# Visual features are processed before navigational affordances in the human brain

Kshitij Dwivedi[1,2], Sari Sadiya[*,2,4], Marta P. Balode[1,3], Gemma Roig[2,5], Radoslaw M. Cichy[1]

[1]Department of Education and Psychology, Freie Universität Berlin, Germany
[2]Department of Computer Science, Goethe University Frankfurt, German
[3]Institute of Neuroinformatics, ETH Zurich and University of Zurich, Switzerland
[4]Frankfurt Institute for Advanced Studies (FIAS), Frankfurt, Germnay
[5]The Hessian Center for Artificial Intelligence (hessian.AI), Darmstadt, Germany
[*]Corresponding Author

## Abstract

To navigate through their immediate environment humans process scene information rapidly. How does the cascade of neural processing elicited by scene viewing to facilitate navigational planning unfold over time? To investigate, we recorded human brain responses to visual scenes with electroencephalography (EEG) and related those to computational models that operationalize three aspects of scene processing (2D, 3D, and semantic information), as well as to a behavioral model capturing navigational affordances. We found a temporal processing hierarchy: navigational affordance is processed later than the other scene features (2D, 3D, and semantic) investigated. This reveals the temporal order with which the human brain computes complex scene information and suggests that the brain leverages these pieces of information to plan navigation.

## 1. Introduction

By looking even only briefly at a scene, we rapidly extract multifaceted pieces of visual information (Fei-Fei et al., 2007; Greene & Oliva, 2009; Potter, 1975; Thorpe et al., 1996) that enable us to navigate through the scene, for example by planning a route through it. How does the brain compute visual information that affords navigational route planning in a scene?

Identifying navigational affordances is a complex computational feat that requires the processing of several different scene features, such as 3-dimensional and semantic scene aspects. For instance, successfully navigating the immediate environment requires localizing obstacles and finding out a way around them, which necessitates 3D scene information. Similarly, semantic scene classification can also benefit route planning as navigating typical basements, balconies, and garages require different procedures. We thus hypothesized that visual features crucial for identifying navigational affordances are processed before processing navigational affordances.

To test the hypothesis, we collected human electroencephalography (EEG) responses to indoor scene images, capturing the temporal order of scene feature processing in the human brain during visual scene perception.

We investigated three types of visual features: 2-dimensional (2D), 3-dimensional (3D) and semantic features. We operationalized the visual features in the indoor scene images as activations of deep neural networks (DNNs) trained to perform respective 2D, 3D and semantic tasks (Zamir et al., 2018). Navigational features are captured using navigational affordance maps (NAM) constructed using human behavioral responses when planning exit routes in natural indoor scene images (Bonner & Epstein, 2018).

We then related the visual and navigational features to EEG data using representational similarity analysis (RSA) (Kriegeskorte et al., 2008) in a time-resolved manner, yielding time courses with which visual representations of particular features emerge. Finally, we compared these features and NAM with EEG, revealing the temporal order in which these features are processed in the human brain.

We found that navigational affordance representations emerged significantly later than visual features. This is consistent with the view that the brain uses 2D, 3D and semantic scene features to facilitate navigation planning.

## 2. Results

We recorded EEG responses from 16 healthy volunteers (7 females, mean age 28.9 ± SD 5.6) to 50 indoor scene images. While viewing the stimuli, participants were asked to assess navigational affordance by imagining the directions of the navigational paths relative to the participant's viewpoint, i.e., whether the paths were leading to the left, the center, or the right (Figure 1A).

We then investigated when representations of visual and navigational features emerge in the human brain by comparing the EEG data to deep neural network (DNN) models and behavioral data operationalizing those features using representational similarity analysis.

For this we first transformed the peri-stimulus EEG responses (from -100 to +800ms with respect to stimulus onset) into representational dissimilarity matrices (RDMs) (Figure 1B) in steps of 10ms. We then created the 2D, 3D, and semantic RDMs using the activations of DNNs trained on 2D, 3D, and semantic tasks (Figure 1C).

To construct the navigational affordance model RDM, subjects were asked to indicate the exit routes starting from the bottom of a scene image presented to them using a computer mouse. Then, probabilistic heatmaps of navigational affordances were created pooling e data across subjects. These heatmaps were transformed into angular histograms that approximate a probabilistic navigational affordance map (NAM) of potential navigational paths radiating from the viewer’s perspective. Pairwise comparison between NAMs resulted in NAM RDMs (Figure 1D).

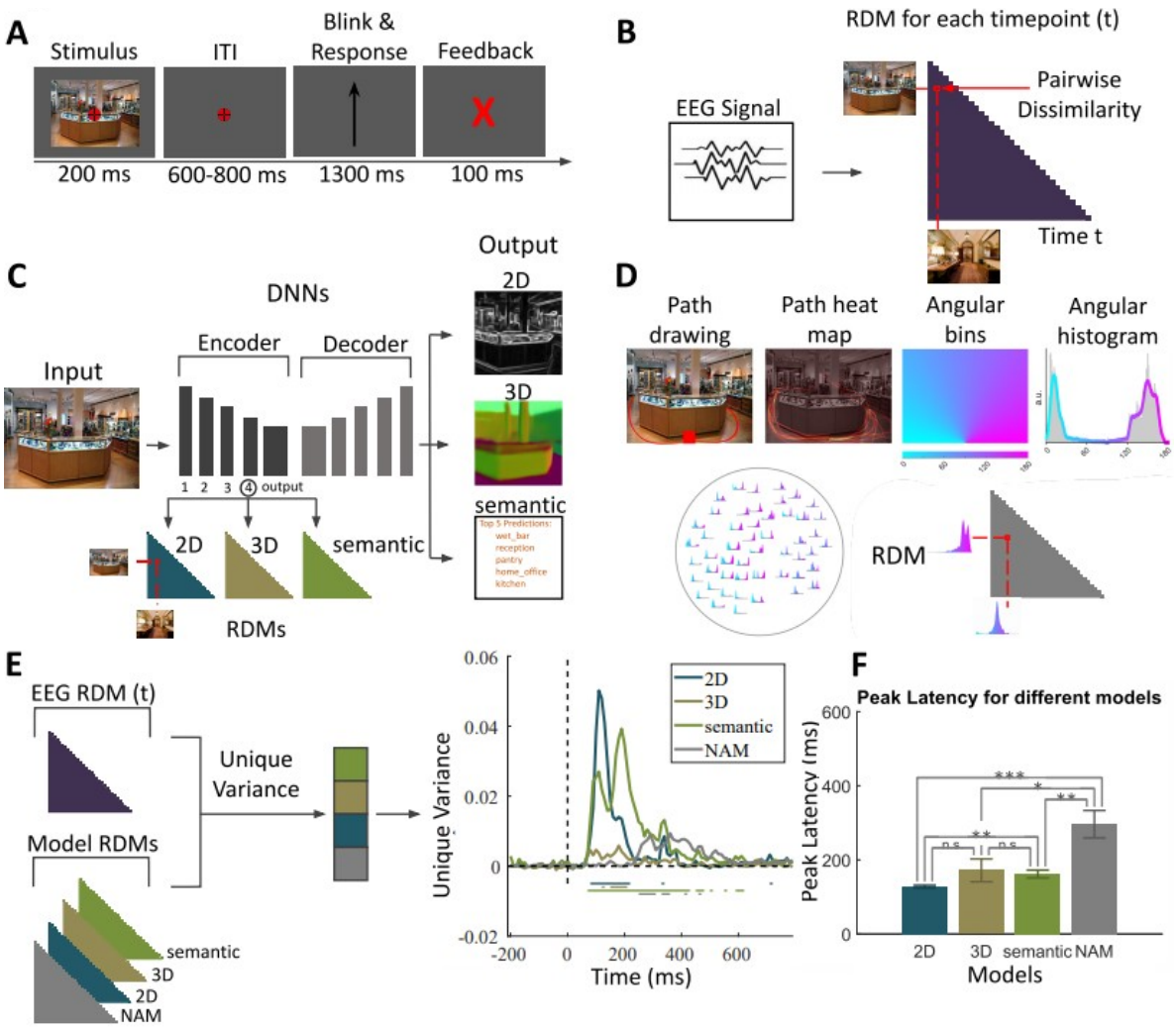


**Figure 1: A) EEG paradigm.** Participants viewed 50 images of indoor scenes and were asked to mentally plan possible exit paths through the scenes. On interspersed catch trials participants had to respond whether the exit path displayed on the screen corresponded to any of the exit paths from the previous trial. **B) EEG RDMs.** We computed RDMs for each EEG time point (every 10 ms from -200 to +800 ms with respect to image onset). **DNN RDMs.** We calculated RDMs from the activations extracted from the 4th block and output layer of a ResNet50 DNN trained on 2D, 3D and semantic tasks. **D) NAM model and RDM** (Bonner and Epstein, 2018). **E) Variance partitioning.** We calculated the unique EEG variance explained by each of the models, revealing different temporal activation patterns. Lines below the plots indicate significant times using t-test (FDR corrected $p < 0.05$). **F) Peak latencies of different models.** Bars indicate the peak latencies for different models. Error bars indicate standard deviation for 16 subjects. Stars above bars indicate significant differences across different models (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, t-test FDR corrected).

Having transformed all the modalities into a common representational space, we performed variance partitioning via regression to find out how much variance of an EEG RDM at a given time point is explained uniquely by the RDMs of any given model (Figure 1E, left panel). For this, we first performed a regression with all model RDMs as the independent variables and EEG RDM as the dependent variable. This determined the variance


4

explained by all the models together ($R^2_{all}$). Then, we performed a second set of regressions, removing the RDMs of a given model (e.g., NAM) from the independent variables to find the variance explained by models leaving out the model of interest ($R^2_{all\text{-}model}$). The unique variance of the EEG RDM explained by the selected model is then calculated as $R^2_{all}$- $R^2_{all\text{-}model}$.

We found all models explained unique variance in EEG, but to different degrees (Figure 1E, right panel). The 2D DNN RDM explained the most variance in EEG (max $R^2$=0.0502), followed by the semantic DNN (max $R^2$=0.0393). The contributions of NAM and 3D DNN RDMs were lower (NAM RDM max $R^2$=0.0094, 3D DNN RDM max $R^2$=0.0058). This provides evidence that all feature representations can be uniquely tracked in our experiment, and allowed us to inspect the time course further.

We observed a temporal pattern in peak timings (Figure 1F). The highest uniquely explained variance by the 2D DNN RDM occurred first at 128.12±3.56ms after stimulus onset, followed by 3D and semantic DNN RDMs peaking at 171.87±30.79ms and 161.87±10.45ms, respectively. The unique variance of the NAM RDM reached its peak significantly later than the other model RDMs at 296.25±37.05ms after stimulus onset. This suggests a hierarchy of scene feature processing leading up to the representation of navigational affordances.

## 3. Discussion

In this study, we investigated the temporal dynamics of scene perception, focusing particularly on the temporal order in which 2D, 3D, semantic features and navigational affordances emerge. We found that the emergence of 2D, semantic and 3D features preceded the emergence of navigational affordance representations.

The early emergence of low-level 2D features followed by high-level semantic features has been previously observed (Harel et al., 2016), in

particular in studies investigating the correspondence of layers of DNNs trained on scene classification tasks with time-resolved human EEG responses (Greene & Hansen, 2018). Equating early layers with low-level visual features and later layers with high-level features yielded a temporal hierarchy, as also observed for object processing (Cichy et al., 2016). In contrast, 3D features have attracted less attention in M/EEG studies and thus the temporal dynamics with which 3D feature representations emerge are less well understood. Although some studies investigated the temporal dynamics of spatial layout (Cichy et al., 2017; Henriksson et al., 2019), spatial layout represents only coarse-grained 3D features such as the size of the scene in the real world and the position of large surfaces, but does not take into account fine-grained 3D information such as the pose of different objects present in the scene. fMRI studies have investigated fine-grained spatial 3D features by investigating the representation of surface normals (Lescroart & Gallant, 2019) and correspondence to DNNs trained to solve 3D tasks (e.g. depth, occlusion) on the Taskonomy dataset (Dwivedi et al., 2021; Wang et al., 2019) . Here we complement these efforts in the temporal dimension by showing that 3D features are processed in parallel with semantic features.

Navigational affordance representation emerged significantly later than 2D, 3D, and semantic representations. This suggests by temporal order that humans leverage those features to process navigational affordances. In contrast to our results, a recent study (Harel et al., 2022) reported both early and late physiological markers of navigational affordances. We note the key differences in both the studies that may have led to different conclusions about navigational affordance processing. First, the images used here are natural and complex, while in Harel et al. 2022, the images were simple and synthetic. Image complexity can influence processing time, potentially due to recurrence (Kar et al., 2019). Second, in our study, subjects were tasked to identify and find their way around obstacles, making them process occlusions and 3D scene information. In Harel et al. (2022), subjects had to count the number of doors, for which processing


6

3D scene information might not be needed. The difference in timing might thus reflect a difference in feature processing as required by the task.

A limitation on the ecological validity of our study is that we used static images to assess temporal aspects of scene perception, whereas in real-world situations, humans would process moving images, especially while navigating through indoor environments. Another aspect that could be addressed in the future is inclusion of additional computational models in studying the temporal dynamics of scene perception. Nevertheless, our findings demonstrate a timeline of hierarchical scene feature processing, suggesting that the visual scene features investigated here support navigational planning.

# 4. Materials and Methods

## 4.1. EEG

### 4.1.1. Participants

We recorded EEG data from 16 healthy volunteers (7 females, mean age 28.9 ± SD 5.6). All participants had normal or corrected-to-normal vision. Participants gave informed consent before the experiment and were provided with monetary compensation. The experiment was conducted in compliance with the Declaration of Helsinki and was approved by the ethics committee of the Freie Universität Berlin.

### 4.1.2. Stimuli

The stimuli were 50 color images of different indoor environments with easily detectable navigational paths originating at the bottom center of each image, previously used in a study by Bonner & Epstein, 2017. The dimensions of all the images were 1024 x 768 pixels and subtended 7° of visual angle in width and 5.25° in height. They were presented on a gray screen with a combination of bull's-eye and crosshair fixation targets (Thaler et al., 2013) positioned centrally.

### 4.1.3. Experimental Paradigm


7

The paradigm was designed to engage the participants in explicit navigational affordance processing of every image. While viewing the stimuli, participants were asked to imagine the directions of the navigational paths relative to the participant's viewpoint, i.e., whether the paths were leading to the left, the center, or the right (Figure 1A).

On each trial Images were presented for 200ms followed by a randomly varying inter-trial interval between 600ms and 800ms. Image presentation trials were ordered in blocks of one to 5 trials in length.

Blocks were followed by the presentation of a catch trial during which participants had to conduct a task meant to ensure that participants remained attentive and processed the images with respect to spatial and navigational aspects. During catch trials an arrow appeared on the screen for 1.3s, during which the participants had to indicate whether an arrow on the screen pointed in the same (congruent) or in a different (incongruent) direction than the navigational path in the previous trial. Participants had to respond by pressing the right arrow key for "yes" (congruent, pointing in the same direction) and the left arrow key for "no" (incongruent, not pointing in the same direction). After the response, feedback was presented for 0.1s, followed by a post-feedback time of 0.2s. The number of congruent and incongruent trials was balanced across the experiment.

Blocks were organized in runs: there were 69 blocks in total (24 blocks of 5 trials, 15 blocks of 3 and 15 blocks of 4 trials, 10 blocks of 2 trials, and 5 of 1 trial) for each run presented in random order. There were 15 runs (6.2 minutes each) in total in the experiment. This design resulted in each image being repeated 75 times.

#### 4.1.4. EEG Recording and pre-processing


8

For all participants, we recorded continuous neural activity with EEG using the Easycap 64-channel standard electrode system and Brainvision actiCHamp amplifier. We followed the 10-10 system for electrode placement. EEG signals were recorded with a sampling rate of 1000Hz and bandpass filtered online between 0.03Hz and 100Hz.

All electrodes were online referenced to the FCz electrode and grounded to the AFz electrodes. We pre-processed the data offline using FieldTrip (Oostenveld et al., 2010) and selected 17 posterior and occipital channels for further analysis. The data was segmented in epochs of -0.2s to 0.8s relative to stimulus onset, baseline-corrected to the average pre-stimulus signal, and down-sampled to 100Hz. Eye blinks and other artifacts were identified with independent component analysis and removed after inspection.

#### 4.1.5. Pairwise Decoding

To determine how well EEG signals can be used to differentiate between the 50 scene images, we calculated the pairwise decoding accuracy score for each image pair at every time point using CoSMoMVPA (Oosterhof et al., 2016). This was done in a time-resolved manner, assessing 100 time points every 0.01s from -0.2 to 0.8s relative to image onset. For every possible pair of image conditions, we partitioned the pre-processed EEG signals across all trial repetitions into training and test data using a leave-one-trial-out cross-validation scheme. We then trained LDA classifiers on all-but-one trials and tested them on the left-out trials. Decoding accuracy scores were averaged across cross-validation folds. To create a grand average time series of EEG decoding accuracy, we calculated the mean decoding accuracy across all pairs.

### 4.2. Behavioral Data

To quantify the navigational affordance features in the 50 experimental images, we used the navigational affordance model (NAM) by (Bonner & Epstein, 2017), which was created using the same set of 50 images.


9

Bonner & Epstein asked participants to draw all possible navigational paths in each image starting from the bottom center of the image. The responses of the participants were aggregated together into heatmaps. Then angular binning was performed to create a navigational affordance histogram by counting the number of pixels in each one-degree bin from 0-180°.

### 4.3. Deep Neural Network models

To assess low-, mid-, and high-level features of indoor scenes, we used activations from 18 pre-trained deep neural network (DNN) models from the Taskonomy Task Bank (Zamir et al., 2018). The Task Bank consists of DNN models that cover various tasks in computer vision, which can be grouped into three categories (2D, 3D, and semantic tasks). Models from the 2D task category process low-level visual features (for example, 2D edges and key points), models from the 3D category process such mid-level features as surface normals and depth, whereas models trained on classification tasks process high-level semantic features.

All models were trained on the Taskonomy dataset (Zamir et al., 2018), which consists of 4.5 million fully annotated images of indoor environments from 600 buildings. The model architecture is built on an encoder and a decoder. The encoder is identical across the 18 task models, being based on ResNet-50 (He et al., 2016) with a compressed convolutional output layer. The decoder architecture differs between the models; for example, 2D and 3D models have a decoder structure of 15 convolution and deconvolution layers, whereas the scene classification model has only two fully connected layers in the decoder because of its low output dimensions.

To ensure comparability across models, for each model, we selected the block4 and the output layer from the encoder as the representative task-specific layers following (Dwivedi et al., 2021).


10

### 4.4. EEG-DNN/Model comparison

To compare the EEG responses with the DNN and behavioral responses we used representational similarity analysis (Kriegeskorte et al., 2008). In RSA, data from different incommensurate sources are related using a common summary of the representational geometry of each source. For this we first computed representational dissimilarity matrices (RDMs) for each model and for the EEG data. RDMs are diagonally symmetric square NxN dimensional matrices (where N is the number of conditions) that summarize the dissimilarity between condition-specific responses in each source space.

Then, we performed a variance partitioning analysis (Legendre, 2008) to estimate the unique variances of the EEG RDMs explained by the model RDM investigated in this work. We detail the RDM construction below.

#### 4.4.2. EEG RDMs

We used pairwise decoding accuracy between image conditions to construct EEG RDMs. The rationale is that the more dissimilar the EEG signals arising from two different images are, the higher the decoding accuracy score for that pair of images will be. RDMs were constructed in a time-resolved manner for each time point (Figure 1 B), yielding 50x50 EEG RDM for each of the 100 time points.

#### 4.4.3. Model RDMs

The NAM RDM was constructed by computing the euclidean distance between navigational affordance histograms of all pairs of images (Figure 1D). We downloaded the precomputed NAM RDMs from https://figshare.com/s/5ff0a04c2872e1e1f416.

For each DNN RDM, we selected block4 and the encoder output layer for creating RDMs. We measured the dissimilarity between any two image representations in the DNN by calculating 1 minus the Pearson correlation


11

distance (1-ρ) between the corresponding layer activations. This resulted in two 50x50 RDMs for each DNN (i.e., the block4 and the output layer RDM). RDMs were aggregated and averaged by DNN group for each task type (2D, 3D, semantic) (Figure 1C), resulting in two RDMs for each task type.

### 4.4.4. Variance Partitioning analysis

Since the DNNs investigated are trained on the same dataset, their RDMs are expected to be correlated. To nevertheless identify the aspects unique to a particular task type, we use variance partitioning with the goal to identify variance uniquely attributable to any one model. To compute the unique variance of a given model, we calculated the difference in variance explained when all the model RDMs are used as independent variables and variance explained when all but current model RDMs are used as independent variables.

We conducted this analysis at every time point separately, i.e. every 10ms from -200 to +800ms relative to image onset. In the regression we used the lower triangular part of the RDM as it describes the representational geometry fully and avoids potential artifacts created by including the diagonal. This resulted in four time series, one for each model (3D, 3D, semantic, navigational affordance) indicating when feature representations corresponding to the model type emerge during visual processing.

### 4.4.5. Statistical analysis

We applied a two-sided t-test to assess the statistical significance of the unique variance explained by different models and peak latencies. We corrected the p-values for multiple comparisons by applying FDR correction with a threshold of 0.05.


12

## Data and Code Availability

The code and data necessary for reproducing the results presented in this paper can be found at https://osf.io/wz4ha/.

## Acknowledgments

This work was supported by the German Research Council grants (CI241/1-1, CI241/3-1, CI241/7-1) awarded to Radoslaw Cichy, the German Research Council grants (FOR 5368 ARENA, RO 6458/2-1, RO 6458/2-1) awarded to Gemma Roig, and a European Research Council grant (ERC-StG-2018-803370) awarded to Radoslaw Cichy. Additional support was provided by the Hessian ministry of Art and Science LOEWE program through the Frankfurt Center for Multi-scale Modeling in the Life Sciences. Computing resources were provided by the high-performance computing facilities at ZEDAT, Freie Universität Berlin (Bennett et al.).

## Author Contributions

K.D., R.C., G.R. designed experiments; M.B. and K.D. performed experiments; M.B. and K.D. analyzed data; K.D. and S.S. prepared code and data for final submission; R.C. and G.R. provided funding; K.D., M.B., G.R. and R.C wrote the paper. All authors discussed the results and contributed toward the manuscript.

## Affiliations

K.D. worked on this manuscript prior to joining Amazon Inc.


13

## References


Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, *114*(18), 4793–4798.

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology*, *14*(4), e1006111.

Cichy, R. M., 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.

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.

Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLOS Computational Biology*, *17*(8), e1009267. https://doi.org/10.1371/journal.pcbi.1009267

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 10. https://doi.org/10.1167/7.1.10

Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Computational Biology*, *14*(7), e1006327. https://doi.org/10.1371/journal.pcbi.1006327

Greene, M. R., & Oliva, A. (2009). The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science*, *20*(4), 464–472.

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multi-faceted EEG investigation. *ENeuro*. https://doi.org/10.1523/ENEURO.0139-16.2016

Harel, A., Nador, J. D., Bonner, M. F., & Epstein, R. A. (2022). Early Electrophysiological Markers of Navigational Affordances in Scenes. *Journal of Cognitive Neuroscience*, *34*(3), 397–410. https://doi.org/10.1162/jocn_a_01810

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. https://doi.org/10.1038/s41593-019-0392-5

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Legendre, P. (2008). Studying beta diversity: Ecological variation partitioning by multiple regression and canonical analysis. *Journal of Plant Ecology*, *1*(1), 3–8. https://doi.org/10.1093/jpe/rtm001

Lescroart, M. D., & Gallant, J. L. (2019). Human scene-selective areas represent 3D configurations of surfaces. *Neuron*, *101*(1), 178–192.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2010). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*, *2011*, e156869. https://doi.org/10.1155/2011/156869

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal



14

Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*. https://www.frontiersin.org/articles/10.3389/fninf.2016.00027

Potter, M. C. (1975). Meaning in Visual Search. *Science*, *187*(4180), 965–966. https://doi.org/10.1126/science.1145183

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. https://doi.org/10.1038/381520a0

Wang, A. Y., Wehbe, L., & Tarr, M. J. (2019). *Neural Taskonomy: Inferring the Similarity of Task-Derived Representations from Brain Activity* [Preprint]. Neuroscience. https://doi.org/10.1101/708016

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). *Taskonomy: Disentangling Task Transfer Learning*. 3712–3722. http://openaccess.thecvf.com/content_cvpr_2018/html/Zamir_Taskonomy_Disentangling_Task_CVPR_2018_paper.html

Bennett, L., Melchers, B., Proppe, B. (2020). Curta: A General-purpose High-Performance Computer at ZEDAT, Freie Universität Berlin. http://dx.doi.org/10.17169/refubium-26754


15