NeuroImage 194 (2019) 12–24

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

The spatiotemporal neural dynamics underlying perceived similarity for real-world objects

Radoslaw M. Cichy[a,b,c,*], Nikolaus Kriegeskorte[d], Kamila M. Jozwik[a], Jasper J.F. van den Bosch[e], Ian Charest[e,f]

[a] *Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany*
[b] *Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany*
[c] *Berlin School of Mind and Brain, Berlin, Germany*
[d] *Department of Psychology, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA*
[e] *MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK*
[f] *School of Psychology, University of Birmingham, Birmingham, UK*

ARTICLE INFO

Keywords:
fMRI
MEG
Object recognition
Perceived similarity
Visual perception

ABSTRACT

The degree to which we perceive real-world objects as similar or dissimilar structures our perception and guides categorization behavior. Here, we investigated the neural representations enabling perceived similarity using behavioral judgments, fMRI and MEG. As different object dimensions co-occur and partly correlate, to understand the relationship between perceived similarity and brain activity it is necessary to assess the unique role of multiple object dimensions. We thus behaviorally assessed perceived object similarity in relation to shape, function, color and background. We then used representational similarity analyses to relate these behavioral judgments to brain activity. We observed a link between each object dimension and representations in visual cortex. These representations emerged rapidly within 200 ms of stimulus onset. Assessing the unique role of each object dimension revealed partly overlapping and distributed representations: while color-related representations distinctly preceded shape-related representations both in the processing hierarchy of the ventral visual pathway and in time, several dimensions were linked to high-level ventral visual cortex. Further analysis singled out the shape dimension as neither fully accounted for by supra-category membership, nor a deep neural network trained on object categorization. Together our results comprehensively characterize the relationship between perceived similarity of key object dimensions and neural activity.

1. Introduction

Perceived similarity refers to the impression of how much one object looks like another. It structures our conscious visual experience when comparing and differentiating objects. It also guides behavior when identifying novel objects by comparing them to other resembling objects or known categories (Nosofsky, 1984; Shepard, 1987; Ashby and Perrin, 1988; Edelman, 1998).

If two objects are perceived to be similar, it could be assumed this is because their neural representations are similar, too (Shepard and Chipman, 1970; Edelman, 1998). However, judgments of similarity must reflect specific perceptual dimensions: two objects sharing similar color can have a very different shape. Thus, when comparing perceived similarity to neural similarity, it is important to characterize the objects dimensions that support and mediate the link.

Previous studies used artificial 2D (Kourtzi and Kanwisher, 2001; Op de Beeck et al., 2001; Kayaert et al., 2005; Haushofer et al., 2008; Drucker and Aguirre, 2009; Wardle et al., 2016), 3D shapes (Op de Beeck et al., 2008b), and faces (Rotshtein et al., 2005; Davidesco et al., 2014) and isolated shape as an object dimension underlying the link between perceived similarity and brain activity. However, they did not assess other important properties that pertinently characterize everyday objects such as function, color, or category membership. In contrast, studies using real-world object stimuli (Edelman, 1998; Weber et al., 2009; Connolly et al., 2012; Mur et al., 2013, 2013; Charest et al., 2014, 2014; Peelen et al., 2014; Bankson et al., 2018) assessed the role of such properties implicitly, but did not explicitly tease apart their respective link between brain activity and perceived similarity.

[*] Corresponding author. Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany.
E-mail address: rmcichy@zedat.fu-berlin.de (R.M. Cichy).

https://doi.org/10.1016/j.neuroimage.2019.03.031
Received 16 July 2018; Received in revised form 25 January 2019; Accepted 13 March 2019
Available online 17 March 2019

Here we investigated the unique link between brain activity and different object properties. For this, we acquired behavioral assessments of perceived similarity for object dimensions shape, color, function and background. We then used representational similarity analysis (Kriegeskorte, 2008; Kriegeskorte and Kievit, 2013) to establish multivariate relationships between perceived similarity judgments to brain data recorded using functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG).

To further elucidate the nature of the link found for each object dimension, we conducted two control analyses. In each, we partialled out model RDMs that capture particular hypotheses about the nature of the underlying link. This tests whether the hypothesis captured in the RDM does fully capture the link between perceived similarity and brain measurement, or not. First, as previous research has shown that artificial deep neural networks (DNNs) predict behavioral assessment of perceived similarity (Kubilius et al., 2016; Peterson et al., 2016), we investigated whether DNNs account for the link between brain activity and perceived similarity, too. Second, category membership has been found to strongly predict perceived similarity and ventral visual cortex high-level representations, as measured with functional Magnetic Resonance Imaging (fMRI) of the infero-temporal cortex (IT; (Mur et al., 2013). We thus further investigated the role of category membership in predicting the relationship between perceived similarity and similarity of activity patterns in the brain.

2. Materials and methods

2.1. Visual stimulus set

The stimulus set consisted of 118 square images of everyday objects, each one from a different entry-level object category, on real-world backgrounds (Fig. 1A). Images were taken from the ImageNet database

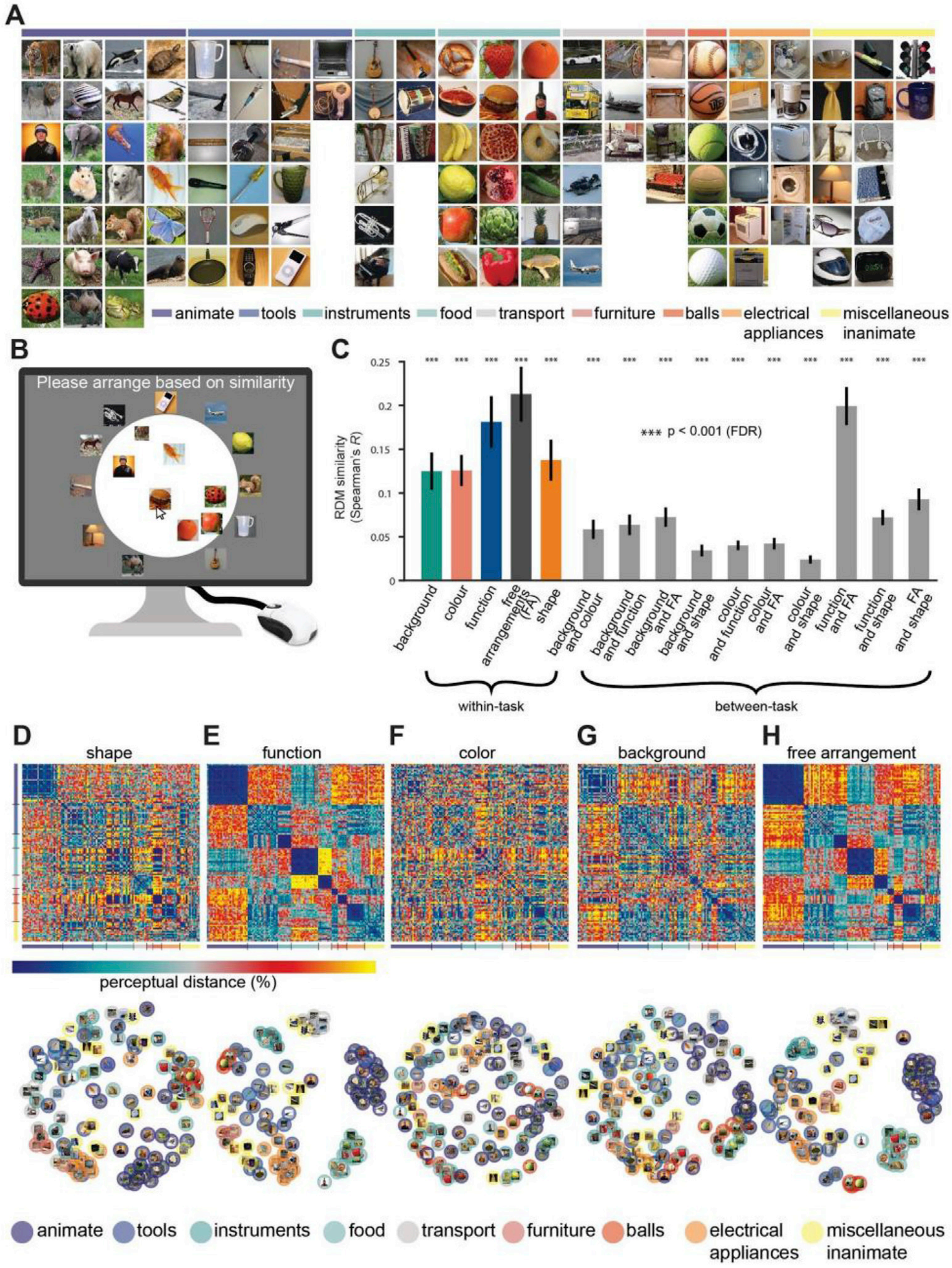

Fig. 1. Stimulus set, and behavioral assessment. A) The stimulus set consisted of 118 images of real-world objects, each from a different entry-level level category, clustering into several supra-level categories. **B**) Multiple arrangement procedures. Participants were asked to arrange images in a 2D arena such that similar objects were put close together, and dissimilar objects were put far apart. **C**) RDM correlations within and between multiple arrangement tasks defined by object dimension to be used for similarity ratings. We computed the correlations between RDMs obtained from each multiple arrangement task across subjects and either within task and across tasks. Overall, RDM correlations were highest within-task (colored bars), and between-task correlations were also high (gray bars). Asterisks indicate significant correlations (all $p < 0.001$, FDR-corrected), and error bars indicate the standard error of the mean. Perceived similarity RDMs resulting from judging objects according to **D**) shape, **E**), function, **F**) color, **G**) background, and freely without detailed instructions (free arrangement). The bottom row shows respective MDS projections into two dimensions. The columns and rows of the RDMs are arranged based on the supra-category color codes used in panel A. MDS solutions indicate category membership according to the same color code.

and cropped to square size. The same stimulus set was used in three separate experiments: behavioral assessment, MEG and fMRI recordings.

2.2. Behavioral ratings of perceived similarity

Five separate pools of participants (total $n = 127$, 66 female, age: mean ± s.d. = 28.96 ± 8.12 years) gave ratings on the perceived similarity of object images according to one of five instructions. Participants were asked to rate similarity by i) shape ($n = 27$) ii) function ($n = 26$), iii) the color and ($n = 27$) iv) the background of the objects ($n = 27$), or iv) freely without detailed instructions ($n = 20$) (i.e. 'free arrangements') in a multiple arrangements (MA) task (Kriegeskorte and Mur, 2012; Charest et al., 2014). In detail, participants were asked to arrange the images on a computer screen inside a white circular arena by using computer mouse drag and drop operations. For example, when participants performed the MA task based on shape, they were instructed to place object images closer together when they had a similar shape, and further apart when their shape was dissimilar (Fig. 1B). The instruction followed the same logic for all 4 MA tasks based on specific object dimensions. The MA task based on free arrangement was different, in that it instructed participants simply to arrange the objects based on their similarity, without specific instructions on what dimensions to use. On the first trial of each MA task, participants arranged all images according to the task's instructions. Subsequent trials consisted of a subset of those objects, which were selected based on an adaptive procedure aimed at 1) minimizing uncertainty for all possible pairs of images (e.g. items that initially were placed very close to each other) and 2) better approximate the high-dimensional perceptual representational space. Often these subsequent trials included items that were placed close together in the initial trial. As subsequent trials included fewer objects, this allowed participants to refine their judgements with distinctions that are more difficult to carry in the context of the whole image set and the limited arena space. This task can efficiently deal with large number of stimuli and obtains reliable similarity judgments within 60 min per participant (please refer to Kriegeskorte and Mur, 2012 for further details on the adaptive selection procedure). The behavioral data was collected either in the behavioral laboratory (for free arrangements) or using the Meadows web-based platform for psychophysical experiments (http://meadows-research.com ; for all other tasks). Online participants were recruited from the Prolific online participant pool (http://www.prolific.ac.uk). The behavioral experiments were conducted according to the Declaration of Helsinki and approved by the local Ethics Committee of the Freie Universität Berlin.

2.3. Neuroimaging experiments: participants and experimental design

MEG and fMRI data have been used in a previous study (Cichy et al., 2016b). Here we provide a summary of the relevant parameters.

Participants ($n = 15$, 5 female, age: mean ± s.d. = 26.6 ± 5.18 years) took part in an fMRI and MEG experiment. These participants were distinct from the participants that took part in the behavioral experiments. All participants were healthy and right handed with normal or corrected-to-normal vision. The stimulus set used in the neuroimaging experiments was identical to the one used in the behavioral experiments. In both the fMRI and MEG experiments, images were presented for 500 ms at the center of a screen (image width and height: 4.0° of visual angle). A gray fixation cross was shown in the middle of the screen throughout the experiment. Further presentation parameters were adjusted to the specific requirements of each imaging technique.

In the MEG experiment, each participant completed one session consisting of 15 runs of 314 s duration each. In each run, every object image was shown twice, with random condition order and a trial onset asynchrony of 0.9–1.0 s. Participants were instructed to respond with an eye blink and a button press to the image of a paper clip shown randomly every 3–5 trials (average = 4 s). Participants were instructed not to blink their eyes at any other times.

In the fMRI experiment, each participant completed two sessions of 9–11 runs of 486 s duration each. In each run, every object image was shown once, and condition order was randomized with an inter-trial interval of 3 s. In addition, 39 null trials (gray background) were interspersed randomly during which only a gray background was presented. Participants were instructed to respond to a change in luminance of the fixation cross with a button press.

2.4. MEG acquisition and preprocessing

MEG signals were recorded with a sampling rate of 1 kHz from 306 channels (204 planar gradiometers, 102 magnetometers, Elekta Neuromag TRIUX, Elekta, Stockholm). Data were filtered online between 0.03 and 330 Hz. We applied temporal source space separation (maxfilter software, Elekta, Stockholm (Taulu et al., 2004; Taulu and Simola, 2006)) before further analyzing the data with the Brainstorm software (Tadel et al., 2011). The data were epoched from −100 ms to +1000 ms with respect to the onset of each trial.

We baseline corrected each trial by subtracting the pre-onset period average from every other point, and smoothed data with a 20-ms sliding window. This resulted in 30 trials for each condition, session, and participant. Following current preprocessing recommendation for multivariate analysis of MEG data (Guggenmos et al., 2018), we noise-normalized MEG data. For this we calculated covariance matrices based on sensor activation patterns of 30 trials, for each condition and time point separately. We then (i) averaged all covariance matrices, (ii) inverted the mean covariance matrix including shrinkage (Ledoit and Wolf, 2004), and (iii) multiplied MEG data for each condition, trial and time point with the inverted covariance matrix.

2.5. fMRI acquisition, preprocessing and analysis

MRI data was acquired using a 3 T TIM Trio scanner (Siemens, Erlangen, Germany) with a 32-channel head coil. Structural images were acquired using a standard T1-weighted sequence (192 sagittal slices, FOV = 256 mm^2, TR = 1900 ms, TE = 2.52 ms, flip angle = 9°). Functional images covering the entire cortex were acquired in runs of 648 vol using a gradient-echo EPI sequence (TR = 750 ms, TE = 30 ms, flip angle = 61°, FOV read = 192 mm, FOV phase = 100% with a partial fraction of 6/8, through-plane acceleration factor 3, bandwidth 1816Hz/Px, resolution = 3 mm^3, slice gap 20%, slices = 33, ascending acquisition).

We processed fMRI data using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) for each participant separately. We realigned and co-registered fMRI data to the T1 structural scan before normalizing it to the standard MNI template. We estimated condition-specific responses using a general linear model (GLM), consisting of regressors of interest based on condition-specific image onsets convolved with a hemodynamic response function, and nuisance regressors based on movement parameters. The estimated condition-specific GLM parameters were converted to *t*-values by contrasting each condition estimate against the implicitly modeled baseline. For each subject, this resulted in 118 condition-specific *t*-value maps.

To analyze fMRI data in a spatially unbiased fashion, we performed a volumetric searchlight analysis in each participant separately (Haynes and Rees, 2005; Kriegeskorte et al., 2006). For each voxel v, we extracted condition-specific *t*-value patterns in a sphere centred at v with a radius of 4 voxels (searchlight at v) and arranged them into fMRI *t*-value pattern vectors for further representational similarity analysis (described below).

2.6. DNN architecture and training

To investigate the degree to which an artificial deep neural network (DNN) trained on object categorization accounts for the link between patterns of brain activity and perceived similarity, we evaluated a DNN used in a previous publication (Cichy et al., 2016a) that is freely available (http://brainmodels.csail.mit.edu/object_dnn.tar.gz). In detail, the DNN had an architecture identical to the network used by Krizhevsky et al. (2012). It consisted of 8 layers: the first five layers were convolutional, the remaining three layers were fully connected. Layers 1 and 2 had three stages: convolution, max pooling and normalization, layers 3–5 had a convolution stage only. The number of units and features are enumerated in Supplementary Table 2. The training of the DNN was carried out as follows. We trained the DNN on 900k images in 683 different object classes from the ImageNet database with roughly equal number of images per category (~1300). The training was done on a GPU using the Caffe toolbox (http://caffe.berkeleyvision.org/) with the following learning parameters: the DNN was trained for 450k iterations, with the initial learning rate set to 0.01 and a step multiple of 0.1 every 100k iterations. The momentum and weight decay were fixed at 0.9 and 0.0005 respectively.

2.7. Representational similarity analysis (RSA) links brain data to perceived similarity ratings

Brain imaging data and perceived similarity ratings were acquired in different multivariate measurement spaces: perceived similarity ratings provided coordinates in 2D areas, MEG yielded sensor activation patterns, and fMRI provided voxel activation patterns in searchlights. The difference in the nature of measurement spaces necessitates additional steps to make data comparable. One way to do so is representational similarity analysis (RSA) (Kriegeskorte, 2008; Kriegeskorte and Kievit, 2013). The basic idea is that if two objects are perceived as similar, and also evoke similar brain activation patterns, then the brain activation patterns and the perception are to be linked. Formally, pairwise similarities (or equivalently: dissimilarities) for a measurement space are ordered in matrices, called representational dissimilarity matrices (RDM), indexed in rows and columns by the compared conditions (here of size 118 × 118 due to the 118 object categories). RDMs are then compared using correlation to establish links between measurement spaces.

An extension of the RSA framework allows further to investigate the importance of factors hypothesized to be important for the established relationship. For this, in the process of correlating one RDM to another, a single RDM or a set of RDMs that models the factor is partialled out. If the partialling out does not abolish the link, this indicates that it cannot be fully explained by the hypothesized factor.

Below we first explicate how RDMs were calculated for (i) fMRI, (ii) MEG, (iii) perceived similarity ratings and (iv) model RDMs for partial correlation RSA of two kinds: a DNN RDM that captures how a deep neural network trained on object categorization processes the stimulus set, and a supra-category RDM that captures category relations. We then give the details of single analyses using RSA.

2.7.1. Construction of perceived similarity rating RDMs

For a given participant and a given task instruction, we reconstructed a similarity rating RDM from the partial RDMs obtained in each trial of the multiple arrangement task. As described above, on the first trial, all objects are arranged by the participant in the circular arena. In subsequent trials, a “lift the weakest” adaptive selection procedure defines a subset of images to present. This procedure utilizes the current information across pairs of objects and samples image subsets for subsequent trials in order to maximize the dissimilarity information across all pairs over the course of the experiment and reflect the multidimensional nature of perceived similarity. These subsequent trials, for which a subset of the objects is arranged, provide partial RDMs. The partial RDMs are combined by estimating each dissimilarity as a weighted average of the scale-adjusted distances in the arrangements in which an item pair was included. The complete algorithm for this “weighted average of iteratively scaled-to-match subset dissimilarity matrices” has been detailed previously (Kriegeskorte and Mur, 2012). This combining of the partial RDMs leads to a full RDM, indexed in rows and columns by the compared conditions (118 × 118).

2.7.2. Construction of fMRI RDMs

For each searchlight, we compared the dissimilarity between fMRI pattern vectors by calculating 1 minus Pearson's R for each pair of conditions, resulting in a 118 × 118 fMRI representational dissimilarity matrix (fMRI RDM). These RDMs were symmetric across the diagonal, and entries were bounded between 0 (no dissimilarity) and 2 (complete dissimilarity). The choice of Pearson's R as dissimilarity measure is current standard in the field, and Pearson's R was found to be a measure that reliably characterizes similarity relations in fMRI data (Walther et al., 2016).

2.7.3. Construction of MEG RDMs

To calculate similarity relations between condition-specific MEG sensor patterns, we first averaged the noise-normalized MEG data for each condition across trials, resulting in one noise-normalized activation pattern per time point and condition. Then, for each time point, we calculated the Euclidean distance pairwise for all pairs of conditions, and assigned it to a matrix of size 118 × 118, with rows and columns indexed by the conditions compared. The matrix was symmetric and unbounded. This procedure yielded one 118 × 118 matrix of Euclidean distances for every time-point, and we refer to it as the MEG representational dissimilarity matrix (MEG RDM). The choice of Euclidean distance as a dissimilarity measure was based on recent recommendations for RSA on MEG data (Guggenmos et al., 2018).

2.7.4. Construction of DNN RDMs

We used the last processing stage of each DNN layer to build layer-specific RDMs. In detail, we determined the activation pattern across all units in a given layer for each image of the stimulus set. For each layer separately, we calculated dissimilarity (1 - Spearman's R) for all condition-specific activation values. In total, this resulted in 8 layer-specific DNN RDMs. The DNN RDM is symmetric across the diagonal and bounded between 0 (no dissimilarity) and 2 (complete dissimilarity).

2.7.5. Construction of supra-category RDMs

While the stimulus set was designed such that every object was from a different entry-level category, as expected for any larger set of objects, multiple objects fell into supra-level categories (e.g. a bear and a dog are different categories, but they are both animate). To investigate the role of supra-category membership in the link between brain activity and perceived similarity ratings, we assessed the supra-category structure of the stimulus set and captured it in supra-category model RDMs.

We identified the following supra-level categories in the stimulus set, guided by semantic divisions in the organization of object knowledge in the human brain observed in neuropsychological research (Warrington and Shallice, 1984; Hart et al., 1985; Damasio, 1990; Martin et al., 1996; Caramazza and Mahon, 2003; Mahon and Caramazza, 2009). Two raters performed category classification independently and discussed unclear cases until consensus was reached with the following results: animate objects (27), and inanimate objects (91), where inanimate are further subdivided into tools (21), food (18), music instruments (9), means of transport (9), electric appliance (11), balls (6), furniture (4) and miscellaneous (14).

To assess the effect of supra-category in model RDMs, we constructed RDMs that model the hypothesis of mean distance differences between subdivisions in two ways. First, for each of the 10 subdivisions, we defined one model RDM that captures the mean effects *within* each subdivision. For this RDM, matrix elements defined by the relevant subdivision were set to 1 (e.g. for within animate: all matrix elements defined by animate objects), and 0 otherwise. Second, for all possible pairs of subdivision comparisons (45), we defined an RDM that captured the mean effect *between* subdivisions. Again, for this model RDM, matrix elements defined by the relevant subdivision were set to 1 (e.g. for animate vs inanimate: all matrix elements defined by one animate and one inanimate object), and 0 otherwise. In total, this resulted in 55 model RDMs (10 within subdivision RDMs, plus 45 between subdivision RDMs for all between supra-category pairs ((10*10)-10)/2 = 45).

2.8. RSA using correlation and partial correlation

2.8.1. Relating brain data to perceived similarity by RSA

All analyses using RSA were conducted independently for each MEG or fMRI participant. We first extracted the lower triangular part of each RDM – excluding the diagonal (Ritchie et al., 2017) – and vectorized them. Second, we correlated each behavioral RDM with the fMRI RDMs (Fig. 2A) and MEG RDMs (Fig, 5A), using Spearman's *R*. For each dimension of perceived similarity, for each fMRI participant ($n = 15$) this yielded a 3D correlation map revealing locations in the brain where perceived similarity ratings and brain activity were related. Similarly, for every MEG participant ($n = 15$), the correlation with the temporally-resolved MEG RDMs yielded one time course indicating the time points during which perceived similarity ratings and brain activity were related. Finally, participant-specific 3D maps and time courses were analyzed statistically.

2.8.2. Partial correlation analyses

We conducted five types of partial correlation analyses correlating brain data RDMs (MEG, fMRI) with a particular similarity rating RDM to characterize the link between perceived similarity and brain activation patterns comprehensively. First, assessing each similarity rating dimension in turn, we partialled out all other similarity rating RDMs (Figs. 3A and 5A). This analysis revealed the unique relationship between brain activation and a given dimension of perceived similarity. Second, we partialled out supra-category RDMs ($M = 55$, where M is the number of RDMs partialled out; Figs. 4A and 6A), revealing the relationship between brain activation and perceived similarity not accounted for by supra-category membership. Third, we partialled out layer-specific DNN RDM ($M = 8$; Figs. 4C and 7C). This revealed the relationship between brain activation and perceived similarity not accounted by features of the DNN. Fourth, motivated by previous research showing that a combination of category effects and DNN features best explains brain activity in the cortical region (IT) believed to be strongly related to perceived similarity (Khaligh-Razavi and Kriegeskorte, 2014), we partialled out the effect of supra-category RDMs and DNN RDMs together ($M = 55 + 8 = 63$; Figs. 4F and 7E). Fifth, we investigated the relationship between brain activation and perceived similarity that is unique to a particular dimension of perceived similarity and not accounted for by either supra-category membership or DNN features. For this, assessing each similarity rating dimension in turn, we partialled out all other similarity rating RDMs, supra-category RDMs and DNN RDMs ($M = 4 + 55+8 = 67$; Figs. 4H and 7G).

2.9. Statistical testing

Results were tested for statistical significance using one-sided sign permutation tests. In short, we randomly flipped the sign (10,000 permutation samples) of subject-specific data (i.e., 3D fMRI correlation maps or MEG time courses) to determine an empirical distribution on the basis of which we determined significant effects at a threshold of $P < 0.05$ for MEG and $P < 0.001$ for fMRI. Next, we used maximum weighted cluster size inference (i.e., the sum of all values in a cluster) with a cluster extent threshold of $P < 0.05$ (Nichols and Holmes, 2002; Pantazis et al., 2005; Cichy et al., 2014). This procedure effectively controls for multiple comparisons in cases where neighboring tests have a meaningful structure, i.e. neighboring voxels in the searchlight analysis and neighboring time points in the MEG analysis, respectively. The cluster-extent threshold was Bonferroni-corrected for multiple comparisons by the number of object dimensions investigated.

To provide estimates of the accuracy of peak latency in MEG RSA results and in peak latency differences between MEG RSA results, we bootstrapped the pool of subjects (1000 bootstraps) and calculated the 95% confidence intervals of the sampled bootstrapped distribution. Finally, to test peak latency differences for statistical significance, we determined *p*-values based on the bootstrap distribution. Results are reported as significant with $p < 0.05$, FDR corrected for multiple comparisons.

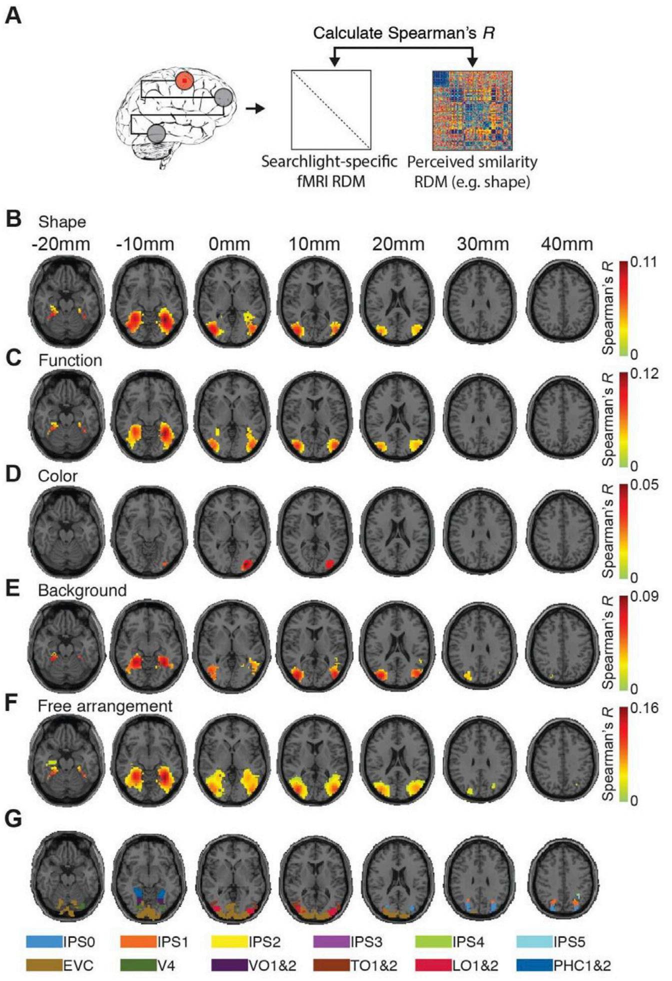

Fig. 2. Location of representations underlying perceived similarity in the human brain. A) Procedure. We used a spatially unbiased searchlight procedure to localize neural representations underlying perceived similarity. For every voxel in the brain, we extracted the activation patterns for the 118 object images in a sphere around the voxel (symbolized by orange circle). We calculated searchlight-specific fMRI RDMs from those patterns. We then correlated (Spearman's *R*) the fMRI RDMs with the perceived similarity RDMs. The results of relating fMRI activation patterns to perceived similarity according to **B**) shape, **C**) function, **D**) color, **E**) background and **F**) in free arrangement. There was a significant relationship between fMRI activation patterns and perceived similarity for all object dimensions ($n = 15$, sign-rank test, cluster-definition threshold $p < 0.001$, cluster size threshold $p < 0.05$ Bonferroni corrected by number of object dimensions assessed (5); results are overlaid on axial slices of a standard T1 in MNI space). **G**) Visualization of ventral and dorsal regions from a probabilistic atlas (Wang et al., 2015).

3. Results

3.1. Behavioral assessment of perceived similarity

We assessed perceived similarity through a multiple arrangement similarity rating task (Fig. 1B) (Kriegeskorte and Mur, 2012) on a set of 118 everyday object stimuli (Fig. 1A) presented on real-world backgrounds. While each stimulus was from a different entry level category (Rosch et al., 1976), as expected for any sizable number of natural objects, our stimulus set consisted of several supra-level categorical

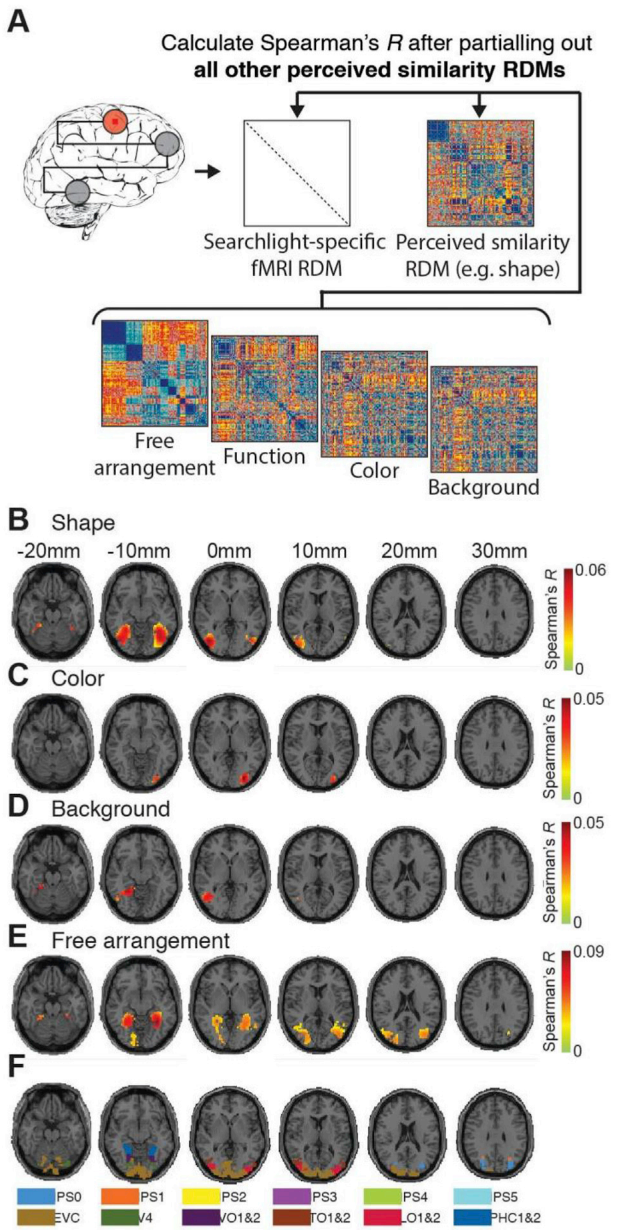

Fig. 3. Location of representations related uniquely to a particular dimension of perceived similarity. A) Procedure. We adapted the searchlight procedure for a partial correlation analysis. We compared (Spearman's *R*) the fMRI RDMs with a perceived similarity RDM for a particular object dimension (e.g. shape) while partialling out the RDMs for all other object dimensions. Significant results were found for **B**) shape, **C**) color, **D**) background and **E**) free arrangement ($n = 15$, sign-rank test, cluster-definition threshold $p < 0.001$, cluster size threshold $p < 0.05$ Bonferroni corrected for number of object dimensions assessed (5); results are overlaid on axial slices of a standard T1 in MNI space). **F**) Visualization of ventral and dorsal regions from a probabilistic atlas (Wang et al., 2015).

divisions: animals (27), tools (21), food (18), electric appliances (11), music instruments (9), means of transport (9), balls (6), furniture (4) and miscellaneous (14).

By necessity, similarity is defined with respect to the dimensions the similarity judgment is based on: for example, two object images might be similar according to one dimension, e.g. color, but dissimilar according to another dimension, e.g. function. However, in real-world objects different properties often concur, e.g. many objects that have similar function have similar shapes. Thus, it is important to assess the degree to which different object properties both uniquely and in common determine perceived similarity. Here, we investigated the perceived similarity of the stimulus set along four pertinent dimensions of real-world objects: i) shape, ii) function, ii) color and iv) background of the object. As any list of properties is unlikely to be exhaustive, to capture other properties we also evaluated perceived similarity by free arrangement, i.e. without giving explicit instructions about which dimensions to use. The rationale is that when participants judge similarity of objects holistically rather than according to a particular property, their similarity judgments reflect the influence of a mixture of object properties in different, unknown proportions. Thus, in total five separate sets of participants ($n = 127$) were asked to judge object image similarity.

This resulted in one perceived similarity representational dissimilarity matrix (RDM) per participant and object dimension, summarizing the pairwise perceived similarity for all condition pairs. The RDMs averaged across participants are shown in Fig. 1C–G. To visualize the main structure in the RDMs, we used multidimensional scaling (MDS) (Kruskal and Wish, 1978; Shepard, 1980). The MDS solutions in two dimensions are plotted below the RDMs (Fig. 1D–H). Visual inspection revealed commonalities and differences in perceived similarity across investigated dimensions. For example, in all cases animate object images were perceived to be different from all other objects, but more so when judging freely or according to the shape or function of objects. To quantify the similarity of perceived similarity ratings and their respective reliability, we calculated the correlation (Spearman's *R*) between and within the five similarity RDMs (Fig. 1C; all values in Supplementary Table 1). We found that RDM correlations were highest within task as expected, indicating the reliability of perceived similarity across subjects. Further, all similarity ratings were significantly positively correlated (sign-rank test, all $p < 0.05$, FDR-corrected). The most strongly correlated RDMs were 'free arrangement' and 'function' ($R = 0.19 \pm 0.016$), and the least correlated were 'color' and 'shape' ($R = 0.024 \pm 0.0024$). An analysis of the coefficient of determinations further revealed that all investigated factors accounted for significant variance, and that function strongly dominated (Supplementary Fig. 1). Together, this comparison quantifies the intuition that while perceived similarity depends on the dimension of the objects judged, in real-world stimuli many dimensions - to varying degrees - concurrently determine perceived similarity.

The behavioral assessment forms the basis for investigation of the link between perceived similarity and neural representations resolved in space and time using fMRI and MEG respectively.

3.2. Representations in ventral visual cortex underlie perceived similarity for all object dimensions

To localize the neural representations underlying perceived similarity, we recorded fMRI data while participants viewed the stimulus set. We then related perceived similarity ratings and fMRI data using representational similarity analysis (RSA) (Kriegeskorte, 2008; Kriegeskorte and Kievit, 2013) in a searchlight analysis (Haynes and Rees, 2005; Kriegeskorte et al., 2006) (Fig. 2A). Conducted for each object dimension separately, this resulted in 3D maps indicating where in the brain local fMRI activity patterns are representationally correspondent to perceived similarity.

We found a significant relationship between fMRI activation patterns and perceived similarity for all object dimensions (Fig. 2B–F). Significant results were found in large patches across ventral and dorsal visual stream, with highest values in high-level ventral visual cortex for all assessed object dimensions. This result further tightens the link between high-level ventral visual cortex and perceived similarity of objects. Further, it extends previous studies that investigated this link using shape or free arrangements to other dimensions of perceived similarity. Finally, it forms a solid basis for more specific analyses of the nature of the link below.

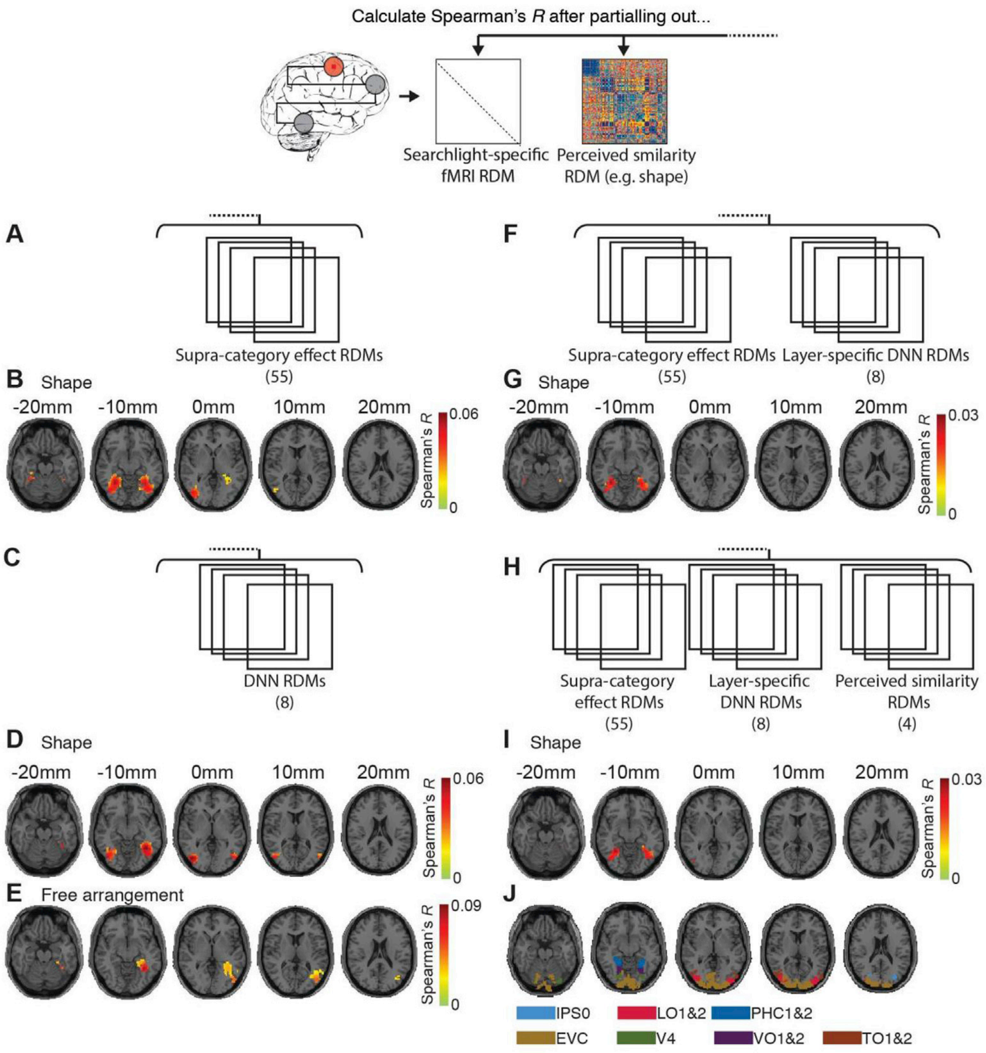
Calculate Spearman's R after partialling out...
Searchlight-specific fMRI RDM
Perceived smilarity RDM (e.g. shape)
A
Supra-category effect RDMs (55)
B Shape
-20mm
-10mm
0mm
10mm
20mm
Spearman's R
0.06
0
C
DNN RDMs (8)
D Shape
E Free arrangement
0.09
F
Supra-category effect RDMs (55)
Layer-specific DNN RDMs (8)
G Shape
0.03
H
Supra-category effect RDMs (55)
Layer-specific DNN RDMs (8)
Perceived similarity RDMs (4)
I Shape
J
IPS0
LO1&2
PHC1&2
EVC
V4
VO1&2
TO1&2

Fig. 4. Supra-category membership and a DNN do not fully account for the link between perceived similarity and brain activity measured with fMRI. In the partial correlation framework, we compared (Spearman's R) the fMRI RDMs with each perceived similarity RDM while partialling out RDMs capturing **A**) supra-category level effects (results in **B**), **C**) layer-specific DNN RDMs (results in D,E), **F**) supra-level category effects and layer-specific DNN RDMs in combination (results in G), as well as **H**) as F and all other perceived similarity RDMs (results in I). No investigated factor fully accounted for the link between perceived similarity by shape and brain activity measured with fMRI. In addition, DNN RDMs did fully account for the perceived similarity by free arrangement and brain activity link ($n = 15$, sign-rank test, cluster-definition threshold $p < 0.001$, cluster size threshold $p < 0.05$ Bonferroni corrected by number of object dimensions assessed (5); results are overlaid on axial slices of a standard T1 in MNI space). **J**) Visualization of ventral and dorsal regions from a probabilistic atlas (Wang et al., 2015).

3.2.1. Representations underlying perceived similarity unique to particular object dimensions are distributed and partly overlapping

Behavioral assessment revealed positive correlations between perceived similarity ratings for different object dimensions (Fig. 1C, gray bars). This raises the question to what degree the observed relationship between the different dimensions of perceived similarity and brain activity is due to aspects common to all object dimensions, or to aspects unique to a particular dimension. To reveal the unique aspects of the relationship we used a partial correlation analysis (Fig. 3A): correlating fMRI RDMs to a particular perceived similarity RDMs we partialled out the effect of all other perceived similarity RDMs. This analysis controls for the effect of the partialled out factor on the relationship between perceived similarity and brain activity.

This partial correlation analysis revealed a positive correlation between representations in ventral visual cortex and perceived similarity for all dimensions, except function (Fig. 3B–E). The strongest correlation for shape, background and free arrangements was consistently found in ventral-medial and lateral-occipital cortex. This result demonstrates that overlapping representations in high-level ventral visual cortex account for aspects of perceived similarity that are unique to specific object dimensions. In contrast, the strongest correlation for color was found in posterior occipital cortex. This indicates that different brain regions underlie the perceived similarity of objects for other stimulus dimensions, suggesting a partly distributed representational scheme. Together, these results reveal a partially overlapping and distributed representational scheme underlying the perceived similarity of objects.

3.2.2. Supra-category membership and a deep neural network (DNN) only partially account for the brain-perceived similarity link

The hypothesis that supra-category membership might be relevant is prompted by the observation that objects belonging to different categories are behaviorally judged to be different, and tend to evoke different brain responses (Rosch et al., 1976; Caramazza and Mahon, 2003; Grill-Spector and Malach, 2004; Op de Beeck et al., 2008a). In particular, Mur et al. (2013) highlighted the role of category membership in accounting for the relation between activation patterns in inferior temporal cortex IT and perceived similarity of freely arranged objects.

The hypothesis that deep neural networks (DNNs) trained on object classification might account for the observed relationship between brain activity and perceived similarity is motivated by the combination of two recent findings. First, DNNs predict visual object-related brain activity better than any previous model class (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and Gerven, 2015; Cichy et al., 2016a). Second, DNNs predict human similarity ratings better than other models, too (Kubilius et al., 2016). If DNNs explain both perceived similarity and a cortical region strongly implicated in underlying perceived similarity, they might also account for the link between the two observed here.

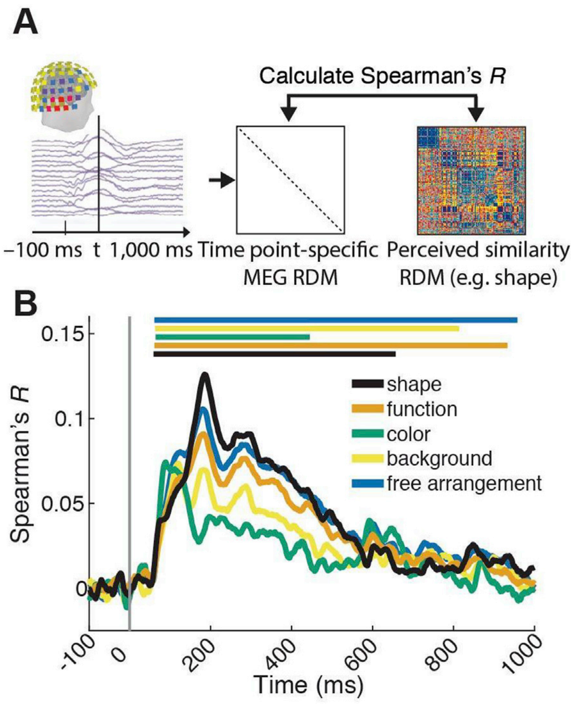

Fig. 5. The temporal evolution of neural representations underlying perceived similarity. A) Procedure. Using a time-resolved scheme, for every millisecond from −100 to +1000 ms with respect to image onset we calculated a RDM from MEG sensor activation patterns. We then compared (Spearman's R) the MEG RDMs with the perceived similarity RDMs (e.g. shape). B) Results. We found a significant relationship between MEG sensor activation patterns and perceived similarity for all dimensions. Significant time points are indicated by lines above result curves ($n = 15$, sign-rank test, cluster-definition threshold $P < 0.05$, cluster size threshold $P < 0.05$, Bonferroni corrected for multiple comparisons (5)). Peak latencies with 95% confidence intervals are listed in Table 1A.

To better understand the link between brain and behavior, we used the partial correlation variant of RSA, partialling out RDMs capturing supra-category membership (Fig. 4A) or DNN RDMs constructed from layer-specific DNN activations to the stimulus set (Fig. 4C). We found that supra-category membership accounted for the correlation between fMRI activation patterns in high-level ventral visual cortex and perceived similarity for all dimensions, with the exception of shape (Fig. 4B). Further, DNN features did account for the correlation with all object dimensions, with the exception of shape (Fig. 4D) and free arrangement (Fig. 4E).

While supra-category membership or features of a DNN might not fully account for the correlation between perceived similarity and neural representations as measured with fMRI, they might do so when combined (Khaligh-Razavi and Kriegeskorte, 2014). A partial-correlation analysis partialling out both factors (Fig. 4F) revealed that they did not fully account for the link between high-level ventral visual cortex and perceived similarity link for the shape dimension (Fig. 6B) either.

Finally, we investigated whether this result also held when considering the correlation between visual representations measured with fMRI and similarity ratings unique to the dimension of object shape. For this, we partialled out the effect of supra-category, DNN features, and all other perceived object dimensions combined (Fig. 4H). The results revealed a significant effect for shape (Fig. 4I) and high-level ventral visual cortex, demonstrating that none of the investigated factors fully accounted for the brain-perceived similarity link.

Together, out results are fourfold: i) they highlight the role of category membership in mediating the relationship between brain responses measured with fMRI and perceived similarity; ii) they show that DNNs account well but incompletely for the link, too; iii) they reinforce the tight link between perceived similarity and high-level ventral visual cortex; and iv) they single out shape as an object dimension whose link to the brain can only partially be explained by the factors investigated here..

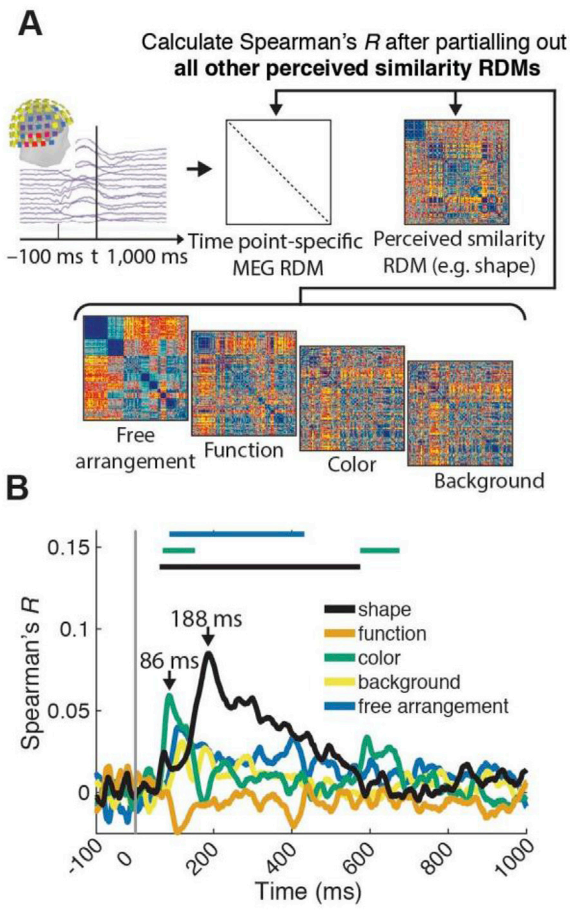

Fig. 6. The temporal evolution of neural representations related uniquely to a particular object dimension. A) Procedure. We adapted the time-resolved RSA procedure to a partial correlation analysis. We compared (Spearman's R) the MEG RDMs with a perceived similarity RDM for a particular object dimension (e.g. shape) while partialling out the RDMs for all other object dimensions. **B**) Results. There was a significant relationship between MEG sensor activation patterns and perceived similarity for shape, color and free arrangement. Significant time points are indicated by lines above result curves ($n = 15$, sign-rank test, cluster-definition threshold $p < 0.05$, cluster size threshold $p < 0.05$ Bonferroni corrected for multiple comparisons (5)). Peak latencies with 95% confidence intervals are listed in Table 1B.

3.3. *Neural representations related to perceived similarity emerge rapidly for dimensions of perceived similarity*

To reveal the temporal emergence of neural representations underlying perceived similarity we recorded MEG data while participants viewed the stimulus set. We then used RSA to relate each dimension of perceived similarity to MEG sensor activation patterns at each time point (Carlson et al., 2013a; Cichy et al., 2014) (Fig. 2A). Note that all following analyses are of analogous rationaly to the fMRI analyses presented before.

We found a significant correlation between MEG RDMs and perceived similarity RDMs for all dimensions (Fig. 5B). The correlations emerged rapidly, peaking earlier than 200 ms for all dimensions (see Table 1A for peak latencies and 95% confidence intervals). This extends previous work revealing the temporal evolution of perceived similarity

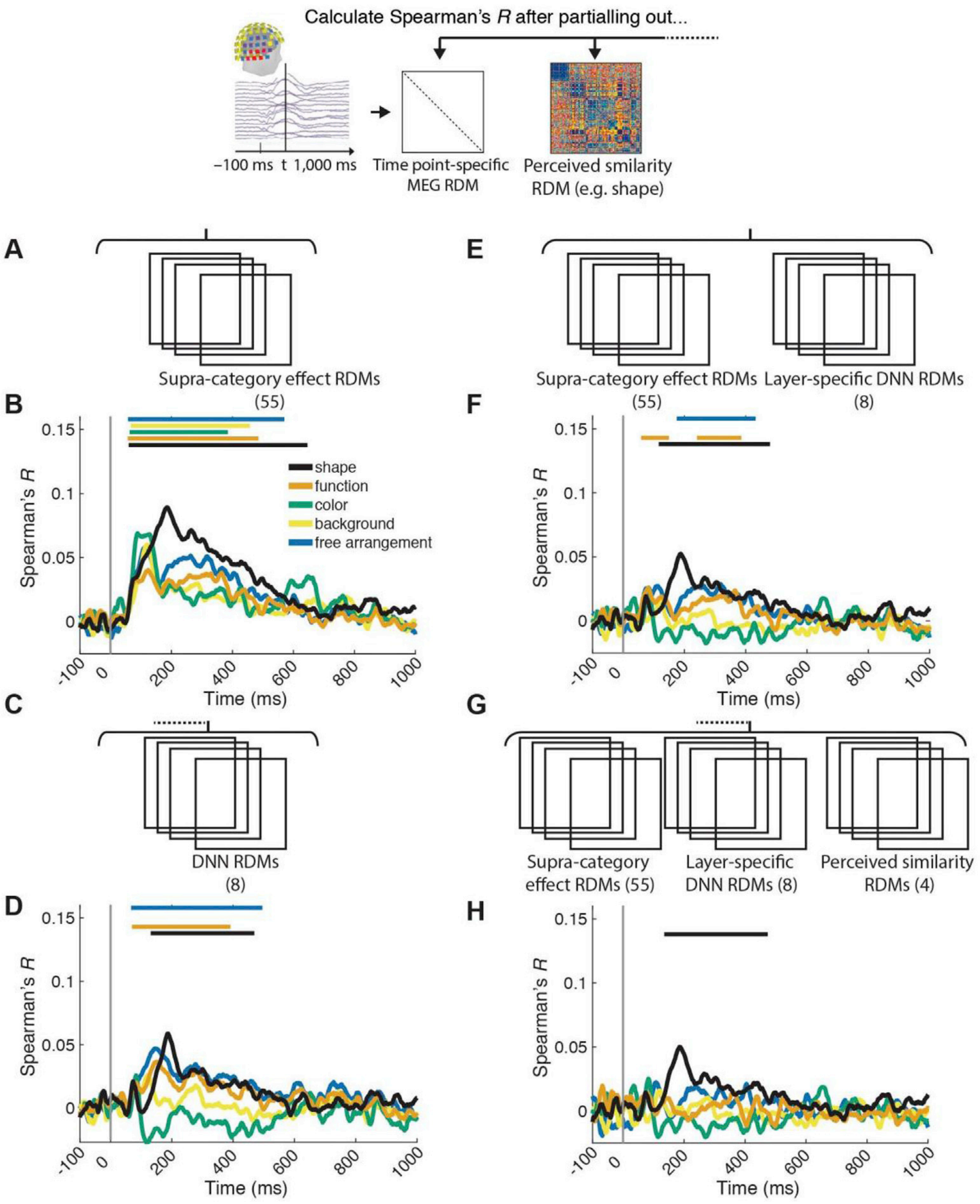

Fig. 7. The role of supra-category membership and ability of DNN features to account for the link between perceived similarity and brain activity measured with MEG. A) Procedure. In the partial correlation framework, we compared (Spearman's R) the fMRI RDMs with each perceived similarity RDM while partialling out RDMs capturing **A)** supra-category level effects (results in **B**), **C)** layer-specific DNN RDMs (results in D), **E)** supra-level category effects and layer-specific DNN RDMs in combination (results in F), as well as **G)** as E and all other perceived similarity RDMs (results in H). Significant time points ($n = 15$, sign-rank test, cluster-definition threshold $p < 0.05$, cluster size threshold $p < 0.05$ Bonferroni corrected for multiple comparisons (5 for B,D; 3 for F,H) are indicated by lines above the result curves. Peak latencies with 95% confidence intervals are listed in Table 1C–F.

representations of synthetic low-level stimuli (Wardle et al., 2016) to complex real-world shapes and a wider set of object dimensions.

3.3.1. *Neural representations uniquely related to a dimension of perceived similarity emerge with distinct dynamics*

To reveal the unique aspects of the relationship between perceived similarity RDMs and visual representations measured with MEG we used a partial correlation analysis (Fig. 6A): correlating MEG RDMs to each perceived similarity RDM, we partialled out all other perceived similarity RDMs. This analysis revealed a relationship between visual representations and object dimensions shape and color, and free arrangement (Fig. 6B).

The analogous fMRI analysis (Fig. 3) suggested that representations in different brain regions account for the relationship between perceived similarity for shape and color: brain regions related to perceived similarity of shape were anterior to those related to color (Fig. 3B and C). If the temporal order with which visual representations emerge is analogous to the order of hierarchical processing stages in the ventral visual system, the relations between perceived similarity and MEG sensor activation patterns should emerge earlier for color than for shape. We tested this prediction by comparing peak latencies (color: 86 ms (82–117); shape: 188 ms (183–202)) and found a significant difference ($n = 15$, $P = 0.0057$, bootstrap test; other peak latencies not significant at $P < 0.05$ Bonferroni corrected; Table 1B for all peak latencies and 95% confidence intervals).

Together, these results unravel the temporal dynamics with which representations uniquely related to particular object dimensions emerge in the human visual system. In particular, they show that these dynamics are different for different object dimensions, with color-related representations emerging before shape-related representations.

3.3.2. *Neither supra-category membership nor a deep neural network (DNN) fully account for the brain - perceived similarity relation*

Analogous to the fMRI-based analysis, we investigated whether supra-category membership or a deep neural network trained on object categorization accounted for the established link between visual representations measured with MEG and perceived similarity. For this we partialled out RDMs capturing supra-category membership (Fig. 7A) or DNN RDMs (Fig. 4C). We found that supra-category membership did not account for the link fully for any object dimension (Fig. 7B). This result contrasts with the fMRI-based analysis (Fig. 4), where supra-level category membership

Table 1

Peak latencies of RSA results relating perceived similarity RDMs with MEG RDMs. RSA relating MEG to A) perceived similarity rating RDMs. **B**) perceived similarity while partialling out the RDMs for all other perceived object dimension RDMs, **C**) perceived similarity ratings partialling out DNN RDMs, **D**) perceived similarity ratings partialling out supra-category RDMs, **E**) perceived similarity ratings partialling out both supra-category RDMs and DNN RDMs, **F**) to perceived similarity ratings partialling out the RDMs for all other perceived object dimensions, DNN RDMs and supra-category RDMs. The numbers are means across participants with 95% confidence intervals (10,000 bootstraps of the participant pool) in parentheses.

Object dimension	Peak (ms)
A) MEG-to-perceived similarity rating RSA	
Background	119 (114–191)
Color	89 (83–130)
Function	183 (168–235)
Free Arrangement	180 (119–188)
Shape	185 (181–191)
B) MEG-to-perceived similarity rating RSA unique to object dimension	
Free arrangement	108 (105–405)
Shape	188 (183–202)
Color	86 (82–117)
C) MEG-to-perceived similarity rating RSA partialling out supra-category RDMs	
Background	120 (114–132)
Color	89 (84–133)
Function	124 (88–365)
Free Arrangement	114 (106–315)
Shape	185 (178–221)
D) MEG-to-perceived similarity rating RSA partialling out DNN RDMs	
Free arrangement	145 (124–278)
Shape	188 (184–199)
Function	150 (131–355)
E) MEG-to-perceived similarity rating RSA partialling out DNN & supra-category RDMs	
Free arrangement	317 (75–364)
Shape	188 (183–300)
Function	83 (74–370)
F) MEG-to-perceived similarity rating RSA unique to particular object dimension partialling out DNN & supra-category RDMs	
Shape	185 (179–300)

accounted for all links between brain activity and perceived similarity except for shape. This result demonstrates the complementary nature of different brain imaging modalities in revealing how neural activity underlies cognitive phenomena such as perceived similarity, and highlights the sensitivity of multivariate analysis on MEG data. Further, DNN features did not fully account for the link for object dimensions shape and function, and in free arrangement (Fig. 7D). This shows that while DNNs are the computational model class best explaining brain activity as measured with MEG during object vision (Cichy et al., 2016a), they do not fully account for the relation between brain activity and perceived similarity along several fundamental object dimensions.

To investigate whether supra-category effects and the DNN together account for the link between visual representations and perceived similarity, we partialled out supra-category membership and DNN RDMs in combination (Fig. 7E). This analysis revealed significant effects for shape, function and in free arrangement (Fig. 7E), similar to the analysis partialling out the DNN alone.

Finally, we determined whether this result also held when considering the link between visual representations measured with MEG and similarity ratings unique to the dimension of object shape. For this we partialling out the effect of supra-category, DNN features and all other perceived object dimensions except the one at hand in combination (Fig. 4F). We found a significant relationship for shape (Fig. 4G).

Together, our results are threefold: i) they show a weaker role of category membership in mediating the link between perceived similarity and visual representations when the latter are measured with MEG rather than fMRI; ii) they show that DNNs account well but not fully for the link; and iii) concurrent with the fMRI-based analysis they single out shape as an object dimension whose link to the brain remains unexplained by the investigated factor here.

4. Discussion

4.1. Summary

In this study, we investigated the relation between perceived similarity of everyday objects and neural representations measured with fMRI and MEG. We found a tight link between perceived similarity and representations in visual cortex and identified the rapid time course with which those representations emerge. By assessing the unique relationship between brain activity patterns and the perceived similarity of the object dimensions shape, function, color, and background, we revealed a partly overlapping and distributed representational scheme: While color-related representations were related to earlier processing stages than shape-related representations considering both fMRI and MEG data, several dimensions were linked to high-level ventral visual cortex. Finally, supra-level category membership and a DNN trained to categorize objects accounted only for part of the observed relationship between brain activity and perceived similarity.

4.2. A distributed and partially overlapping representational scheme underlies perceived similarity of everyday objects

A novel contribution of our study is the discovery of a partially overlapping and distributed representational code for different dimensions of perceived similarity. Neural representations were *overlapping* in space and time, in that i) for object dimensions shape, background and those assessed in free arrangement, representations were co-localized in high-level ventral visual cortex, and ii) for shape and free arrangement exhibited similar temporal dynamics. Neural representations were *distributed*, in that representations of perceived color similarity preceded representations of perceived shape similarity both in the processing hierarchy of the ventral visual pathway and in time.

The current study builds on previous research linking perceived similarity and brain activity, using three innovative methodological components. First, previous studies accessed the role of single dimensions of similarity (e.g. shape) (Kourtzi and Kanwisher, 2001; Op de Beeck et al., 2001, 2008b; Kayaert et al., 2005; Haushofer et al., 2008; Drucker and Aguirre, 2009), or all dimensions together implicitly when assessing free arrangements (Edelman, 1998; Rotshtein et al., 2005; Weber et al., 2009; Connolly et al., 2012; Mur et al., 2013, 2013; Davidesco et al., 2014; Peelen et al., 2014; Bankson et al., 2018). Here, we assessed the unique role of multiple dimensions of similarity explicitly. Second, we extended fMRI analysis from a region-of-interest (Mur et al., 2013) to a searchlight-based approach (Connolly et al., 2012; Groen et al., 2018), allowing for spatially unbiased analysis. Finally, we extended the analysis of the temporal emergence of visual representations underlying perceived similarity from artificial shapes to real-world objects (Wardle et al., 2016).

Our results have implications for our understanding of the general representational scheme of the ventral visual system and its role in visual perception. For one, by showing that partly distributed and overlapping representations underlie our subjective perception of the visual world, our results support the notion that distributed object representations may in fact used by the brain in object perception (Haxby et al., 2001; O'Toole et al., 2005; Cichy et al., 2012), rather than being purely epiphenomenal effects (Reddy and Kanwisher, 2007; Williams et al., 2007; de-Wit et al., 2016). Second, our results support the idea that high-level ventral visual cortex represents objects according to a multitude of properties simultaneously, such as the ones investigated here as well as size (Konkle and Oliva, 2012), material (Hiramatsu et al., 2011), category (Reddy and Kanwisher, 2006), function (Mahon et al., 2007), eccentricity (Hasson et al., 2002) and retinal location (Schwarzlose et al., 2008). All those properties might be organized in overlapping feature maps (Op de Beeck et al., 2008a).

4.3. Visual representations underlying perceived similarity emerge rapidly

A previous study investigated the temporal emergence of representations underlying perceived similarity for synthetics stimuli consisting of arranged Gabor patches using MEG (Wardle et al., 2016). Our results concur with Wardle at al. in that representations underlying perceived similarity emerge rapidly, with peaks below 200ms. Our results go beyond this previous work in two ways. First, we clarified the temporal emergence of representations underlying perceived similarity for real-world object images, thus probing the visual system with stimulus material that is closer to real-world experience. Second, by distinguishing between pertinent object dimensions when assessing perceived similarity, we demonstrated that the timing with which neural representations underlying perceived similarity depends on the object dimension. Together, our results illuminate the rapid timing with which neural representations underlying perceived similarity emerge in the human brain.

4.4. The role of supra-category membership in mediating the link between brain activity and perceived similarity

While our results in general are consistent with the notion that supra-category membership partly mediates the link between perceived similarity and brain activity (Mur et al., 2013), our fMRI and MEG-based results yield a more complex picture. In the fMRI results, supra-category membership had a prominent role in accounting for the brain-perceived similarity link, only leaving the perceived similarity by shape unexplained. In contrast, in the MEG results, accounting for supra-category membership did not abolish the link for any object dimension, suggesting a much more modest mediating role for category membership.

There are several ways in which this divergence in results can be explained. One possibility is that fMRI signals in ventral visual cortex might more strongly reflect category membership than MEG signals originating from this region. Consequently, accounting for category membership might more strongly affect fMRI than MEG measurements. Interestingly, a recent study revealed a discrepancy between MEG and fMRI results consistent with this hypothesis (Proklova et al., 2017). Proklova et al. (2017) reported that when animate and inanimate stimuli are matched perceptually, brain responses to those stimuli in high-level ventral visual cortex measured with fMRI reflected the categorical divide, whereas MEG patterns did not. Another possibility is that MEG and fMRI in this study predominantly reflect different neuronal sources, e.g. different brain regions (Agam et al., 2011). As the fMRI results indicate mainly high level ventral visual cortex as the source of neuronal representations underlying similarity judgements, this would suggest that fMRI might be blind to the source of the signals dominating MEG here. Future studies are required to answer these open questions.

4.5. A deep neural network trained on object categorization does not fully account for the brain-perceived similarity link

DNNs trained on object categorization are currently the best predictors of brain activity (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and Gerven, 2015; Cichy et al., 2016a) and perceived similarity (Kubilius et al., 2016; Peterson et al., 2016). This prompts the question whether such a DNN can explain the link between perceived similarity and brain activity. Our results indicate that it can do so only partly, further highlighting the gap between DNNs and the human brain (Kriegeskorte, 2015; Yamins and DiCarlo, 2016; Kietzmann et al., 2017). Two ways in which perceived similarity judgments could be used to narrow this gap have been proposed. For one, transformations to the DNN representations may increase the fit to perceived similarity (Peterson et al., 2016) and neural representations (Khaligh-Razavi et al., 2017). Another possibility is to improve DNNs as models of human cognition or neural visual representations by directly influencing their training procedure to better approximate perceived similarity by so-called representational transfer learning (McClure and Kriegeskorte, 2016). Perceived similarity judgments of object shape – the dimension for which the brain-perceived similarity relation remained unaccounted for by DNNs – is a particularly fruitful candidate.

4.6. The link between brain activity and perceived similarity as an index of behavioral relevance

Perceived similarity is arguably fundamental to categorization behavior: if two objects look alike, they are likely to be categorized similarly and thus afford similar behavior. (Rosch et al., 1976; Nosofsky, 1984; Tversky and Hemenway, 1984; Shepard, 1987; Ashby and Perrin, 1988). This suggests that neural representations that account for perceived similarity are suitable to be read out by the brain in categorization and thus to guide adaptive behavior.

Our results suggest that this hypothesis might hold across varying contexts and consistent across participants. In our analyses the link between perceived similarity and brain activity was established across different task contexts (target detection in brain imaging and similarity ratings in behavior), this strongly suggests that the identified representations underlie perceived similarity and may thus guide behavior in other task contexts as well (Bracci et al., 2017; Kay and Yeatman, 2017). Furthermore, the identified neural activity likely generalizes across subjects, as participants differed across experiments.

However, as the link between neural activity and behavior via similarity ratings is indirect, future studies that directly compare binary classification tasks (Newsome et al., 1989; Britten et al., 1996; Thorpe et al., 1996; Grill-Spector et al., 2000; VanRullen and Thorpe, 2001; Philiastides and Sajda, 2006; Williams et al., 2007; Ratcliff et al., 2009; Carlson et al., 2013b; Ritchie et al., 2015) with perceived similarity ratings and their respective link to behavior are necessary. Further, as task impacts object representations in occipitotemporal and parietal cortex (Çukur et al., 2013; Harel et al., 2014; Erez and Duncan, 2015; Bracci et al., 2017; Hebart et al., 2018; Nastase et al., 2017; Vaziri-Pashkam and Xu, 2017), an experimental setup in which participants perform perceived similarity judgments during brain imaging might reveal task-specific representations missed here. Future experiments that tackle the complex experimental challenges of such a complex experimental setup (Woolgar et al., 2014; Bankson et al., 2018) are needed.

4.7. Dissociations between factors accounting for perceived similarity and brain patterns

A recent study relating perceived similarity, brain data and other models found a dissociation between the factors function and DNN features in their contribution to perceived similarity and brain representations: while function best explained perceived scene similarity, DNN features best accounted for neural activity in scene-selective cortex (Groen et al., 2018). Our results concur with Groen et al. (2018) in that function accounted for most variance in perceived similarity of objects (Supplementary Fig. 1) but diverge in that the DNN features did not best account for neural activity, but rather shape judgments. How is this divergence to be explained? Note that Groen at el. did not assess perceived similarity of shape, and thus it is an open question whether perceived similarity by shape would explain their data as good as DNN features. Similarly, we assessed function by perceived similarity, whereas Groen et al. (2018) used a function model based on human assigned labels of action. Further research is required to assess the role of factors such as shape and function when operationalized in different ways.

Last, please note that the finding of function accounting for most variance in perceived similarity ratings in our study must interpreted with care, as it is at least in part likely due to the nature of the stimulus set. Each stimulus was from a different category and given that function is strongly related to category (Greene et al., 2016) this might have biased participants to strongly rely on function in their ratings. For other stimulus sets this might not be so: e.g. for a stimulus set consisting of exemplars from only one category the role of function in the free arrangement should be much reduced. Further research controlling for category membership and function are necessary to answer these open questions. One particular interesting approach could be the use of a stimulus set consisting of artificial shapes that does not have an a priori structure in terms of real-world categories or functions.

Acknowledgments

This work was funded by the German Research Foundation (DFG, CI241/1-1) to R.M.C. and by the European Research Council (ERC, 680906) to I.C and N.K. MEG and fMRI data were collected at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research, MIT. Computing resources were provided by the high-performance computing facilities at ZEDAT, Freie Universität Berlin. We thank Martin Hebart and Matthias Guggenmos for comments on the manuscript, and Seyed-Mahdi Khaligh-Razavi for support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.03.031.

References

Agam, Y., Hämäläinen, M.S., Lee, A.K.C., Dyckman, K.A., Friedman, J.S., Isom, M., Makris, N., Manoach, D.S., 2011. Multimodal neuroimaging dissociates hemodynamic and electrophysiological correlates of error processing. Proc. Natl. Acad. Sci. Unit. States Am. 108, 17556–17561.

Ashby, F.G., Perrin, N.A., 1988. Toward a unified theory of similarity and recognition. Psychol. Rev. 95, 124–150.

Bankson, B.B., Hebart, M.N., Groen, I.I.A., 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.

Bracci, S., Daniels, N., Op de Beeck, H., 2017. Task context overrules object- and category-related representational content in the human parietal cortex. Cerebr. Cortex 27, 310–321.

Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S., Movshon, J.A., 1996. A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis. Neurosci. 13, 87–100.

Caramazza, A., Mahon, B.Z., 2003. The organization of conceptual knowledge: the evidence from category-specific semantic deficits. Trends Cognit. Sci. 7, 354–361.

Carlson, T., Tovar, D.A., Alink, A., Kriegeskorte, N., 2013a. Representational dynamics of object vision: the first 1000 ms. J. Vis. 13, 1–19.

Carlson, T.A., Ritchie, J.B., Kriegeskorte, N., Durvasula, S., Ma, J., 2013b. RT for object categorization is predicted by representational distance. J. Cogn. Neurosci. 1–11.

Charest, I., Kievit, R.A., Schmitz, T.W., Deca, D., Kriegeskorte, N., 2014. Unique semantic space in the brain of each beholder predicts perceived similarity. Proc. Natl. Acad. Sci. U. S. A. 111, 14565–14570.

Cichy, R.M., Heinzle, J., Haynes, J.-D., 2012. Imagery and perception share cortical representations of content and location. Cerebr. Cortex 22, 372–380.

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

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

Cichy, R.M., Pantazis, D., Oliva, A., 2016b. Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. Cerebr. Cortex (8), 3563–3579.

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618.

Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L., 2013. Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. 16, 763.

Damasio, A.R., 1990. Category-related recognition defects as a clue to the neural substrates of knowledge. Trends Neurosci. 13, 95–98.

Davidesco, I., Zion-Golumbic, E., Bickel, S., Harel, M., Groppe, D.M., Keller, C.J., Schevon, C.A., McKhann, G.M., Goodman, R.R., Goelman, G., Schroeder, C.E., Mehta, A.D., Malach, R., 2014. Exemplar selectivity reflects perceptual similarities in the human fusiform cortex. Cerebr. Cortex 24, 1879–1893.

de-Wit, L., Alexander, D., Ekroll, V., Wagemans, J., 2016. Is neuroimaging measuring information in the brain? Psychon. Bull. Rev. 23, 1415–1428.

Drucker, D.M., Aguirre, G.K., 2009. Different spatial scales of shape similarity representation in lateral and ventral LOC. Cerebr. Cortex 19, 2269–2280.

Edelman, S., 1998. Representation is representation of similarities. Behav. Brain Sci. 21, 449–467.

Erez, Y., Duncan, J., 2015. Discrimination of visual categories based on behavioral relevance in widespread regions of frontoparietal cortex. J. Neurosci. 35, 12383–12393.

Grill-Spector, K., Kushnir, T., Hendler, T., Malach, R., 2000. The dynamics of object-selective activation correlate with recognition performance in humans. Nat. Neurosci. 3, 837–843.

Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annu. Rev. Neurosci. 27, 649–677.

Greene, M.R., Baldassano, C., Esteva, A., Beck, D.M., Fei-Fei, L., 2016. Visual scenes are categorized by function. J. Exp. Psychol. Gen. 145, 82–94.

Groen II, 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 Tsao DY. eLife 7 e32962.

Güçlü, U., Gerven, MAJ van, 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.

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

Harel, A., Kravitz, D.J., Baker, C.I., 2014. Task context impacts visual object processing differentially across the cortex. Proc. Natl. Acad. Sci. Unit. States Am. 111, 962–971.

Hart Jr., J., Berndt, R.S., Caramazza, A., 1985. Category-specific naming deficit following cerebral infarction. Nature 316, 439–440.

Hasson, U., Levy, I., Behrmann, M., Hendler, T., Malach, R., 2002. Eccentricity bias as an organizing principle for human high-order object areas. Neuron 34, 479–490.

Haushofer, J., Livingstone, M.S., Kanwisher, N., 2008. Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. PLoS Biol. 6, e187.

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

Haynes, J.-D., Rees, G., 2005. Predicting the stream of consciousness from activity in human visual cortex. Curr. Biol. 15, 1301–1307.

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.

Hiramatsu, C., Goda, N., Komatsu, H., 2011. Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. Neuroimage 57, 482–494.

Kay, K.N., Yeatman, J.D., 2017. Bottom-up and top-down computations in word- and face-selective cortex. eLife 6 e22341.

Kayaert, G., Biederman, Irving, Hans, P., de Beeck, Op, Vogels, Rufin, 2005. Tuning for shape dimensions in macaque inferior temporal cortex. Eur. J. Neurosci. 22, 212–224.

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., Kriegeskorte, N., 2017. Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. J. Math. Psychol. 76, 184–197.

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

Kietzmann, T.C., McClure, P., Kriegeskorte, N., 2017. Deep Neural Networks In Computational Neuroscience. bioRxiv:133504.

Konkle, T., Oliva, A., 2012. A real-world size organization of object responses in occipitotemporal cortex. Neuron 74, 1114–1124.

Kourtzi, Z., Kanwisher, N., 2001. Representation of perceived object shape by the human lateral occipital complex. Science 293, 1506–1509.

Kriegeskorte, N., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4.

Kriegeskorte, N., 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu Rev Vis Sci 1, 417–446.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868.

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cognit. Sci. 17, 401–412.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105.

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

Kruskal, J.B., Wish, M., 1978. Multidimensional Scaling. SAGE.

Kubilius, J., Bracci, S., Beeck, HPO de, 2016. Deep neural networks as a computational model for human shape sensitivity. PLoS Comput. Biol. 12 e1004896.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivar. Anal. 88, 365–411.

Mahon, B.Z., Caramazza, A., 2009. Concepts and categories: a cognitive neuropsychological perspective. Annu. Rev. Psychol. 60, 27–51.

Mahon, B.Z., Milleville, S.C., Negri, G.A.L., Rumiati, R.I., Caramazza, A., Martin, A., 2007. Action-related properties shape object representations in the ventral stream. Neuron 55, 507–520.

Martin, A., Wiggs, C.L., Ungerleider, L.G., Haxby, J.V., 1996. Neural correlates of category-specific knowledge. Nature 379, 649–652.

McClure, P., Kriegeskorte, N., 2016. Representational distance learning for deep neural networks. Front. Comput. Neurosci. 10, 131.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P.A., Kriegeskorte, N., 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. Front. Psychol. 4.

Nastase, S.A., Connolly, A.C., Oosterhof, N.N., Halchenko, Y.O., Guntupalli, J.S., Visconti di Oleggio Castello, M., Gors, J., Gobbini, M.I., Haxby, J.V., 2017. Attention selectively reshapes the geometry of distributed semantic representation. Cerebr. Cortex 27, 4277–4291.

Newsome, W.T., Britten, K.H., Movshon, J.A., 1989. Neuronal correlates of a perceptual decision. Nature 341, 52–54.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

Nosofsky, R.M., 1984. Choice, similarity, and the context theory of classification. J. Exp. Psychol. Learn. Mem. Cogn. 10, 104–114.

Op de Beeck, H., Wagemans, J., Vogels, R., 2001. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nat. Neurosci. 4, 1244–1252.

Op de Beeck, H.P., Haushofer, J., Kanwisher, N.G., 2008a. Interpreting fMRI data: maps, modules and dimensions. Nat. Rev. Neurosci. 9, 123–135.

Op de Beeck, H.P., Torfs, K., Wagemans, J., 2008b. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. J. Neurosci. 28, 10111–10123.

O'Toole, A.J., Jiang, F., Abdi, H., Haxby, J.V., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. J. Cogn. Neurosci. 17, 580–590.

Pantazis, D., Nichols, T.E., Baillet, S., Leahy, R.M., 2005. A comparison of random field theory and permutation methods for the statistical analysis of MEG data. Neuroimage 25, 383–394.

Peelen, M.V., He, C., Han, Z., Caramazza, A., Bi, Y., 2014. Nonvisual and visual object shape representations in occipitotemporal cortex: evidence from congenitally blind and sighted adults. J. Neurosci. 34, 163–170.

Peterson, J.C., Abbott, J.T., Griffiths, T.L., 2016. Adapting Deep Network Features to Capture Psychological Representations. arXiv:160802164.

Philiastides, M.G., Sajda, P., 2006. Temporal characterization of the neural correlates of perceptual decision making in the human brain. Cereb Cortex N Y N 16, 509–518, 1991.

Proklova, D., Kaiser, D., Peelen, M., 2017. MEG Sensor Patterns Reflect Perceptual but Not Categorical Similarity of Animate and Inanimate Objects. 238584.

Ratcliff, R., Philiastides, M.G., Sajda, P., 2009. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proc. Natl. Acad. Sci. Unit. States Am. 106, 6539–6544.

Reddy, L., Kanwisher, N., 2006. Coding of visual objects in the ventral stream. Curr. Opin. Neurobiol. 16, 408–414.

Reddy, L., Kanwisher, N., 2007. Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. Curr. Biol. 17, 2067–2072.

Ritchie, J.B., Bracci, S., Op de Beeck, H., 2017. Avoiding illusory effects in representational similarity analysis: what (not) to do with the diagonal. Neuroimage 148, 197–200.

Ritchie, J.B., Tovar, D.A., Carlson, T.A., 2015. Emerging object representations in the visual system predict reaction times for categorization. PLoS Comput. Biol. 11 e1004316.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P., 1976. Basic objects in natural categories. Cogn. Psychol. 8, 382–439.

Rotshtein, P., Henson, R.N.A., Treves, A., Driver, J., Dolan, R.J., 2005. Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nat. Neurosci. 8, 107–113.

Schwarzlose, R.F., Swisher, J.D., Dang, S., Kanwisher, N., 2008. The distribution of category and location information across object-selective regions in human visual cortex. Proc. Natl. Acad. Sci. U. S. A. 105, 4447–4452.

Shepard, R.N., 1980. Multidimensional scaling, tree-fitting, and clustering. Science 210, 390–398.

Shepard, R.N., 1987. Toward a universal law of generalization for psychological science. Science 237, 1317–1323.

Shepard, R.N., Chipman, S., 1970. Second-order isomorphism of internal representations: shapes of states. Cogn. Psychol. 1, 1–17.

Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., Leahy, R.M., 2011. Brainstorm: a user-friendly application for MEG/EEG analysis. Comput. Intell. Neurosci. 1–13, 2011.

Taulu, S., Kajola, M., Simola, J., 2004. Suppression of interference and artifacts by the signal space separation method. Brain Topogr. 16, 269–275.

Taulu, S., Simola, J., 2006. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. 51, 1759.

Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.

Tversky, B., Hemenway, K., 1984. Objects, parts, and categories. J. Exp. Psychol. Gen. 113, 169–193.

VanRullen, R., Thorpe, S.J., 2001. The time course of visual processing: from early perception to decision-making. J. Cogn. Neurosci. 13, 454–461.

Vaziri-Pashkam, M., Xu, Y., 2017. Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. J. Neurosci. 37, 8767–8782.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., Diedrichsen, J., 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage 137, 188–200.

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

Wardle, S.G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., Carlson, T.A., 2016. Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. Neuroimage 132, 59–70.

Warrington, E.K., Shallice, T., 1984. Category specific semantic impairments. Brain J Neurol 107 (Pt 3), 829–854.

Weber, M., Thompson-Schill, S.L., Osherson, D., Haxby, J., Parsons, L., 2009. Predicting judged similarity of natural categories from their neural representations. Neuropsychologia 47, 859–868.

Williams, M.A., Dang, S., Kanwisher, N.G., 2007. Only some spatial patterns of fMRI response are read out in task performance. Nat. Neurosci. 10, 685–686.

Woolgar, A., Golland, P., Bode, S., 2014. Coping with confounds in multivoxel pattern analysis: what should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. Neuroimage 98, 506–512.

Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356–365.

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. Unit. States Am. 111, 8619–8624.