Behavioral/Cognitive

Cortical Encoding of Spatial Structure and Semantic Content in 3D Natural Scenes

Riikka Mononen,[1,2] **Toni Saarela,**[3] **Jaakko Vallinoja,**[1,2] **Maria Olkkonen,**[3] **and Linda Henriksson**[1,2]

[1]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo FI-00076, Finland, [2]MEG Core, Aalto NeuroImaging, Aalto University, Espoo FI-00076, Finland, and [3]Department of Psychology, University of Helsinki, Helsinki FI-00014, Finland

Our visual system enables us to effortlessly navigate and recognize real-world visual environments. Functional magnetic resonance imaging (fMRI) studies suggest a network of scene-responsive cortical visual areas, but much less is known about the temporal order in which different scene properties are analyzed by the human visual system. In this study, we selected a set of 36 full-color natural scenes that varied in spatial structure and semantic content that our male and female human participants viewed both in 2D and 3D while we recorded magnetoencephalography (MEG) data. MEG enables tracking of cortical activity in humans at millisecond timescale. We compared the representational geometry in the MEG responses with predictions based on the scene stimuli using the representational similarity analysis framework. The representational structure first reflected the spatial structure in the scenes in time window 90–125 ms, followed by the semantic content in time window 140–175 ms after stimulus onset. The 3D stereoscopic viewing of the scenes affected the responses relatively late, from ∼140 ms from stimulus onset. Taken together, our results indicate that the human visual system rapidly encodes a scene's spatial structure and suggest that this information is based on monocular instead of binocular depth cues.

Key words: **3D; depth; human visual system; MEG; scene; spatial structure**

Significance Statement

Our visual system enables us to recognize and navigate our visual surroundings seemingly effortlessly, but what exactly happens in our brains remains poorly understood. With the help of time-resolved brain imaging (magnetoencephalography), we found that the brain first encodes the spatial structure of a scene (e.g., cluttered or navigable) before its semantic content (e.g., a car park or farm). Brain imaging studies typically use 2D pictures as stimuli. Here we asked whether binocular disparity, a depth cue which arises from our two eyes seeing the scene from slightly different angles, aids the coding of the spatial structure. Our results suggest that this 3D depth cue plays little role in the rapid, initial sensing of our spatial surroundings.

Introduction

Humans understand complex real-world visual environments without much effort. Scene perception engages specialized cortical areas that include the occipital place area (OPA; Grill-Spector, 2003; Dilks et al., 2013) and the parahippocampal place area (PPA; Epstein and Kanwisher, 1998; for a review, see Epstein and Baker, 2019). Recent brain imaging studies suggest complementary roles in visual scene processing for these regions. PPA has a central role in scene categorization, that is, in recognizing the environment or place that we are in, whereas OPA is critical for visually guided navigation, that is, in processing the spatial structure and navigational affordances of the visible environment (for reviews, see Epstein and Baker, 2019; Dilks et al., 2022). Most previous brain imaging studies on scene processing have used functional magnetic resonance imaging (fMRI), and therefore, much less is known about the temporal order in which different scene properties are analyzed by the human visual system (for reviews, see Epstein and Baker, 2019; Bartnik and Groen, 2023). Disentangling the contributions of different scene features in cortical processing is challenging (Greene and Hansen, 2020), and overall, scene processing is unlikely to happen as a clear temporal cascade from low- to high-level scene properties (Ramkumar et al., 2016; Groen et al., 2017). Recently, interest has turned to how rapidly navigational features are encoded by the human visual system (Harel et al., 2022; Dwivedi et al., 2024).

Scenes can be understood and analyzed along many dimensions (Malcolm et al., 2016). Anderson et al. (2021) recently

Received Nov. 14, 2023; revised Nov. 25, 2024; accepted Dec. 24, 2024.

Author contributions: R.M., T.S., M.O., and L.H. designed research; R.M. and J.V. performed research; T.S., M.O., and L.H. contributed unpublished reagents/analytic tools; L.H. analyzed data; L.H. wrote the paper.

This work was supported by Aalto Brain Center. We thank Aalto NeuroImaging staff, especially Mia Illman and Tuomas Tolvanen, for assistance with the measurements and for technical support, and Matti Stenroos for providing scripts and help for performing the spatial whitening of the MEG signals. We acknowledge the computational resources provided by the Aalto Science-IT project.

The authors declare no competing financial interests.

Correspondence should be addressed to Linda Henriksson at linda.henriksson@aalto.fi.

https://doi.org/10.1523/JNEUROSCI.2157-23.2024

identified category systems that humans naturally use to classify real-world scenes along three dimensions: visual appearance, spatial structure, and semantic content. In a subsequent behavioral study, they studied the temporal order in which these dimensions are used in scene discrimination, with largely similar results for spatial structure and semantic content (Anderson et al., 2022). In the present study, we address this question more directly by using magnetoencephalography (MEG), which enables tracking of cortical activity in humans at millisecond timescale (for a review, see Baillet, 2017).

Little is known about how the human visual system analyzes real-world 3D environments. Most brain imaging studies on scene processing use 2D pictures as visual stimuli. The 3D spatial structure of a scene can be inferred from the 2D stimulus images, but the vivid impression of depth provided by disparity that arises from our two eyes seeing the world from slightly different viewpoints is, however, difficult to convey with 2D pictorial cues (Wheatstone, 1838; Vishwanath, 2023). Few previous brain imaging studies have investigated how binocular depth cues in natural scenes affect cortical responses (Duan et al., 2018, 2021).

In the present study, participants viewed a set of full-color natural scenes that varied in spatial structure and semantic content (Fig. 1). Our study participants viewed the same real-world scenes both as 2D images and in 3D as a stereo-pair of images. Using representational similarity analysis (RSA; Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013), we investigate the temporal order in which brain responses reflect the spatial structure and the semantic content of the scenes. Moreover, we evaluate the discriminability of the responses between the 2D and 3D viewing conditions. Our aim was to reveal the temporal order in which the spatial structure and the semantic content are encoded by the human visual system and whether disparity depth cues aid in coding the spatial structure.

Materials and Methods

Participants. Twenty healthy volunteers (10 females; age range, 21–37; 17 right-handed, 1 ambidextrous) participated in this study. All participants had normal or corrected to normal visual acuity, normal color vision (tested with Ishihara, 38 plates) and normal stereo vision (tested with TNO, 19th edition, threshold value 120 arcsec or lower was required; Piano et al., 2016). Participants were tested also for stereo anomaly by asking them to report the direction of the stereo percept in TNO (is the target closer or further from the background). Ethical approval for the research was obtained from Aalto University Ethics Committee. Participants gave written informed consent before participating in the study.

Scene stimuli and experimental design. The stimulus image set (Fig. 1*A*) consisted of 36 full-color stereo-pairs of natural scenes from the Southampton-York Natural Scenes (SYNS) database (Adams et al., 2016). The selection of the image set was based on human-generated labels for semantic category and spatial structure (Anderson et al., 2020, 2021). The six semantic category labels were nature, residence, farm, car park (or commercial), road and beach, of which the first four were the most frequently selected category labels for our selected set of scene images (Fig. 1*B*). The four spatial structure labels were cluttered (or pointy), closed off, flat, and navigable routes (or tunnel). The scene images also had labels for visual appearance (dark, bright, blue, green, brown; Anderson et al., 2021).

The scene stimuli were presented to a polarization-preserving back-projection screen with a PROPixx DLP LED projector (VPixx Technologies; resolution, 1,920 × 1,080; refresh rate, 120 Hz, 60 Hz per eye) combined with an active circular polarizer (DepthQ, Lightspeed Design). Participants viewed the screen with passive polarizing glasses at a viewing distance of 1.65 m. The size of the projected image was 49.5 × 32.5 cm, subtending ~17 × 11 degrees of visual angle. The spectral power distributions and chromaticities of the projector primaries were measured with a Photo Research PR-655 SpectraScan Spectroradiometer. Mean luminance of the scene images was set to 37.8 cd/m^2. Stimuli were presented using MATLAB (MathWorks) with the Psychtoolbox-3 extensions (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). Stereoscopic stimuli were presented with the temporally interleaved stereo mode of Psychtoolbox's PsychImaging function.

The MEG experiment included three viewing conditions: 2D, 3D, and reversed-stereo viewing. In the 2D condition, the same image was shown to both eyes. In the 3D condition, a stereo image pair was shown to the left and right eyes. In the reversed-stereo viewing condition, the right eye image of the stereo image pair was shown to the left eye and vice versa for the left eye image.

One experimental run consisted of 216 stimulus trials and 14 task trials. Each of the 36 scene images was shown two times in each of the

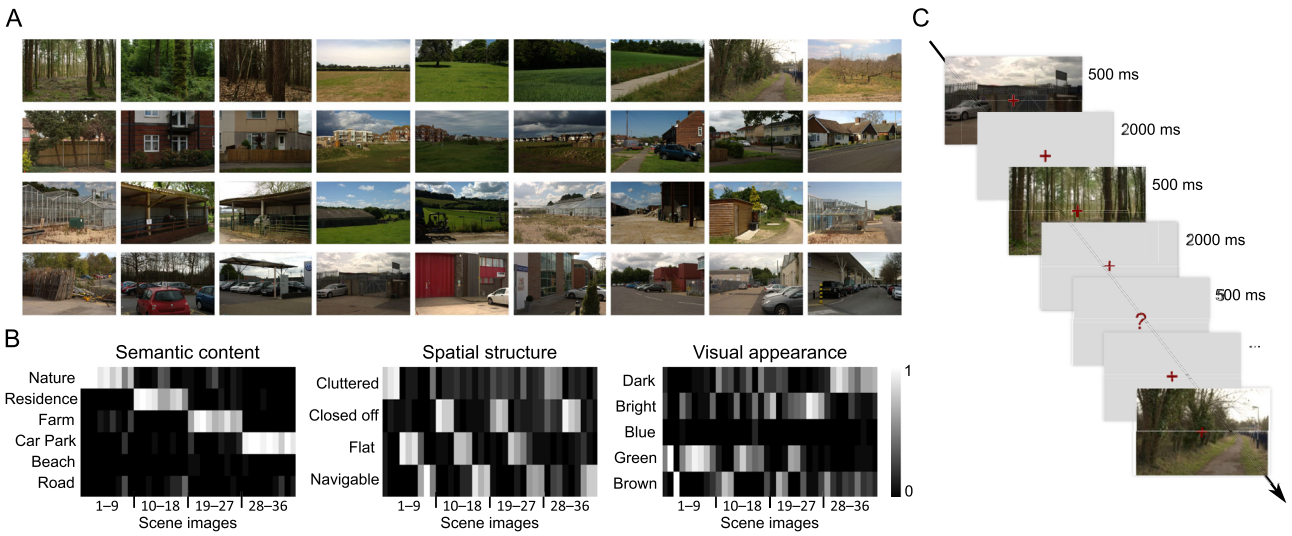


Figure 1. Scene stimuli. ***A***, A set of 36 scene images were selected from the SYNS database (Adams et al., 2016). ***B***, We selected the scenes so that they vary both in their semantic content and spatial structure based on previously published behavioral data on these same images (Anderson et al., 2020, 2021). Grayscale indicates the proportion of selected label for each of the 36 scenes, separately for semantic, spatial structure, and visual appearance categorization tasks (Anderson et al., 2020). ***C***, During the MEG experiment, the scene images were presented for 500 ms with a 2,000 ms interstimulus interval. When the scene was replaced by a question mark, the participant's task was to respond whether the two previous scenes could have been taken from the same place.

three different viewing conditions (108 unique stimulus trials). The duration of one trial was 2.5 s: 500 ms of the stimulus presentation and 2 s interstimulus interval (Fig. 1*C*). Participants were instructed to keep their gaze at a red fixation dot throughout the experimental runs and pay attention to the scenes. Between stimulus images, the red fixation dot was shown on a uniform gray background of the same mean luminance as the images. During task trials, a question mark was shown instead of an image, and the participants' task was to respond with a finger lift whether the two previous scene images could have been taken from the same place. The order of the images, conditions, and task trials was randomized and was different for each participant. The duration of one run was $(216 + 14) \times 2.5\ \text{s} = 575\ \text{s}$. Each participant completed five experimental runs, except for one participant who completed only four runs due to feeling unwell. In addition, a 50-s rest run, during which participants were instructed to maintain fixation at the fixation dot on the gray background, was collected before the task runs. All runs were completed in one measurement session.

MEG data acquisition and preprocessing. MEG data were recorded in a magnetically shielded room with a whole-scalp 306-channel MEG device (MEGIN Oy) at the MEG Core (Aalto NeuroImaging, Aalto University School of Science). The device comprises 102 triple-sensor elements, with one magnetometer and two orthogonal planar gradiometers at each location. The signals were sampled at 1,000 Hz using a recording passband of 0.03–330 Hz. The position of the participant's head with respect to the MEG sensors was tracked throughout the experiment using five head position indicator coils. Horizontal and vertical electro-oculograms (EOGs) were recorded with the same passband and sampling rate as applied for the MEG data.

The MEG data were preprocessed using spatiotemporal signal-space separation implemented in the MaxFilter software (MEGIN Oy) to suppress magnetic interference of external sources and to compensate for head movement (Taulu and Simola, 2006). Eyeblink and heartbeat-related artifacts were removed by applying independent component analysis (ICA) using the FastICA algorithm (Hyvärinen and Oja, 2000) as implemented in the MNE-Python software package (Gramfort et al., 2013). EOG signals were used as a reference to find eye-related artifacts. For one participant, EOG signal was not available, and the blink-related components were manually selected. Similarly, components related to heartbeat were manually identified based on the topography and time course of the IC components. The data were low-pass filtered at 40 Hz using the MNE-Python software package (Gramfort et al., 2013). Next, the data were spatially whitened based on a covariance matrix estimated from the 50 s rest MEG data collected before the task runs using custom MATLAB code, similarly as done in Kurki et al. (2022). Whitening with dimensionality reduction was applied to reduce redundancy in the signals that originates from the MEG sensors having overlapping sensitivity profiles for the underlying cortical sources and from the MaxFilter preprocessing that reconstructs the sensor signals using truncated series expansions. The dimensionality of the data was reduced from 306 (number of channels) to 66–71, which corresponds to the rank of the data after MaxFilter and ICA processing. Epochs were extracted from the continuous data from 200 ms before to 1,000 ms after stimulus onset and were baseline corrected from −200 to 0 ms.

Representational similarity analysis and statistical analysis. Representational similarity analysis (RSA; Kriegeskorte et al., 2008) was performed using the RSA Toolbox (Nili et al., 2014) and custom MATLAB code. The MEG representational dissimilarity matrices (MEG-RDMs) were constructed separately for each time point using cross-validated Euclidean distance estimates:

$$d^2_{\text{Euc,cv}}(r_j, r_i) = (r_j - r_i)^T_A (r_j - r_i)_B,$$

where i and j refer to two different conditions (different scene stimuli and/or different viewing conditions), r_j and r_i refer to response patterns for the conditions i and j across the whitened sensor space, and A and B refer to two independent splits of the data. Leave-one-run-out cross-validation folds were used to create unbiased estimates (Walther et al., 2016; Guggenmos et al., 2018; Arbuckle et al., 2019). Results were averaged across the folds. The dimensions of an RDM were $108 \times 108 \times 1{,}201$, corresponding to the 36 scenes in three different viewing conditions and 1,201 time points (Fig. 2*A*). The three $36 \times 36 \times 1{,}201$ RDMs corresponding to the three different viewing conditions were extracted from this RDM for further analysis (Fig. 2*A*,*B*). In addition, the effect of the viewing condition (2D, 3D, reversed-stereo) was evaluated from the off-center diagonals corresponding to the distance estimate between the same scene under two different viewing conditions (Fig. 2*A*).

The replicability of the representational structure in the MEG-RDMs at each time point was evaluated by comparing each individual participant's MEG-RDM with the average of the other participants' MEG-RDMs using Kendall's tau-a rank correlation (Nili et al., 2014). The significance was tested using signed-rank test across the correlations

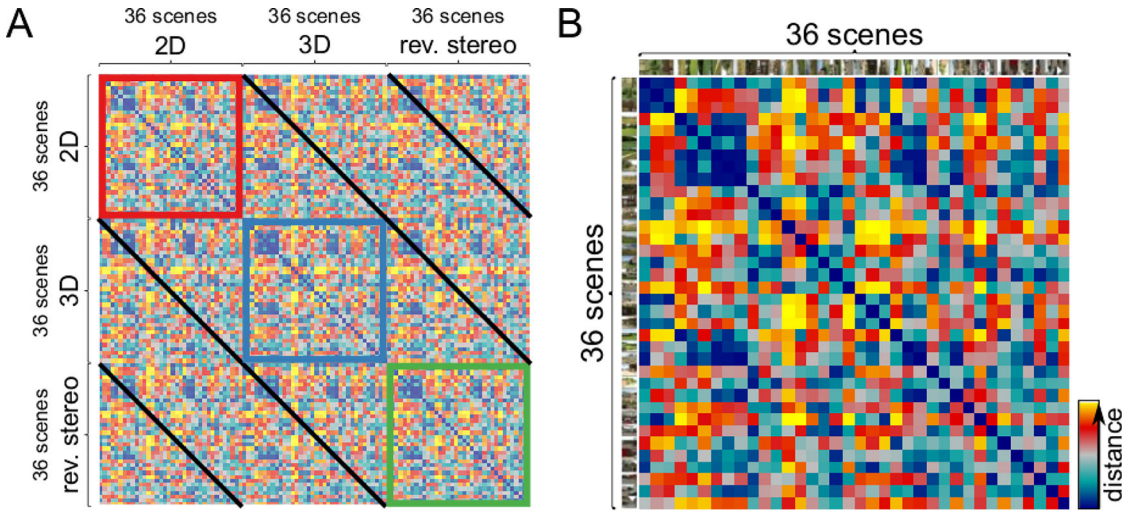


Figure 2. MEG-RDMs. ***A***, The discriminability of the 36 scenes was evaluated from the MEG response patterns for each pair in three different viewing conditions: 2D, 3D, and reversed-stereo (depicted with red, blue, and green outlines). The off-center diagonals (depicted with black lines) correspond to the discriminability of the same scene under two different viewing conditions. The analyses were done separately for each time point and individual, and the results were averaged across participants. For visualization purposes, the representational dissimilarity matrices (RDMs) are rank transformed (Nili et al., 2014). ***B***, An example MEG-RDM averaged across viewing conditions at 100 ms from stimulus onset is shown. See Movie 1 for time-varying RDM and visualization using multidimensional scaling.

values, and multiple testing across time points was accounted for by controlling the false discovery rate (FDR). The average RDM replicability can also be interpreted as the (lower bound of the) noise ceiling (Nili et al., 2014).

Generalization between 2D and 3D viewing conditions was analyzed with distance estimates

$$d^2_{\text{Euc,2D3Dgen}}(r_j, r_i) = (r_j^{\text{2D}} - r_i^{\text{2D}})_A^T (r_j^{\text{3D}} - r_i^{\text{3D}})_B,$$

where 2D and 3D refer to the different viewing conditions, i and j refer to two different scenes, r_j and r_i refer to the response patterns for these scenes across the whitened sensor space, and A and B refer to independent splits of the data based on the experimental runs. The distance estimates were averaged across the leave-one-run-out cross-validation folds. The replicability of and the average scene discriminability in the resulting 2D/3D-generalized MEG-RDMs were compared with results obtained with MEG-RDMs constructed from response patterns during 3D viewing.

Candidate model RDMs (Fig. 3) were constructed from the human-generated ratings for visual appearance, spatial structure, and semantic content (Fig. 1*B*; Anderson et al., 2021). To derive category labels that humans naturally use, Anderson et al. (2021) had participants sort images from the SYNS dataset (Adams et al., 2016) into discrete categories separately by the type of place (semantic task), by their 3D depth structure (spatial structure task), and by their 2D appearance while ignoring the 3D structure (visual appearance task). They used a data-driven clustering method to derive the representative category labels

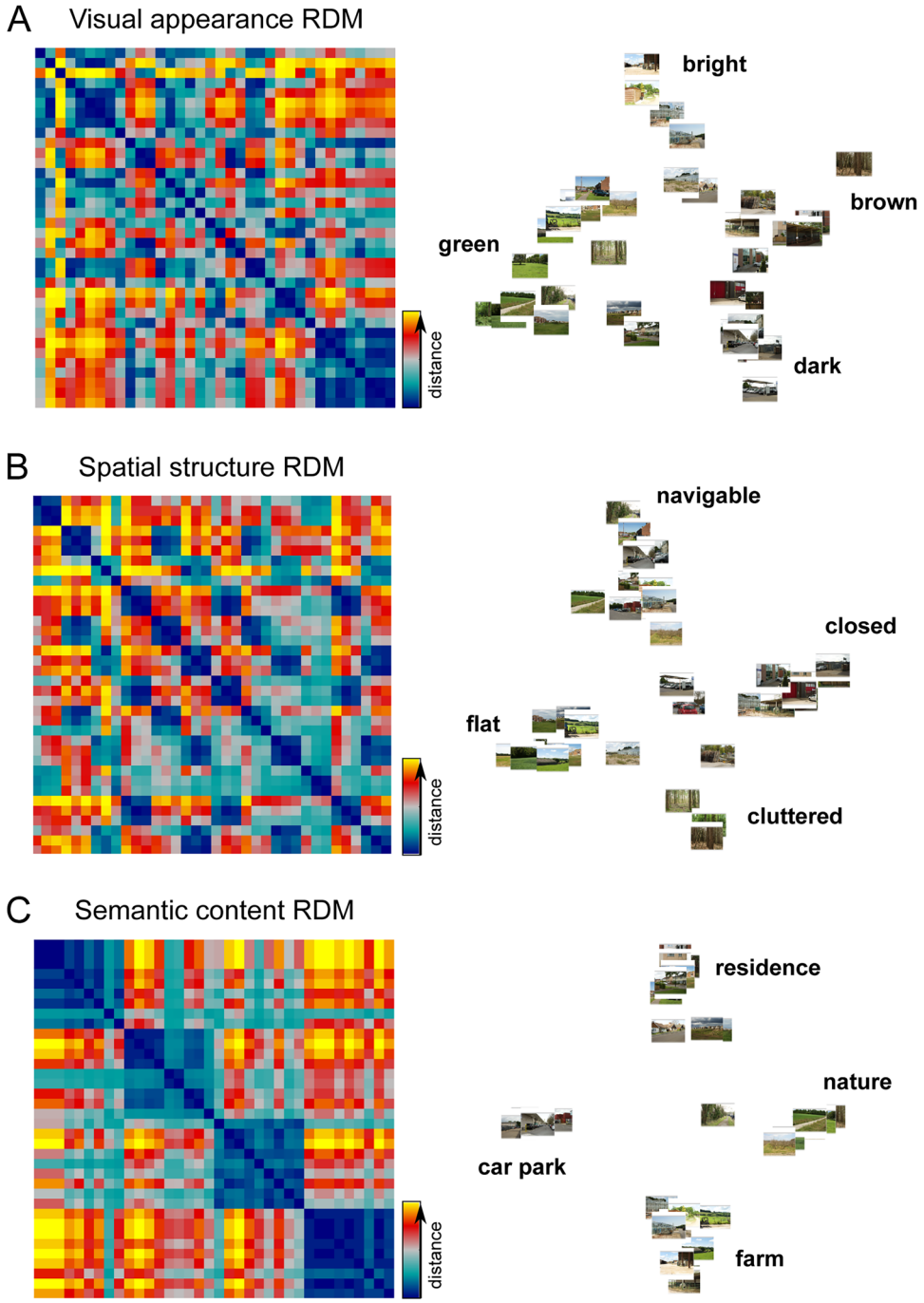


Figure 3. Candidate model RDMs and their multidimensional scaling visualizations for (***A***) visual appearance, (***B***) spatial structure, and (***C***) semantic content. RDMs were constructed from the human-generated ratings (Fig. 1*B*) by Anderson et al. (2021). Clusters are named for visualization purposes based on the representative category labels.

for each task. Here we used the ratings from their Experiment 2, where they had validated the category labels with 20 new participants for each of the three tasks (Anderson et al., 2020, 2021). For visual appearance, the scenes had been categorized as dark, bright, blue, green, or brown. For spatial structure, the categories were cluttered (or pointy), closed off, flat, and navigable routes (or tunnel). The semantic category labels were nature, residence, farm, car park, road, and beach. We constructed the RDMs separately for visual appearance, spatial structure, and semantic content based on the Euclidean distance between the category ratings of each pair of scenes. Figure 3 shows these candidate model RDMs and their visualizations using multidimensional scaling (metric stress).

MEG-RDMs were also compared with predictions based on low-level visual features as captured by the GIST descriptor (Oliva and Torralba, 2001), mean depth from LiDAR data provided with the SYNS dataset (Adams et al., 2016), and mean chromaticity of each scene in CIE1931 x–y coordinates. MEG-RDMs were compared with the candidate model RDMs using Kendall's tau-a rank correlation. The relatedness was tested using signed-rank test across the single-subject RDM correlations. Multiple testing across time points was accounted for by controlling the FDR.

Multiple candidate models were fitted together using non-negative least-squares regression (Khaligh-Razavi and Kriegeskorte, 2014). The unique contribution of a candidate model was evaluated by comparing the explained variance (tau-a^2) of the full combined model to the explained variance of the model, where the candidate model had been left out (Khaligh-Razavi and Kriegeskorte, 2014; Henriksson et al., 2019). The fitting was cross-validated using a leave-one-participant-out approach.

Behavioral data acquisition and analysis. After the MEG measurement, each participant completed a similarity judgment task for the 36 scene images using a multi-arrangement task (Kriegeskorte and Mur, 2012). The images were presented on a computer screen, and participants were asked to spatially arrange the images based on their similarity in a circular arena. They were not given any specific instructions on what features to pay attention to when arranging the scenes. The first image set included all 36 images. Following trials included subsets of the full image set based on the adaptive algorithm for efficient acquisition of the large number of pairwise dissimilarity judgments (Kriegeskorte and Mur, 2012). The task duration was limited to 15 min. Participants completed on average 10 trials (range, 6–20) and spent on average 5 min (range, 3–8.5 min) to arrange the complete set of images on the first trial.

For each participant, the representational dissimilarity matrix constructed based on the scene similarity judgments (behavior-RDM) was compared with the candidate model-RDMs for visual appearance, spatial structure, and semantic content using Kendall's tau-a rank correlation. The relatedness of a candidate model-RDM and behavior-RDM was tested using signed-rank test across the single-subject correlations. The upper and lower bounds of the noise ceiling were estimated using the RSA toolbox (Nili et al., 2014). The unique and shared contributions of the models in explaining the behavior-RDM were evaluated using commonality analysis (Lescroart et al., 2015; Tarhan et al., 2021). Non-negative least-squares regression was used to fit the models, and the result was cross-validated across subjects. The unique contribution of a model, here as an example for the visual model, was calculated as follows:

$$\text{unique\_var}_{\text{vis}} = \text{tau-a}^2_{\text{vis,spa,sem}} - \text{tau-a}^2_{\text{spa,sem}},$$

where $\text{tau-a}^2_{\text{vis,spa,sem}}$ is the explained variance of the full model and $\text{tau-a}^2_{\text{spa,sem}}$ is the explained variance of the model, where the visual model has been left out. Variance shared by all three models was calculated as follows:

$$\begin{aligned}\text{shared\_var}_{\text{vis,spa,sem}} = {} & \text{tau-a}^2_{\text{vis}} + \text{tau-a}^2_{\text{spa}} + \text{tau-a}^2_{\text{sem}} - 2 \cdot \text{tau-a}^2_{\text{vis,spa,sem}} \\ & + \text{unique\_var}_{\text{vis}} + \text{unique\_var}_{\text{spa}} + \text{unique\_var}_{\text{sem}},\end{aligned}$$

and the variance shared between two models, e.g., visual and semantic, was calculated as follows:

$$\begin{aligned}\text{shared\_var}_{\text{vis, sem}} = {} & \text{tau-a}^2_{\text{vis}} + \text{tau-a}^2_{\text{sem}} - \text{tau-a}^2_{\text{vis,sem}} \\ & - \text{shared\_var}_{\text{vis,spa,sen}}.\end{aligned}$$

The MEG-RDMs were also compared with the average behavior-RDM using Kendall's tau-a rank correlation similarly to how the comparison was done with the other candidate models.

Data availability. Behavioral-RDMs, MEG-RDMs, candidate model RDMs, and the scene stimuli are uploaded to the Open Science Foundation repository at https://osf.io/jp26k/.

Results

Behavioral scene similarity judgments emphasize semantic content

We selected 36 natural scenes from the SYNS database (Adams et al., 2016) as the stimuli for the MEG experiment (Fig. 1*A*). The selection of the scenes was based on human-generated labels for semantic category and spatial structure (Anderson et al., 2021). We aimed for a scene stimulus set that varies both in semantic content and spatial structure (Fig. 1*B*; e.g., a navigable nature scene, a flat nature scene, a flat residential scene). The scene images also had labels for visual appearance (dark, bright, blue, green, brown; Anderson et al., 2021). First, we ask to what extent our study participants used these dimensions when their behavioral task was to arrange the scenes by their similarity (Fig. 4*A*). The scene similarity judgments were collected after the MEG experiment, and the participants were not given any specific instructions on what features to pay attention to.

Figure 4*B* shows the group-averaged representational dissimilarity matrix (RDM) based on the behavioral scene similarity judgments. The candidate model RDM based on the semantic content ratings (Fig. 3*C*) best explained the behavioral RDM (mean rank correlation coefficient tau-a of 0.26; $p < 0.001$, one-tailed signed rank test across the 20 participants; Fig. 4*C*). In other words, our participants' judgments on the scene similarity emphasized the semantic categories: nature, residence, farm, and car park. The second-best model was based on the visual appearance (tau-a = 0.19; $p < 0.001$), and the spatial structure had the worst model fit (tau-a = 0.06; $p < 0.001$; Fig. 4*C*).

The unique contribution of each of the models in explaining the behavioral data was evaluated using commonality analysis (Lescroart et al., 2015; Tarhan et al., 2021). The semantic content accounted for most of the unique variance (tau-a^2 = 0.045; Fig. 4*D*), the visual appearance accounted for some of the unique variance (tau-a^2 = 0.010), and these models also shared some variance (tau-a^2 = 0.020). The spatial structure did not account for any unique variance.

Taken together, our study participants used the semantic and visual appearance dimensions when judging the scene similarity.

Time-resolved response discriminability for the scene stimuli in the MEG data

We recorded MEG data from the participants while they viewed the 36 scene stimuli during 2D, 3D, and reversed-stereo viewing conditions. A scene image evokes a complex MEG response that is difficult to interpret directly. Therefore, we applied representational similarity analysis (RSA; Kriegeskorte et al., 2008) to capture the relative distinctiveness of the responses between the scene stimuli.

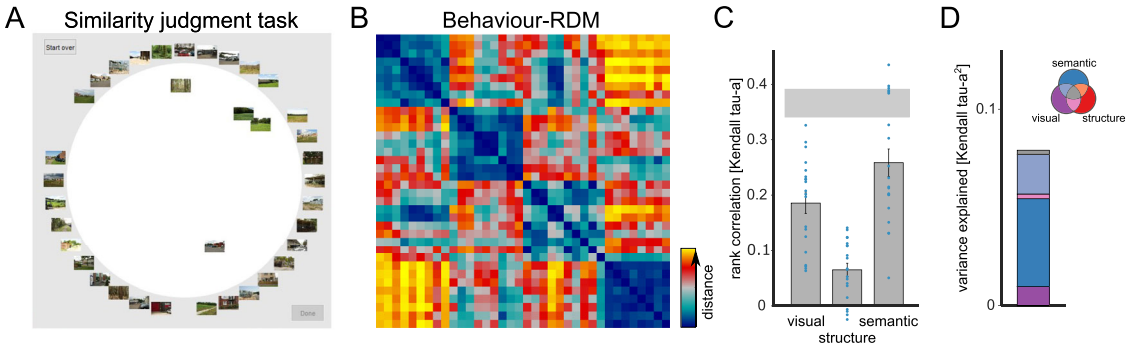


Figure 4. Behavioral task: scene similarity judgments. ***A***, Participants judged the similarity of the 36 scenes by arranging them in a circular arena (Kriegeskorte and Mur, 2012). This task was completed after the MEG data collection. ***B***, A representational dissimilarity matrix (RDM) was constructed based on the similarity judgments (Behavior-RDM). Here, the group-average result is shown. ***C***, Rank correlations between the behavioral RDM and the candidate model RDMs for visual appearance, spatial structure, and semantic content are shown. Blue dots show the individual data ($N = 20$), and gray horizontal bar indicates the noise ceiling. ***D***, The unique and shared variance of the behavioral data explained by the combinations of the three models are shown. Venn diagram illustrates the meaning of each color.

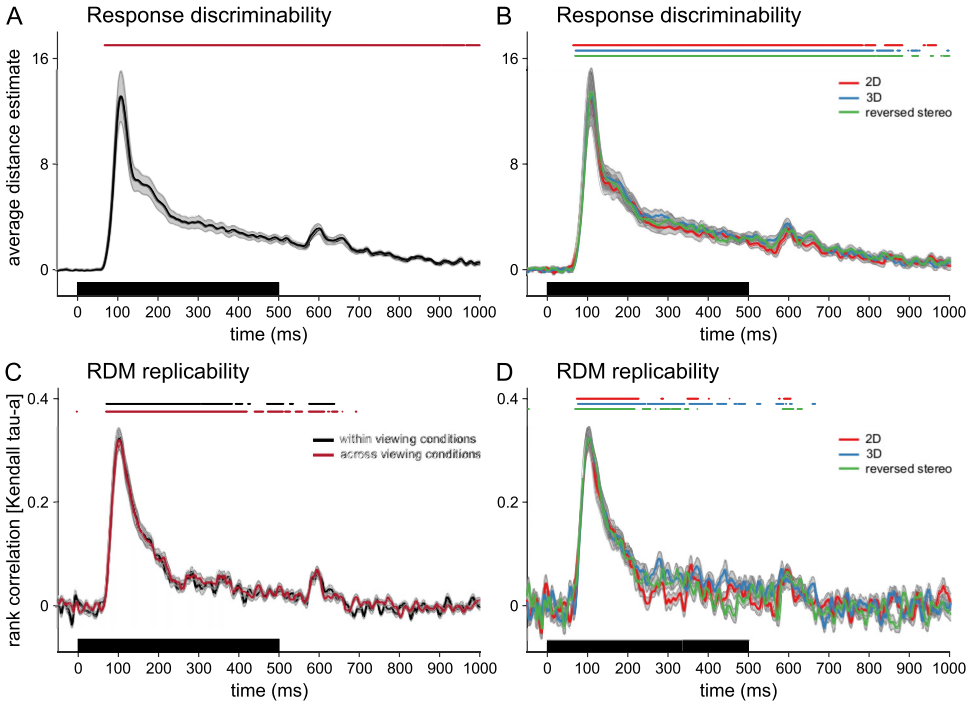


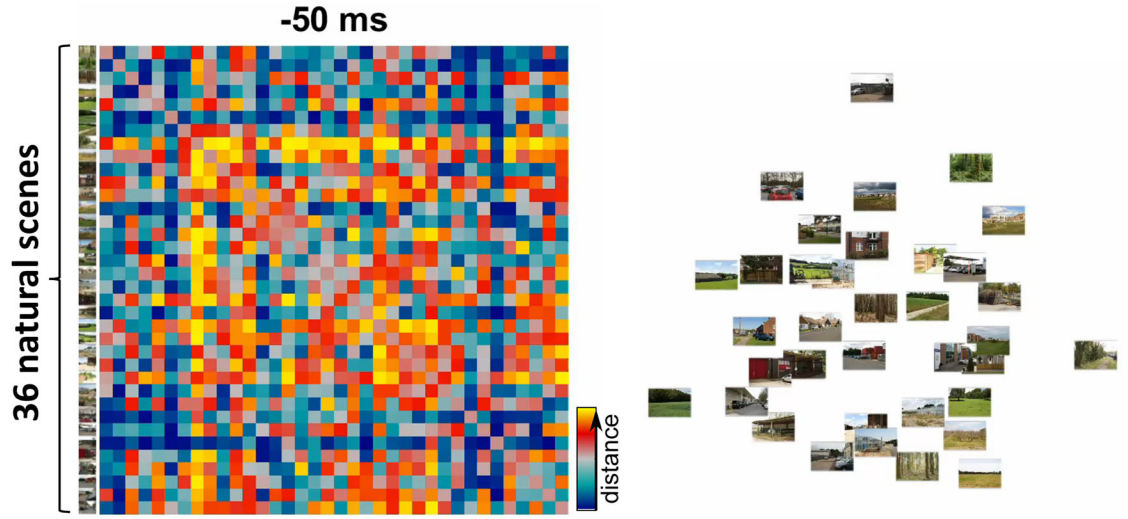
Figure 5. Time-resolved scene decoding and RDM replicability. ***A***, The time course of average response discriminability is shown. Shaded region indicates the standard-error-of-the-mean (SEM) across participants. Significant time points are indicated with red (FDR of 0.01; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points). The black bar indicates the stimulus on-period (0–500 ms). ***B***, Response discriminability results are shown separately for the 2D (shown in red), 3D (blue), and reversed-stereoscopic (green) viewing conditions. ***C***, The replicability of the representational structure as calculated from the rank correlation between RDMs from different participants is shown. Replicability was assessed both within the viewing conditions (shown in black) and across the viewing conditions (shown in red). Significant time points are indicated by black and red lines correspondingly (FDR of 0.01; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points; stimulus onset at 0 ms). There is no significant difference between the across and within RDM replicabilities. ***D***, RDM replicability is shown separately for the three viewing conditions in different colors.

The MEG-RDMs (Fig. 2) were constructed based on cross-validated Euclidean distance between the evoked responses of each pair of the scene stimuli. In each cell of an RDM, a systematically positive distance estimate indicates reliable difference between the response patterns corresponding to the two scenes. To evaluate how well the MEG signals discriminated between the scene stimuli, we averaged all pairwise distances in the lower triangular of each RDM. On average, the scene stimuli elicited distinct response patterns from ~70 ms from the stimulus onset (peak at ~110 ms; Fig. 5*A*). The results were highly similar between 2D, 3D, and reversed-stereo viewing conditions (Fig. 5*B*).

Furthermore, an RDM (Fig. 2*B*) captures which scenes are represented more similarly and which have more distinct representations. The dynamic representational structure in the group-average MEG-RDM is shown in Movie 1. We can study the reliability of the emerging structure in the MEG-RDMs by comparing the RDMs computed for individual participants. The MEG-RDMs showed replicable structure between participants from ~70 ms from stimulus onset (Fig. 5*C*; peak at ~100 ms). The results were highly similar in the three viewing conditions (Fig. 5*D*). Figure 5*C* shows the average RDM replicability across viewing conditions, where, for example, an individual participant's MEG-RDM during 2D viewing was compared with the average of the other participants' MEG-RDMs during 3D viewing, and this was repeated for all combinations of the viewing conditions. There was no significant difference between the across and within RDM replicabilities, and hence, when we next compare the MEG-RDMs with different model predictions, the results for different viewing conditions are averaged.

Spatial structure and semantic content explain complementary parts of the MEG-RDMs

Next, we aim to interpret the emerging structure in the MEG-RDMs by comparing them with candidate model RDMs based on the visual appearance, spatial structure, and semantic



Movie 1. Time-varying MEG-RDM. The dynamic representational structure in the group-average MEG-RDM is shown and visualized using multidimensional scaling. [View online]

content of the scene stimuli (Fig. 3). Figure 6 shows the average rank correlation between each model and the MEG-RDMs. All models showed significant correlation with the MEG-RDMs (onset ~80 ms from stimulus onset). The best model fit (peak at ~110 ms from stimulus onset) was obtained with the spatial structure model, which was constructed based on human ratings for category labels: cluttered, closed off, flat, and navigable routes (Anderson et al., 2021).

The unique contribution of each of the three candidate models was evaluated by fitting the models to the MEG-RDMs together and evaluating whether including a model significantly improved the fit (Khaligh-Razavi and Kriegeskorte, 2014; Henriksson et al., 2019). The results are shown in Figure 6*D*,*E*. The candidate model based on the spatial structure had a unique contribution to the MEG-RDM in time window 90–125 ms, followed by a unique contribution of the semantic content in time window 140–175 ms. The visual appearance model did not explain any unique variance.

We also considered competing candidate models at explaining the MEG-RDMs. First, the Gist model (Oliva and Torralba, 2001) has previously been successful in explaining early MEG responses for scene stimuli (Henriksson et al., 2019). For our present data, the Gist model explained structure in the MEG-RDMs from ~90 ms from stimulus onset, but the correlation was much lower compared with the other models (Fig. 7*A*). Next, the mean distance information in each scene was calculated from the LiDAR depth data provided with the scene dataset (Adams et al., 2016). The similarity in scene distance (near–far) correlated with the MEG-RDMs with a similar time course as the spatial structure (peak at ~110 ms), though not providing an equally good fit (Fig. 7*B*). Mean chromaticity of each scene was also calculated, and the Euclidean distance in CIExy color space showed significant correlation with MEG-RDMs in time window 110–190 ms (Fig. 7*C*).

Finally, we compared the MEG-RDMs with the average RDM from the behavioral scene similarity judgments (behavior-RDM). The behavior-RDM was significantly correlated with the MEG-RDMs from ~75 ms onward (Fig. 7*D*), though not providing an equally good fit in the early time window as the spatial structure. Overall, the behavior-RDM had a similar time course with the model based on the visual appearance and the semantic content, which was expected based the result that our study participants' scene similarity judgments were well captured by these dimensions (Fig. 4), and the behavior-RDM correlated with these model-RDMs (Fig. 7*E*). When fitted to the MEG-RDMs together, the average behavior-RDM did not explain any unique variance that would not have been explained by the visual appearance and semantic content RDMs.

Taken together, the MEG-RDMs first reflected dimensions related to the spatial structure of the scenes followed by the semantic content and perceptual similarity.

Generalization across 2D and 3D viewing conditions

As our participants viewed the scenes under different viewing conditions, we can ask how well our results generalize across the viewing condition and if there is any superiority in the 3D viewing condition. To ask the question whether the differences in response patterns to a pair of scenes during 2D viewing generalize to 3D viewing, we built the 2D/3D generalized MEG-RDMs by cross-validating the distance estimate using responses from different viewing conditions. Figure 8*A* compares the MEG-RDM replicability between the 2D/3D generalized MEG-RDMs with the MEG-RDMs during 3D viewing. Interestingly, the 2D/3D generalized MEG-RDMs showed a better replicability in time windows 70–170 and 190–200 ms compared with the MEG-RDMs constructed from the responses during 3D viewing. That is, generalization across viewing conditions reduced variability in the representational structure at the group level. When compared with the candidate model RDMs, the 2D/3D generalized MEG-RDMs showed an overall similar result to the results obtained within the viewing conditions (compare Figs. 8*B*, 6*E*) with the unique contribution of the spatial structure (90–130 ms) being followed by semantic content (145–170 ms).

Figure 8*C* compares how well on average the scene pairs could be discriminated from the MEG signals when generalized across the 2D and 3D viewing conditions and when evaluated within the 3D viewing condition. The results show superiority of the 3D viewing condition from ~150 ms from stimulus onset with on average more distinct scene responses during 3D viewing compared with the 2D/3D generalized results.

Compared with 2D viewing, the 3D viewing made the scene responses more distinct from ~150 ms onward (Fig. 8*C*) but appeared not to significantly affect the emerging representational geometry of our selection of scene stimuli (Figs. 5*C*, 8).

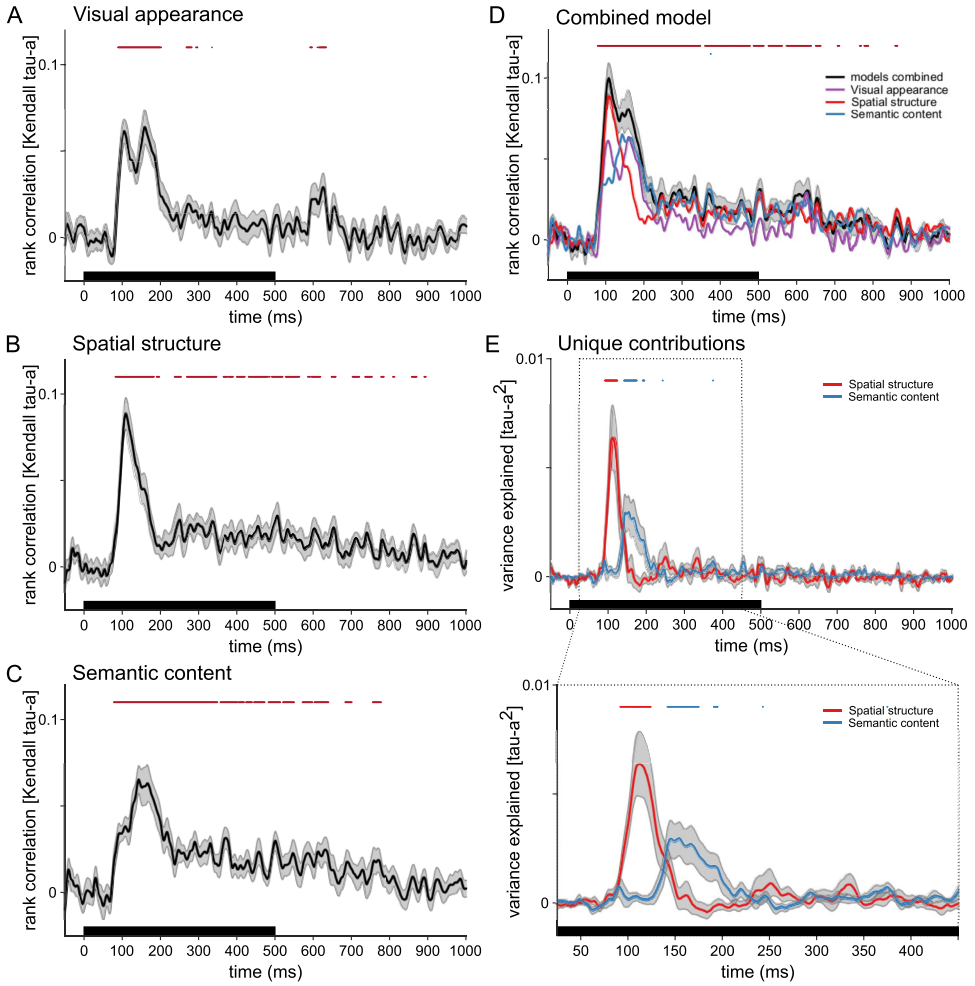


Figure 6. The contributions of visual appearance, spatial structure, and semantic content in explaining the MEG-RDMs. ***A***, The time course of the mean rank correlation between the MEG-RDMs and the candidate model based on visual appearance of the scenes is shown. Shaded region indicates the standard-error-of-the-mean (SEM) across participants. Significant time points are indicated with red (FDR of 0.05; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points). The black bar indicates the stimulus on-period (0–500 ms). ***B***, The time course of the mean rank correlation between the MEG-RDMs and the candidate model based on spatial structure of the scenes is shown. ***C***, The time course of the mean rank correlation between the MEG-RDMs and the candidate model based on semantic content of the scenes is shown. ***D***, Rank correlations with each of the three models are shown in different colors (visual appearance in purple; spatial structure in red; semantic content in blue). The unique contribution of each of these candidate models was evaluated by fitting the models to the MEG-RDMs together (black line). The significant time points for the combined model are indicated with red (FDR of 0.05; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points). ***E***, The unique variance explained by the spatial and semantic content are shown, with a zoom-in to the early time window shown below the full time course. Including the spatial structure model significantly improved the model fit in time window 90–125 ms. Including the semantic content model significantly improved the model fit in time window 140–175 ms. The visual appearance model did not account for any unique variance. Significant time points are indicated with red and blue, corresponding to the two candidate models (FDR of 0.05; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points).

Difference between scenes viewed in 2D versus 3D

Finally, we directly ask whether we can discriminate the viewing condition from the MEG data. Figure 9*A* shows the average response discriminability for the same scene between 3D and 2D viewing conditions. The viewing condition had a significant effect on the response discriminability from ∼140 ms from stimulus onset (peak at ∼180 ms). A similar result was obtained when comparing the 2D viewing condition with reversed-stereo condition (Fig. 9*B*). The response discriminability between the 3D and reversed-stereo viewing conditions showed a similar onset but the average response discriminability was lower compared with the comparison with the 2D viewing condition (Fig. 9*C*).

Taken together, the stereoscopic viewing of the scenes affected the responses relatively late with onset latency ∼140 ms (Fig. 9). This timing overlaps with the time window where the scenes' semantic content explained the MEG-RDMs, whereas the spatial structure of the scenes explained the MEG-RDMs already in an earlier time window of 90–125 ms (Fig. 6).

Discussion

The human visual system analyses complex real-world visual scenes rapidly, enabling us to smoothly navigate and interact with our visual surroundings. The aim of the current study was to characterize how the cortical encoding of a scene's spatial structure and semantic content unfold over time and to examine the role of 3D disparity cues on the cortical responses of real-world scenes. We found that information related to the spatial structure of the scenes (cluttered, closed off, flat, and navigable) emerged in the MEG responses earlier than information related to the semantic content (nature, residence, farm, car park). Interestingly, the 3D disparity cues from stereoscopic viewing of the scenes affected the responses relatively late, in a time window overlapping with the processing of the semantic content instead of the spatial structure of the scene.

The finding on the rapid encoding of a scene's spatial structure is consistent with previous M/EEG studies (Cichy et al., 2017; Lowe et al., 2018; Henriksson et al., 2019; Harel et al., 2022; see, however, Dwivedi et al., 2024). Lowe et al. (2018) found that the discrimination of open versus closed natural scenes was possible from EEG signals within 100 ms from stimulus onset. Cichy et al. (2017) measured MEG responses for gray-scale scenes that differed in physical size and clutter level, and they reported that neural representations of clutter level peaked at 107 ms from stimulus onset whereas representations of scene size emerged later. In our recent study, we measured both fMRI and MEG responses to artificial natural scenes that had a complete set of all combinations of the five scene-bounding elements (floor, ceiling, walls; Henriksson et al., 2019). The fMRI results showed that the scene-responsive cortical area OPA encodes

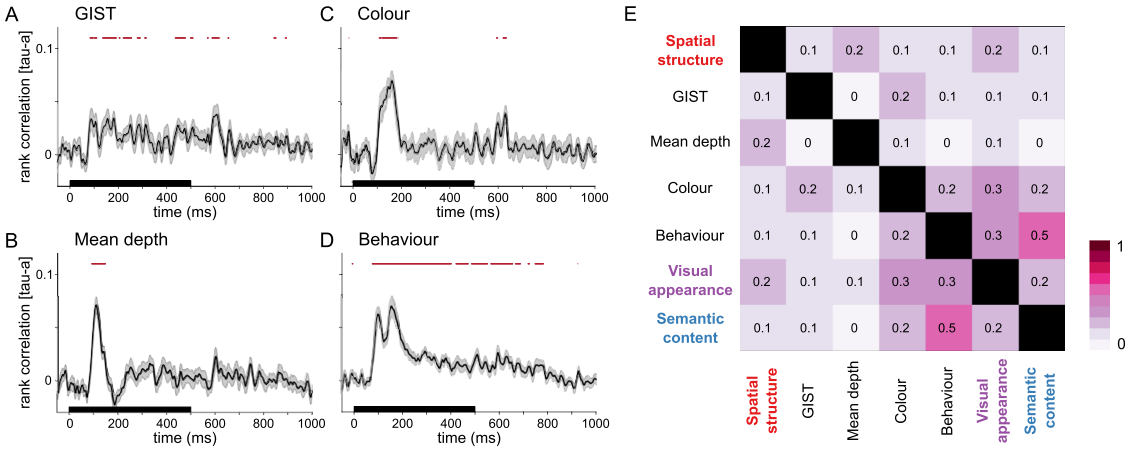


Figure 7. Results from alternative models. The time course of the mean rank correlation between the MEG-RDMs and the alternative candidate models based on (***A***) low-level features as captured by the GIST descriptor, (***B***) scene distance (near–far) as calculated from the LiDAR data provided with the scene images, (***C***) mean chromaticity of the scenes, and (***D***) behavioral scene similarity judgments are shown. Shaded region indicates the standard-error-of-the-mean (SEM) across participants. Significant time points are indicated with red (FDR of 0.05; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points). The black bar indicates the stimulus on-period (0–500 ms). ***E***, Pairwise rank correlations of all model-RDMs are shown.

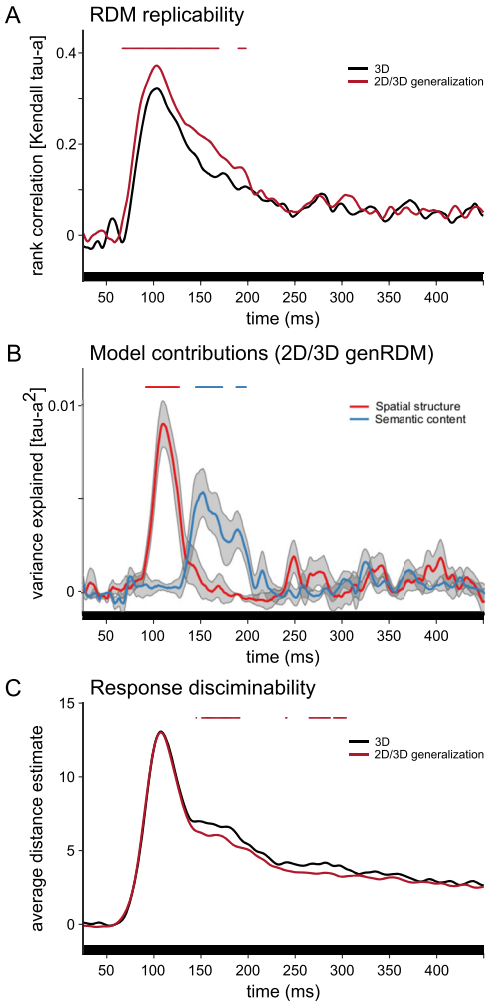


Figure 8. Generalization across viewing conditions. ***A***, The replicability of the MEG-RDMs during 3D viewing (shown in black) are compared with the replicability of the MEG-RDMs generalized across 2D and 3D viewing conditions (shown in red). Time points with significant difference are indicated with red (FDR of 0.05; *p* values computed with two-tailed signed rank test across the 20 participants, FDR adjusted across time points). ***B***, Spatial structure (shown in red) followed by semantic content (shown in blue) explained unique variance when the three candidate models (Fig. 3) were fitted together to the 2D/3D generalized MEG-RDMs (FDR of 0.05; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points; compare with Fig. 5*E*). ***C***, The average scene discriminability during 3D viewing (shown in black) is compared with the average scene discriminability generalized between 2D and 3D viewing (shown in red). Time points with significant difference are indicated with red (FDR of 0.05; *p* values computed with two-tailed signed rank test across the 20 participants, FDR adjusted across time points).

scene-layout with invariance to changes in surface texture (see also Lescroart and Gallant, 2019), and complementary MEG evidence indicated that this texture-invariant scene-layout representation emerges within ∼100 ms (Henriksson et al., 2019). Harel et al. (2022) studied cortical encoding of navigational affordances by measuring EEG responses while participants viewed pictures of computer-generated room scenes that had different numbers of doors. They found that the amplitude of the P2 ERP was modulated by the navigability of the scenes and that the number of doors and their locations could be decoded from the EEG responses at even earlier time windows. Altogether, the rapid cortical encoding of a scene's spatial structure and navigational affordances supports smooth navigation through the immediately visible environment.

Our study participants only used semantic content and visual appearance when making behavioral judgments of the scene similarity, but cortical responses were evidently affected by the spatial structure of the scenes. Similarly, Groen et al. (2018) used a multi-arrangement task similar to ours with scene stimuli and found a dissociation between behavioral and fMRI results. Their behavior-RDM was best captured by a functional model of potential actions in a scene whereas their fMRI-RDMs from scene-responsive areas were better captured by mid-level layers of a deep neural network (DNN) model. Though our behavioral result could reflect participants' implicit tendency to interpret the similarity judgment task as a categorization task and therefore to arrange the scenes in terms of their basic level category membership (e.g., beach, mountain; Tversky and Hemenway, 1983), the dissociation of the behavioral and early brain responses could also reflect a more profound difference between the spatial and semantic dimensions. Previously, Greene and Oliva (2009a,b)

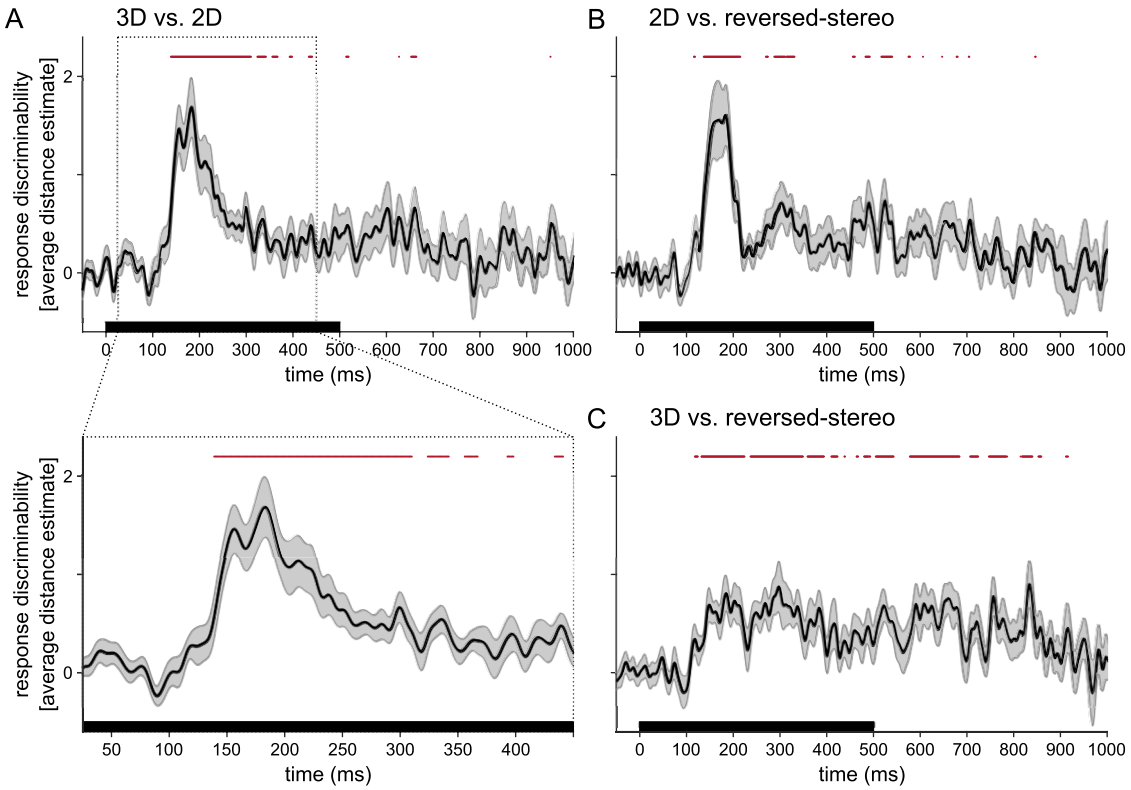


Figure 9. Scenes viewed in 2D versus 3D. ***A***, The average discriminability between the same scene viewed in 2D and stereoscopically in 3D is shown, with a zoom-in to the early time window shown below the full time course. Shaded region indicates the standard-error-of-the-mean (SEM) across participants. Significant time points are indicated with red (FDR of 0.05; *p* values computed with one-tailed signed rank test across the 20 participants, FDR adjusted across time points). The black bar indicates the stimulus on-period (0–500 ms). ***B***, Average discriminability between a scene viewed in 2D and in reversed-stereo viewing condition. ***C***, Average discriminability between a scene viewed in 3D and in reversed-stereo viewing condition.

showed that human observers perceive spatial dimensions, such as the mean depth or openness of a scene, from a shorter presentation time compared with the basic level category and concluded that spatial aspects precede category information in scene recognition. Furthermore, they argued that the entry level for communication about scenes may still be the basic level category. Together, these findings suggest that cortical processing of a scene's spatial structure occurs rapidly and likely automatically.

The three cortical scene-responsive areas OPA, PPA, and retrosplenial complex (RSC) have been suggested to be dissociated based on whether their processing is automatic or deliberate (for a review, see Dilks et al., 2022). By comparing passive and active navigation tasks in simple and complex environments, a recent fMRI study showed the RSC to be involved in deliberate navigation through the broader spatial environment, whereas the scene-responsive cortical area OPA was shown to support navigation through the local visual environment automatically (Suzuki et al., 2021). The central role of OPA in the automatic extraction of navigational affordances of the immediate environment was also demonstrated in the fMRI study by Bonner and Epstein (2017) and was later attributed to purely feedforward computations (Bonner and Epstein, 2018). Interestingly, a recent study where participants were explicitly asked to identify navigational paths in scene images during EEG data collection suggested that this information is processed significantly later compared with more basic scene features, such as edges or depth, or the scene's semantic content (Dwivedi et al., 2024). In the results presented here, the timing of the effect of scenes' spatial structure on the MEG responses together with the negligible effect of the spatial structure on the behavioral results are consistent with the view that our visual system, most likely OPA, rapidly and automatically encodes spatial structure and navigability of our local visual surroundings. An open question remains how this information is transformed and used while acting in the local visual environment in comparison with passive viewing.

Few previous M/EEG studies have studied how binocular depth cues in natural scenes affect cortical responses (Fischmeister and Bauer, 2006, Duan et al., 2018, 2021). Fischmeister and Bauer (2006) presented different scenes in stereoscopic, 2D and scrambled modes during EEG recordings with the aim to localize the cortical areas activated when switching, for example, from 2D to stereoscopic viewing. Their main conclusions were on the feasibility of the use of stereoscopic natural images as stimuli in EEG experiments with no definite conclusions on the location or timing of the stereoscopic depth cues. More recently, Duan et al. (2018, 2021) applied a method called reliable-component analysis to EEG data obtained during viewing of stereo-image pairs and 2D images of outdoor scenes. They reported a difference between the 2D and 3D scenes starting as early as 95 ms from stimulus onset and that the differential response between the 2D and 3D scenes was affected by the high-level scene structure (Duan et al., 2018). In a subsequent study, Duan et al. (2021) independently varied the monocular scene content and the stereoscopic cues in natural scene stimuli. They reported an interaction with the monocular scene content and the stereoscopic depth cues starting from 150 ms with scene-specific and perceptually predictive responses (Duan et al., 2021). This timing is consistent with what we found, namely, that stereoscopic viewing of the scenes affected the responses with the onset latency ~140 ms.

In our results, the encoding of the scene's spatial structure precedes the encoding of the semantic content and the sensitivity

to stereoscopic depth cues. The temporal overlap between the semantic content and stereoscopic depth cues suggests that the stereoscopic depth cues may not be critical for encoding of a scene's spatial structure but rather benefit object segmentation. Moreover, the similarity in the timing of the decoding results between the 2D versus 3D and 2D versus reversed-stereo viewing condition of the same scene further suggests that the disparity cues are related to figure-and-ground separation. This interpretation is consistent with previous studies on the advantage of stereoscopic depth cues on object recognition (Adams et al., 2019; Anderson et al., 2022).

Ultimately, we are interested in how the human visual system understands and enables the smooth interaction with our natural 3D environments. In the present study, we used a relatively small number of picture stimuli typical to brain imaging experiments. The stereoscopic depth cues brought vividness to the scene stimuli but had nonetheless surprisingly little effect on the cortical responses. A possible limitation here is the use of outdoor natural images, taken from a distance, as the stimuli. Binocular cues contribute to depth judgments at far distances (Palmisano et al., 2010; McCann et al., 2018), but they are most effective in the space within "grasping distance" (Cutting and Vishton, 1995). Hence, future studies could complement the present findings with close-up views of scenes. Such reachable-scale views of environments were recently suggested to have distinct cortical representations from objects and navigable environments (Josephs and Konkle, 2020) and could also show distinct response profile to different depth cues. Furthermore, an exciting direction for future research is to mimic real-world 3D visual environments and realistic visual tasks using virtual reality during brain imaging. Virtual environments would allow controlled manipulation of different depth cues and spatial layouts. Such stimulus sets would likely facilitate to more directly tease apart the roles of different monocular and binocular cues for object segmentation and overall layout perception compared with the relatively small number of real-world scenes that were tested in the present study. Challenges with naturalistic virtual reality experiments lie, however, both in the implementation of the experimental setups and in the analysis and interpretation of the complex data.

References

Adams WJ, Elder JH, Graf EW, Leyland J, Lugtigheid AJ, Muryy A (2016) The Southampton-York Natural Scenes (SYNS) dataset: statistics of surface attitude. Sci Rep 6:35805.

Adams WJ, Graf EW, Anderson M (2019) Disruptive coloration and binocular disparity: breaking camouflage. Proc Biol Sci 286:20182045.

Anderson MD, Elder JH, Graf EW, Adams WJ (2022) The time-course of real-world scene perception: spatial and semantic processing. Iscience 25:105633.

Anderson M, Graf E, Elder JH, Ehinger K, Adams W (2020) Dataset for: category systems for real-world scenes. Available at: https://eprints.soton.ac.uk/446699/

Anderson MD, Graf EW, Elder JH, Ehinger KA, Adams WJ (2021) Category systems for real-world scenes. J Vis 21:8.

Arbuckle SA, Yokoi A, Pruszynski JA, Diedrichsen J (2019) Stability of representational geometry across a wide range of fMRI activity levels. Neuroimage 186:155–163.

Baillet S (2017) Magnetoencephalography for brain electrophysiology and imaging. Nat Neurosci 20:327–339.

Bartnik CG, Groen II (2023) Visual perception in the human brain: how the brain perceives and understands real-world scenes. In: *Oxford research encyclopedia of neuroscience*. Oxford, UK: Oxford University Press.

Bonner MF, Epstein RA (2017) Coding of navigational affordances in the human visual system. Proc Natl Acad Sci U S A 114:4793–4798.

Bonner MF, Epstein RA (2018) Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. PLoS Comput Biol 14:e1006111.

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436.

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

Cutting JE, Vishton PM (1995) Perceiving layout and knowing distances: the integration, relative potency, and contextual use of different information about depth. In: *Perception of space and motion* (Epstein W, Rogers S, eds), pp 69–117. San Diego: Academic Press.

Dilks DD, Julian JB, Paunov AM, Kanwisher N (2013) The occipital place area is causally and selectively involved in scene perception. J Neurosci 33:1331–1336.

Dilks DD, Kamps FS, Persichetti AS (2022) Three cortical scene systems and their development. Trends Cogn Sci 26:117–127.

Duan Y, Thatte J, Yaklovleva A, Norcia AM (2021) Disparity in context: understanding how monocular image content interacts with disparity processing in human visual cortex. Neuroimage 237:118139.

Duan Y, Yakovleva A, Norcia AM (2018) Determinants of neural responses to disparity in natural scenes. J Vis 18:21.

Dwivedi K, Sadiya S, Balode MP, Roig G, Cichy RM (2024) Visual features are processed before navigational affordances in the human brain. Sci Rep 14:5573.

Epstein RA, Baker CI (2019) Scene perception in the human brain. Annu Rev Vis Sci 5:373–397.

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

Fischmeister FPS, Bauer H (2006) Neural correlates of monocular and binocular depth cues based on natural images: a LORETA analysis. Vision Res 46:3373–3380.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L (2013) MEG and EEG data analysis with MNE-python. Front Neurosci 267:1–13.

Greene MR, Hansen BC (2020) Disentangling the independent contributions of visual and conceptual features to the spatiotemporal dynamics of scene categorization. J Neurosci 40:5283–5299.

Greene MR, Oliva A (2009a) The briefest of glances: the time course of natural scene understanding. Psychol Sci 20:464–472.

Greene MR, Oliva A (2009b) Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cogn Psychol 58:137–176.

Grill-Spector K (2003) The neural basis of object perception. Curr Opin Neurobiol 13:159–166.

Groen II, Greene MR, Baldassano C, Fei-Fei L, Beck DM, Baker CI (2018) Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7:e32962.

Groen IIA, Silson EH, Baker CI (2017) Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. Philos Trans R Soc Lond B Biol Sci 372:20160102.

Guggenmos M, Sterzer P, Cichy RM (2018) Multivariate pattern analysis for MEG: a comparison of dissimilarity measures. Neuroimage 173:434–447.

Harel A, Nador JD, Bonner MF, Epstein RA (2022) Early electrophysiological markers of navigational affordances in scenes. J Cogn Neurosci 34:397–410.

Henriksson L, Mur M, Kriegeskorte N (2019) Rapid invariant encoding of scene layout in human OPA. Neuron 103:161–171.

Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13:411–430.

Josephs EL, Konkle T (2020) Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. Proc Natl Acad Sci U S A 117:29354–29362.

Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 10:e1003915.

Kleiner M, Brainard D, Pelli D (2007) What's new in psychtoolbox-3? Perception 36:1.

Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412.

Kriegeskorte N, Mur M (2012) Inverse MDS: inferring dissimilarity structure from multiple item arrangements. Front Psychol 3:245.

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

Kurki I, Hyvärinen A, Henriksson L (2022) Dynamics of retinotopic spatial attention revealed by multifocal MEG. Neuroimage 263:119643.
Lescroart MD, Gallant JL (2019) Human scene-selective areas represent 3D configurations of surfaces. Neuron 101:178–192.
Lescroart MD, Stansbury DE, Gallant JL (2015) Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front Comput Neurosci 9:135.
Lowe MX, Rajsic J, Ferber S, Walther DB (2018) Discriminating scene categories from brain activity within 100 milliseconds. Cortex 106:275–287.
Malcolm GL, Groen II, Baker CI (2016) Making sense of real-world scenes. Trends Cogn Sci 20:843–856.
McCann BC, Hayhoe MM, Geisler WS (2018) Contributions of monocular and binocular cues to distance discrimination in natural scenes. J Vis 18:12.
Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS Comput Biol 10:e1003553.
Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42:145–175.
Palmisano S, Gillam B, Govan DG, Allison RS, Harris JM (2010) Stereoscopic perception of real depths at large distances. J Vis 10:19.
Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10:437–442.
Piano ME, Tidbury LP, O'Connor AR (2016) Normative values for near and distance clinical tests of stereoacuity. Strabismus 24:169–172.
Ramkumar P, Hansen BC, Pannasch S, Loschky LC (2016) Visual information representation and rapid-scene categorization are simultaneous across cortex: an MEG study. Neuroimage 134:295–304.
Suzuki S, Kamps FS, Dilks DD, Treadway MT (2021) Two scene navigation systems dissociated by deliberate versus automatic processing. Cortex 140:199–209.
Tarhan L, De Freitas J, Konkle T (2021) Behavioral and neural representations en route to intuitive action understanding. Neuropsychologia 163:108048.
Taulu S, Simola J (2006) Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys Med Biol 51: 1759–1768.
Tversky B, Hemenway K (1983) Categories of environmental scenes. Cogn Psychol 15:121–149.
Vishwanath D (2023) From pictures to reality: modelling the phenomenology and psychophysics of 3D perception. Philos Trans R Soc Lond B Biol Sci 378:20210454.
Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J (2016) Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage 137:188–200.
Wheatstone C (1838) XVIII. Contributions to the physiology of vision.—Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. Philos Trans R Soc Lond B Biol Sci 128:371–394.