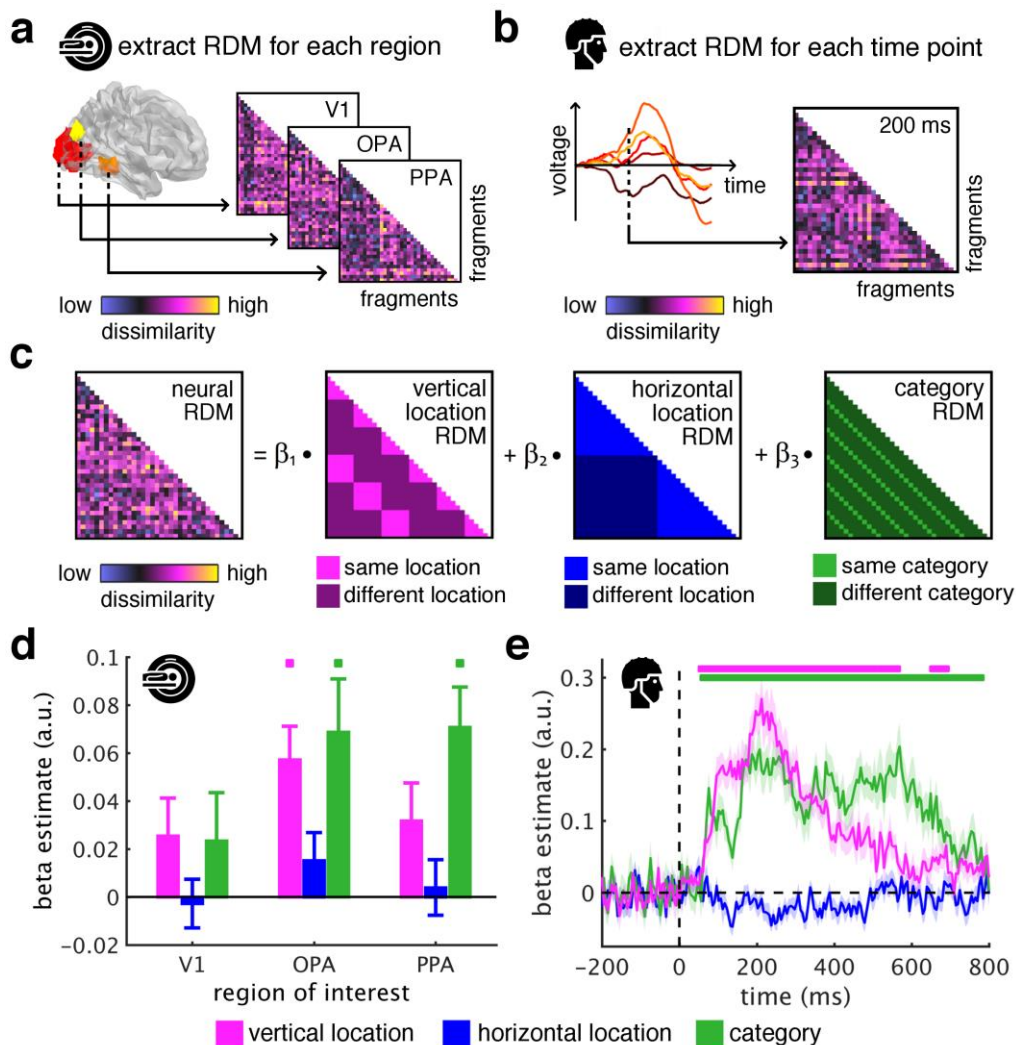
136 The key observation was that the fragments' vertical location predicted neural  
137 representations in OPA ( $t[29]=4.12$ ,  $p<0.001$ ,  $p_{\text{corr}}<0.05$ ), but not in V1 and PPA (test  
138 statistics for all analyses and ROIs are reported in Supplementary file 1) (Fig. 2d) and  
139 between 55ms and 685ms (peak:  $t[19]=9.03$ ,  $p<0.001$ ,  $p_{\text{corr}}<0.05$ ) (Fig. 2e). This vertical-  
140 location organization was consistent across the first and second half of the experiments  
141 (see Figure 2 – Figure Supplement 6) and across all pairwise comparisons along the  
142 vertical axis (see Figure 2 – Figure Supplement 7). No effects were observed for horizontal  
143 location, consistent with more rigid spatial scene structure in the vertical dimension  
144 (Mandler and Parker, 1976). This result provides a first characterization of where and  
145 when incoming information is organized in accordance with scene schemata: in OPA and  
146 rapidly after stimulus onset, scene fragments are sorted according to their origin within the  
147 environment.

148 The schema-based organization co-exists with a prominent scene-category  
149 organization: In line with previous findings (Lowe et al., 2018; Walther et al., 2009),  
150 category was accurately predicted in OPA ( $t[29]=3.12$ ,  $p=0.002$ ,  $p_{\text{corr}}<0.05$ ) and PPA



151 (t[29]=4.26,  $p < 0.001$ ,  $p_{\text{corr}} < 0.05$ ) (Fig. 2d), and from 60ms to 775ms (peak: t[19]=6.39,  
 152  $p < 0.001$ ,  $p_{\text{corr}} < 0.05$ ) (Fig. 2e).

153



154

155 **Fig. 2: Spatial schemata determine cortical representations of fragmented scenes. a,**  
 156 To test where and when the visual system sorts incoming sensory information by spatial  
 157 schemata, we first extracted spatially (fMRI) and temporally (EEG) resolved neural  
 158 representational dissimilarity matrices (RDMs). In the fMRI, we extracted pairwise neural  
 159 dissimilarities of the fragments from response patterns across voxels in the occipital place  
 160 area (OPA), parahippocampal place area (PPA), and early visual cortex (V1). **b,** In the  
 161 EEG, we extracted pairwise dissimilarities from response patterns across electrodes at  
 162 every time point from -200ms to 800ms with respect to stimulus onset. **c,** We modelled the  
 163 neural RDMs with three predictor matrices, which reflected their vertical and horizontal  
 164 positions within the full scene, and their category (i.e., their scene or origin). **d,** The fMRI  
 165 data revealed a vertical-location organization in OPA, but not V1 and PPA. Additionally,

166 the fragment's category predicted responses in both scene-selective regions. **e**, The EEG  
167 data showed that both vertical location and category predicted cortical responses rapidly,  
168 starting from around 100ms. These results suggest that the fragments' vertical position  
169 within the scene schema determines rapidly emerging representations in scene-selective  
170 occipital cortex. Significance markers represent  $p < 0.05$  (corrected for multiple  
171 comparisons). Error margins reflect standard errors of the mean. In further analysis, we  
172 probed the flexibility of this schematic coding mechanism (Fig. 3).

173

174 To efficiently support vision in dynamic natural environments, schematic coding  
175 needs to be flexible with respect to visual properties of specific scenes. The absence of  
176 vertical location effects in V1 indeed highlights that schematic coding is not tied to the  
177 analysis of simple visual features. To more thoroughly probe this flexibility, we additionally  
178 conducted three complementary analyses (Fig. 3).

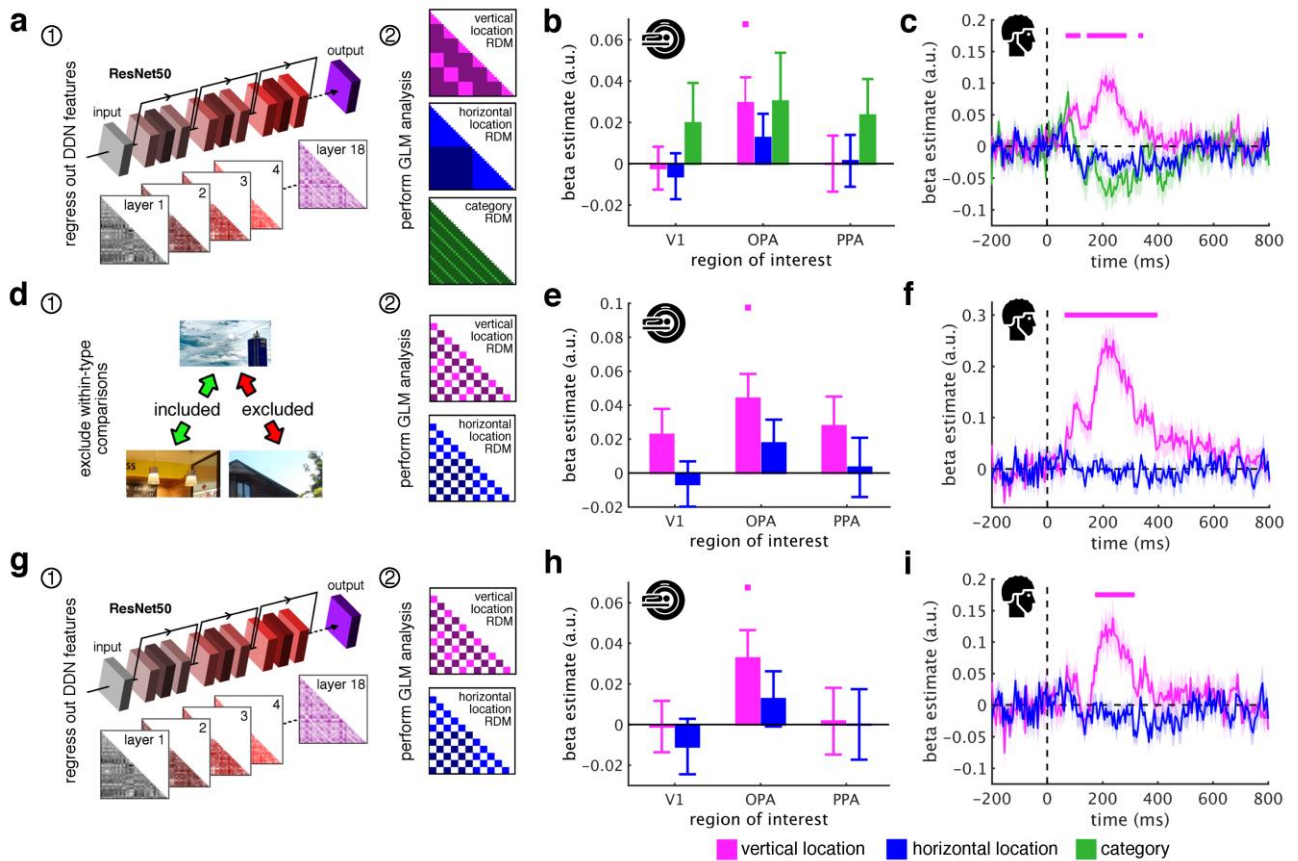
179 First, we tested whether schematic coding is tolerant to stimulus features relevant  
180 for visual categorization. Categorization-related features were quantified using a deep  
181 neural network (DNN; ResNet50), which extracts such features similarly to the brain (Wen  
182 et al., 2018). We removed DNN features by regressing out layer-specific RDMs  
183 constructed from DNN activations (see Materials and Methods for details) (Fig. 3a);  
184 subsequently, we re-estimated location and category information.

185 After removing DNN features, category information was rendered non-significant in  
186 both fMRI and EEG signals. When directly comparing category information before and  
187 after removing the DNN features, we found reduced category information in PPA  
188 ( $t[29]=2.48$ ,  $p=0.0096$ ,  $p_{\text{corr}} < 0.05$ ) and OPA ( $t[29]=1.86$ ,  $p=0.036$ ,  $p_{\text{corr}} > 0.05$ ), and a strong  
189 reduction of category information across time, from 75ms to 775ms (peak  $t[19]=13.0$ ,  
190  $p < 0.001$ ,  $p_{\text{corr}} < 0.05$ ). Together, this demonstrates that categorization-related brain  
191 activations are successfully explained by DNN features (Cichy et al., 2016, 2017; Groen et  
192 al., 2018; Güclü and van Gerven, 2015; Wen et al., 2018), indicating the appropriateness  
193 of our DNN for modelling visual brain activations. Despite the suitability of our DNN model

194 for modelling categorical brain responses, vertical location still accounted for the neural  
195 organization in OPA ( $t[29]=2.37$ ,  $p=0.012$ ,  $p_{\text{corr}}<0.05$ ) (Fig. 3b) and between 75ms and  
196 335ms (peak:  $t[19]=5.06$ ,  $p<0.001$ ,  $p_{\text{corr}}<0.05$ ) (Fig. 3c). Similar results were obtained using  
197 a shallower feed-forward DNN (see Figure 3 – Figure Supplement 1). This result suggests  
198 that schematic coding cannot be explained by categorization-related features extracted by  
199 DNN models.

200 DNN features are a useful control for flexibility towards visual features, because  
201 they cover both low-level and high-level visual features, explaining variance across fMRI  
202 regions and across EEG processing time (see Figure 3 – Figure Supplement 2; see also  
203 Cichy et al., 2016; Güclü & van Gerven, 2015). However, to more specifically control for  
204 low-level features, we used two commonly employed low-level control models: pixel  
205 dissimilarity and GIST descriptors (Oliva and Torralba, 2001). These models neither  
206 explained the vertical location organization nor the category organization in the neural data  
207 (see Figure 3 – Figure Supplement 3). Finally, as an even stronger control of the low-level  
208 features encoded in V1, we used the neural dissimilarity structure in V1 (i.e., the neural  
209 RDMS) as a control model, establishing an empirical neural measure of low-level features.  
210 With V1 housing precise low-level feature representations, this measure should very well  
211 capture the features extracted during the early processing of simple visual features.  
212 However, removing the V1 dissimilarity structure did neither abolish the schematic coding  
213 effects in the OPA nor in the EEG data (see Figure 3 – Figure Supplement 3). This shows  
214 that even if we had control models that approximated V1 representations extremely well –  
215 as well as the V1 representations approximate themselves – these models could not  
216 explain vertical location effects in downstream processing. Together, these results provide  
217 converging evidence that low-level feature processing cannot explain the schematic  
218 coding effects reported here.

219



220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

**Fig. 3: Schematic coding operates flexibly across visual and conceptual scene properties.** **a**, To determine the role of categorization-related visual features in this schematic organization, we regressed out RDMs obtained from 18 layers along the ResNet50 DNN before repeated the three-predictor general linear model (GLM) analysis (Fig. 2c). **b/c**, Removing DNN features abolished category information in fMRI and EEG signals, but not vertical location information. **d**, To test for generalization across different scene types, we restricted location predictor RDMs to comparisons across indoor and outdoor scenes. Due to this restriction, category could not be modelled. **e/f**, In this analysis, vertical location still predicted neural organization in OPA and from 70ms. **g**, Finally, we combined the two analyses: we first regressed out DNN features prior and then modelled the neural RDMs using the restricted predictor RDMs (d). **h**, In this analysis, we still found significant vertical location information in OPA. **i**, Notably, vertical location information in the EEG signals was delayed to after 180ms, suggesting that at this stage schematic coding becomes flexible to visual and conceptual attributes. Significance markers represent  $p < 0.05$  (corrected for multiple comparisons). Error margins reflect standard errors of the mean.

238           Second, we asked whether schematic coding operates flexibly across visually  
239 diverse situations. To test this explicitly we restricted RDMs to comparisons between  
240 indoor and outdoor scenes, which vary substantially in visual characteristics (Oliva and  
241 Torralba, 2003) (Fig. 3d).

242           Vertical location still predicted cortical organization in OPA ( $t[29]=3.05$ ,  $p=0.002$ ,  
243  $p_{\text{corr}}<0.05$ ) (Fig. 3e) and from 70ms to 385ms (peak:  $t[19]=7.47$ ,  $p<0.001$ ,  $p_{\text{corr}}<0.05$ ) (Fig.  
244 3f). The generalization across indoor and outdoor scenes indicates that schematic coding  
245 operates similarly across radically different scenes, suggesting that the mechanism can  
246 similarly contextualize information across different real-life situations.

247           Finally, for a particularly strong test of flexibility, we tested for schematic coding  
248 after removing both DNN features and within-category comparisons (Fig. 3g). In this  
249 analysis, OPA representations were still explained by the fragments' vertical location  
250 ( $t[29]=2.38$ ,  $p=0.012$ ,  $p_{\text{corr}}<0.05$ ) (Fig. 3h). Notably, early schema effects were rendered  
251 non-significant, while vertical location still predicted representations after 180ms (peak:  
252  $t[19]=4.41$ ,  $p<0.001$ ,  $p_{\text{corr}}<0.05$ ) (Fig. 3i), suggesting a high degree of flexibility emerging at  
253 that time. Interestingly, across all analyses, vertical location information was exclusively  
254 found in OPA and always peaked shortly after 200ms (see Supplementary file 2),  
255 suggesting that schematic coding occurs during early perceptual analysis of scenes.

256

257

258 **DISCUSSION**

259 Together, our findings characterize a novel neural mechanism for contextualizing  
260 fragmented inputs during naturalistic vision. The mechanism exploits schemata to sort  
261 sensory inputs into meaningful representations of the environment. This sorting occurs  
262 during perceptual scene analysis in scene-selective OPA and within the first 200ms of  
263 vision, and operates flexibly across changes in visual properties.

264 That schema-based coding can be localized to OPA is consistent with the region's  
265 important role in visual scene processing. Transcranial magnetic stimulation studies  
266 suggest that OPA activation is crucial for various scene perception tasks, such as scene  
267 discrimination (Dilks et al., 2013; Ganaden et al., 2013), navigating through scenes (Julian  
268 et al., 2016) and anticipating upcoming scene information (Gandolfo and Downing, 2019).  
269 Functional MRI work suggest that computations in the OPA include the analysis of spatial  
270 scene layout (Dillon, et al., 2018; Henriksson et al., 2019) and the parsing of local scene  
271 elements like objects and local surfaces (Kamps et al., 2016). Future studies are needed  
272 to clarify which of these computations mediate the schema-based coding described here.

273 As the current study is limited to a small set of scenes, more research is needed to  
274 explore whether schema-based coding generalizes to more diverse contents. It is  
275 conceivable that schema-based coding constitutes a more general coding strategy that  
276 may generalize to other visual contents (such as faces; Henriksson et al., 2015) and non-  
277 visual processing domains: when sensory information is fragmented and spatial  
278 information is unreliable, the brain may use schematic information to contextualize sensory  
279 inputs. This view is in line with Bayesian theories of perception where the importance of  
280 prior information for perceptual inference grows with the noisiness and ambiguity of the  
281 sensory information at hand (Ernst and Banks, 2002; Kersten et al., 2004).

282           The schema-based sorting of scene representations provides a mechanism for  
283 efficient communication between perceptual and cognitive systems: when scene  
284 information is formatted with respect to its role in the environment, it can be efficiently read  
285 out by downstream processes. This idea is consistent with the emerging view that cortical  
286 representations depend on functional interactions with the environment (Bonner and  
287 Epstein, 2017; Groen et al., 2018; Malcolm et al., 2016; Peelen and Downing, 2017).  
288 Under this view, formatting perceptual information according to real-world structure may  
289 allow cognitive and motor systems to efficiently read out visual information that is needed  
290 for different real-world tasks (e.g., immediate action versus future navigation). As the  
291 schema-based sorting of scene information happens already during early scene analysis,  
292 many high-level processes have access to this information.

293           Lastly, our results have implications for computational modelling of vision. While  
294 DNNs trained on categorization accurately capture the representational divide into different  
295 scene categories, they cannot explain the schema-based organization observed in the  
296 human visual system. Although this does not mean that visual features extracted by DNN  
297 models in principle are incapable of explaining schema-based brain representations, our  
298 results highlight that current DNN models of categorization do not use real-world structure  
299 in similar ways as the human brain. In the future, augmenting DNN training procedures  
300 with schematic information (Katti et al., 2019) may improve their performance on real-world  
301 tasks and narrow the gap between artificial and biological neural networks.

302           To conclude, our findings provide the first spatiotemporal characterization of a  
303 neural mechanism for contextualizing fragmented visual inputs. By rapidly organizing  
304 visual information according to its typical role in the world, this mechanism may contribute  
305 to the optimal use of perceptual information for guiding efficient real-world behaviors, even  
306 when sensory inputs are incomplete or dynamically changing.

307

308 **MATERIALS AND METHODS**

309

<b>Key Resources Table</b>				
<b>Reagent type (species) or resource</b>	<b>Designation</b>	<b>Source or reference</b>	<b>Identifiers</b>	<b>Additional information</b>
software, algorithm	CoSMoMVPA	Oosterhof et al., 2016	RRID:SCR_014519	For data analysis
software, algorithm	fieldtrip	Oostenveld et al., 2011	RRID:SCR_004849	For EEG data preprocessing
software, algorithm	MATLAB	Mathworks Inc.	RRID:SCR_001622	For stimulus delivery and data analysis
software, algorithm	Psychtoolbox 3	Brainard, 1997	RRID:SCR_002881	For stimulus delivery
software, algorithm	SPM12	www.fil.ion.ucl.ac.uk/spm/software/spm12/	RRID:SCR_007037	For fMRI data preprocessing

310

311 *Participants*

312 Thirty adults (mean age 23.9 years,  $SD=4.4$ ; 26 females) completed the fMRI experiment and  
313 twenty (mean age 24.0 years,  $SD=4.3$ ; 15 females) completed the EEG experiment. All  
314 participants had normal or corrected-to-normal vision. They all provided informed consent  
315 and received monetary reimbursement or course credits for their participation. All  
316 procedures were approved by the ethical committee of the Department of Education and  
317 Psychology at Freie Universität Berlin (reference 140/2017) and were in accordance with  
318 the Declaration of Helsinki.

319 *Stimuli*



320 The stimulus set (Fig. 1a) consisted of fragments taken from three images of indoor  
321 scenes (bakery, classroom, kitchen) and three images of outdoor scenes (alley, house,  
322 farm). Each image was split horizontally into two halves, and each of the halves was  
323 further split vertically in three parts, so that for each scene six fragments were obtained.  
324 Participants were not shown the full scene images prior to the experiment.

### 325 Experimental design

326 The fMRI and EEG designs were identical, unless otherwise noted. Stimulus presentation  
327 was controlled using the Psychtoolbox (Brainard, 1997; RRID:SCR\_002881). In each trial,  
328 one of the 36 fragments was presented at central fixation (7° horizontal visual angle) for  
329 200ms (Fig. 1b). Participants were instructed to maintain central fixation and  
330 categorize each stimulus as an indoor or outdoor scene image by pressing one of two  
331 buttons.

332 In the fMRI experiment, the inter-trial interval was kept constant at 2,300ms,  
333 irrespective of the participant's response time. In the EEG experiment, after each response  
334 a green or red fixation dot was presented for 300ms to indicate response correctness;  
335 participants were instructed to only blink after the feedback had occurred. Trials were  
336 separated by a fixation interval randomly varying between 1500ms and 2000ms.

337 In the fMRI, participants performed six identical runs. Within each run, each of the 36  
338 scene fragments was shown four times, resulting in 144 trials. Additionally, each run contained 29  
339 fixation trials, where only the central fixation dot was shown. Runs started and ended with brief  
340 fixation periods; the total run duration was 7:30 minutes. In the EEG, each of the 36 fragments  
341 was presented 40 times during the experiment, for a total of 1440 trials, divided into 10  
342 runs. Three participants performed a shorter version of the experiment, with only 20  
343 repetitions of each image (720 trials in total).

344 In both experiments, participants performed very well in the indoor/outdoor  
345 categorization task (fMRI: 94% correct, 658ms mean response time, EEG: 96%, 606ms).

346 Differences in task difficulty across fragments were not related to the neural effects of  
347 interest (Figure 2 – Figure Supplement 8).

#### 348 fMRI recording and preprocessing

349 MRI data was acquired using a 3T Siemens Tim Trio Scanner equipped with a 12-channel head  
350 coil. T2\*-weighted gradient-echo echo-planar images were collected as functional volumes  
351 (TR=2s, TE=30ms, 70° flip angle, 3mm<sup>3</sup> voxel size, 37 slices, 20% gap, 192mm FOV, 64×64 matrix  
352 size, interleaved acquisition). Additionally, a T1-weighted image (MPRAGE; 1mm<sup>3</sup> voxel size) was  
353 obtained as a high-resolution anatomical reference. During preprocessing, the functional volumes  
354 were realigned and coregistered to the T1 image, using MATLAB (RRID:SCR\_014519) and SPM12  
355 (www.fil.ion.ucl.ac.uk/spm/; RRID:SCR\_014519).

#### 356 fMRI region of interest definition

357 We restricted our analyses to three regions of interest (ROIs). We defined scene-selective  
358 occipital place area (OPA; Dilks et al., 2013) and parahippocampal place area (PPA; Epstein and  
359 Kanwisher, 1998) using a functional group atlas (Julian et al., 2012). As a control region, we  
360 defined early visual cortex (V1) using a probabilistic atlas (Wang et al., 2015). All ROIs were  
361 defined in standard space and then inverse-normalized into individual-participant space. For each  
362 ROI, we concatenated the left- and right-hemispheric masks and performed analyses on the joint  
363 ROI.

#### 364 EEG recording and preprocessing

365 The EEG was recorded using an EASYCAP 64-channel system and a Brainvision  
366 actiCHamp amplifier. The electrodes were arranged in accordance with the standard 10-10  
367 system. The data was recorded at a sampling rate of 1000Hz and filtered online between  
368 0.03Hz and 100Hz. All electrodes were referenced online to the Fz electrode. Offline  
369 preprocessing was performed in MATLAB, using the FieldTrip toolbox (Oostenveld et al.,  
370 2011; RRID:SCR\_004849). The continuous EEG data were epoched into trials ranging from

371 200ms before stimulus onset to 800ms after stimulus onset, and baseline corrected by  
372 subtracting the mean of the pre-stimulus interval for each trial and channel separately.  
373 Trials containing movement-related artefacts were automatically identified and removed  
374 using the default automatic rejection procedure implemented in Fieldtrip. Channels  
375 containing excessive noise were removed based on visual inspection. Blinks and eye  
376 movement artifacts were identified and removed using independent components analysis  
377 and visual inspection of the resulting components. The epoched data were down-sampled  
378 to 200Hz.

### 379 Representational Similarity Analysis

380 To model the representational structure of the neural activity related to our stimulus set,  
381 we used representational similarity analysis (RSA; Kriegeskorte et al., 2008). We first  
382 extracted neural RDMs separately for the fMRI and EEG experiments, and then used the  
383 same analyses to model their organization. To retrieve the fragments' position within the  
384 original scene, as well their scene category, we used a regression approach, where we  
385 modeled neural dissimilarity as a linear combination of multiple predictors (Proklova et al.,  
386 2016, 2019).

### 387 *Constructing neural dissimilarity – fMRI*

388 For the fMRI data, we used cross-validated correlations as a measure of pairwise neural  
389 dissimilarity. First, patterns for each ROI were extracted from the functional images  
390 corresponding to the trials of interest. After shifting the activation time course by 3 TRs (i.e., 6s,  
391 accounting for the hemodynamic delay), we extracted voxel-wise activation values for each trial,  
392 from the TR that was closest to the stimulus onset on this trial (for results across 6 TRs with  
393 respect to trial onset, see Figure 2 – Figure Supplement 2). To account for activation  
394 differences between runs, the mean activation across conditions was subtracted from each  
395 voxel's values, separately for each run. For each ROI, response patterns across voxels were used

396 to perform multivariate analyses using the CoSMoMVPA toolbox (Oosterhof et al., 2016;  
397 RRID:SCR\_014519). Then, for each TR separately, we performed correlation-based (Haxby et al.,  
398 2001) multi-voxel pattern analyses (MVPA) for each pair of fragments. These analyses were cross-  
399 validated by repeatedly splitting the data into two equally-sized sets (i.e., half of the runs per set).  
400 For this analysis, we correlated the patterns across the two sets, both within-condition (i.e., the  
401 patterns stemming from the two same fragments and from different sets) and between-  
402 conditions (i.e., the patterns stemming from the two different fragments and from different sets).  
403 These correlations were Fisher-transformed. Then, we subtracted the within- and between-  
404 correlations to obtain a cross-validated correlation measure, where above-zero values reflect  
405 successful discrimination. This procedure was repeated for all possible splits of the six runs.  
406 Performing this MVPA for all pairs of fragments yielded a  $36 \times 36$  representational dissimilarity  
407 matrix (RDM) for each ROI. RDMs' entries reflected the neural dissimilarity between pairs of  
408 fragments (the diagonal remained empty).

#### 409 *Constructing neural dissimilarity – EEG*

410 For the EEG data, we used cross-validated classification accuracies as a measure of pairwise  
411 neural dissimilarity. We thus constructed RDMs across time by performing time-resolved  
412 multivariate decoding analyses (Contini et al., 2017). RDMs were built by computing pair-  
413 wise decoding accuracy for all possible combinations of the 36 stimuli, using the  
414 CoSMoMVPA toolbox (Oosterhof et al., 2016). As we expected the highest classification in  
415 sensors over visual cortex (Battistoni et al., 2018; Kaiser et al., 2016), only 17 occipital and  
416 posterior sensors (O1, O2, Oz, PO3, PO4, PO7, PO8, POz, P1, P2, P3, P4, P5, P6, P7,  
417 P8, Pz) were used in this analysis. We report results for other electrode groups in Figure 2  
418 – Figure Supplement 3-5. For each participant, classification was performed separately for  
419 each time point across the epoch (i.e., with 5ms resolution). The analysis was performed  
420 in a pair-wise fashion: Linear discriminant analysis classifiers were always trained and

421 tested on data from two conditions (e.g., the middle left part of the alley versus the top  
422 right part of the farm), using a leave-one-trial-out partitioning scheme. The training set  
423 consisted of all but one trials for each of the two conditions, while one trial for each of the  
424 two conditions was held back and used for classifier testing. This procedure was repeated  
425 until every trial was left out once. Classifier performance was averaged across these  
426 repetitions. The pairwise decoding analysis resulted in a 36-by-36 neural RDM for each  
427 time point. A schematic description of the RDM construction can be found in Figure 2 –  
428 Figure Supplement 1.

#### 429 *Location and category predictors*

430 We predicted the neural RDMs in a general linear model (GLM; see below) with three  
431 different predictor RDMs (36×36 entries each) (Fig. 2c): In the vertical location RDM, each  
432 pair of conditions is assigned either a value of 0, if the fragments stem from the same  
433 vertical location, or the value 1, if they stem from different vertical locations (for results with  
434 an alternative predictor RDM using Euclidean distances see Figure 2 – Figure Supplement  
435 9). In the horizontal location RDM, each pair of conditions is assigned either a value of 0, if  
436 the fragments stem from the same horizontal location, or a value of 1, if they stem from  
437 different horizontal locations. In the category RDM, each pair of conditions is assigned  
438 either a value of 0, if the fragments stem from the same scene, or a value of 1, if they stem  
439 from different scenes.

440 In an additional analysis, we sought to eliminate properties specific to either the  
441 indoor or outdoor scenes, respectively. We therefore constructed RDMs for horizontal and  
442 vertical location information which only contained comparisons between the indoor and  
443 outdoor scenes. These RDMs were constructed in the same way as explained above, but  
444 all comparisons within the same scene type of scene were removed (Fig. 3d).

#### 445 *Modelling neural dissimilarity*

446 To reveal correspondences between the neural data and the predictor matrices, we used  
447 GLM analyses. Separately for each ROI (fMRI) or time point (EEG), we modelled the  
448 neural RDM as a linear function of the vertical location RDM, the horizontal location RDM,  
449 and the category RDM. Prior to each regression, the neural RDMs and predictor RDMs  
450 were vectorized by selecting all lower off-diagonal elements – the rest of the entries,  
451 including the diagonal, was discarded. Values for the neural RDMs were z-scored.  
452 Separately for each subject and each time point, three beta coefficients (i.e., regression  
453 weights) were estimated. By averaging across participants, we obtained time-resolved  
454 beta estimates for each predictor, showing how well each predictor explains the neural  
455 data over time.

456 Furthermore, we performed an additional GLM analysis with a vertical location  
457 predictor and a horizontal location predictor, where comparisons within indoor- and  
458 outdoor-scenes were removed (Fig. 3d-f); these comparisons were also removed from the  
459 criterion. Using the same procedure as in the previous GLM analysis, we then estimated  
460 the beta coefficients for each predictor at each time point, separately for each subject. For  
461 this analysis, a category RDM could not be constructed, as all comparisons of fragments  
462 from the same scene were eliminated.

#### 463 *Controlling for deep neural network features*

464 To control for similarity in categorization-related visual features, we used a deep neural  
465 network (DNN) model. DNNs have recently become the state-of-the-art model of visual  
466 categorization, as they tightly mirror the neural organization of object and scene  
467 representations (Cichy et al., 2016, 2017; Cichy and Kaiser, 2019; Groen et al., 2018;  
468 Güclü and van Gerven, 2015; Wen et al., 2018). DNNs are similar to the brain as they are  
469 trained using excessive training material while dynamically adjusting the “tuning” of their  
470 connections. Here, we used a DNN that has been trained to categorize images (see  
471 below) on a large number of images and categories, therefore providing us with a high-

472 quality model of how visual features are extracted for efficient categorization. By  
473 comparing DNNs activations and brain responses to the scene fragments, we could  
474 quantify to which extent features routinely extracted for categorization purposes account  
475 for schema-based coding in the human visual system.

476 In a two-step approach, we re-performed our regression analysis after removing the  
477 representational organization emerging from the DNN. First, we used a regression model  
478 to remove the contribution of the dissimilarity structure in the DNN model. This model  
479 included one predictor for each layer extracted from the DNN (i.e., one RDM for each  
480 processing step along the DNN). Estimating this model allowed us to remove the neural  
481 organization explained by the DNN while retaining what remains unexplained (in the  
482 regression residuals). Second, we re-ran the previous regression analyses (see above),  
483 but now the residuals of the DNN regression were used as the regression criterion, so that  
484 only the organization that remained unexplained by the DNN was modeled.

485 As a DNN model, we used a pre-trained version (trained on image categorization for  
486 the ImageNet challenge) of the ResNet50 model (He et al., 2016), as implemented in  
487 MatConvNet (Vedaldi and Lenc, 2015). This model's deeper, residual architecture  
488 outperforms shallower models in approximating visual cortex organization (Wen et al.,  
489 2018). ResNet50 consists of 16 blocks of residual layer modules, where information both  
490 passes through an aggregate of layers within the block, and bypasses the block; then the  
491 residual between the processed and the bypassing information is computed. Additionally,  
492 ResNet50 has one convolutional input layer, and one fully-connected output layer. Here, to  
493 not inflate the number of intercorrelated predictor variables, we only used the final layer of  
494 each residual block, and thus 18 layers in total (16 from the residual blocks, and the input  
495 and output layers). For each layer, an RDM was built using 1-correlation between the  
496 activations of all nodes in the layer, separately for each pair of conditions. For regressing  
497 out the DNN RDMs, we added one predictor for each available RDM. In Figure 3 – Figure

498 Supplement 1, we show that an analysis using the AlexNet architecture (Krizhevsky et al.,  
499 2012) yields comparable results; in Figure 3 – Figure Supplement 2, we additionally  
500 provide information about the DNN model fit across regions and time points.

#### 501 Statistical testing

502 For the fMRI data, we tested the regression coefficients against zero, using one-tailed, one-  
503 sample t-tests (i.e., testing the hypothesis that coefficients were greater than zero). Multiple-  
504 comparison correction was based on Bonferroni-corrections across ROIs. A complete report  
505 of all tests performed on the fMRI data can be found in Supplementary file 1. For the EEG  
506 data, we used a threshold-free cluster enhancement procedure (Smith and Nichols, 2009)  
507 to identify significant effects across time. Multiple-comparison correction was based on a  
508 sign-permutation test (with null distributions created from 10,000 bootstrapping iterations)  
509 as implemented in CoSMoMVPA (Oosterhof et al., 2016). The resulting statistical maps  
510 were thresholded at  $Z > 1.64$  (i.e.,  $p < .05$ , one-tailed against zero). Additionally, we report  
511 the results of one-sided t-tests for all peaks effects. To estimate the reliability of onset and  
512 peak latencies we performed bootstrapping analyses, which are reported in  
513 Supplementary Items 2/3.

#### 514 Data availability

515 Data are publicly available on OSF ([DOI.ORG/10.17605/OSF.IO/H3G6V](https://doi.org/10.17605/OSF.IO/H3G6V)).



516

517 **ACKNOWLEDGEMENTS**

518 D.K. and R.M.C. are supported by Deutsche Forschungsgemeinschaft (DFG) grants  
519 (KA4683/2-1, CI241/1-1, CI241/3-1). R.M.C. is supported by a European Research  
520 Council Starting Grant (ERC-2018-StG).

521

522 **COMPETING INTERESTS**

523 The authors declare no competing interests.

524

525

526 **REFERENCES**

- 527 C. Baldassano, A. Esteva, L. Fei-Fei, D. M. Beck, Two distinct scene processing networks  
528 connecting vision and memory. *eNeuro* **3**, ENEURO.0178-16.2016 (2016).
- 529 M. Bar, The proactive brain: memory for predictions. *Phil. Trans. Royal Soc. B Biol. Sci.*  
530 **364**, 1235-1243 (2009).
- 531 F. C. Barlett, Remembering: a study in experimental and social psychology. (Cambridge  
532 University Press, 1932).
- 533 E. Battistoni, D. Kaiser, C. M. Hickey, M. V. Peelen, The time course of spatial attention  
534 during naturalistic visual search. *Cortex*, doi.org/10.1016/j.cortex.2018.11.018  
535 (2018).
- 536 I. Biederman, R. J. Mezzanotte, J. C. Rabinowitz, Scene perception: detecting and judging  
537 objects undergoing relational violations. *Cogn. Psychol.* **14**, 143-177 (1982).
- 538 M. F. Bonner, R. A. Epstein, Coding of navigational affordances in the human visual  
539 system. *Proc. Natl. Acad. Sci. USA* **114**, 4793-4798 (2017).
- 540 D. H. Brainard, The psychophysics toolbox. *Spat. Vis.* **10**, 433-436 (1997).
- 541 W.F. Brewer, J.C. Treyns, The role of schemata in memory for places. *Cogn. Psychol.*  
542 **13**, 207-230 (1981).
- 543 R. M. Cichy, D. Kaiser, Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**,  
544 305-317 (2019).
- 545 R. M. Cichy, A. Khosla, D. Pantazis, A. Oliva, Dynamics of scene representations in the  
546 human brain revealed by magnetoencephalography and deep neural networks.  
547 *Neuroimage* **153**, 346-358 (2017).
- 548 R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural  
549 networks to spatio-temporal cortical dynamics of human visual object recognition  
550 reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).

551 R. M. Cichy, D. Pantazis, A. Oliva, Resolving human object recognition in space and time.  
552 *Nat. Neurosci.* **17**, 455-462 (2014).

553 E. W. Contini, S. G. Wardle, T. A. Carlson, Decoding the time-course of object recognition  
554 in the human brain: from visual features to categorical decisions. *Neuropsychologia*  
555 **105**, 165-176 (2017).

556 J. L. Davenport, M. C. Potter, Scene consistency in object and background perception.  
557 *Psychol. Sci.* **15**, 559-564 (2004).

558 D. D. Dilks, J. B. Julian, A. M. Paunov, N. Kanwisher, The occipital place area is causally  
559 and selectively involved in scene perception. *J. Neurosci.* **33**, 1331-1336 (2013).

560 R. A. Epstein, "Neural systems for visual scene recognition" in Scene vision, M. Bar, K.  
561 Keveraga, Eds. (MIT Press, 2014).

562 R. A. Epstein, N. Kanwisher, A cortical representation of the local visual environment.  
563 *Nature* **392**, 598-601 (1998).

564 M. O. Ernst, M. S. Banks, Humans integrate visual and haptic information in a  
565 statistically optimal fashion. *Nature* **415**, 429-433 (2002).

566 R. E. Ganaden, C. R. Mullin, J. K. Steeves, Transcranial magnetic stimulation to the  
567 transverse occipital sulcus affects scene but not object processing. *J. Cogn.*  
568 *Neurosci.* **25**, 961-968 (2013).

569 M. Gandolfo, P. E. Downing, Causal evidence for expression of perceptual expectations in  
570 category-selective extrastriate regions. *Curr. Biol.*,  
571 doi.org/10.1016/j.cub.2019.06.024 (2019).

572 I. I. A. Groen, M. R. Greene, C. Baldassano, L. Fei-Fei, D. M. Beck, C. I. Baker, Distinct  
573 contributions of functional and deep neural network features to representational  
574 similarity of scenes in human brain and behavior. *eLife* **7**, e32962 (2018).

575 U. Güçlü, M. A. van Gerven, Deep neural networks reveal a gradient in the complexity  
576 of neural representations across the ventral stream. *J. Neurosci.*, **35**, 10005-  
577 10014 (2015).

578 A. Harel, I. I. A. Groen, D. J. Kravitz, L. Y. Deouell, C. I. Baker, The temporal dynamics of  
579 scene processing: A multifaceted EEG investigation. *eNeuro* **3**, ENEURO.0139-  
580 16.2016 (2016).

581 J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, P. Pietrini, Distributed and  
582 overlapping representations of faces and objects in ventral temporal cortex. *Science*  
583 **293**, 2425-2430 (2001).

584 K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition.  
585 *Proceedings of the IEEE Conference on Computer Vision and Pattern*  
586 *Recognition, Las Vegas, NV, USA*, 770-778 (2016).

587 J. Henderson, Gaze control as prediction. *Trends Cogn. Sci.* **21**, 15-23 (2017).

588 L. Henriksson, M. Mur, N. Kriegeskorte, Faciotopy – A face-feature map with face-like  
589 topography in the human occipital face area. *Cortex* **72**, 156-167 (2015).

590 L. Henriksson, M. Mur, N. Kriegeskorte, Rapid invariant encoding of scene layout in  
591 human OPA. *Neuron* **103**, 161-171.

592 J. B. Julian, E. Fedorenko, J. Webster, N. Kanwisher, An algorithmic method for  
593 functionally defining regions of interest in the ventral visual pathway. *Neuroimage*  
594 **60**, 2357-2364 (2012).

595 J. B. Julian, J. Ryan, R. H. Hamilton, R. A. Epstein, The occipital place area is causally  
596 involved in representing environmental boundaries during navigation. *Curr. Biol.* **26**,  
597 1104-1109 (2016).

598 D. Kaiser, G. Häberle, R. M. Cichy, Cortical sensitivity to natural scene structure. *bioRxiv*,  
599 doi.org/10.1101/613885 (2019b).

600 D. Kaiser, N. N. Oosterhof, M. V. Peelen, The neural dynamics of attentional selection in  
601 natural scenes. *J. Neurosci.* **36**, 10522-10528 (2016).

602 D. Kaiser, T. Stein, M. V. Peelen, Object grouping based on real-world regularities  
603 facilitates perception by reducing competitive interactions in visual cortex. *Proc.*  
604 *Natl. Acad. Sci. U.S.A.* **111**, 11217-11222 (2014).

605 D. Kaiser, G. L. Quek, R. M. Cichy, M. V. Peelen, Object vision in a structured world.  
606 *Trends Cogn. Sci.* **23**, 672-685 (2019a).

607 F. S. Kamps, J. B. Julian, J. Kubilius, N. Kanwisher, D. D. Dilks, The occipital place area  
608 represents the local elements of scenes. *Neuroimage* **132**, 417-424 (2016).

609 I. Kant, Kritik der reinen Vernunft. (Johann Friedrich Hartknoch, 1781).

610 H. Katti, M. V. Peelen, S. P. Arun, Machine vision benefits from human contextual  
611 expectations. *Sci. Rep.* **9**, 2112 (2019).

612 D. Kersten, P. Mamassian, A. Yuille, Object perception as Bayesian inference. *Annu.*  
613 *Rev. Psychol.* **55**, 271-304 (2004).

614 N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis – connecting  
615 the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).

616 A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep  
617 convolutional neural networks. *Advances in neural information processing*  
618 *systems*, **25**, 1097-1105 (2012).

619 M. X. Lowe, J. Rajsic, S. Ferber, D. B. Walther, Discriminating scene categories from brain  
620 activity within 100 milliseconds. *Cortex* **106**, 275-287 (2018).

621 M. X. Lowe, J. Rajsic, J. P. Gallivan, S. Ferber, J. S. Cant, Neural representation of  
622 geometry and surface properties in object and scene perception. *Neuroimage* **157**,  
623 586-597.

624 G. L. Malcolm, I. I. A. Groen, C. I. Baker, Making sense of real-world scenes. *Trends*  
625 *Cogn. Sci.* **20**, 843-856 (2016).

626 J. M. Mandler, *Stories, scripts and scenes: aspects of schema theory.* (L. Erlbaum, 1984).

627 J. M. Mandler, N. S. Johnson, Some of the thousand words a picture is worth. *J. Exp.*  
628 *Psychol. Hum. Learn. Mem.* **2**, 529–540 (1976).

629 J. M. Mandler, R. E. Parker, Memory for descriptive and spatial information in complex  
630 pictures. *J. Exp. Psychol. Hum. Learn. Mem.* **2**, 38-48 (1976).

631 M. Minsky, “A framework for representing knowledge” in *The psychology of computer*  
632 *vision*, P. Winston, Ed. (McGraw-Hill, 1975).

633 A. Oliva, A. Torralba, Modelling the shape of the scene: a holistic representation of the  
634 spatial envelope. *Int. J. Comput. Vis.* **42**, 145-175 (2001).

635 A. Oliva, A. Torralba, Statistics of natural image categories. *Network* **14**, 391-412 (2003).

636 R. Oostenveld, P. Fries, E. Maris, J. M. Schoffelen, FieldTrip: Open source software for  
637 advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput.*  
638 *Intell. Neurosci.* **2011**, 156869 (2011).

639 N. N. Oosterhof, A. C. Connolly, J. V. Haxby, CoSMoMVPA: Multi-modal multivariate  
640 pattern analysis of neuroimaging data in Matlab/GNU Octave. *Front. Neuroinform.*  
641 **10**, 20 (2016).

642 M. V. Peelen, P. E. Downing, Category selectivity in human visual cortex: Beyond visual  
643 object recognition. *Neuropsychologia* **105**, 177-183 (2017).

644 J. Piaget, *The language and thought of the child.* (Keagan Paul, Trench, Trubner & Co,  
645 1926).

646 D. Proklova, D. Kaiser, M. V. Peelen, Disentangling representations of object shape and  
647 object category in human visual cortex: the animate-inanimate distinction. *J. Cogn.*  
648 *Neurosci.* **28**, 680-692 (2016).

649 D. Proklova, D. Kaiser, M. V. Peelen, MEG sensor patterns reflect perceptual but not  
650 categorical similarity of animate and inanimate objects. *Neuroimage* **192**, 167-177  
651 (2019).

652 D. E. Rumelhart, "Schemata: the building blocks of cognition" in Theoretical issues in  
653 reading comprehension, R. J. Spiro et al., Eds. (L. Erlbaum, 1980).

654 S. M. Smith, T. E. Nichols, Threshold-free cluster enhancement: addressing problems of  
655 smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*  
656 **44**, 83-98 (2009).

657 T. Stein, D. Kaiser, M. V. Peelen, Interobject grouping facilitates visual awareness. *J. Vis.*  
658 *15*, 10 (2015).

659 A. Torralba, A. Oliva, M. S. Castelhana, J. M. Henderson, Contextual guidance of eye  
660 movements and attention in real-world scenes: the role of global features in  
661 objects search. *Psychol. Rev.* **113**, 766-786 (2006).

662 A. Vedaldi, K. Lenc, MatConvNet – convolutional neural networks for Matlab. *Proceedings*  
663 *of the ACM International Conference on Multimedia* (2015).

664 M. L.-H. Võ, S. E. P. Boettcher, D. Draschkow, Reading scenes: How scene grammar  
665 guides attention and aids perception in real-world environments. *Curr. Opin.*  
666 *Psychol.* **29**, 205-210 (2019).

667 D. B. Walther, E. Caddigan, L. Fei-Fei, D. M. Beck, Natural scene categories revealed in  
668 distributed patterns of activity in the human brain. *J. Neurosci.* **29**, 10573-10581  
669 (2009).

670 L. Wang, R. E. Mruzek, M. J. Arcaro, S. Kastner, Probabilistic maps of visual topography  
671 in human cortex. *Cereb. Cortex* **25**, 3911-3931 (2015).

672 H. Wen, J. Shi, W. Chen, Z. Liu, Deep residual network predicts cortical representation  
673 and organization of visual features for rapid categorization. *Sci. Rep.* **8**, 3752  
674 (2018).

675 J. M. Wolfe, M. L.-H. Võ, K. K. Evans, M. R. Greene, Visual search in scenes involves  
676 selective and nonselective pathways. *Trends Cogn. Sci.* **15**, 77-84 (2011).

677

678 **SUPPLEMENTARY INFORMATION**

679

680 **A neural mechanism for contextualizing fragmented inputs during naturalistic vision**

681 *Kaiser, Turini, & Cichy*

682 *page*

683 Supplementary Figures:

684 *Figure 1 – Figure Supplement 1: MDS visualization of neural RDMS* 34

685 *Figure 2 – Figure Supplement 1: Details on neural dissimilarity construction* 35

686 *Figure 2 – Figure Supplement 2: fMRI response time courses* 36

687 *Figure 2 – Figure Supplement 3: Pairwise decoding across EEG electrode groups* 37

688 *Figure 2 – Figure Supplement 4: RSA using central EEG electrodes* 38

689 *Figure 2 – Figure Supplement 5: RSA using anterior EEG electrodes* 39

690 *Figure 2 – Figure Supplement 6: Vertical location effects across experiment halves* 40

691 *Figure 2 – Figure Supplement 7: Pairwise comparisons along the vertical axis* 41

692 *Figure 2 – Figure Supplement 8: Controlling for task difficulty* 42

693 *Figure 2 – Figure Supplement 9: Categorical versus Euclidean vertical*  
694 *location predictors* 43

695 *Figure 3 – Figure Supplement 1: AlexNet as a model of visual categorization* 44

696 *Figure 3 – Figure Supplement 2: DNN model fit* 45

697 *Figure 3 – Figure Supplement 3: Low-level control models* 46

698

699 Supplementary Items:

700 *Supplementary file 1: Complete statistical report for fMRI results* 47

701 *Supplementary file 2: Estimating EEG peak latencies* 48

702 *Supplementary file 3: Estimating EEG onset latencies* 49



703

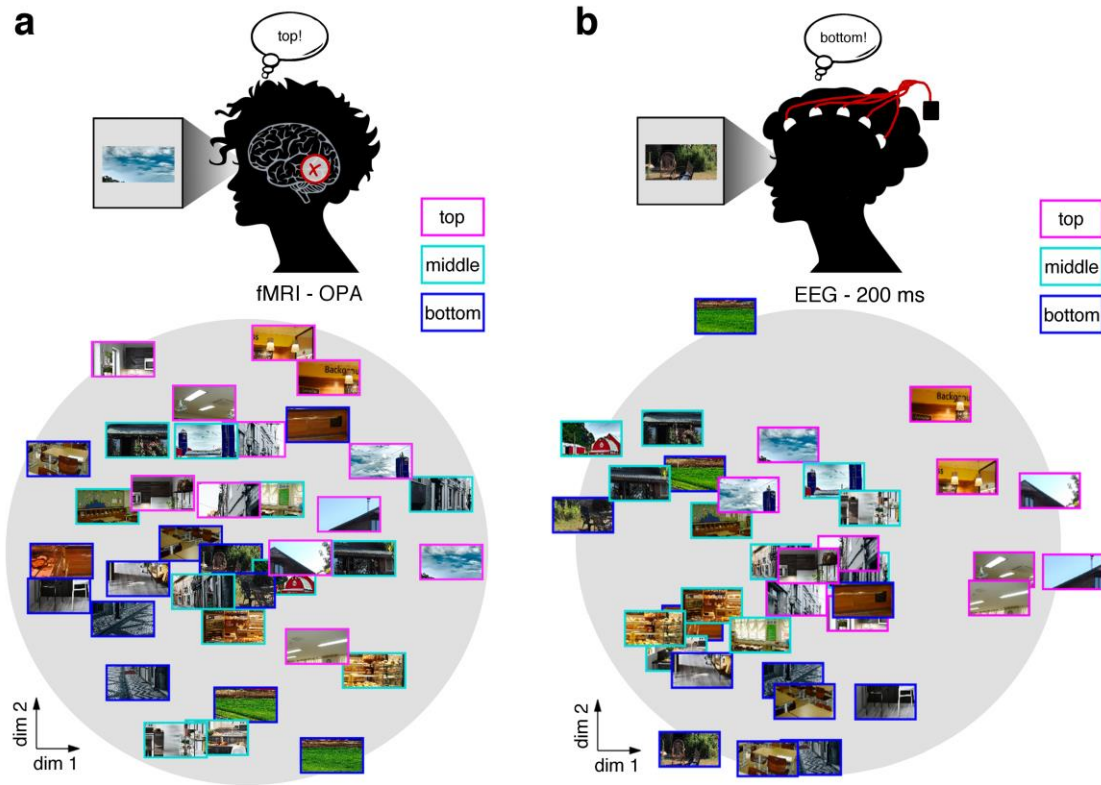
704 Supplementary Videos:

705 *Video 1: Time-resolved MDS visualization of EEG RDMs*

50

706

707 Figure 1 – Figure Supplement 1



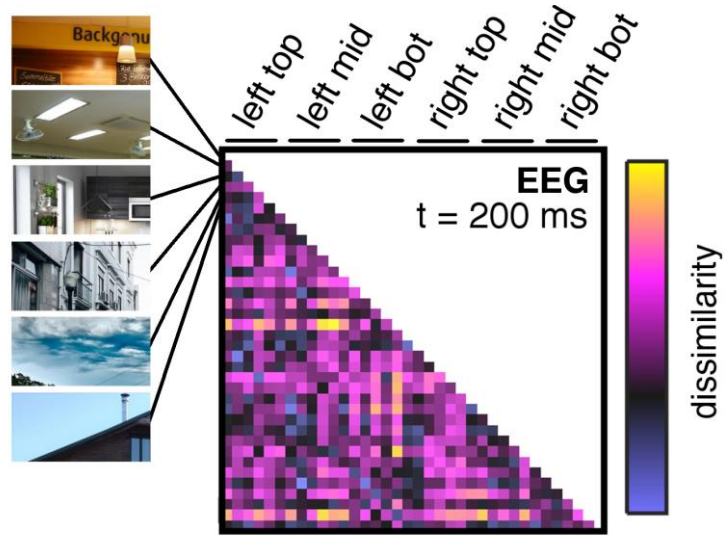
708

709 **MDS visualization of neural RDMs.** a/b, A multi-dimensional scaling (MDS) of the  
710 fragments' neural similarity in OPA (a) and after 200ms of processing (b) revealed a  
711 sorting according to vertical location, which was visible in a two-dimensional solution. This  
712 visualization suggests that schemata are a prominent organizing principle for  
713 representations in OPA and after 200ms of vision. A time-resolved MDS for the EEG data  
714 can be found in Video 1.

715

716

717 Figure 2 – Figure Supplement 1

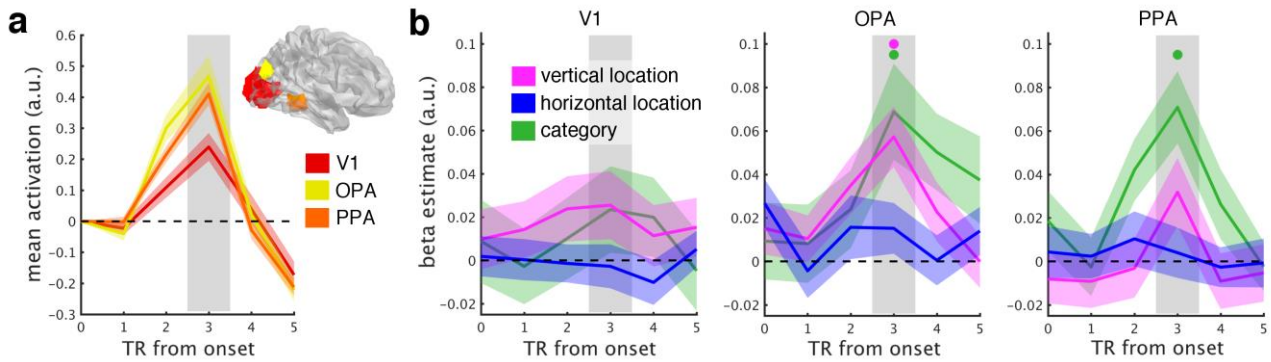


718

719 **Details on neural dissimilarity construction.** Pairwise neural dissimilarity values were  
720 into representational dissimilarity matrices (RDMs), so that for every time point one 36X36  
721 matrix containing estimates of neural dissimilarity was available. Here, an example RDM at  
722 200ms post-stimulus is shown, which exemplifies the ordering of fragment combinations  
723 for all RDMs.

724

725 *Figure 2 – Figure Supplement 2*

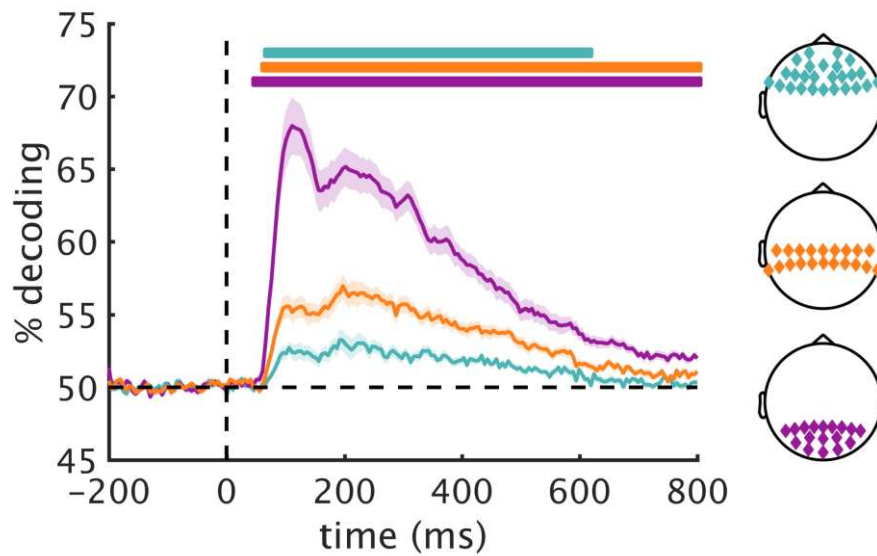


726

727 **fMRI response time courses.** a, Functional MRI data were analyzed in three regions of  
728 interest (here shown on the right hemisphere): primary visual cortex (V1), occipital place  
729 area (OPA), and parahippocampal place area (PPA). Each of these ROIs showed reliable  
730 net responses to the fragments, peaking 3 TRs after stimulus onset. The activation time  
731 courses were baseline-corrected by subtracting the activation from the first two TRs. b,  
732 GLM analysis across the response time course. Most prominently after 3 TRs, the neural  
733 organization in OPA was explained by the fragments' vertical location, reflecting a neural  
734 coding in accordance with spatial schemata. Additionally, scene category predicted neural  
735 organization in OPA and PPA. Error margins reflect standard errors of the mean.  
736 Significance markers represent  $p < 0.05$  (corrected for multiple comparisons across ROIs).

737

738 *Figure 2 – Figure Supplement 3*

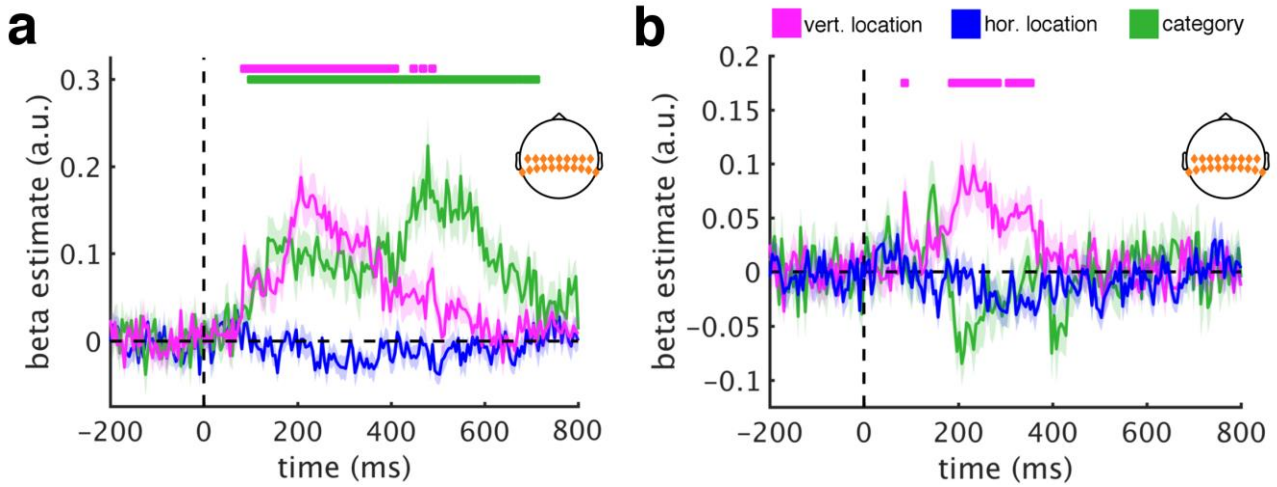


739

740 **Pairwise decoding across electrode groups.** Based on previous studies on multivariate  
741 decoding of visual information, we restricted our main analysis to a group of posterior  
742 electrodes (where we expected the strongest effects). For comparison, we also analyzed  
743 data in central and anterior electrode groups. The central group consisted of 20 electrodes  
744 (C3, TP9, CP5, CP1, TP10, CP6, CP2, Cz, C4, C1, C5, TP7, CP3, CPz, CP4, TP8, C6,  
745 C2, T7, T8) and the anterior group consisted of 26 electrodes (F3, F7, FT9, FC5, FC1,  
746 FT10, FC6, FC2, F4, F8, Fp2, AF7, AF3, AFz, F1, F5, FT7, FC3, FCz, FC4, FT8, F6, F2,  
747 AF4, AF8, Fpz). RDMs were constructed in an identical fashion to the posterior group used  
748 for the main analyses (Figure 2 – Figure Supplement 1). We computed general  
749 discriminability of the 36 scene fragments in the three groups by averaging all off-diagonal  
750 elements of the RDMs. As expected, the resulting time courses of pair-wise discriminability  
751 revealed the strongest overall decoding in the posterior group, followed by the central and  
752 anterior groups. RSA results for these electrodes are found in Figure 2 – Figure  
753 Supplements 4/5. Significance markers represent  $p < 0.05$  (corrected for multiple  
754 comparisons). Error margins reflect standard errors of the mean.

755

756 *Figure 2 – Figure Supplement 4*

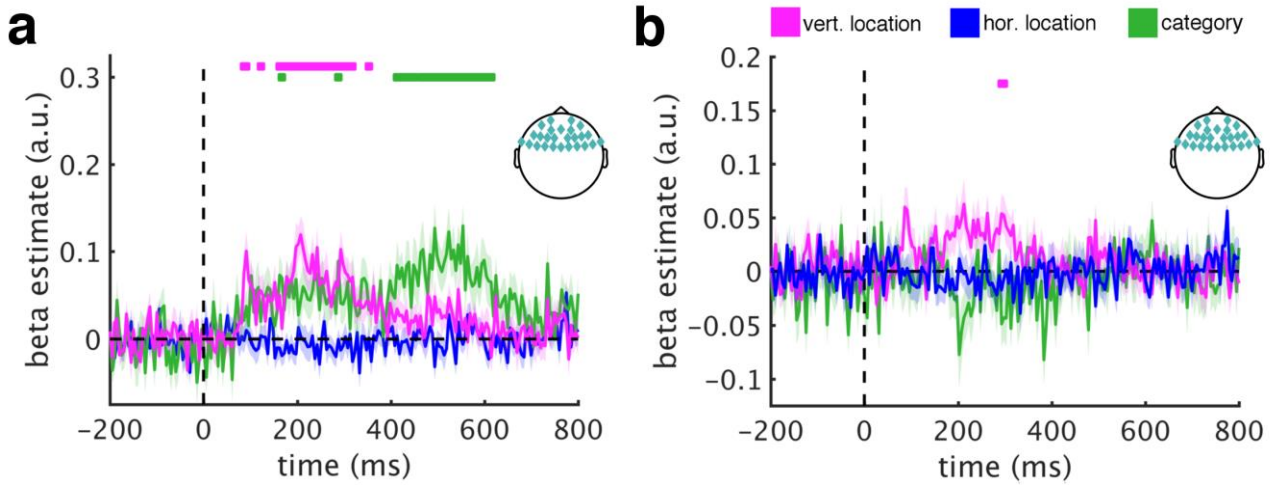


757

758 **RSA using central electrodes. a/b**, Repeating the main RSAs for the central electrode  
759 group yielded a similar pattern as the posterior group, revealing both vertical location  
760 information (from 85ms to 485ms) and category information (from 100ms to 705ms). **c/d**,  
761 Removing DNN features abolished category information, but not vertical location  
762 information, most prominently between 185ms and 350ms. This result is consistent with  
763 the schematic coding observed for posterior signals. Significance markers represent  
764  $p < 0.05$  (corrected for multiple comparisons). Error margins reflect standard errors of the  
765 mean.

766

767 *Figure 2 – Figure Supplement 5*

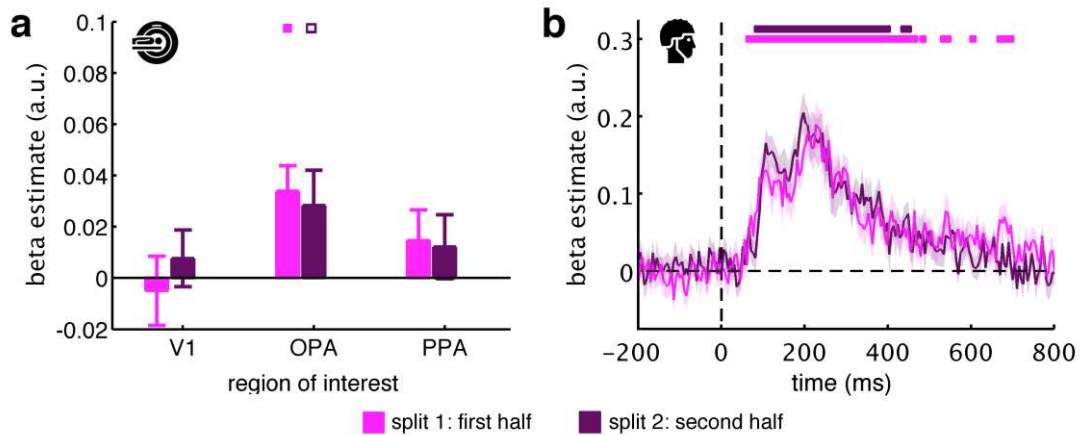


768

769 **RSA using anterior electrodes. a/b**, Also responses recorded from the anterior group  
770 yielded both vertical location information (from 85ms to 350ms) and category information  
771 (from 165ms to 610ms). **c/d**, In contrast to the other electrode groups, removing DNN  
772 features rendered location and category information insignificant, suggesting that they are  
773 not primarily linked to sources in frontal brain areas. This observation also excludes  
774 explanations based on oculomotor confounds. Significance markers represent  $p < 0.05$   
775 (corrected for multiple comparisons). Error margins reflect standard errors of the mean.

776

777 *Figure 2 – Figure Supplement 6*



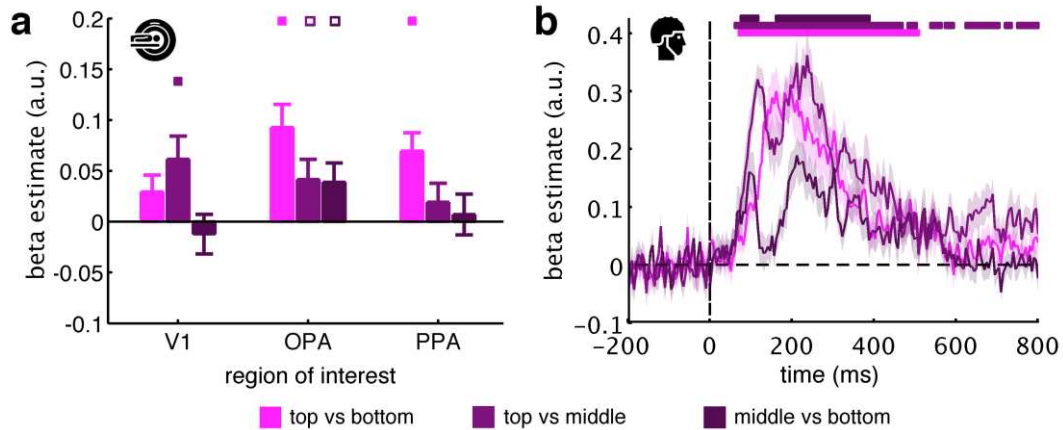
778

779 **Vertical location effects across experiment halves.** We interpret the vertical location  
780 organization in the neural data as reflecting prior schematic knowledge about scene  
781 structure. Alternatively, however, the vertical location organization could in principle result  
782 from learning the composition of the scenes across the experiment. In the latter case, one  
783 would predict that vertical location effects should primarily occur late in the experiment  
784 (e.g., in the second half), and less so towards the beginning (e.g., in the first half). To test  
785 this, we split into halves both the fMRI data (three runs each) and the EEG data (first  
786 versus second half of trials) and for each half modeled the neural data as a function of the  
787 vertical and horizontal location and category predictors. **a**, For the fMRI data, we found  
788 significant vertical location information in the OPA for in the first half ( $t[29]=3.46$ ,  $p<0.001$ ,  
789  $p_{\text{corr}}<0.05$ ) and a trending effect for the second half ( $t[29]=2.07$ ,  $p=0.024$ ,  $p_{\text{corr}}>0.05$ ). No  
790 differences between the splits were found in any region (all  $t[29]<0.90$ ,  $p>0.37$ ). **b**, For the  
791 EEG data, we also found very similar results for the two splits, with no significant  
792 differences emerging at any time point. Together, these results suggest that the vertical  
793 location organization cannot solely be explained by extensive learning over the course of  
794 the experiment. Significance markers represent  $p<0.05$  (corrected for multiple  
795 comparisons). Empty markers represent  $p<0.05$  (uncorrected). Error margins reflect  
796 standard errors of the mean.



797

798 *Figure 2 – Figure Supplement 7*



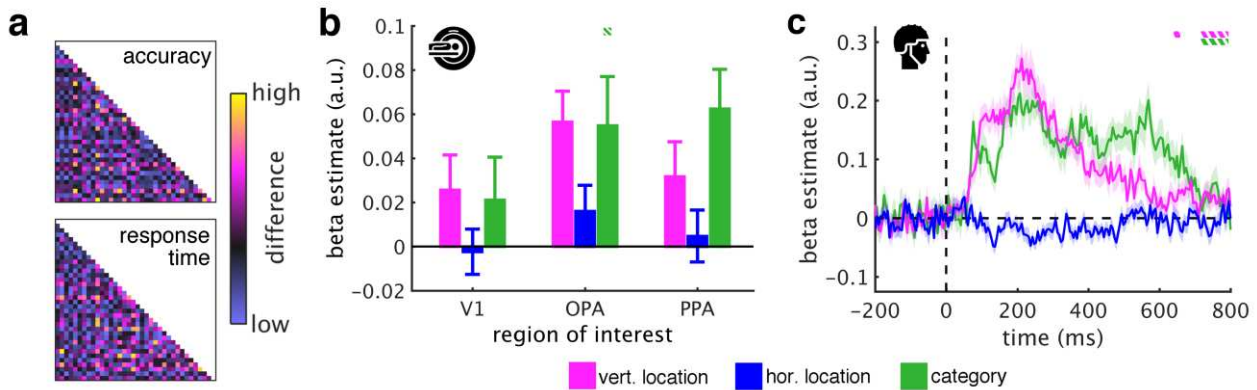
799

800 **Pairwise comparisons along the vertical axis.** To test whether vertical location  
801 information can be observed across all three vertical bins, we modelled the neural data as  
802 a function of the fragments' vertical location, now separately for each pairwise comparison  
803 along the vertical axis (i.e., top versus bottom, top versus middle, and middle versus  
804 bottom). **a**, For the fMRI data, we only found consistent evidence for vertical location  
805 information in the OPA: top versus bottom ( $t[29]=4.10$ ,  $p<0.001$ ,  $p_{\text{corr}}<0.05$ ), top versus  
806 middle ( $t[29]=2.13$ ,  $p=0.021$ ,  $p_{\text{corr}}>0.05$ ), middle versus bottom ( $t[29]=2.06$ ,  $p=0.024$ ,  
807  $p_{\text{corr}}>0.05$ ). Although the effect was numerically bigger for top versus bottom, we did not  
808 find a significant difference between the three pairwise comparisons in OPA ( $F[2,58]=2.71$ ,  
809  $p=0.075$ ). **b**, For the EEG data, we found significant vertical location information for all  
810 three comparisons. Here, the middle-versus-bottom comparison yielded the weakest  
811 effect, which was significantly smaller than the effect for top versus bottom from 120ms  
812 and 195ms and significantly smaller than the effect for top versus middle from 110ms to  
813 285ms. Together, these results suggest that schematic coding can be observed  
814 consistently across the different comparisons along the vertical axis, although  
815 comparisons including the top fragments yielded stronger effects. Significance markers  
816 represent  $p<0.05$  (corrected for multiple comparisons). Empty markers represent  $p<0.05$   
817 (uncorrected). Error margins reflect standard errors of the mean.

818

819

820 *Figure 2 – Figure Supplement 8*

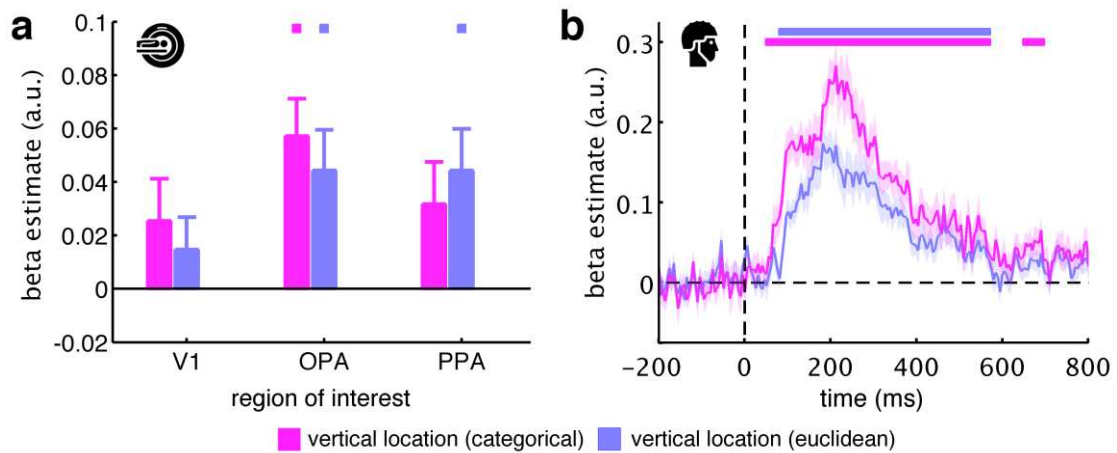


821

822 **Controlling for task difficulty.** **a**, To control for task difficulty effects in the indoor/outdoor  
823 classification task, we computed paired t-tests between all pairs of fragments, separately  
824 for their associated accuracies and response times. We then constructed two predictor  
825 RDMs that contained the t-values of the pairwise tests between the fragments: For each  
826 pair of fragments, these t-values corresponded to dissimilarity in task difficulty (e.g.,  
827 comparing two fragments associated with similarly short categorization response times  
828 would yield a low t-value, and thus low dissimilarity). This was done separately for the  
829 fMRI and EEG experiments (matrices from the EEG experiment are shown). The accuracy  
830 and response time RDMs were mildly correlated with the category RDM (fMRI: accuracy:  
831  $r=0.10$ , response time:  $r=0.15$ ; EEG: accuracy:  $r=0.17$ , response time:  $r=0.16$ ), but not with  
832 the vertical location RDM (fMRI: both  $r<0.01$ , EEG: both  $r<0.01$ ). After regressing out the  
833 task difficulty RDMs, we found highly similar vertical location and category information as  
834 in the previous analyses (Fig. 3b/c). **b**, In the fMRI, only category information in OPA was  
835 significantly reduced when task difficulty was accounted for. **c**, In the EEG, towards the  
836 end of the epoch – when participants responded – location and category information were  
837 decreased. This shows that the effects of schematic coding – emerging around 200ms  
838 after onset – cannot be explained by differences in task difficulty. The dashed significance  
839 markers represent significantly reduced information (compared to the main analyses, Fig.  
840 3b/c) at  $p<0.05$  (corrected for multiple comparisons).

841

842 *Figure 2 – Figure Supplement 9*



843

844 **Categorical versus Euclidean vertical location predictors.** We defined our vertical

845 location predictor as categorical, assuming that top, middle, and bottom fragments are

846 coded distinctly in the human brain. An alternative way of constructing the vertical location

847 predictor is in terms of the fragments' Euclidean distances, where fragments closer

848 together along the vertical axis (e.g., top and middle) are represented more similarly than

849 fragments further apart (e.g., top and bottom). **a**, For the fMRI data, we found that the

850 categorical and Euclidean predictors similarly explained the neural data, with no statistical

851 differences between them (all  $t[29] < 1.15$ ,  $p > 0.26$ ). **b**, For the EEG data, we found that both

852 predictors explained the neural data well. However, the categorical predictor revealed

853 significantly stronger vertical location information from 75ms to 340ms, suggesting that, at

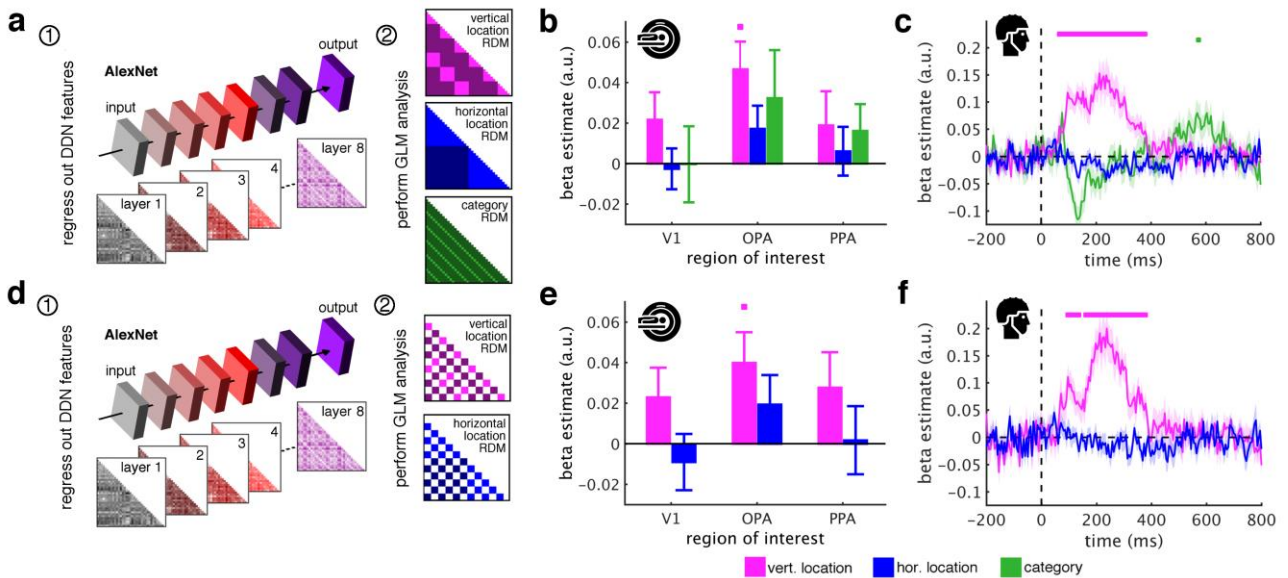
854 least in the EEG data, the differentiation along the vertical axis is more categorical in

855 nature. Significance markers represent  $p < 0.05$  (corrected for multiple comparisons). Error

856 margins reflect standard errors of the mean.

857

858 *Figure 3 – Figure Supplement 1*

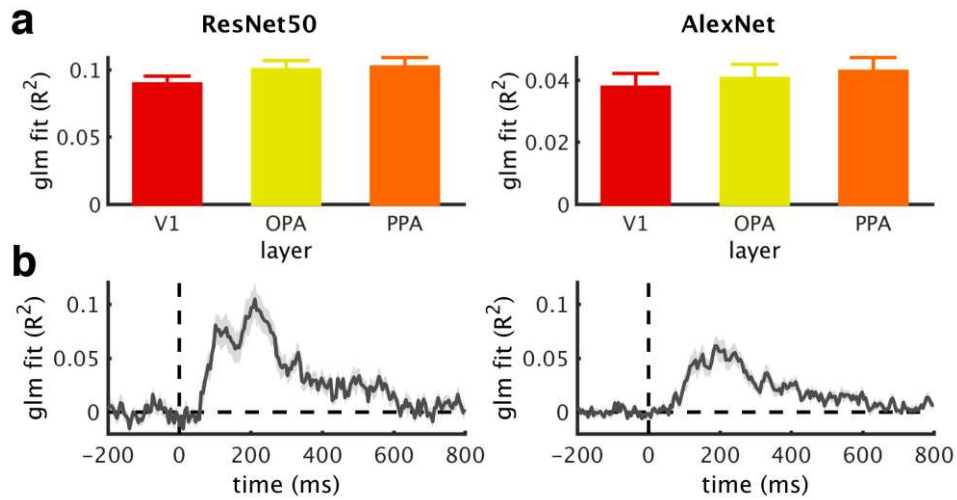


859

860 **AlexNet as a model of visual categorization.** **a**, In addition to the ResNet50 DNN, we  
 861 also used the more widely used AlexNet DNN architecture (pretrained on the ImageNet  
 862 dataset, implemented in the MatConvNet toolbox) as a model for visual categorization.  
 863 AlexNet consists of 5 convolutional and 3 fully-connected layers. We created 8 RDMs,  
 864 separately for each layer of the DNN. **b/c**, Removing the AlexNet DNN features rendered  
 865 category information non-significant in fMRI and EEG signals. However, we still found  
 866 vertical location information in OPA and from 65ms to 375ms. **c-e**, When additionally  
 867 restricting the analysis to comparisons between indoor and outdoor scenes, the fragments'  
 868 vertical location still predicted neural activations in OPA and from 95ms to 375ms. In sum,  
 869 these results are highly similar to the results obtained with the ResNet50 model (Fig.  
 870 3b/c/h/i). Significance markers represent  $p < 0.05$  (corrected for multiple comparisons).  
 871 Error margins reflect standard errors of the mean.

872

873 *Figure 3 – Figure Supplement 2*



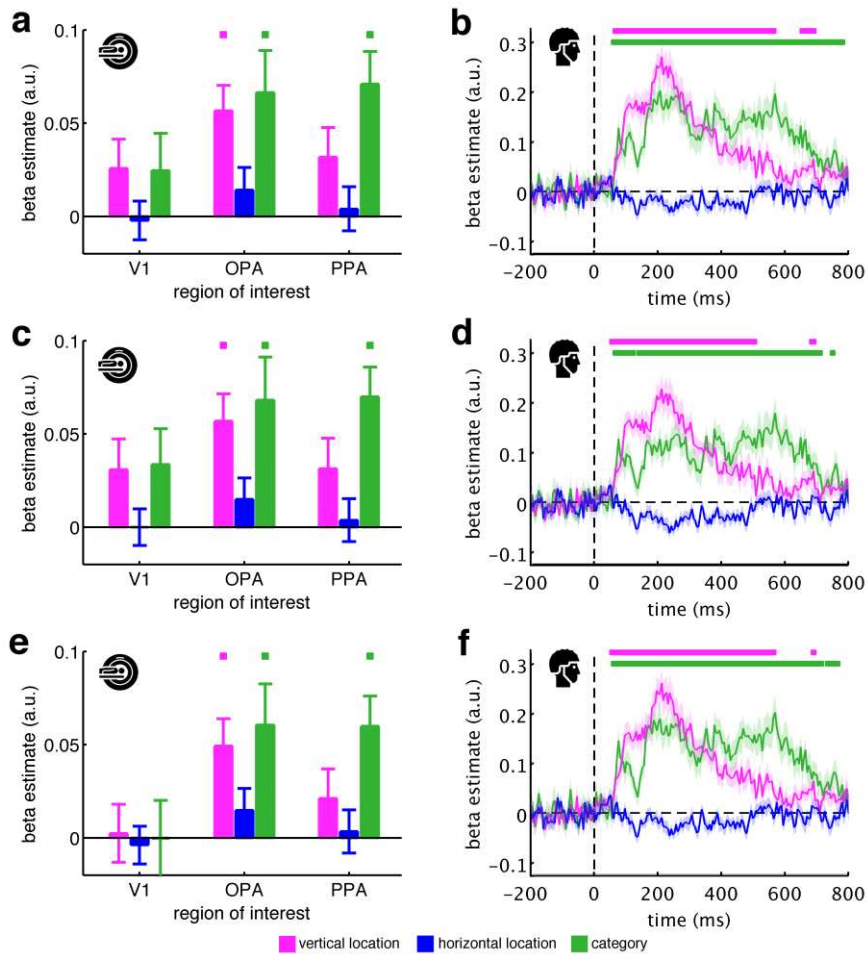
874

875 **DNN model fit. a/b,** Goodness of fit ( $R^2$ ) across ROIs (a) and time (b) of the GLMs used to  
876 regress out DNN features, obtained from ResNet50 (left) or AlexNet (right). For the EEG  
877 time series, mean  $R^2$  across the baseline period were subtracted. Note that GLMs based  
878 on the ResNet50 RDMs had more predictor variables, which may contribute to their better  
879 fit. Error bars represent standard errors of the mean.

880

881

882 *Figure 3 – Figure Supplement 3*



883

884 **Low-level control models.** We used three control models that explicitly account for low-  
 885 level visual features: a pixel-dissimilarity model, GIST descriptors, and the fragments'  
 886 neural dissimilarity in V1. Critically, all three models did not account for the fragments'  
 887 vertical location organization. Moreover, unlike the DNN models, the low-level models  
 888 were also unable to account for the fragments' categorical organization. **a/b**, Results after  
 889 regressing out the pixel dissimilarity model, which captured the fragments' pairwise  
 890 dissimilarity in pixel space (i.e., 1- the correlation of their pixel values). **c/d**, Results after  
 891 regressing out the GIST model, which captured the fragments' pairwise dissimilarity in  
 892 GIST descriptors (i.e., in their global spatial envelope). **e/f**, Results after regressing out the  
 893 V1 model, which captured the fragments' pairwise neural dissimilarity in V1 (i.e., the  
 894 averaged RDM across participants) and thereby provides a brain-derived measure of low-  
 895 level feature similarity. Significance markers represent  $p < 0.05$  (corrected for multiple  
 896 comparisons). Error margins reflect standard errors of the mean.

897

898 Supplementary file 1

899 **Complete statistical report for fMRI results.** The table shows test statistics and p-values  
900 for all tests performed in the fMRI experiment (Fig. 2/3). Values reflect t-tests one-sided t-  
901 tests against zero. All p-values are uncorrected; in the main manuscript, only tests  
902 surviving Bonferroni-correction across the three ROIs (marked in color) are considered  
903 significant.

904

905 Supplementary file 2

906 **Estimating peak latencies.** The table shows means and standard deviations (in brackets)  
907 of peak latencies in ms for vertical location and category information in the main analyses  
908 (Fig. 2/3). To estimate the reliability of peaks and onsets (Supplementary file 3) of location  
909 and category information in the key analyses, we conducted a bootstrapping analysis. For  
910 this analysis, we choose 100 samples of 20 randomly chosen datasets (with possible  
911 repetitions). For each random sample, we computed peak and onset latencies; we then  
912 averaged the peak and onset latencies across the 100 samples. Peak latencies were  
913 defined as the highest beta estimate in the time course. Notably, the peak latency of  
914 vertical location information remained highly stable across analyses.



915

916 Supplementary file 3

917 **Estimating onset latencies.** The table shows means and standard deviations (in  
918 brackets) of onset latencies in ms for vertical location and category information in the main  
919 analyses (Fig. 2/3). Onset latencies were quantified using the bootstrapping logic  
920 explained above (Supplementary file 2). Onsets were defined by first computing TFCE  
921 statistics for each random sample, with multiple-comparison correction based on 1,000 null  
922 distributions. The onset latency for each sample was then defined as the first occurrence  
923 of three consecutive time points reaching significance ( $p < 0.05$ , corrected for multiple  
924 comparisons).

925

926 Video 1

927 **Time-resolved MDS visualization of the neural RDMs.** To directly visualize the  
928 emergence of schematic coding from the neural data, we performed a multi-dimensional  
929 scaling (MDS) analysis, where the time-resolved neural RDMs (averaged across  
930 participants) were projected onto a two-dimensional space. The RDM time series was  
931 smoothed using a sliding averaging window (15ms width). Computing MDS solutions  
932 across time yielded a movie (5ms resolution), where fragments travel through an arbitrary  
933 space, eventually forming a meaningful organization. Notably, around 200ms, a division  
934 into the three vertical locations can be observed. The movie is attached to this file  
935 (time\_resolved\_mds.mov).