Manuscript

Working title: “The link between visual representations and behavior in human scene perception”

Abbreviated title: “Linking brain and behavior in scene perception”

Authors:

Johannes J.D. Singer[1,2*], Agnessa Karapetian[1,3], Martin N. Hebart[2,4¶], Radoslaw M. Cichy[1¶]

Affiliations:

[1]*Department of Education and Psychology, Freie Universität Berlin, Germany*
[2]*Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*
[3]*Charité – Universitätsmedizin Berlin, Einstein Center for Neurosciences Berlin, Berlin, Germany*
[4]*Department of Medicine, Justus-Liebig-Universität Gießen, Germany*
* Corresponding author
Email: johannes.singer@arcor.de
¶ These authors contributed equally

Funding information:

This work was supported by a Max Planck Research Group grant (M.TN.A.NEPF0009) of the Max Planck Society awarded to MNH, a European Research Council grant (ERC-StG-2021-101039712) awarded to MNH, the German Research Council grants (CI241/1-1, CI241/3-1, CI241/7-1) awarded to RMC, and a European Research Council grant (ERC-StG-2018-803370) awarded to RMC.

Conflict of interest statement:

The authors declare no competing financial interests.

Acknowledgements:

We thank Marleen Haupt, Alessandro Gifford and Tony Carricarte for comments on the manuscript. Computing resources were provided by the high-performance computing facilities at ZEDAT, Freie Universität Berlin. Some of the figures used in this paper have been designed using images from Flaticon.com

Number of pages: 31
Number of figures: 5
Abstract - Number of words: 246
Introduction - Number of words: 682
Discussion - Number of words: 1491

1. Abstract

Scene recognition is a core sensory capacity that enables humans to adaptively interact with their environment. Despite substantial progress in the understanding of the neural representations underlying scene recognition, it remains unknown how these representations translate into behavior given different task demands. To address this, we aimed to identify behaviorally relevant scene representations, to characterize them in terms of their underlying visual features, and to reveal how they vary given different tasks. We recorded fMRI data while human participants viewed manmade and natural scenes and linked brain responses to behavior in one of two tasks acquired in a separate set of subjects: a manmade/natural categorization task or an orthogonal task on fixation. First, we found correlations between scene categorization response times (RTs) and scene-specific brain responses, quantified as the distance to a hyperplane derived from a multivariate classifier, in occipital and ventral-temporal, but not parahippocampal cortex. This suggests that representations in early visual and object-selective cortex are relevant for scene categorization. Next, we revealed that mid-level visual features, as quantified using deep convolutional neural networks, best explained the relationship between scene representations and behavior, indicating that these features are read out in scene categorization. Finally, we observed opposite patterns of correlations between brain responses and RTs in the categorization and orthogonal task, suggesting a critical influence of task on the behavioral relevance of scene representations. Together, these results reveal the spatial extent, content, and task-dependence of the visual representations that mediate behavior in complex scenes.

2. Significance statement

Humans rapidly process scene information, allowing them to flexibly categorize and adaptively react to their immediate environment. Here, we sought to determine how the neural representations of scene information translate into adaptive behavior given different task demands. We show that scene representations in early visual and object-selective brain regions are relevant for categorization behavior. Further, we reveal that visual features at an intermediate level of complexity underlie those behaviorally relevant representations. Finally, we demonstrate that depending on the task demands scene representations may facilitate or interfere with behavior. By characterizing the spatial extent, content, and task-dependence of behaviorally relevant scene representations, these findings elucidate the link between neural representations and adaptive perceptual decisions in complex scenes.

Keywords:

scene perception - perceptual decision-making - decoding - fMRI - visual features

2. Introduction

To successfully interact with the environment, humans need to translate sensory information into appropriate actions. The visual system plays a crucial role in this process by extracting visual features from the environment and integrating them into increasingly complex representations through a series of hierarchically organized brain regions in the ventral visual stream (Epstein & Baker, 2019; Grill-Spector & Weiner, 2014; Op de Beeck et al., 2008). While these representations must ultimately serve as the basis for different behavioral goals, their relevance for adaptive behavior in complex scenes is poorly understood. In particular, it remains unknown **i)** where in the brain scene representations relevant for behavior emerge, **ii)** what visual features these representations capture, and **iii)** whether the relevance of these representations for behavior varies given different behavioral goals.

Previous studies have used diverse methods to identify visual representations of simple and complex stimuli that are relevant for behavior (DiCarlo & Maunsell, 2005; Majaj et al., 2015; Philiastides et al., 2006; Philiastides & Sajda, 2006). One such method particularly suited for complex real-world stimuli is the neural distance-to-bound approach (Ritchie & Carlson, 2016), which links visual representations in the brain to behavioral responses via the distance of brain responses from a hyperplane in a high-dimensional response space estimated by a multivariate classifier. Using this approach, behaviorally relevant object representations have been identified in early visual as well as high-level object selective regions (Carlson et al., 2014; Grootswagers et al., 2018; Ritchie & de Beeck, 2019). A recent study has extended these insights to representations of complex scenes, demonstrating that behaviorally relevant scene representations arise in a time window from 100-200 ms after stimulus onset (Karapetian et al., 2023). However, where in the brain such scene representations emerge remains unknown.

Understanding the relevance of scene representations for behavior entails not only identifying when and where behaviorally relevant representations emerge but also characterizing them in terms of their underlying visual features. Prior research has suggested that representations in scene-selective regions capture a variety of visual features, ranging from low to high level of complexity (MacEvoy & Epstein, 2011; Stansbury et al., 2013; Watson et al., 2014). Yet, scene categorization can be accomplished using only low-level (Oliva & Torralba, 2001) or mid-level visual features (Renninger & Malik, 2004). This highlights that not all visual features that are captured by scene representations might be required for scene categorization behavior and raises the question of what visual features underlie behaviorally relevant scene representations.

While the relevance of object and scene representations for behavior has been demonstrated for categorization tasks (Contini et al., 2021; Grootswagers et al., 2018; Karapetian et al., 2023; Ritchie & de Beeck, 2019), this relevance might change given task demands that do not align with the represented information. For instance, when engaged in an orthogonal task, viewing scenes can impair

performance (Greene & Fei-Fei, 2014; Reeder et al., 2015; Seidl-Rathkopf et al., 2015; Wyble et al., 2013), which suggests that scene representations interfere with behavior in certain tasks. However, how the behavioral relevance of scene representations changes for tasks other than categorization remains unknown.

Here, we aimed to identify behaviorally relevant scene representations in the brain, to characterize them in terms of their underlying visual features, and to investigate how they vary given different behavioral tasks. For this, we linked fMRI data from human participants viewing scene images to behavioral responses from a previous study (Karapetian et al., 2023) where participants performed either a manmade/natural categorization task on the same scene images or an orthogonal task on the fixation cross. To identify behaviorally relevant scene representations in the brain, we first localized scene category representations using multivariate decoding (Haynes & Rees, 2006) and then determined which of these representations are relevant for categorization behavior by employing the neural distance-to-bound approach (Ritchie & Carlson, 2016). Next, to elucidate the nature of the behaviorally relevant representations, we determined what type of visual features, quantified as activations from different layers of deep neural networks, best explained behaviorally relevant representations. Finally, to investigate how the behavioral relevance of scene representations varies with the task, we related scene representations to behavior in either a categorization task or an orthogonal task on fixation.

3. Materials and Methods

3.1. Participants

30 healthy adults with normal or corrected-to-normal vision participated in the present study. All participants provided their written informed consent before taking part in the study and were compensated for their time. One participant was excluded from the analyses due to incidental findings consistent with a recognized neurological disorder, resulting in a final sample of 29 participants (mean age = 24.4, SD=3.7, 21 female, 8 male). The study was approved by the ethics committee of Freie Universität Berlin in accordance with the Declaration of Helsinki.

3.2. Experimental stimuli

We used 60 individual scene images from the validation set of the large-scale scene dataset Places365 (Zhou et al., 2018) (see Fig. 1A). Half of the images depicted manmade scenes and the other half natural scenes. The images were further subdivided into 6 categories (beach, canyon, forest, apartment building, bedroom, highway), with 10 exemplars for each category. To standardize the size and aspect ratio of the stimuli, all images were center cropped and resized to 480x480 pixels

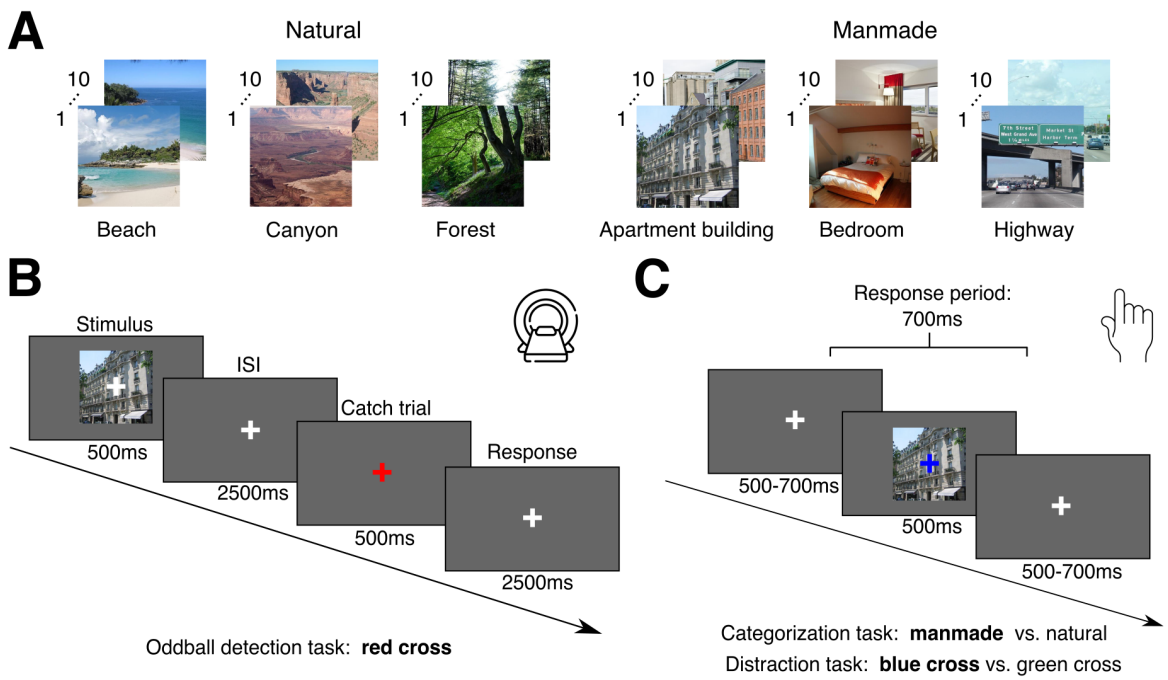

Figure 1. Stimulus set and experimental paradigm. A) Stimulus set used in the experiment. We used 60 scene images from the validation set of the Places365 dataset (Zhou et al., 2018). Half of the stimuli depicted manmade and the other half natural scenes and spanned 6 categories: beach, canyon, forest, apartment building, bedroom, highway. **B) fMRI paradigm.** In a given trial, a scene image was presented for 500ms overlaid with a white fixation cross, followed by an interstimulus interval (ISI) of 2500ms. In 20% percent of the trials the fixation cross turned red instead of the stimulus presentation and participants were instructed to press a button. **C) Behavioral paradigm.** Behavioral data was acquired with a different set of participants in a previous experiment (Karapetian et al., 2023). In a given trial, a scene image was presented for 500ms, overlaid with a blue or green fixation cross (chosen randomly), followed by the presentation of a white fixation cross for a variable time between 500-700ms. In separate blocks, participants were either instructed to report if a given scene image was a manmade or natural scene (categorization task) or if the color of the fixation cross was green or blue (distraction task).

3.3. Experimental design and procedure

3.3.1. fMRI experimental paradigm

During the main fMRI experiment, participants were presented with individual scene images while fixating. Stimuli were presented for 500ms at 12 degrees of visual angle (width & height), overlaid with a central white fixation cross subtending 1 degree of visual angle (Fig. 1B). This was followed by an interstimulus interval of 2,500ms. In 20% of the trials, the fixation cross turned red instead of a stimulus presentation, and the participants were tasked to respond with a button press. Stimulus order was pseudo-randomized within a given run, avoiding immediate repetition of the same stimulus. Each participant completed either 8 or 10 runs with each run lasting 7min 46.5s. In a given run each stimulus was presented twice, resulting in 16 or 20 stimulus repetitions in total for a given participant.

3.3.2. Functional localizer task

To define regions of interest (ROIs), participants completed a functional localizer run at the beginning of the recording session. The localizer consisted of 15s blocks of objects, scrambled objects and scenes (not used in the main experiment) interleaved with 7.5s blocks of only the fixation cross on background as baseline. The images were displayed at a size of 12 degrees of visual angle, at the center of the screen for 400ms, followed by a 350ms presentation of the fixation cross. Participants were instructed to maintain fixation on the fixation cross and to press a button in case the same image was presented in two consecutive trials. In total, the localizer run included 8 blocks of each image type, resulting in a duration of 7min 22.5s. The order of the blocks was pseudo-randomized, avoiding immediate repetition of the same type of block.

3.4. fMRI acquisition, preprocessing and univariate analysis

3.4.1. fMRI acquisition

We collected MRI data using a Siemens Magnetom Prisma Fit 3T system (Siemens Medical Solutions, Erlangen, Germany) with a 64-channel head coil. Structural scans were acquired using a standard T1-weighted sequence (TR=1.9s, TE=2.52ms, number of slices: 176, FOV=256mm, voxel size=1.0mm isotropic, flip angle=9°). Functional images were acquired using a multiband 3 sequence with partial brain coverage (TR=1s, TE=33.3ms, number of slices: 39, voxel size: 2.49x2.49mm, matrix size=82x82, FOV=204mm, flip angle=70°, slice thickness=2.5mm, acquisition order=interleaved, inter-slice gap=0.25mm). The acquisition volume fully covered the occipital and temporal lobes. Due to a technical update of the scanner the voxel size as well as the FOV was slightly changed for the sequence used in the localizer experiment for 20 out of the 30 participants (voxel size: 2.5x2.5mm, FOV=205mm).

3.4.2. fMRI preprocessing

We preprocessed the fMRI data using SPM12 utilities (https://www._l.ion.ucl.ac.uk/spm/) and custom scripts in MATLAB R2021a (www.mathworks.com).

We realigned all functional images to the first image of each run, slice-time corrected them and co-registered them to the anatomical image. Further, based on the functional images and tissue probability maps for the white matter and cerebrospinal fluid, we estimated noise components using the aCompCor method (Behzadi et al., 2007) implemented in the TAPAS PhysIO toolbox (Kasper et al., 2017). Finally, we smoothed the functional images of the localizer run with a Gaussian kernel (FWHM=5). The functional images of the experimental runs were not smoothed.

3.4.3. fMRI univariate analysis

We used a general linear model (GLM) to model the fMRI responses to each scene image in a given run. As the regressors of interest, we entered the onsets and durations of each of the 60 scene images, convolved with a hemodynamic response function (HRF). As nuisance regressors, we entered the noise components and the movement parameters and their first and second order derivatives. In order to account for task- and region-specific variability in the HRF (Polimeni & Lewis, 2021) we employed an HRF-fitting procedure as described in (Prince et al., 2022). For this, we repeated the GLM fitting 20 times, each time convolving all of the regressors of interest with a different HRF obtained from an open-source library of HRFs derived from the Natural Scenes Dataset (Allen et al., 2022). After fitting all the GLMs, we extracted the beta parameter estimates for the scene image regressors from the GLM with the HRF that had resulted in the minimum mean residual for a given voxel. Please note that this approach does not introduce any positive bias to multivariate decoding analyses, since it only focuses on maximizing the overall fit to the data without using any condition-specific information. This procedure resulted in 60 beta maps (one for each scene image) for each run and participant.

For the localizer experiment, we used a separate GLM to model the fMRI responses. Onsets and durations of the blocks of objects, scrambled objects and scenes defined regressors that were convolved with the canonical HRF. We only included movement parameters as nuisance regressors in this GLM. For localizing functionally defined brain areas, we computed three contrasts: scrambled > objects to localize early visual brain areas, objects > scrambled to localize object-selective cortex, and scenes > objects to localize scene-selective cortex. This yielded three *t*-maps for each participant.

3.4.4. Region-of-interest (ROI) definition

As ROIs, we defined early visual cortex (EVC) i.e. V1, V2, and V3, as well as object-selective lateral occipital complex (LOC) and scene-selective parahippocampal cortex (PPA). For the definition of all ROIs we followed a two step procedure. First, we used masks based on a brain atlas with anatomical criteria for EVC (Glasser et al., 2016) and masks based on functional criteria for LOC and PPA (Julian et al., 2012). We transformed these masks into the individual subject space. Next, we computed the overlap between the subject-specific masks and the corresponding *t*-maps from the localizer experiment and only retained the overlapping voxels with *p*-values smaller than 0.0001. For EVC, we used the scrambled > objects *t*-map, for LOC we used the objects > scrambled *t*-map and for PPA we used the scenes > objects *t*-map. Finally, we excluded voxels that overlapped between any of the ROIs. This resulted in one EVC, LOC and PPA ROI mask for each subject.

3.5. Multivariate decoding of scene category information

To determine the amount of scene category information present in the fMRI response patterns we used multivariate decoding. For this, we trained and tested linear Support Vector Machine (SVM) classifiers (Chang & Lin, 2011) to distinguish if a given fMRI response pattern belonged to a manmade or a natural scene. For selecting train and test data for the classifiers, we used two different approaches: an ROI-based method targeting predefined regions and a spatially unbiased searchlight method for further specifying the spatial extent of local effects (Haynes et al., 2007; Kriegeskorte et al., 2006). We conducted all analyses separately for each subject and in the subject's native anatomical space.

We formed pattern vectors based on the beta values from the voxels in a given ROI or searchlight. For this, we assigned all but four beta patterns for each scene image to the train set and the remaining four beta patterns to the test set. Please note that each beta pattern was based on data from a separate run, thereby avoiding potential false positives due to carry-over effects (Mumford et al., 2014). In order to improve the signal-to-noise ratio, for a given scene image we averaged betas from multiple runs into pseudo betas (Stehr et al., 2023). For the train set we averaged two betas into one pseudo beta and for the test set we averaged all four betas into one pseudo beta. Depending on whether participants finished 8 or 10 main experimental runs, this resulted in either 2 or 3 pseudo betas per scene image for the train set and one pseudo beta for the test set.

To increase the robustness of the results, we repeated the splitting of the data into train and test sets and the pseudo beta averaging 1,000 times while randomly shuffling the order of the betas. The resulting decoding accuracies were averaged across repetitions.

For the ROI-based method we iterated this procedure across ROIs and for the searchlight-based method across searchlights. This resulted in one decoding accuracy for every ROI and one searchlight decoding map for every subject. For later group-level statistical analyses, we normalized the searchlight decoding maps to the MNI template brain.

3.6. Behavioral data

In order to identify behaviorally relevant scene representations, we linked the neural data recorded in the present study to behavioral data from 30 participants recorded in a previous study (Karapetian et al., 2023). In short, participants were presented with the same scene images as used in the current study and performed either a manmade/natural categorization task on the stimuli or an orthogonal color discrimination task on the fixation cross while EEG was recorded.

The experiment consisted of 20 blocks, 10 per task, and included at least 30 trials per scene image per block. In each trial, a stimulus was presented for 500ms overlaid with a green or blue (randomly assigned) fixation cross, followed by a presentation of a white fixation cross for a variable time window between 500 to 700ms. Participants were instructed at the beginning of each block to either report if the presented stimulus was a manmade or a natural scene or to report the color of the fixation cross, as accurately and as quickly as possible.

We first averaged the response time (RT) data from the correctly answered trials separately for the categorization and the fixation cross task for each subject and then obtained the mean RT for each scene image and each task across participants. On average, for a given subject 23.2 (*SD* = 6.0) trials were included for each scene for the categorization task and 26.0 (*SD* = 1.46) for the orthogonal task. This resulted in one mean RT for each scene image and each task.

3.7. Distance-to-bound analysis

We used the neural distance-to-bound approach (Carlson et al., 2014; Ritchie et al., 2015; Ritchie & Carlson, 2016) to determine if scene information represented in fMRI response patterns is behaviorally relevant (see Fig. 2A). The neural distance-to-bound approach links the information in brain patterns to behavior by predicting a relationship between RTs and distances of individual brain responses to a criterion in the high-dimensional neural response space. The concept of a criterion is based on signal detection theory (Green & Swets, 1966) and can be formulated in high-dimensional spaces as a hyperplane that is estimated when using multivariate decoding. The approach assumes a negative relationship between distances of individual brain response patterns to the hyperplane and RTs: points close to the hyperplane have weak sensory evidence and are difficult to categorize, leading to longer RTs. Vice versa, points far from the hyperplane have strong sensory evidence and can be easily categorized, resulting in short RTs. If this predicted relationship holds true for observed brain response patterns and behavioral responses, then it is assumed that information represented in these brain patterns is relevant for behavior.

To test the predicted relationship between neural distances to the hyperplane and RTs, we obtained distances for every scene image using the hyperplanes estimated with the same decoding procedure as described above. We iterated this procedure over ROIs and searchlights, resulting in a vector with 60 values (one for each scene image) for each ROI, and searchlight. Finally, we correlated the vectors of distances with the vector of mean RTs for each ROI and searchlight using Pearson's correlation. This yielded distance-RT correlations for each ROI, searchlight and subject.

3.8. Model-based distance-to-bound analysis

To examine what type of visual features best explains behaviorally relevant scene representations in the brain, we used the neural distance-to-bound approach in combination with deep neural network (DNN) modeling and commonality analysis (Mood, 1971; Reichwein Zientek & Thompson, 2006). The basic rationale (see Fig. 4A-C) involved first extracting activations from different DNN architectures and layers as an approximation of visual feature representations at different levels of complexity (Bankson et al., 2018; Groen et al., 2018; Reddy et al., 2021; Xie et al., 2020). The assumption that these activations approximate a gradient of feature complexity is based on demonstrations of a hierarchical correspondence between representations in DNNs and the human brain (Cichy et al., 2016; Güçlü & Gerven, 2015). Moreover, as processing advances in the network, representations undergo increasingly more non-linear transformations, further supporting the claim of increasing feature complexity in DNNs. Next, in order to link neural network activations, brain response patterns and behavioral RTs, we derived distances to the hyperplane based on the

neural network activations. Finally, to determine which activations best explained behaviorally relevant scene representations, we estimated the shared variance between model distances, neural distances and RTs using commonality analysis.

In detail, as models we used the ResNet-50, ResNet-18 (He et al., 2015), AlexNet (Krizhevsky et al., 2012) and DenseNet161 (Huang et al., 2018) architectures, pre-trained on the Places365 dataset (Zhou et al., 2018) (https://github.com/CSAILVision/places365). We chose to examine different DNN architectures to ensure that a given pattern of results is not idiosyncratic to a given architecture but can be generalized to a given hierarchical level regardless of the specific architecture. We extracted activations for 1,200 images from the validation set of Places365 (Zhou et al., 2018) as well as for our experimental stimuli. The Places365 images were sampled from 80 categories (half manmade, half natural), including the six categories from our stimulus set, and contained 15 images per category. For the extraction we focussed on a selection of layers including all pooling layers and the last fully connected layer for AlexNet, the output of all residual blocks and the last fully connected layer for the ResNets, as well as the first pooling layer, the output of all the DenseBlocks and the last fully connected layer for DenseNet161. For computational efficiency, we reduced the network activations for every layer except for the fully connected layers to a dimensionality of 1,000 by using PCA on the activations for the 1,200 images from the validation set of Places365 and applying the parameters to these activations as well as the activations for our experimental stimuli.

Next, we trained SVM classifiers on a manmade/natural classification task using the reduced activations for the 1,200 Places365 validation images for every layer and network separately. Subsequently, we tested the trained SVM classifiers on the reduced activations for our 60 experimental stimuli and derived a distance to the hyperplane for each scene image. This resulted in 60 distances for each layer and network.

Finally, using commonality analysis we determined the common variance between the network distances, neural distances and behavioral RTs. In commonality analysis, the common variance that can be explained in a given outcome variable by two predictor variables is defined as the amount of variance explained by both predictors in the outcome variable minus the unique contribution of each of the predictors. In simplified form this term can be written as: $C(AB)=R^2_{y.A}+R^2_{y.B}-R^2_{y.AB}$, where R^2 is the explained variance in a multiple regression model with the mean RTs as outcome variable (y) and either neural distances (A), network distances (B) or both (AB) as predictor variables. We fitted the corresponding multiple regression models and computed the commonality based on the R^2 values, resulting in shared variance estimates for each network, layer, ROI and subject.

3.9. Statistical analyses

For statistical testing we used non-parametric sign permutation tests at the group-level (Nichols & Holmes, 2002). In essence, we obtained null distributions for a statistic (decoding accuracies, distance-RT correlations) by randomly permuting the sign of the results at the participant level 10,000 times. Next, we obtained *p*-values for the observed data by comparing their statistic to that of the null distribution. We used one-sided tests for decoding accuracies, as well as two-sided tests for distance-RT correlations and differences between decoding accuracies.

To correct for multiple comparisons, we used two different approaches. In the case of only a limited number of tests (i.e., < 10) such as multiple ROIs or neural network layers, we used the Benjamini-Hochberg FDR-correction without dependency (Benjamini & Hochberg, 1995). When applying a large number of tests such as for testing across searchlights (i.e., ~ 100,000), we used a cluster-based correction (Maris & Oostenveld, 2007). For this, we first thresholded the *p*-values from the non-parametric sign permutation tests at $p<0.001$. Then we clustered the thresholded *p*-values by spatial adjacency, and computed the maximum cluster size for each permutation. Next, we determined the *p*-value for each cluster in the observed data by comparing the cluster size of a given cluster to the maximum cluster size distribution. Finally, we thresholded the cluster *p*-values at $p<0.05$.

To compute 95% confidence intervals for the hierarchical level, i.e. the layer index where there was the peak R^2 value obtained by the commonality analysis, we used bootstrapping. First, we took 100,000 random samples with replacement from the participant-specific R^2 values. We computed the mean over participants for each bootstrap sample and detected the index of the layer with the peak R^2 value across network layers. Finally, we used the 2.5% and 97.5% percentiles of the bootstrap distribution as the lower and upper bound of confidence intervals.

3.10. Data and code availability

The raw fMRI data is available in BIDS format on OpenNeuro (https://openneuro.org/datasets/ds004693). The beta maps obtained from the GLM, the behavioral data, the distances derived from the DNNs, as well as all first-level and group-level results are available via OSF (https://osf.io/y8tx2/). All code used for the first-level and group-level analyses in this study is provided via Github (https://github.com/Singerjohannes/visdecmak).

4. Results

4.1. Scene representations in occipital and ventral-temporal but not parahippocampal cortex are negatively correlated to categorization RTs

To identify scene category representations in the brain we used multivariate decoding. For this, we trained SVM classifiers on the fMRI data to predict if a given brain activity pattern belonged to a manmade or a natural scene and tested the classifier on left-out data. We first performed this analysis for EVC, LOC and PPA. We found decoding accuracies significantly above chance in all ROIs (Fig. 2B, $p<0.001$, sign-permutation test, FDR-corrected), suggesting the presence of scene category representations in these regions as expected from their central role in processing complex visual stimuli (Epstein & Baker, 2019; Grill-Spector & Weiner, 2014; Op de Beeck et al., 2008). To further uncover scene category representations beyond our predefined ROIs, we performed spatially-unbiased searchlight decoding (Haynes & Rees, 2006; Kriegeskorte et al., 2006). We found that scene category was decoded with accuracies significantly above chance ($p<0.05$, cluster-based permutation test) throughout the ventral and dorsal visual stream, with peaks in

posterior and lateral parts of the occipital cortex and decreases towards anterior parts of cortex (Fig. 3A). Together, these results suggest a widespread presence of scene category representations as candidates for categorization behavior-relevant representations along both the ventral and dorsal stream (Walther et al., 2009, 2011).

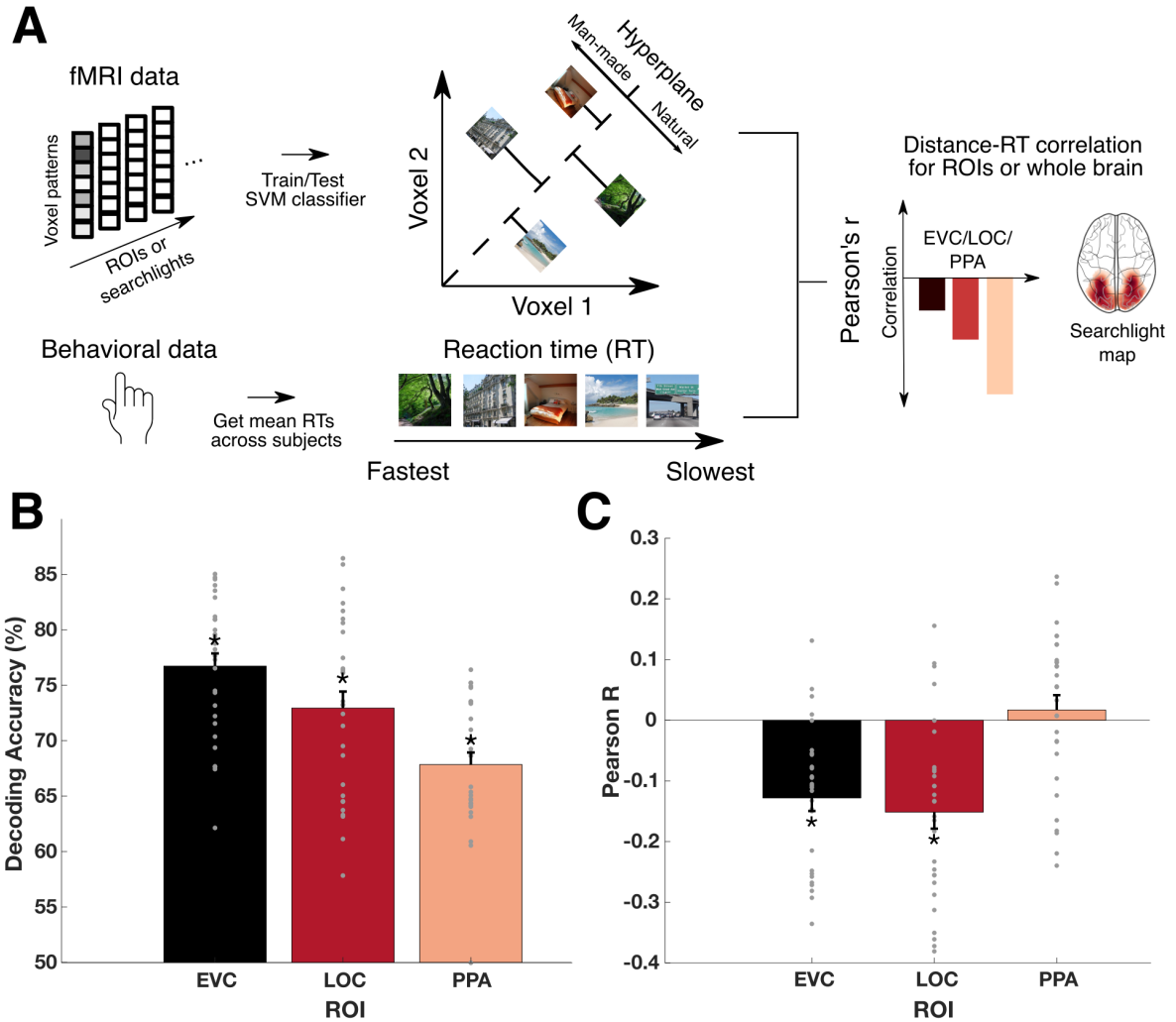

Figure 2. Scene category representations and behaviorally relevant scene representations in EVC, LOC and PPA. A) Neural distance-to-bound approach for identifying behaviorally relevant scene representations. For each subject, we derived neural distances from the fMRI response patterns by training SVM classifiers on part of the fMRI data and obtaining scene-specific distances from the hyperplane of the classifier for the left-out fMRI data. Next, we obtained mean RTs (in a manmade/natural categorization task or an orthogonal task on the fixation cross) across participants for each scene image and linked these RTs to the neural distances using Pearson's correlation. We iterated this procedure over ROIs or searchlights, resulting in ROI-specific correlation values or searchlight correlation maps. Negative correlations between neural distances and RTs at a specific location in the brain indicate that the representations at this location are relevant for behavior. **B) Decoding of scene category in EVC, LOC and PPA.** Scene category could be decoded with accuracies significantly above chance in EVC, LOC and PPA. **C) Correlations between behavioral RTs and neural distances in EVC, LOC and PPA.** There were negative correlations between behavioral RTs and neural distances in EVC, LOC but not PPA. Error bars depict the standard error of the mean across participants. Stars above or below the bars indicate significant results.

Having identified scene category representations in the brain, we sought to determine to what extent these representations are relevant for categorization behavior. For this, we applied the neural distance-to-bound approach (Ritchie & Carlson, 2016; for visualization see Fig. 2A). In short, we first obtained mean RTs for the manmade/natural categorization task across participants for each scene image recorded in a previous experiment (Karapetian et al., 2023) and derived neural distances for each scene image from the SVM classifiers trained on the fMRI response patterns recorded in this study. Next, we correlated the neural distances and behavioral RTs across the 60 scene images and repeated this procedure across ROIs and searchlights. Given a negative relationship between neural distances and RTs at a given locus in the brain, it is assumed that representations in this area are relevant for behavior.

We found negative distance-RT correlations in EVC, LOC (both $p<0.001$, sign-permutation test, FDR-corrected, Fig. 2B) but not in PPA ($p=0.487$, sign-permutation test, FDR-corrected), suggesting that scene representations in EVC and LOC are relevant for categorization behavior, without positive evidence for a role of PPA.

Searchlight analysis further revealed significant negative distance-RT correlations ($p<0.05$, cluster-based permutation test, Fig. 3B) at the border between occipital and ventral temporal cortex and between occipital and posterior parietal cortex, but not in parahippocampal cortex. This corroborates the ROI-based analyses and further specifies the locus of behaviorally relevant scene representations in a spatially unbiased fashion.

Surprisingly, we also found significant positive distance-RT correlations ($p<0.05$, cluster-based permutation test), which were confined to the right occipital cortex only. A positive correlation between neural distances and RTs violates the predictions of the neural distance-to-bound approach and suggests that a scene representation with a strong category signal leads to a slow RT in the task and vice versa. This implies interference between scene representations in the occipital cortex and behavior in the categorization task.

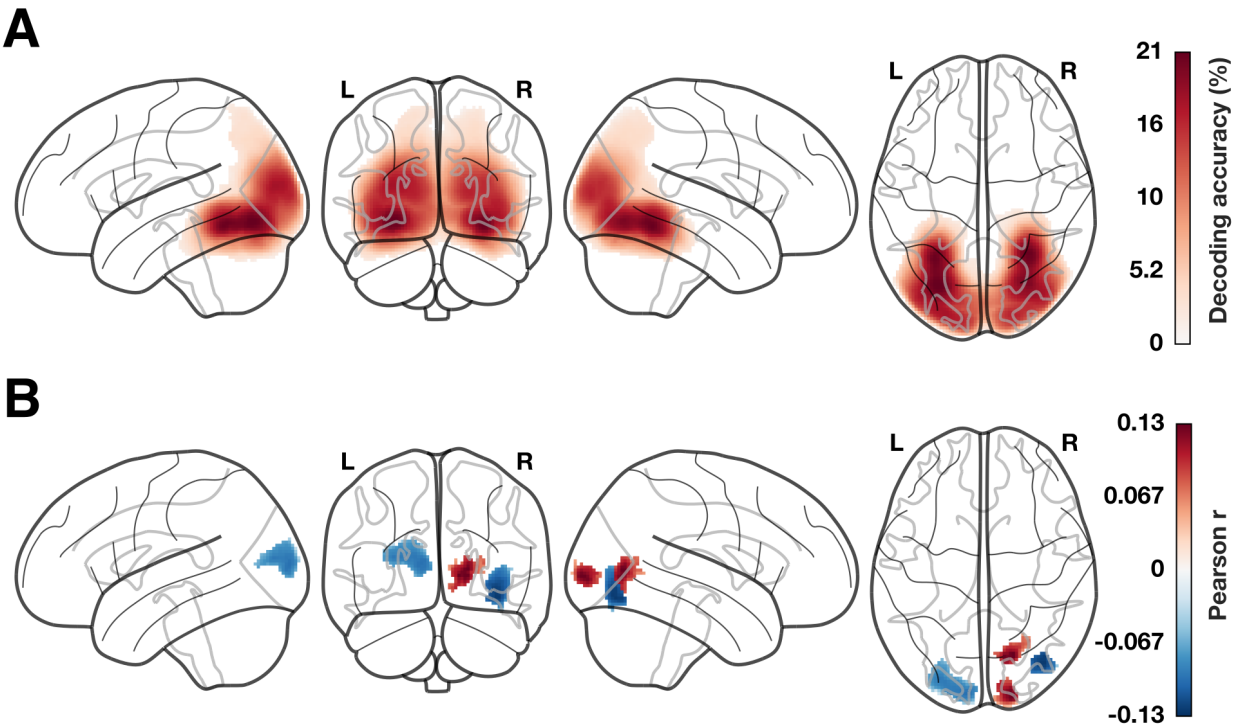

Figure 3. Scene category representations and behaviorally relevant scene representations across the brain. A) Decoding of scene category in the visual cortex. Spatially-unbiased searchlight decoding revealed significant decoding accuracies that were most pronounced in posterior and lateral parts of occipital cortex, with decreasing accuracies towards anterior parts of ventral-temporal cortex and posterior-parietal cortex. **B) Correlations between behavioral RTs and neural distances in the visual cortex.** Iterating the correlation across searchlights showed negative distance-RT correlations that were strongest at the border between occipital and ventral-temporal cortex as well as at the border between occipital and posterior parietal cortex. There were additional significant positive correlations which were strongest in the right occipital cortex.

4.2. Features derived from intermediate neural network layers best explain behaviorally relevant scene representations in the visual cortex

While our findings so far suggest that a subset of scene representations in the visual cortex are relevant for categorization behavior, they leave open what type of visual features underlies these behaviorally relevant scene representations. We investigated this question in terms of feature complexity. As a proxy for low- to high complexity visual features, we used deep neural network activations extracted from different layers (for similar approaches see: (Bankson et al., 2018; Greene & Hansen, 2020; Groen et al., 2018; Reddy et al., 2021; Xie et al., 2020)) and asked to what extent these activations can explain the link between scene representations and behavioral responses (for a visualization of the procedure see Fig. 4A-C). We linked network activations to RTs and fMRI data using the neural distance-to-bound approach (Ritchie & Carlson, 2016) and determined which layer's activations best explain the shared variance between RTs and fMRI data using commonality analysis

(Mood, 1971; Reichwein Zientek & Thompson, 2006). We focussed on EVC and LOC since we found significant distance-RT correlations only there.

In EVC we found significant R^2 values for most of the networks and layers (all $p<0.026$, sign-permutation test, FDR-corrected, Fig. 4D) except for the first layer in ResNet50 ($p=0.051$, sign-permutation test, FDR-corrected), ResNet18 ($p=0.296$, sign-permutation test, FDR-corrected) and DenseNet161 ($p=0.318$, sign-permutation test, FDR-corrected). In addition, for ResNet18 there were significant negative R2 values in the last layer indicating that the distances in this layer did not share variance with the neural or behavioral distances but rather suppressed some of the shared variance between brain and behavior. In LOC we found significant R^2 values for most networks and layers (all $p<0.001$, sign-permutation test, FDR-corrected, Fig. 4E) except for the first layer in DenseNet161 ($p=0.773$, sign-permutation test, FDR-corrected). Similarly to EVC, R2 values for the last layer in ResNet18 were negative, indicating that the distances in this layer did not contribute to explaining the shared variance between brain and behavior. Together, this suggests that in both EVC and LOC visual features from most hierarchical levels, excluding very early and late stages, explain a part of the variance that is shared between brain and behavior.

Next, we determined which visual features explain the shared variance most strongly between brain and behavior by determining the layers with peak shared variance. We found that the shared variance in EVC peaked in intermediate layers for most networks except for AlexNet (peak layer and bootstrap 95% CIs: ResNet18 = 4; [3, 4], ResNet50 = 4; [3, 4], AlexNet = 2; [2, 4], DenseNet161 = 4; [3, 5], Fig. 4D). In LOC, the shared variance peaked in intermediate layers for all networks (peak layer and bootstrap 95% CIs: ResNet18 = 4; [4, 4], ResNet50 = 4; [3, 4], AlexNet = 4; [4, 4], DenseNet161 = 4; [4, 4], Fig. 4E). These results demonstrate that mid-level visual features best explain behaviorally relevant scene representations. Furthermore, the layers best explaining the link between brain and behavior in EVC and LOC showed substantial overlap, potentially indicating a common format of behaviorally relevant scene representations across these regions.

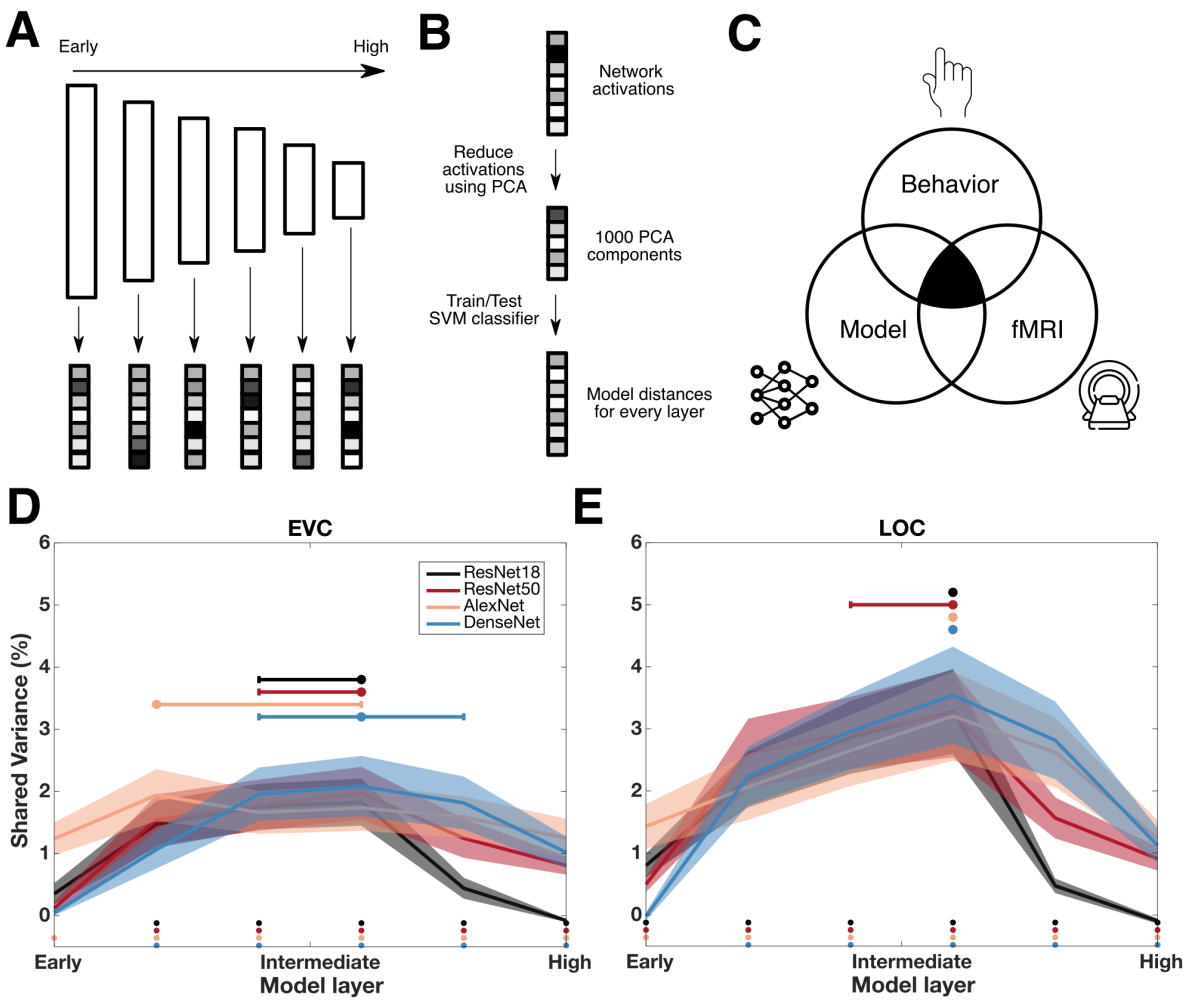

Figure 4. Visual features underlying behaviorally relevant scene representations. A) Extraction of activations from various deep neural network layers. As an approximation of features of different complexity from low- to high-level, we extracted activations for 1200 scene images (half manmade, half natural) from the validation set of Places365 as well as for our experimental stimuli from various deep neural network architectures and layers. **B) Deriving scene-specific distances from neural network activations.** For linking the neural network activations to distances based on fMRI data and behavioral RTs we first reduced the activations using PCA for every layer separately resulting in 1,000 PCA components per scene image. Next, for every layer and network separately, we trained SVM classifiers on a manmade/natural classification task using the Places365 activations and then tested the classifiers on the activations for our experimental stimuli. This yielded distances from the hyperplane for each of our experimental stimuli and every layer and network. **C) Commonality analysis approach.** We asked to what extent model distances for a given network and layer explain the shared variance between distances based on fMRI data and behavioral RTs. For this, we assessed the shared variance between neural distances, model distances and behavioral RTs using commonality analysis. **D) Shared variance for each network and layer in EVC.** We found significant positive R^2 values in all of the layers and networks except for the first layer in ResNet50, ResNet18 and DenseNet161 and the last layer in ResNet18. R^2 values peaked in intermediate layers for all networks except for AlexNet. **E) Shared variance for each network and layer in LOC.** R^2 values were positive and significant in all networks and layers except for the first layer in DenseNet161 and the last layer in ResNet18. For all networks R^2 values peaked in intermediate layers. Colored dots below the lines indicate significant layers. Shaded areas represent the SEM across participants for each layer. Horizontal error bars depict the 95% confidence intervals of the

peak layer index. No horizontal error bar for a given layer indicates that the 95% confidence interval included only the value of the peak layer index.

4.3. Scene representations show opposite patterns of correlation with behavior in a categorization and an orthogonal task

While we identified and characterized scene representations relevant for one particular type of task, i.e. categorization behavior, their impact on behavior might differ for other tasks. To investigate this, we determined the behavioral relevance of scene representations for a distraction task orthogonal to categorization. For this, we correlated scene-specific distances to RTs from a distraction task on which participants judged the color of the fixation cross, analogously to how we identified behaviorally relevant scene representations for the categorization task.

In contrast to the negative correlations for the categorization task, we found positive distance-RT correlations in EVC (p=0.004, sign-permutation test, FDR-corrected) and LOC (p<0.001, sign-permutation test, FDR-corrected), but not in PPA (p=0.073, sign-permutation-test, FDR-corrected, Fig. 5A). Searchlight analysis further revealed positive distance-RT correlations that were most pronounced at the border between occipital and ventral-temporal cortex (p<0.05, cluster-based permutation test, Fig. 5B). These positive correlations are contrary to the predictions of the neural distance-to-bound approach (Ritchie & Carlson, 2016) and demonstrate that scene representations with a strong category signal are associated with slow responses in the distraction task and vice versa for scene representations with a weak category signal and speeded responses. This suggests that scene representations at the border between occipital and ventral-temporal cortex interfere with behavior in the distraction task.

Given these opposing patterns of correlations for the distraction and the categorization task, we asked if the same representations that showed a positive correlation with behavior in the distraction task exhibited a negative correlation with behavior in the categorization task, or if these representations were distinct from each other. For this, we computed the overlap between the voxels with significant negative correlations with RTs in the categorization task (see Fig. 3B in red) and the voxels with significant positive correlations with RTs in the distraction task (see Fig. 5B in red). Interestingly, there was a partial overlap between the significant voxels for the categorization task and the significant voxels for the distraction task (see Fig. 5B in blue), indicating that a subset of scene representations that are relevant for categorization behavior showed an inverse relationship with behavior in the distraction task.

A possible explanation for this observed inverse relationship might be that scene representations that are relevant for categorization behavior evoke a strong category signal in the brain which takes away processing resources from the distraction task, thereby slowing the RT. Given this interference between scene representations and performance in the categorization task we expected to observe a similar relationship between the RTs in the categorization task and the distraction

task, namely that scenes that are solved faster in the categorization task lead to slower RTs in the distraction task and vice versa. To test this, we correlated the RTs from the categorization task with the RTs from the distraction task. We found a negative correlation between the RTs of the two tasks (r = -0.367, p = 0.004; Fig. 5C), indicating that scene images that are solved fast in the categorization task are associated with long RTs when presented during the distraction task and vice versa. This suggests that scene processing interferes with performance in the distraction task, corroborating the interference effect between scene representations and behavior in the distraction task.

In sum, these results provide evidence that a subset of scene representations in the visual cortex are relevant for behavior even in tasks beyond categorization. Yet, the relevance of these representations for behavior differed for the distraction task and the categorization task. While scene representations facilitate categorization behavior, they interfere with behavior in the distraction task. This demonstrates that the task context critically affects the behavioral relevance of scene representations.

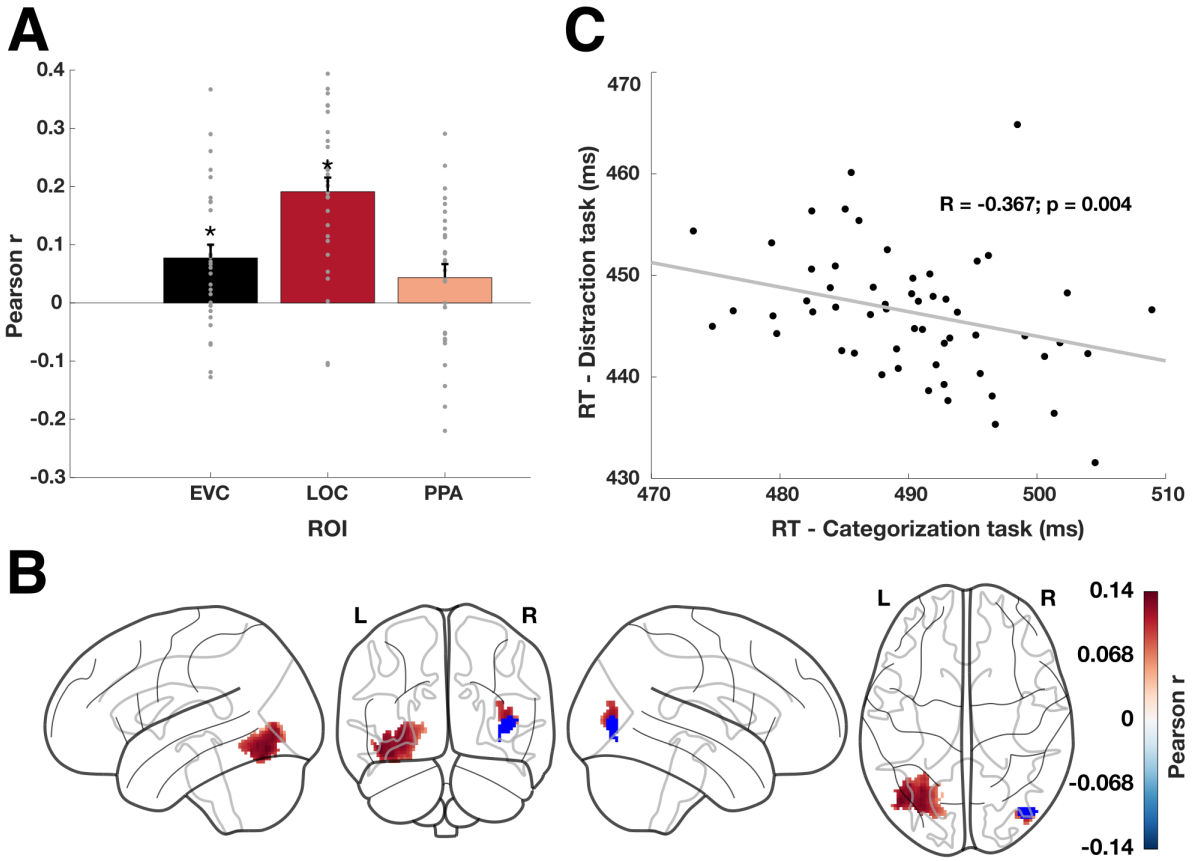

Figure 5. **Effects of task on the behavioral relevance of scene representations. A) Correlations between RTs in the distraction task and neural distances in EVC, LOC and PPA.** We found significant positive correlations in EVC, LOC but not PPA. Error bars depict the SEM across participants. Stars above the bars indicate significant results. **B) Correlations between RTs in the distraction task and neural distances in the visual cortex.** We found positive correlations that were strongest at the border between occipital and ventral-temporal cortex. There was a partial overlap between the significant negative clusters of distance-RT correlations for the categorization task and the positive clusters in the distraction task, colored in blue here. **C) Correlations between**

mean RTs for the categorization and the distraction task. We found a negative correlation between RTs in the distraction and the categorization task.

5. Discussion

In the present study we identified and characterized behaviorally relevant scene representations as well as their dependence on the task by relating fMRI responses to behavioral RTs using the neural distance-to-bound approach (Ritchie & Carlson, 2016). We revealed three key findings. First, while we could decode scene category along both the ventral and dorsal streams, neural distances were negatively correlated to categorization RTs only at the border between occipital and ventral-temporal cortex, but not in parahippocampal cortex. This suggests that despite a widespread presence of scene category representations only a subset of these representations was relevant for categorization behavior. Second, distances derived from intermediate layers of deep neural networks best explained the shared variance between brain and behavior, suggesting that mid-level visual features best account for behaviorally relevant scene representations. Finally, we observed opposing patterns of correlation between neural distances and RTs for the distraction task and the categorization task. While for categorization RTs there was a negative correlation, suggesting facilitation of behavior, for distraction RTs we found a positive correlation, suggesting interference with behavior. This demonstrates that the task context critically affects the behavioral relevance of scene representations.

5.1. Behaviorally relevant scene representations emerge at the border between occipital and ventral-temporal but not in parahippocampal cortex

By employing the neural distance-to-bound approach (Ritchie & Carlson, 2016), we identified scene representations relevant for categorization behavior at the border between occipital and ventral-temporal but not in parahippocampal cortex. These findings align with object recognition studies (Carlson et al., 2014; Grootswagers et al., 2018; Ritchie et al., 2015; Ritchie & de Beeck, 2019) showing behaviorally relevant representations in both early and high-level visual cortex. However, our findings challenge the view that information for categorizing natural images is only read out from high-level visual cortex (Majaj et al., 2015) and suggest that representations from both early and high-level visual cortex might be read out in perceptual decision-making (Birman & Gardner, 2019; Jagadeesh & Gardner, 2021).

Our findings complement a recent characterization of behaviorally relevant scene representations over time (Karapetian et al., 2023) by spatially localizing these representations in the brain. The presence of behaviorally relevant scene representations in LOC, but not PPA, is in line with studies emphasizing the role of LOC in scene recognition (Linsley & MacEvoy, 2014; MacEvoy & Epstein, 2011; Stansbury et al., 2013). However, it conflicts with the pivotal role of PPA in scene recognition (Aguirre et al., 1998; Epstein & Kanwisher, 1998) and with findings of behaviorally relevant representations in PPA (Groen et al., 2018; King et al., 2019).

One possible explanation for this discrepancy is that in our study, participants might have relied on detecting objects rather than the spatial layout of a scene for categorization. Since the representation of objects is more strongly associated with LOC than with PPA (Park et al., 2011), and spatial layout is more strongly associated with PPA than LOC, relying on objects during scene categorization might render representations in LOC relevant and in PPA irrelevant for behavior. Further studies examining categorization behavior while controlling different types of visual information in scenes are needed to determine what type of visual information drives categorization behavior in scenes.

Surprisingly, we found a positive correlation between neural distances in the right occipital cortex and RTs in the categorization task. These findings are not captured by the rationale of the neural distance-to-bound approach (Ritchie & Carlson, 2016), which assumes a negative relationship between neural distances and RTs, where high distances are associated with fast RTs and vice versa. Instead, we observed the opposite: high distances were associated with slow RTs and vice versa, suggesting interference between scene representations and behavior in the categorization task. This interference is hard to reconcile with the role of the occipital cortex in visual processing. However, these positive correlations might be spurious and influenced by a bias in the classifier's hyperplane towards a specific category (e.g. manmade, natural). Such biases in the distance-RT correlations towards one category of a given category division (e.g. animate over inanimate) have been reported previously (Carlson et al., 2014; Grootswagers et al., 2017, 2018; Karapetian et al., 2023; Ritchie et al., 2015). Fully understanding this phenomenon requires simulations of different data regimes in combination with an in-depth geometrical analysis of the estimated hyperplane and its relationship to individual data points, which is an exciting avenue for future studies.

5.2. Mid-level visual features underlie behaviorally relevant scene representations in the visual cortex

We found that mid-level visual features best explained the shared variance between neural distances and RTs in two brain regions: EVC and LOC. These results parallel object recognition studies that highlight the importance of mid-level visual features for categorizing objects (Eberhardt et al., 2016) and for the organization of object representations in occipito-temporal cortex (Long et al., 2018). Along with the localization of behaviorally relevant representations in EVC and LOC, but not PPA, this suggests that object information significantly influenced scene categorization. However, our findings also conflict with previous studies, which showed that high-level conceptual features best explain variance in behavioral similarity judgments for scenes and objects (Greene & Hansen, 2020; King et al., 2019). One potential reason for this divergence is that similarity judgments might be based on different visual features than categorization. While categorization might depend on object information related to mid-level visual features (Eberhardt et al., 2016; Long et al., 2018), judging the similarity of scenes might involve high-level features related to

the semantics of a scene. Contrasting different characterizations of behavior in response to scenes and their relationship to brain data will be a fruitful path towards better understanding the relevance of distinct types of visual features for various behavioral goals.

Interestingly, we observed considerable similarity in the pattern of explained variance across types of visual features in EVC and LOC. This suggests that behaviorally relevant representations in these regions share a common format. This similarity might be explained by feedback from higher level visual areas that shapes visual representations in early visual areas rendering them behaviorally relevant. Recent findings support this notion, indicating that task-relevant information in lower level visual regions emerges only late in time after feedback from higher level visual regions has affected lower level visual areas (Sexton & Love, 2022). Due to the temporal sluggishness of the BOLD response, we cannot directly examine the temporal dynamics of behaviorally relevant representations. Future studies could use techniques such as backward masking (Fahrenfort et al., 2007) to experimentally manipulate feedback processing and investigate its impact on behaviorally relevant representations in the visual cortex directly.

5.3. Task context critically affects behavioral relevance of scene representations

We found opposing patterns of correlation between neural distances and RTs in the categorization task and the distraction task. This suggests distinct relationships between scene representations and behavior depending on the task. In the categorization task, strong category signals were associated with fast RTs and vice versa, suggesting a facilitative relationship between scene representations and behavior. In contrast, for the distraction task, strong category signals were associated with slow RTs and vice versa, indicating interference between scene representations and behavior. This interference could be due to automatic processing of the content of a scene (Greene & Fei-Fei, 2014) which might have interfered with the representation of the fixation cross color. Alternatively, attention might have been differentially captured by the scenes and diverted away from the fixation cross, thereby impairing performance in the distraction task (Reeder et al., 2015; Seidl-Rathkopf et al., 2015; Wyble et al., 2013). While our findings cannot dissociate between these alternatives, they highlight the importance of scene recognition as a core cognitive process which cannot be easily suppressed.

5.4. Limitations

Several experimental factors potentially limit the generalizability of our findings. First, we focussed solely on manmade/natural categorization behavior, and other types of categorization might involve different visual representations and visual features (Contini et al., 2021; Grootswagers et al., 2018). Second, our choice of task in the fMRI experiment might have limited the emergence of behaviorally relevant representations. Participants performed a task on the fixation cross in the fMRI

experiment which differed from the categorization or the distraction task in the behavioral experiment (Karapetian et al., 2023). Even though previous studies have shown that task engagement is not necessary for emergence of behaviorally relevant visual representations in the occipito-temporal cortex (Carlson et al., 2014; Grootswagers et al., 2018), particularly representations in parietal or frontal brain regions are affected by the task (Bracci et al., 2017; Hebart et al., 2018; Vaziri-Pashkam & Xu, 2017). Thus, engaging participants in the same task in the fMRI and behavioral experiment could have expanded the detectable behaviorally relevant representations.

5.5. Conclusion

Together, our findings reveal the spatial extent of the visual representations underlying categorization behavior for real-world scenes, identify mid-level visual features as the main contributor to these behaviorally relevant representations, and suggest that the task context critically affects behavioral relevance of scene representations. These results contribute to the understanding of the neural mechanisms and visual features enabling adaptive perceptual decisions in complex real-world environments.

6. References

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An Area within Human Ventral Cortex Sensitive to "Building" Stimuli: Evidence and Implications. *Neuron*, *21*(2), 373–383. https://doi.org/10.1016/S0896-6273(00)80546-2

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

Bankson, B. B., Hebart, M. N., Groen, I. I., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, *178*, 172–182.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Birman, D., & Gardner, J. L. (2019). A flexible readout mechanism of human sensory representations. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-11448-7

Bracci, S., Daniels, N., & Op de Beeck, H. (2017). Task Context Overrules Object- and Category-Related Representational Content in the Human Parietal Cortex. *Cerebral Cortex*, *27*(1), 310–321. https://doi.org/10.1093/cercor/bhw419

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, *26*(1), 132–142. https://doi.org/10.1162/jocn_a_00476

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27:1-27:27.

https://doi.org/10.1145/1961189.1961199

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

Contini, E. W., Goddard, E., & Wardle, S. G. (2021). Reaction times predict dynamic brain representations measured with MEG for only some object categorisation tasks. *Neuropsychologia*, *151*, 107687. https://doi.org/10.1016/j.neuropsychologia.2020.107687

DiCarlo, J. J., & Maunsell, J. H. R. (2005). Using Neuronal Latency to Determine Sensory–Motor Processing Pathways in Reaction Time Tasks. *Journal of Neurophysiology*, *93*(5), 2974–2986. https://doi.org/10.1152/jn.00508.2004

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How Deep is the Feature Analysis underlying Rapid Visual Categorization? *Advances in Neural Information Processing Systems*, *29*. https://papers.nips.cc/paper/2016/hash/42e77b63637ab381e8be5f8318cc28a2-Abstract.html

Epstein, R., & Baker, C. I. (2019). Scene Perception in the Human Brain. *Annual Review of Vision Science*, *5*, 373–397. https://doi.org/10.1146/annurev-vision-091718-014809

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), Article 6676. https://doi.org/10.1038/33402

Fahrenfort, J. J., Scholte, S. H., & Lamme, V. A. F. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, *19*(9). https://doi.org/10.1162/jocn.2007.19.9.1488

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. *Signal*

detection theory and psychophysics., xi, 455–xi, 455.

Greene, M. R., & Fei-Fei, L. (2014). Visual categorization is automatic and obligatory: Evidence from Stroop-like paradigm. *Journal of Vision*, *14*(1), 14. https://doi.org/10.1167/14.1.14

Greene, M. R., & Hansen, B. C. (2020). Disentangling the Independent Contributions of Visual and Conceptual Features to the Spatiotemporal Dynamics of Scene Categorization. *Journal of Neuroscience*, *40*(27), 5283. https://doi.org/10.1523/JNEUROSCI.2088-19.2020

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), Article 8. https://doi.org/10.1038/nrn3747

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife*, *7*, e32962. https://doi.org/10.7554/eLife.32962

Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, *179*, 252–262. https://doi.org/10.1016/j.neuroimage.2018.06.022

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., & Carlson, T. A. (2017). Asymmetric Compression of Representational Space for Object Animacy Categorization under Degraded Viewing Conditions. *Journal of Cognitive Neuroscience*, *29*(12), 1995–2010. https://doi.org/10.1162/jocn_a_01177

Güçlü, U., & Gerven, M. A. J. van. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading Hidden Intentions in the Human Brain. *Current Biology*, *17*(4), 323–328. https://doi.org/10.1016/j.cub.2006.11.072

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *eLife*, *7*, e32816. https://doi.org/10.7554/eLife.32816

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). *Densely Connected Convolutional Networks* (arXiv:1608.06993). arXiv. https://doi.org/10.48550/arXiv.1608.06993

Jagadeesh, A. V., & Gardner, J. L. (2021, Dezember 15). *V1- and IT-like representations are directly accessible to human visual perception*. SVRHM 2021 Workshop @ NeurIPS. https://openreview.net/forum?id=ec7BWId59zF

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage*, *60*(4), 2357–2364.

Karapetian, A., Boyanova, A., Pandaram, M., Obermayer, K., Kietzmann, T. C., & Cichy, R. M. (2023). Empirically identifying and computationally modelling the brain-behaviour relationship for human scene categorization. *bioRxiv*, 2023.01.22.525084. https://doi.org/10.1101/2023.01.22.525084

Kasper, L., Bollmann, S., Diaconescu, A. O., Hutton, C., Heinzle, J., Iglesias, S., Hauser, T. U., Sebold, M., Manjaly, Z.-M., Pruessmann, K. P., & Stephan, K. E. (2017). The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *Journal of Neuroscience Methods*, *276*, 56–72. https://doi.org/10.1016/j.jneumeth.2016.10.019

King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, *197*, 368–382.

https://doi.org/10.1016/j.neuroimage.2019.04.079

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 1097–1105.

Linsley, D., & MacEvoy, S. P. (2014). Evidence for participation by object-selective visual cortex in scene category judgments. *Journal of Vision*, *14*(9), 19. https://doi.org/10.1167/14.9.19

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024. https://doi.org/10.1073/pnas.1719616115

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), Article 10. https://doi.org/10.1038/nn.2903

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, *35*(39), 13402–13418. https://doi.org/10.1523/JNEUROSCI.5181-14.2015

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mood, A. M. (1971). Partitioning Variance in Multiple Regression Analyses as a Tool For Developing Learning Models. *American Educational Research Journal*, *8*(2), 191–202. https://doi.org/10.3102/00028312008002191

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, *103*, 130–138.

https://doi.org/10.1016/j.neuroimage.2014.09.026

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Op de Beeck, H. P., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fMRI data: Maps, modules and dimensions. *Nature Reviews Neuroscience*, *9*(2), Article 2. https://doi.org/10.1038/nrn2314

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, *31*(4), 1333–1340. https://doi.org/10.1523/JNEUROSCI.3885-10.2011

Philiastides, M. G., & Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex (New York, N.Y. : 1991)*, *16*(4), 509–518. https://doi.org/10.1093/cercor/bhi130

Philiastides, Ratcliff, R., & Sajda, P. (2006). Neural Representation of Task Difficulty and Decision Making during Perceptual Categorization: A Timing Diagram. *Journal of Neuroscience*, *26*(35), 8965. https://doi.org/10.1523/JNEUROSCI.1655-06.2006

Polimeni, J. R., & Lewis, L. D. (2021). Imaging faster neural dynamics with fast fMRI: A need for updated models of the hemodynamic response. *How high spatiotemporal resolution fMRI can advance neuroscience*, *207*, 102174. https://doi.org/10.1016/j.pneurobio.2021.102174

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife*, *11*, e77599. https://doi.org/10.7554/eLife.77599

Reddy, L., Cichy, R. M., & VanRullen, R. (2021). Representational Content of Oscillatory

Brain Activity during Object Recognition: Contrasting Cortical and Deep Neural Network Hierarchies. *eNeuro*, *8*(3), ENEURO.0362-20.2021. https://doi.org/10.1523/ENEURO.0362-20.2021

Reeder, R. R., van Zoest, W., & Peelen, M. V. (2015). Involuntary attentional capture by task-irrelevant objects that match the search template for category detection in natural scenes. *Attention, Perception, & Psychophysics*, *77*(4), 1070–1080. https://doi.org/10.3758/s13414-015-0867-8

Reichwein Zientek, L., & Thompson, B. (2006). Commonality Analysis: Partitioning Variance to Facilitate Better Understanding of Data. *Journal of Early Intervention*, *28*(4), 299–307. https://doi.org/10.1177/105381510602800405

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*(19), 2301–2311. https://doi.org/10.1016/j.visres.2004.04.006

Ritchie, J. B., & Carlson, T. A. (2016). Neural Decoding and “Inner” Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. *Frontiers in Neuroscience*, *10*, 190. https://doi.org/10.3389/fnins.2016.00190

Ritchie, J. B., & de Beeck, H. O. (2019). Using neural distance to predict reaction time for categorizing the animacy, shape, and abstract properties of objects. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-49732-7

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. *PLOS Computational Biology*, *11*(6), e1004316. https://doi.org/10.1371/journal.pcbi.1004316

Seidl-Rathkopf, K. N., Turk-Browne, N. B., & Kastner, S. (2015). Automatic guidance of attention during real-world visual search. *Attention, Perception, & Psychophysics*, *77*(6), 1881–1895. https://doi.org/10.3758/s13414-015-0903-8

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28), eabm2219. https://doi.org/10.1126/sciadv.abm2219

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for

the representation of scene categories in human visual cortex. *Neuron*, *79*(5), 1025–1034. https://doi.org/10.1016/j.neuron.2013.06.034

Stehr, D. A., Garcia, J. O., Pyles, J. A., & Grossman, E. D. (2023). Optimizing multivariate pattern classification in rapid event-related designs. *Journal of Neuroscience Methods*, *387*, 109808. https://doi.org/10.1016/j.jneumeth.2023.109808

Vaziri-Pashkam, M., & Xu, Y. (2017). Goal-Directed Visual Processing Differentially Impacts Human Ventral and Dorsal Visual Representations. *Journal of Neuroscience*, *37*(36), 8767–8782. https://doi.org/10.1523/JNEUROSCI.3392-16.2017

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *Journal of Neuroscience*, *29*(34), 10573–10581. https://doi.org/10.1523/JNEUROSCI.0559-09.2009

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(23), 9661–9666. https://doi.org/10.1073/pnas.1015666108

Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *NeuroImage*, *99*, 402–410. https://doi.org/10.1016/j.neuroimage.2014.05.045

Wyble, B., Folk, C., & Potter, M. C. (2013). Contingent attentional capture by conceptually relevant images. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 861–871. https://doi.org/10.1037/a0030517

Xie, S., Kaiser, D., & Cichy, R. M. (2020). Visual Imagery and Perception Share Neural Representations in the Alpha Frequency Band. *Current Biology*, *30*(13), 2621-2627.e5. https://doi.org/10.1016/j.cub.2020.04.074

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.

https://doi.org/10.1109/TPAMI.2017.2723009