Article:

Kaiser, Daniel 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

A neural mechanism for contextualizing fragmented inputs during naturalistic vision

Daniel Kaiser[1,2,*], Jacopo Turini[2,3], Radoslaw M. Cichy[2,4,5]

[1] *Department of Psychology, University of York, York, UK*
[2] *Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany*
[3] *Institute of Psychology, Goethe-Universität Frankfurt, Frankfurt am Main, Germany*
[4] *Berlin School of Mind and Brain, Humboldt-Universität Berlin, Berlin, Germany*
[5] *Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany*

**Correspondence to:*
Dr. Daniel Kaiser
Department of Psychology
University of York
Heslington, York
YO10 5DD, UK
danielkaiser.net@gmail.com

ABSTRACT

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

IMPACT STATEMENT

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

INTRODUCTION

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

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

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

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

Recently however, it has become clear that scene schemata not only organize memory contents, but also the contents of perception. For example, knowledge about the structure of the world can be used to generate predictions about a scene's content (Bar, 2009; Henderson, 2017), or to efficiently organize the concurrent representation of multiple scene elements (Kaiser et al., 2019a; Kaiser et al., 2019b). This position is reinforced by behavioral studies demonstrating a beneficial role of schema-congruent naturalistic stimuli across a variety of perceptual tasks, such as visual detection (Biederman et al., 1982; Davenport and Potter, 2004; Stein et al., 2015) and visual search (Kaiser et al., 2014; Torralba et al., 2006; Võ et al., 2019).

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

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

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

RESULTS

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

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

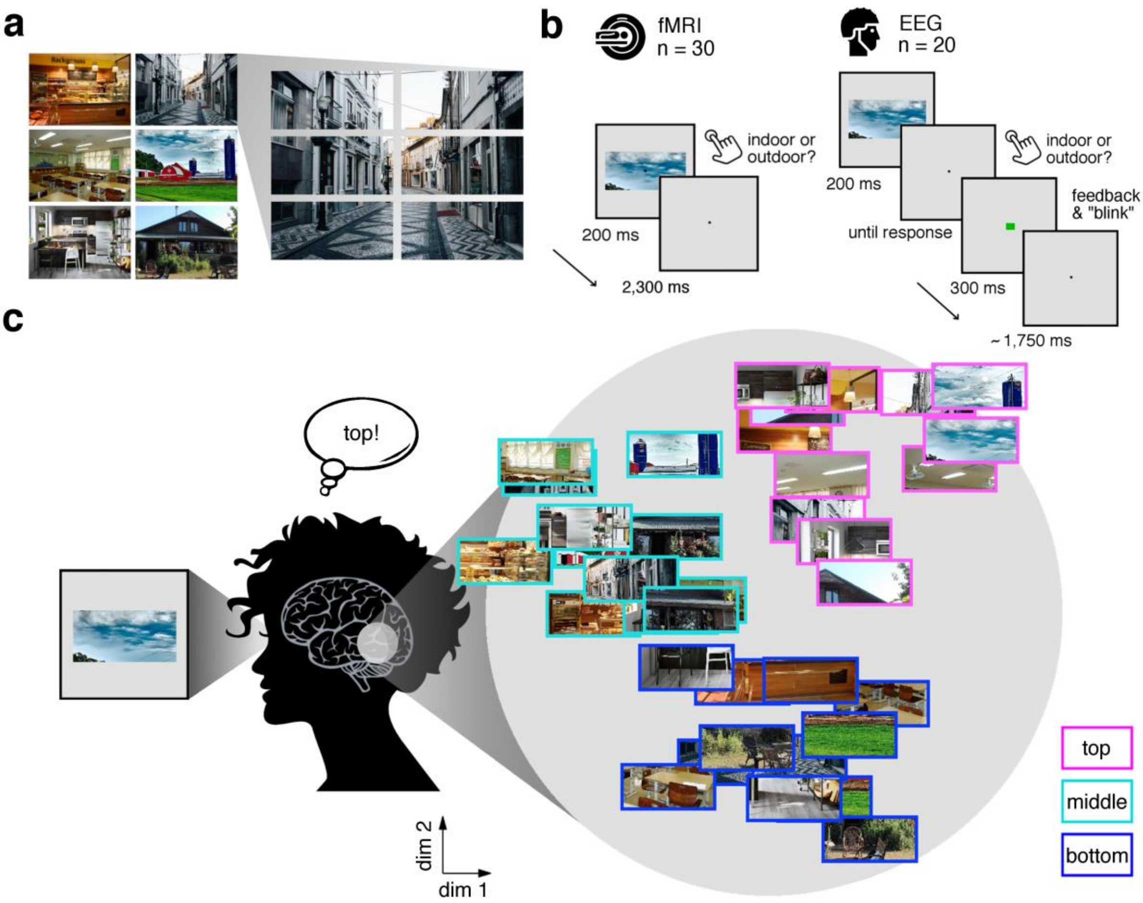


Fig. 1: Experimental design and rationale of schema-based information sorting. a, The stimulus set consisted of six natural scenes (three indoor, three outdoor). Each scene was split into six rectangular fragments. **b**, During the fMRI and EEG recordings, participants performed an indoor/outdoor categorization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

We then quantified schema effects using separate model RDMs for horizontal and vertical locations (Fig. 2c). These location RDMs reflected whether pairs of fragments shared the same location or not. We additionally constructed a category model RDM, which reflected whether pairs of fragments stemmed from the same scene or not.

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

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

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

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

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

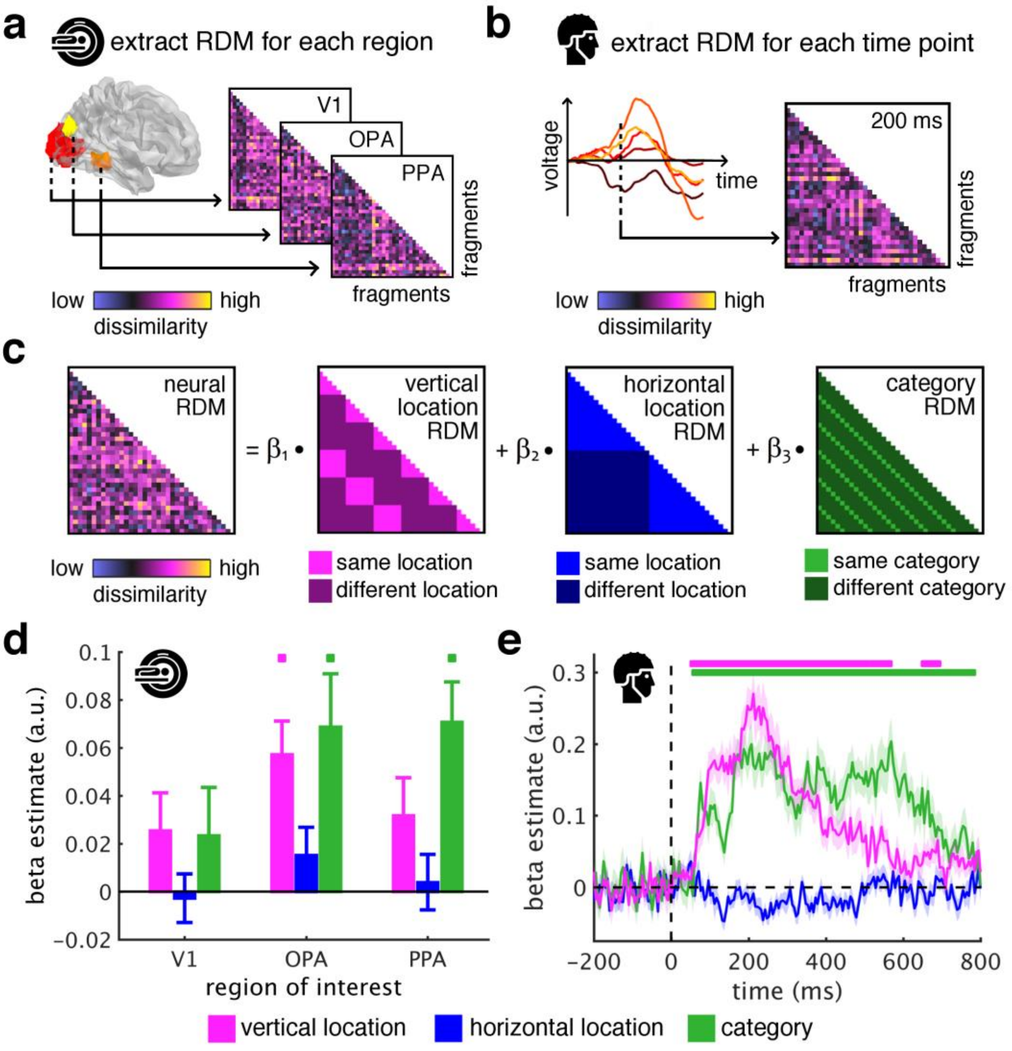


Fig. 2: Spatial schemata determine cortical representations of fragmented scenes. a, To test where and when the visual system sorts incoming sensory information by spatial schemata, we first extracted spatially (fMRI) and temporally (EEG) resolved neural representational dissimilarity matrices (RDMs).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

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

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

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

After removing DNN features, category information was rendered non-significant in both fMRI and EEG signals. When directly comparing category information before and after removing the DNN features, we found reduced category information in PPA ($t[29]=2.48$, $p=0.0096$, $p_{corr}<0.05$) and OPA ($t[29]=1.86$, $p=0.036$, $p_{corr}>0.05$), and a strong reduction of category information across time, from 75ms to 775ms (peak $t[19]=13.0$, $p<0.001$, $p_{corr}<0.05$). Together, this demonstrates that categorization-related brain activations are successfully explained by DNN features (Cichy et al., 2016, 2017; Groen et al., 2018; Güçlü and van Gerven, 2015; Wen et al., 2018), indicating the appropriateness of our DNN for modelling visual brain activations. Despite the suitability of our DNN model for modelling categorical brain responses, vertical location still accounted for the neural organization in OPA (t[29]=2.37, p=0.012, $p_{corr}<0.05$) (Fig. 3b) and between 75ms and 335ms (peak: t[19]=5.06, p<0.001, $p_{corr}<0.05$) (Fig. 3c). Similar results were obtained using a shallower feed-forward DNN (see Figure 3 – Figure Supplement 1). This result suggests that schematic coding cannot be explained by categorization-related features extracted by DNN models.

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

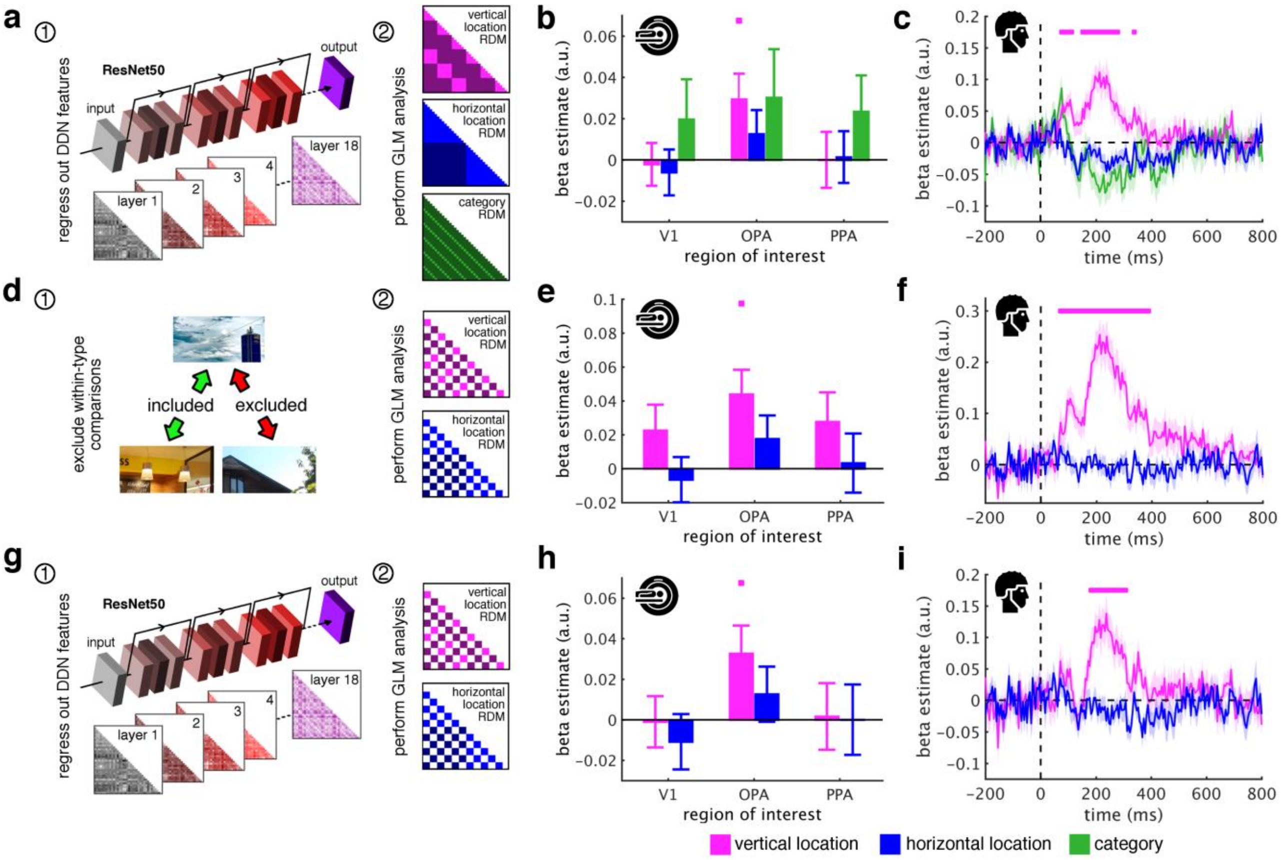


Fig. 3: Schematic coding operates flexibly across visual and conceptual scene properties. a, To determine the role of categorization-related visual features in this schematic organization, we regressed out RDMs obtained from 18 layers along the ResNet50 DNN before repeated the three-predictor general 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.

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

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

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

DISCUSSION

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

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

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

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

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

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

MATERIALS AND METHODS

Key Resources Table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
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

Participants

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

Stimuli

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

Experimental design

The fMRI and EEG designs were identical, unless otherwise noted. Stimulus presentation was controlled using the Psychtoolbox (Brainard, 1997; RRID:SCR_002881). In each trial, one of the 36 fragments was presented at central fixation (7° horizontal visual angle) for 200ms (Fig. 1b). Participants were instructed to instructed to maintain central fixation and categorize each stimulus as an indoor or outdoor scene image by pressing one of two buttons.

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

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

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

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

fMRI recording and preprocessing

MRI data was acquired using a 3T Siemens Tim Trio Scanner equipped with a 12-channel head coil. T2*-weighted gradient-echo echo-planar images were collected as functional volumes (TR=2s, TE=30ms, 70° flip angle, 3mm^3 voxel size, 37 slices, 20% gap, 192mm FOV, 64×64 matrix size, interleaved acquisition). Additionally, a T1-weighted image (MPRAGE; 1mm^3 voxel size) was obtained as a high-resolution anatomical reference. During preprocessing, the functional volumes were realigned and coregistered to the T1 image, using MATLAB (RRID:SCR_014519) and SPM12 (www.fil.ion.ucl.ac.uk/spm/; RRID:SCR_014519).

fMRI region of interest definition

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

EEG recording and preprocessing

The EEG was recorded using an EASYCAP 64-channel system and a Brainvision actiCHamp amplifier. The electrodes were arranged in accordance with the standard 10-10 system. The data was recorded at a sampling rate of 1000Hz and filtered online between 0.03Hz and 100Hz. All electrodes were referenced online to the Fz electrode. Offline preprocessing was performed in MATLAB, using the FieldTrip toolbox (Oostenveld et al., 2011; RRID:SCR_004849). The continuous EEG data were epoched into trials ranging from

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

Representational Similarity Analysis

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

Constructing neural dissimilarity – fMRI

For the fMRI data, we used cross-validated correlations as a measure of pairwise neural dissimilarity. First, patterns for each ROI were extracted from the functional images corresponding to the trials of interest. After shifting the activation time course by 3 TRs (i.e., 6s, accounting for the hemodynamic delay), we extracted voxel-wise activation values for each trial, from the TR that was closest to the stimulus onset on this trial (for results across 6 TRs with respect to trial onset, see Figure 2 – Figure Supplement 2). To account for activation differences between runs, the mean activation across conditions was subtracted from each voxel's values, separately for each run. For each ROI, response patterns across voxels were used to perform multivariate analyses using the CoSMoMVPA toolbox (Oosterhof et al., 2016; RRID:SCR_014519). Then, for each TR separately, we performed correlation-based (Haxby et al., 2001) multi-voxel pattern analyses (MVPA) for each pair of fragments. These analyses were cross-validated by repeatedly splitting the data into two equally-sized sets (i.e., half of the runs per set). For this analysis, we correlated the patterns across the two sets, both within-condition (i.e., the patterns stemming from the two same fragments and from different sets) and between-conditions (i.e., the patterns stemming from the two different fragments and from different sets). These correlations were Fisher-transformed. Then, we subtracted the within- and between-correlations to obtain a cross-validated correlation measure, where above-zero values reflect successful discrimination. This procedure was repeated for all possible splits of the six runs. Performing this MVPA for all pairs of fragments yielded a 36×36 representational dissimilarity matrix (RDM) for each ROI. RDMs' entries reflected the neural dissimilarity between pairs of fragments (the diagonal remained empty).

Constructing neural dissimilarity – EEG

For the EEG data, we used cross-validated classification accuracies as a measure of pairwise neural dissimilarity. We thus constructed RDMs across time by performing time-resolved multivariate decoding analyses (Contini et al., 2017). RDMs were built by computing pair-wise decoding accuracy for all possible combinations of the 36 stimuli, using the CoSMoMVPA toolbox (Oosterhof et al., 2016). As we expected the highest classification in sensors over visual cortex (Battistoni et al., 2018; Kaiser et al., 2016), only 17 occipital and posterior sensors (O1, O2, Oz, PO3, PO4, PO7, PO8, POz, P1, P2, P3, P4, P5, P6, P7, P8, Pz) were used in this analysis. We report results for other electrode groups in Figure 2 – Figure Supplement 3-5. For each participant, classification was performed separately for each time point across the epoch (i.e., with 5ms resolution). The analysis was performed in a pair-wise fashion: Linear discriminant analysis classifiers were always trained and tested on data from two conditions (e.g., the middle left part of the alley versus the top right part of the farm), using a leave-one-trial-out partitioning scheme. The training set consisted of all but one trials for each of the two conditions, while one trial for each of the two conditions was held back and used for classifier testing. This procedure was repeated until every trial was left out once. Classifier performance was averaged across these repetitions. The pairwise decoding analysis resulted in a 36-by-36 neural RDM for each time point. A schematic description of the RDM construction can be found in Figure 2 – Figure Supplement 1.

Location and category predictors

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

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

Modelling neural dissimilarity

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

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

Controlling for deep neural network features

To control for similarity in categorization-related visual features, we used a deep neural network (DNN) model. DNNs have recently become the state-of-the-art model of visual categorization, as they tightly mirror the neural organization of object and scene representations (Cichy et al., 2016, 2017; Cichy and Kaiser, 2019; Groen et al., 2018; Güclü and van Gerven, 2015; Wen et al., 2018). DNNs are similar to the brain as they are trained using excessive training material while dynamically adjusting the “tuning” of their connections. Here, we used a DNN that has been trained to categorize images (see below) on a large number of images and categories, therefore providing us with a highquality model of how visual features are extracted for efficient categorization. By comparing DNNs activations and brain responses to the scene fragments, we could quantify to which extent features routinely extracted for categorization purposes account for schema-based coding in the human visual system.

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

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

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

Statistical testing

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

Data availability

Data are publicly available on OSF (DOI.ORG/10.17605/OSF.IO/H3G6V).

ACKNOWLEDGEMENTS

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

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

C. Baldassano, A. Esteva, L. Fei-Fei, D. M. Beck, Two distinct scene processing networks connecting vision and memory. *eNeuro* **3**, ENEURO.0178-16.2016 (2016).

M. Bar, The proactive brain: memory for predictions. *Phil. Trans. Royal Soc. B Biol. Sci.* **364,** 1235-1243 (2009).

F. C. Barlett, Remembering: a study in experimental and social psychology. (Cambridge University Press, 1932).

E. Battistoni, D. Kaiser, C. M. Hickey, M. V. Peelen, The time course of spatial attention during naturalistic visual search. *Cortex*, doi.org/10.1016/j.cortex.2018.11.018 (2018).

I. Biederman, R. J. Mezzanotte, J. C. Rabinowitz, Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* **14**, 143-177 (1982).

M. F. Bonner, R. A. Epstein, Coding of navigational affordances in the human visual system. *Proc. Natl. Acad. Sci. USA* **114,** 4793-4798 (2017).

D. H. Brainard, The psychophysics toolbox. *Spat. Vis.* **10,** 433-436 (1997).

W.F. Brewer, J.C. Treyens, The role of schemata in memory for places. *Cogn. Psychol.* **13,** 207-230 (1981).

R. M. Cichy, D. Kaiser, Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**, 305-317 (2019).

R. M. Cichy, A. Khosla, D. Pantazis, A. Oliva, Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage* **153,** 346-358 (2017).

R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6,** 27755 (2016).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SUPPLEMENTARY INFORMATION

A neural mechanism for contextualizing fragmented inputs during naturalistic vision

Kaiser, Turini, & Cichy

page

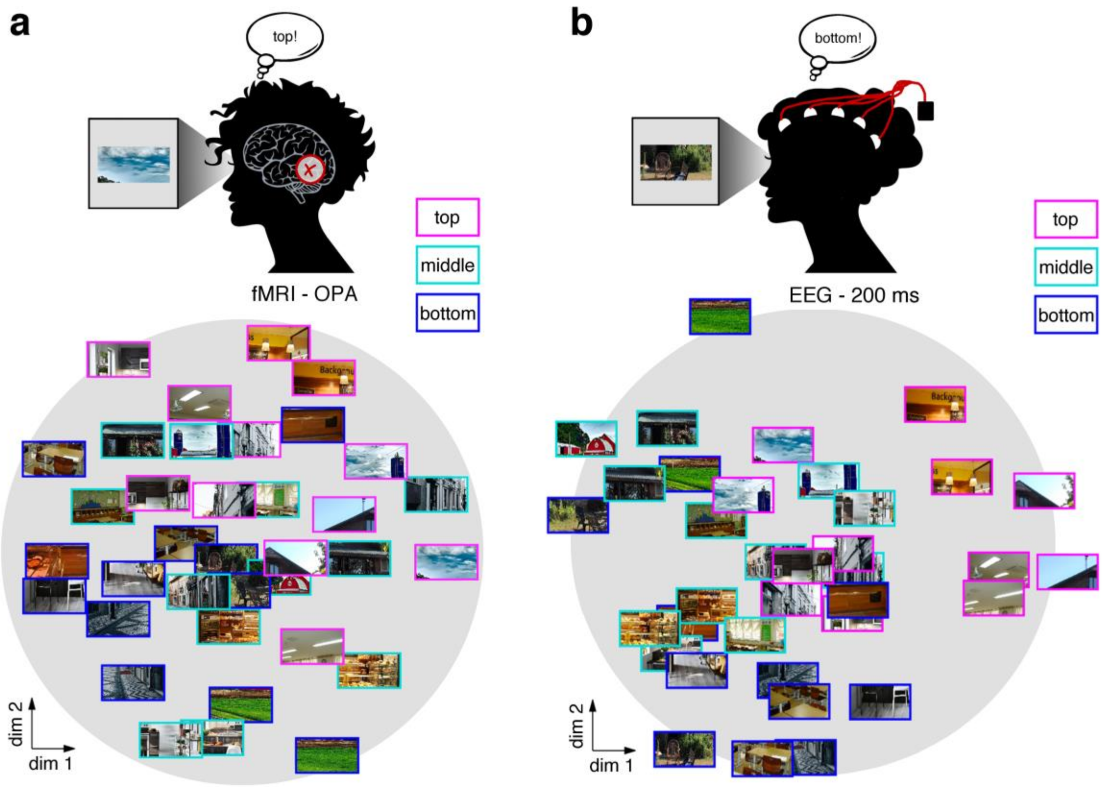
Supplementary Figures:

Supplementary Items:

Supplementary Videos:

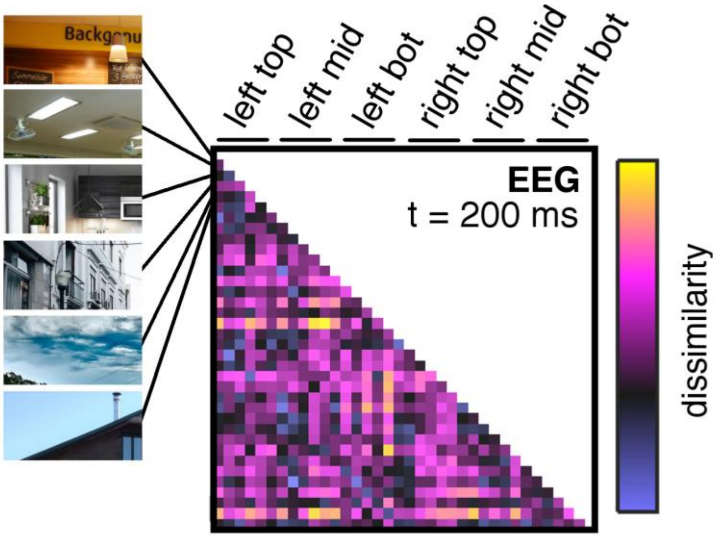
Video 1: Time-resolved MDS visualization of EEG RDMs *50*

Figure 1 – Figure Supplement 1



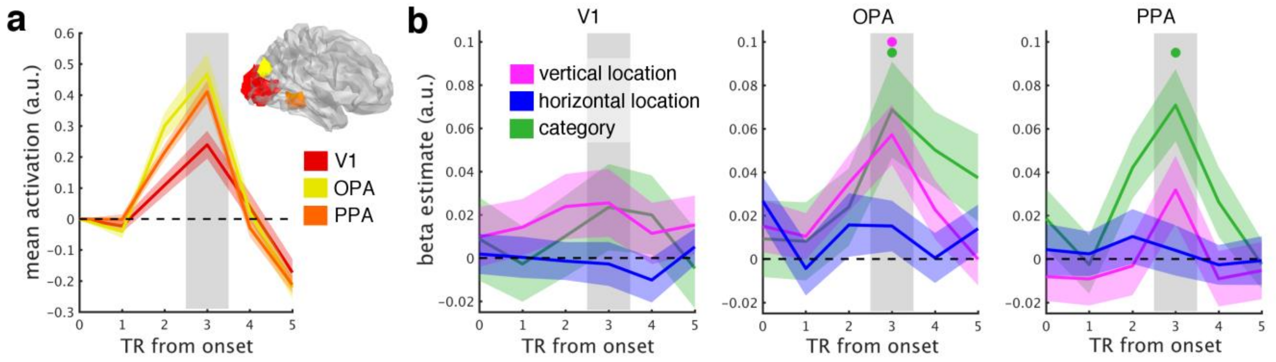
MDS visualization of neural RDMs. a/b, A multi-dimensional scaling (MDS) of the fragments' neural similarity in OPA (a) and after 200ms of processing (b) revealed a sorting according to vertical location, which was visible in a two-dimensional solution. This visualization suggests that schemata are a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

Figure 2 – Figure Supplement 1



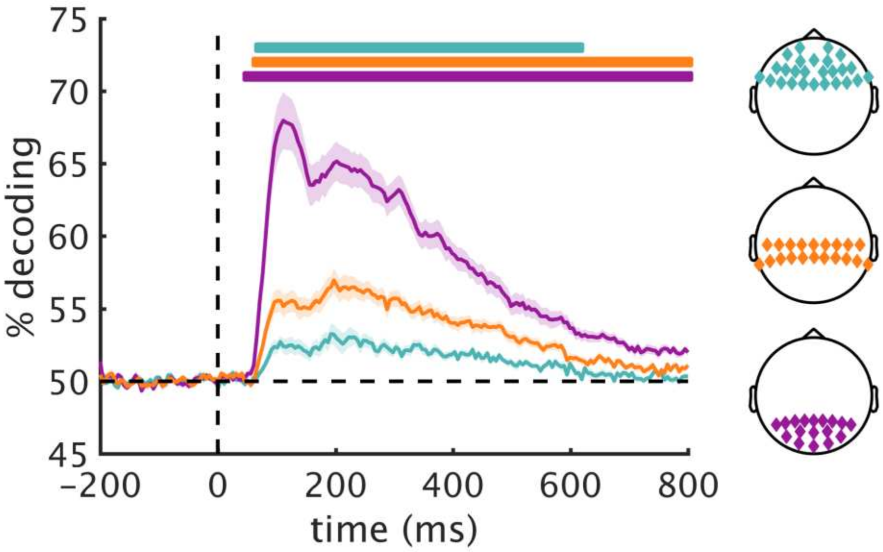
Details on neural dissimilarity construction. Pairwise neural dissimilarity values were into representational dissimilarity matrices (RDMs), so that for every time point one 36X36 matrix containing estimates of neural dissimilarity was available. Here, an example RDM at 200ms post-stimulus is shown,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

Figure 2 – Figure Supplement 2



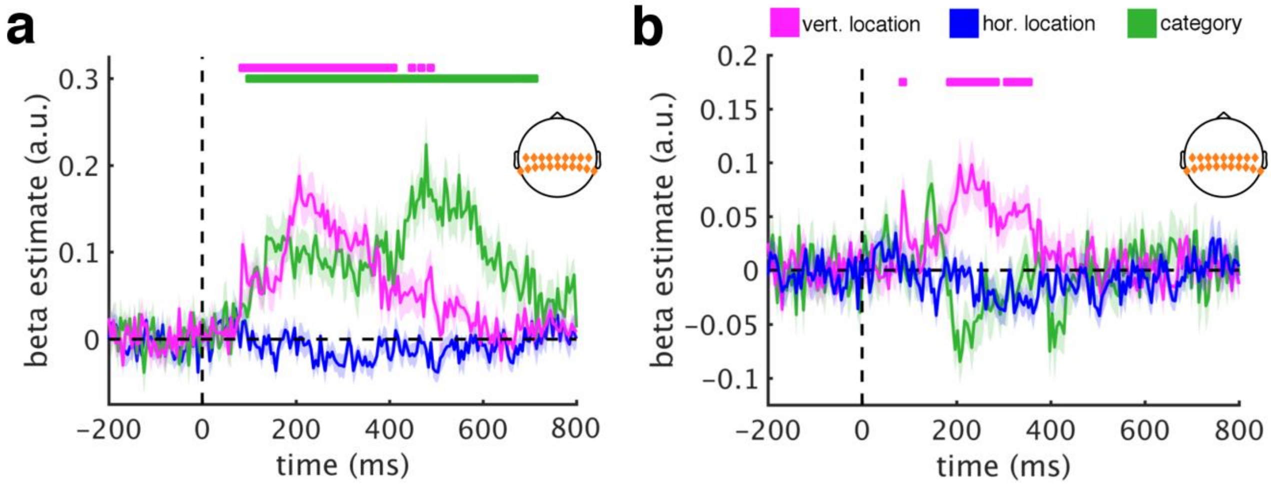
fMRI response time courses. a, Functional MRI data were analyzed in three regions of interest (here shown on the right hemisphere): primary visual cortex (V1), occipital place area (OPA), and parahippocampal place area (PPA). Each of these ROIs showed reliable net responses to the fragments, peaking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

Figure 2 – Figure Supplement 3



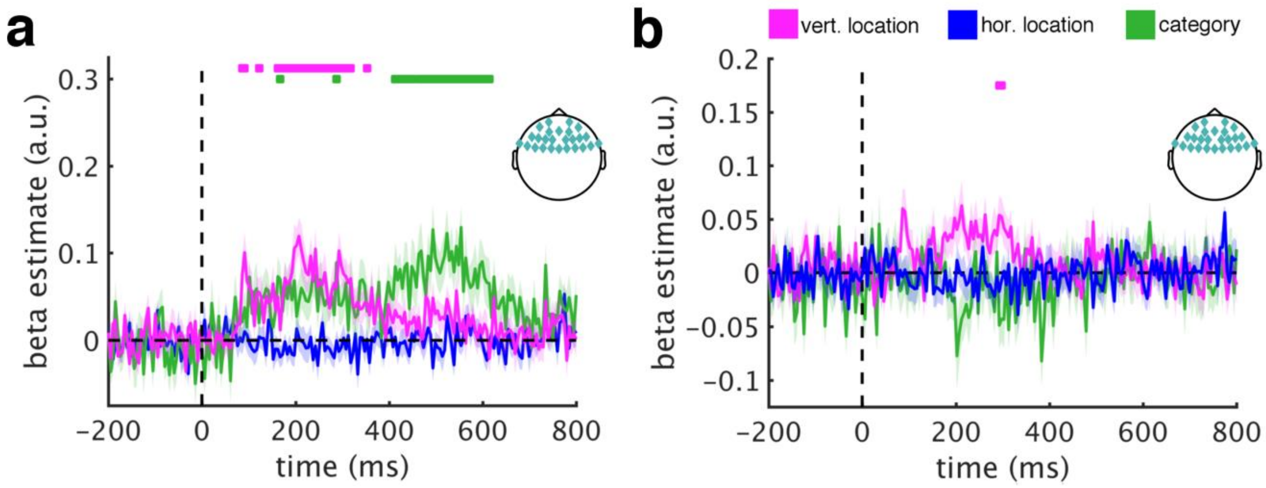
Pairwise decoding across electrode groups. Based on previous studies on multivariate decoding of visual information, we restricted our main analysis to a group of posterior electrodes (where we expected the strongest effects). For comparison, we also analyzed data in central and anterior electrode groups.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

Figure 2 – Figure Supplement 4



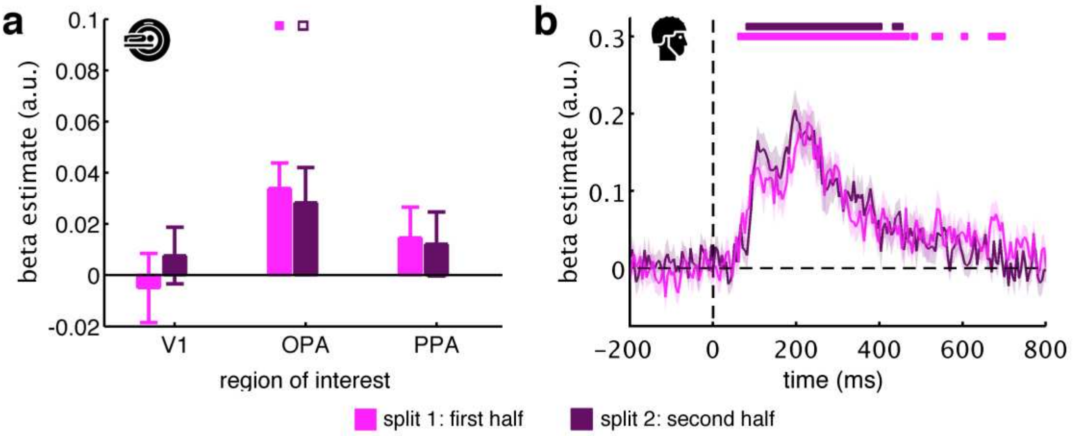
RSA using central electrodes. a/**b**, Repeating the main RSAs for the central electrode group yielded a similar pattern as the posterior group, revealing both vertical location information (from 85ms to 485ms) and category information (from 100ms to 705ms). **c**/**d**, Removing DNN features abolished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

Figure 2 – Figure Supplement 5



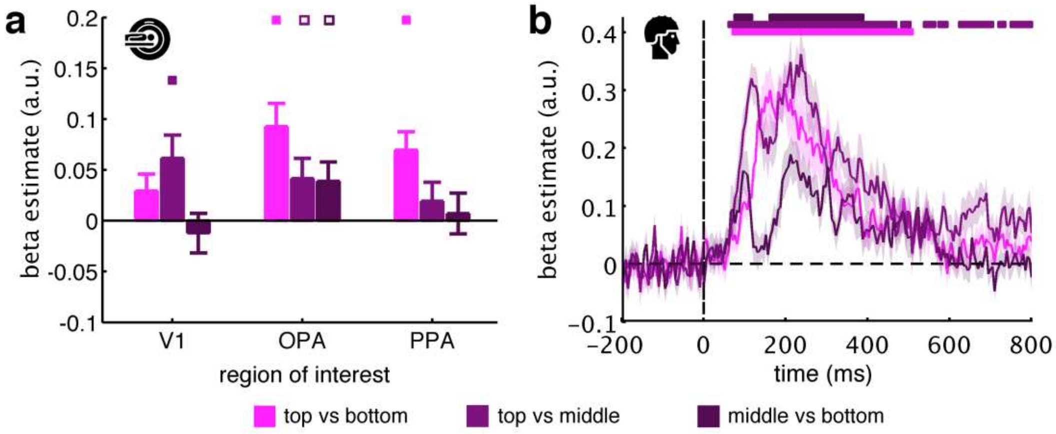
RSA using anterior electrodes. a/**b**, Also responses recorded from the anterior group yielded both vertical location information (from 85ms to 350ms) and category information (from 165ms to 610ms). **c**/**d**, In contrast to the other electrode groups, removing DNN features rendered location and category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

Figure 2 – Figure Supplement 6



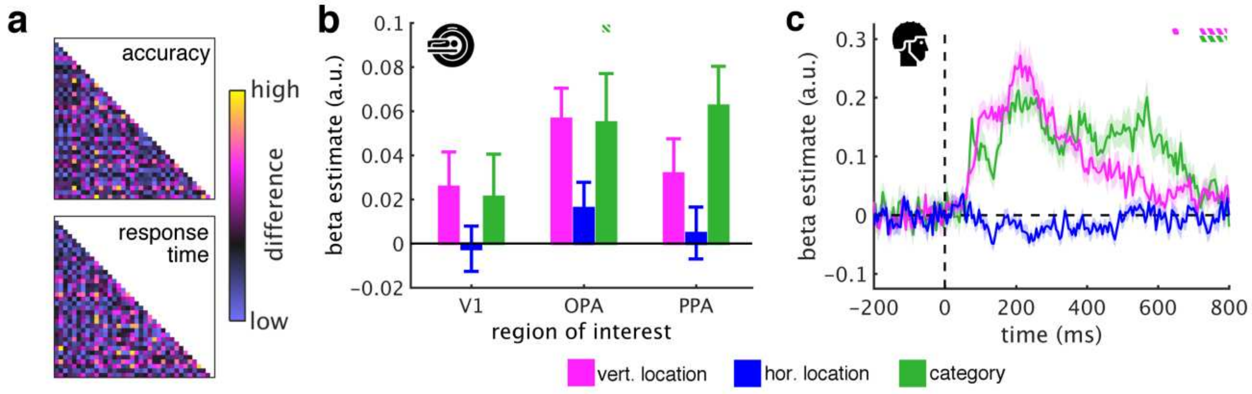
Vertical location effects across experiment halves. We interpret the vertical location organization in the neural data as reflecting prior schematic knowledge about scene structure. Alternatively, however, the vertical location organization could in principle result from learning the composition of the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

Figure 2 – Figure Supplement 7



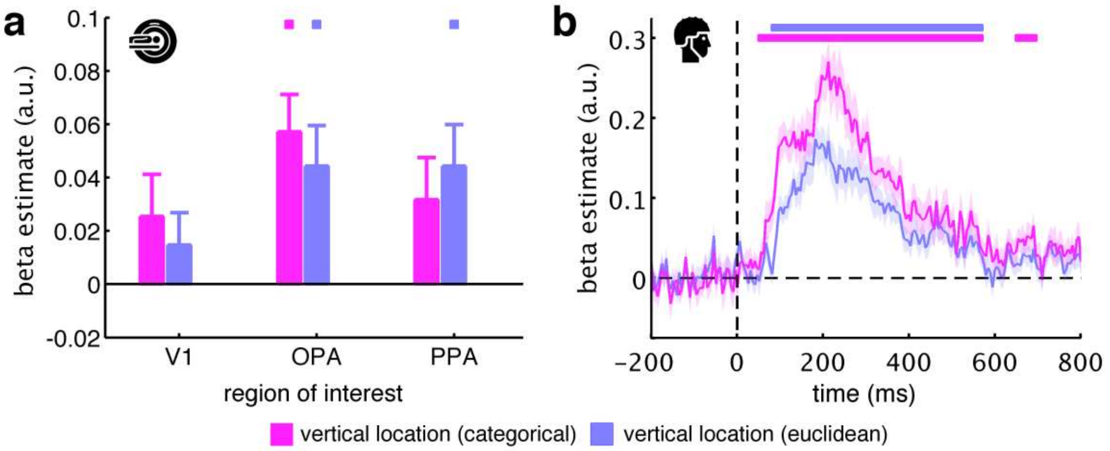
Pairwise comparisons along the vertical axis. To test whether vertical location information can be observed across all three vertical bins, we modelled the neural data as a function of the fragments' vertical location, now separately for each pairwise comparison along the vertical axis (i.e., top versus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

Figure 2 – Figure Supplement 8



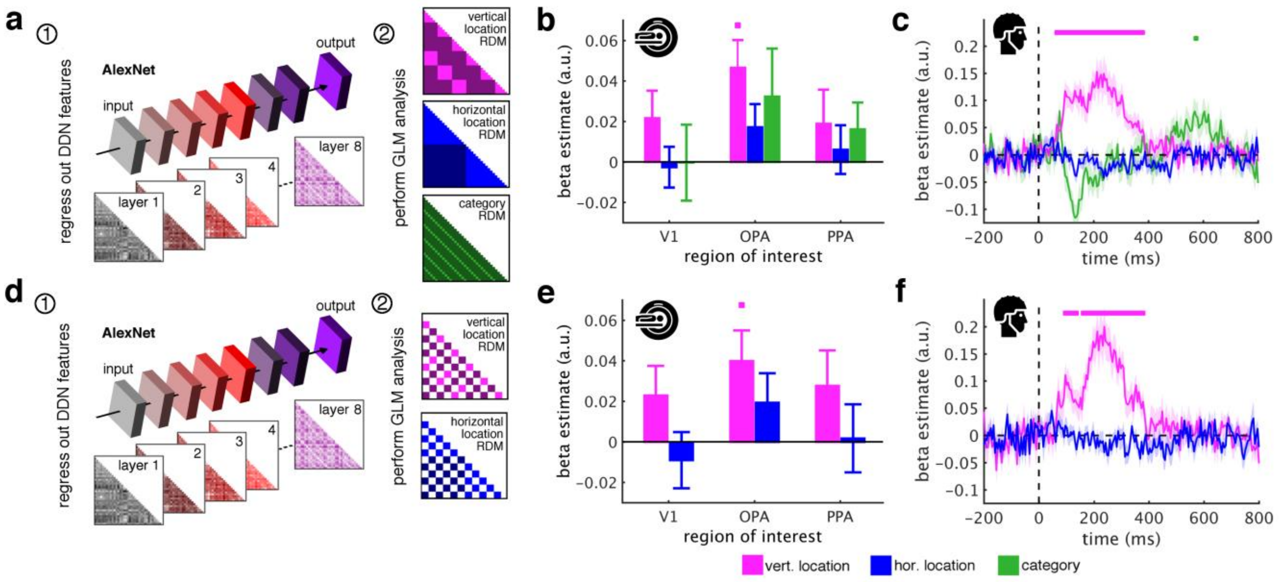
Controlling for task difficulty. a, To control for task difficulty effects in the indoor/outdoor classification task, we computed paired t-tests between all pairs of fragments, separately for their associated accuracies and response times. We then constructed two predictor RDMs that contained the t-values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

Figure 2 – Figure Supplement 9



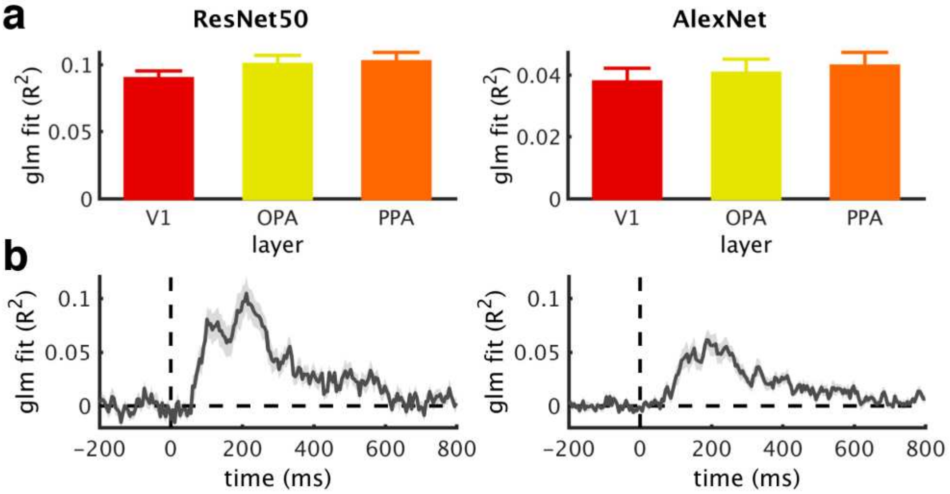
Categorical versus Euclidean vertical location predictors. We defined our vertical location predictor as categorical, assuming that top, middle, and bottom fragments are coded distinctly in the human brain. An alternative way of constructing the vertical location predictor is in terms of the fragments' Euclidean distances, where fragments closer together along the vertical axis (e.g., top and middle) are represented more similarly than fragments further apart (e.g., top and bottom). **a**, For the fMRI data, we found that the categorical and Euclidean predictors similarly explained the neural data, with no statistical differences between them (all t[29]<1.15, p>0.26). **b**, For the EEG data, we found that both predictors explained the neural data well. However, the categorical predictor revealed significantly stronger vertical location information from 75ms to 340ms, suggesting that, at least in the EEG data, the differentiation along the vertical axis is more categorical in nature. Significance markers represent p<0.05 (corrected for multiple comparisons). Error margins reflect standard errors of the mean.

Figure 3 – Figure Supplement 1



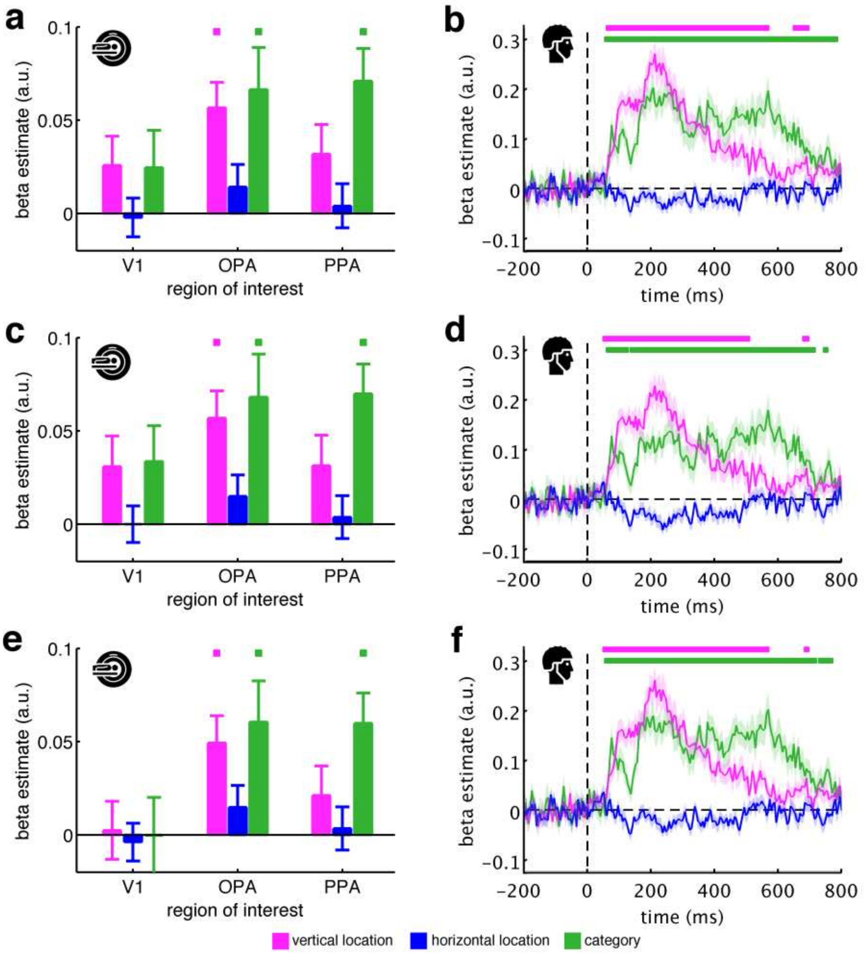
AlexNet as a model of visual categorization. a, In addition to the ResNet50 DNN, we also used the more widely used AlexNet DNN architecture (pretrained on the ImageNet dataset, implemented in the MatConvNet toolbox) as a model for visual categorization. AlexNet consists of 5 convolutional and 3 fully-connected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

Figure 3 – Figure Supplement 2



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

Figure 3 – Figure Supplement 3



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

Supplementary file 1

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

Supplementary file 2

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

<u>Supplementary file 3</u>

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

Video 1

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