Objects in familiar constellations evoke high-level representations of real-world size

Genevieve Quek[1,2,*], Alexandra Theodorou[2,3], Marius V. Peelen[2]

[1]The MARCS Institute for Brain, Behaviour and Development,
Western Sydney University, Sydney, Australia

[2]Donders Institute for Brain, Cognition and Behaviour,
Radboud University, Nijmegen, The Netherlands

[3]Department of Psychology,
University of California, Davis, CA, USA

*Corresponding author

Conflict of interest statement:

The authors declare no competing financial interests.

Acknowledgements:

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 725970). The authors thank Angeliki Karaiskou and Manon Maarseveen for their assistance with data collection.

Abstract

High-level vision is frequently studied at the level of either individual objects or full scenes. An intermediate level of visual organisation that has received less attention is the “object constellation”, defined here as a familiar configuration of contextually-associated objects (e.g., *plate* + *spoon*). Recent work has shown that information from multiple objects can be integrated to support observers’ high-level understanding of a “scene”. Here we used EEG to test when the visual system integrates information across objects to support representations of real-world size. To this end, we briefly presented masked object constellations consisting of either large (e.g., *chair* + *table*) or small (e.g., *plate* + *spoon*) object silhouettes, while independently varying retinal size. As a control, observers also saw each object silhouette presented in isolation. Behavioural results showed that observers recognized the objects’ real-world size more easily when the silhouettes appeared in pairs than when they appeared in isolation. Representational similarity analysis of EEG data revealed that neural activity patterns captured information about the real-world size of object constellations from ~200 ms after stimulus onset. This representation was stronger for, and specific to, object pairs as compared to single objects, and remained significant after regressing out visual similarity models derived from computational models. These results reveal the neural time course of real-world size extracted from familiar constellations of objects. More generally, they provide evidence for inter-object facilitation of visual processing, leading to a qualitatively different high-level representation of object pairs than single objects.

Introduction

A dominant perspective in cognitive neuroscience holds objects and scenes to be fundamentally distinct entities, whose visual processing is subserved by separate, parallel neural pathways known as object-selective cortex and scene-selective cortex (Epstein & Baker, 2019; Epstein & Kanwisher, 1998; Malach et al., 1995). In the real world, however, objects and scenes are harder to separate: As we move through space, for example, individual items traverse the continuum between objects and scenes, such that where a table viewed from a distance seems to be an *object*, with just a few steps, it comprises something much more like a *scene*, serving as a backdrop for smaller items placed on its surface. Moreover, many common 'scenes' in daily life are meaningfully defined almost entirely by the specific constellations of objects they contain (e.g., a knife and fork either side of a plate, or a sofa facing a television). Our encounters with such multi-object arrangements are so ubiquitous that the visual system appears to be remarkably sensitive to statistical regularities within these groups, with both the likelihood of objects' co-occurrence and their configural arrangement influencing the way we search for and remember familiar objects (Biederman, 1972; Biederman, Glass, & Stacy, 1973; Kaiser, Stein, & Peelen, 2015; Võ, Boettcher, & Draschkow, 2019).

Notably, higher-level statistical regularities between objects also appear to influence object processing itself (Kaiser, Quek, Cichy, & Peelen, 2019). For example, behavioural studies have shown that objects are recognized faster and more accurately when viewed together with one or more contextually associated objects (Auckland, Cave, & Donnelly, 2007; Davenport, 2007). Furthermore, ambiguous objects that appear together with an associated object in a familiar configuration (e.g., a keyboard in front of a monitor) are recognized more accurately than the same objects shown in isolation or in an unfamiliar configuration (Bar & Ullman, 1996; Green & Hummel, 2006). In the present study, we use time-sensitive decoding of electroencephalography (EEG) data to examine the time course of this inter-object facilitation effect.

Previous EEG studies investigating the time course of contextual associations between objects, or between objects and scenes, measured the neural response to semantic violations by contrasting incongruent vs. congruent associations. This contrast gives rise to reliable event-related potentials (ERPs; Ganis & Kutas, 2003; Mudrik, Lamy, & Deouell, 2010; Quek & Peelen, 2020; Võ & Wolfe, 2013), including the domain-general N400 over centro-parietal electrodes (Kutas & Federmeier, 2011). Notably, these studies used clearly visible objects, as the interest was in measuring when neural signals reflected the violation. In contrast to these studies, here we were interested in how the processing of *ambiguous* objects is facilitated by object context.

A similar question has been addressed for objects disambiguated by global scene context (Brandman & Peelen, 2017; Wischnewski & Peelen, 2021). In those studies, an object was degraded such that it was difficult to recognize when viewed in isolation, but readily identifiable in the context of its original scene. Multivariate decoding analysis showed that the visual representation of the object's category (animate or inanimate) was facilitated by scene context from around 300 ms after stimulus onset. Subsequent fMRI and TMS studies revealed that this facilitation resulted from an interaction between processing in scene-selective and object-selective cortex (Brandman & Peelen, 2017; Wischnewski & Peelen, 2021). One possibility is that inter-object facilitation follows the same time course as scene-object facilitation, for example because both scene-object and object-object facilitation result from the same feedback mechanism based on the activation of a common conceptual "schema" (Bar, 2004). Alternatively, however, we may expect inter-object facilitation to be faster than scene-object facilitation, as interactions are between representations within the same object processing pathway rather than between separate object and scene processing pathways. Furthermore, fMRI studies have provided evidence for integrative responses to object pairs in anterior parts of object-selective cortex when viewed in familiar spatial relations (Baldassano, Beck, & Fei-Fei, 2017; Kaiser & Peelen, 2018), providing another mechanism for inter-object facilitation (Peelen, Berlot, & de Lange, 2023).

To reveal the time course of inter-object contextual facilitation, we took advantage of an object property that is rapidly and automatically retrieved following the recognition of objects – namely, real-world size (Hagen, Zhao, Moonen, Ulken, & Peelen, 2023; Konkle & Oliva, 2012a). Since real-world size is a large-scale organising principle within the visual system (Coutanche & Koch, 2018; He et al., 2013; Khaligh-Razavi, Cichy, Pantazis, & Oliva, 2018; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b; Long, Yu, & Konkle, 2018; Wang, Janini, & Konkle, 2022) that is readily decoded using EEG or MEG (Cichy, Khosla, Pantazis, & Oliva, 2017; Khaligh-Razavi et al., 2018; Wang et al., 2022), tracking the emergence of real-world size representations is an ideal proxy through which to examine how an object's processing is facilitated by its context.

To manipulate real-world size, we constructed pairs of contextually associated objects with small or large canonical real-world size (e.g., cup + saucer = constellation of small objects encountered in a small-scale scene; desk + chair = constellation of large objects encountered in a large-scale navigable scene). To increase ambiguity and reduce the likelihood that observers could access the real-world size of these object pairs by recognising only one of the component objects, we obscured object identity by presenting the objects in silhouette against phase-scrambled noise. These sparse visual displays have the advantage of eliminating many of the low- and mid-level featural cues for spatial scale contained in naturalistic scenes (; (e.g., texture, contour junctions, openness, perspective; Greene & Oliva, 2009, 2010), such that the real-world size of these stimuli should be primarily accessible via successful recognition of the object identities (Hagen et al., 2023). Importantly, if object context facilitates this recognition (Bar, 2004; Bar & Ullman, 1996; Biederman et al., 1973; Biederman, Mezzanotte, & Rabinowitz, 1982), real-world size should be inferred more effectively when the objects appear in their contextually-associated pairs than when the same objects are shown in isolation. The comparison between object pairs and the same objects presented individually allows us to measure how observers infer real-world size from object constellations while controlling for any residual visual feature differences between large and small scales.

In Experiment 1, online observers (n = 75) made explicit behavioural judgements about the real-world size of objects as presented in contextually-associated pairs and in isolation. In Experiment 2, we recorded 64-channel EEG in a non-overlapping group of observers (n = 41) as they completed an orthogonal 1-back task on those same object pair and single object stimuli. We examined how well real-world size accounted for the evolving representational structure of the neural response (Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008), while taking various other visual control models into account. To anticipate our results, we found that observers judged the real-world size of objects much more accurately when the objects appeared in contextually-associated pairs than when the constituent objects appeared in isolation. EEG decoding revealed the putative time course of this effect: Real-world size was represented uniquely and more strongly for object pairs than single objects by 200 ms of visual processing time.

Methods

All experimental procedures were carried out with the approval of the Radboud University Faculty of Social Sciences Ethics Committee (ECSW2017–2306-517).

Participants

For Experiment 1 (behavioural), we recruited 75 online participants (40 female, 33 male, 2 not answered) aged between 18-35 years (M = 26.31, SD = 4.89) via Prolific (www.prolific.co). All were living in the Netherlands, had normal or corrected-to-normal vision, and had prior experience in online research participation (>10 studies). For Experiment 2 (EEG), we recruited 41 English-speaking individuals (11 males) aged between 18 and 35 years (M = 23.46, SD = 3.88) from the local region of Nijmegen, the Netherlands. All had normal or corrected-to-normal vision, with no history of neurological illness/concussion/brain surgery/migraine. We discarded EEG data from one participant who terminated the experiment early due to discomfort, and from 6 participants with very high noise/muscle artefact in the signal, for a final sample n = 34 in Experiment 2. This sample size provides 80% power to detect a medium-sized effect (Cohen's d = 0.5) with α=0.05. Individuals in both experiments were monetarily compensated for their participation.

Stimuli

Stimuli were identical across the two experiments and were based on 20 pairs of contextually-associated objects, each denoting either a small- or large-scale object constellation (Figure 1A). 'Small' displays contained pairs of desktop-sized objects that could fit into a shoebox, typically with grasp/reach affordances (e.g., bottle + glass). 'Large' displays contained pairs of furniture-sized objects, typically with navigational affordances (e.g., table + chair). All objects appeared in silhouette, embedded in a noise mask, with no other meaningful visual information (e.g., background). These sparse visual displays served to eliminate many of the inherent low- and mid-level visual differences between naturalistic scenes of different spatial scales.

Inspired by previous work showing that relative spatial positioning modulates the extent to which observers group/integrate concurrently-presented objects (Bar & Ullman, 1996; Kaiser et al., 2019), we varied the relative sizing of the objects within each pair as depicted in Figure 1B. We began with the stimuli shown in

Figure 1A, wherein the proportional relative sizing of the object silhouettes within each pair implied them to be at the same distance or depth. We then created an identical 'depth-match' version of the pair at a decreased retinal size (2:1 size decrease), and recombined the elements of these two versions to create two different 'depth-mismatch' pairs in which the objects' disproportionate retinal sizes (e.g., large pillow + small bedframe) implied they were at different distances from the viewer. To ensure that the pixel distance between the two objects in a depth-mismatch pair was equivalent to the pixel distance between their depth-match counterparts, we randomly labelled the objects in each pair as object 1 and object 2. We then utilized the smaller distance whenever object 1 was the small object in the depth-mismatch pair and the large distance when it was the larger object. We maintained the surface alignment of the two-object scene as much as possible across versions, e.g., since desk and chair both rest on the floor, we aligned these items at their bottom edge regardless of whether their depths were matched or mismatched.

In addition to the four versions of each object pair, observers also saw the small and large retinal size versions of every silhouetted object in isolation (Figure 1C). The position of each isolated object at a given retinal size was always halfway between where that object had appeared as part of the depth-match and depth-mismatch pair. This made for a total of 80 object pairs and 80 single objects (10 exemplars x 4 versions representing each level of real-world size). There were 248 phase scrambled noise images, two of which were randomly selected to serve as masks on each trial.

Trial structure was near identical across Experiments 1 and 2 (Figure 1D): The trial began with a central fixation cross (duration of 500 ms in Experiment 1; between 200-500 ms in Experiment 2). Following the fixation period, a phase-scrambled noise mask appeared for 400 ms, then the silhouette target appeared at 40% contrast embedded within the mask for 100 ms. The trial concluded with a different phase-scrambled backward mask that lasted 300 ms. In both experiments, the trial sequence appeared against a uniform grey background.

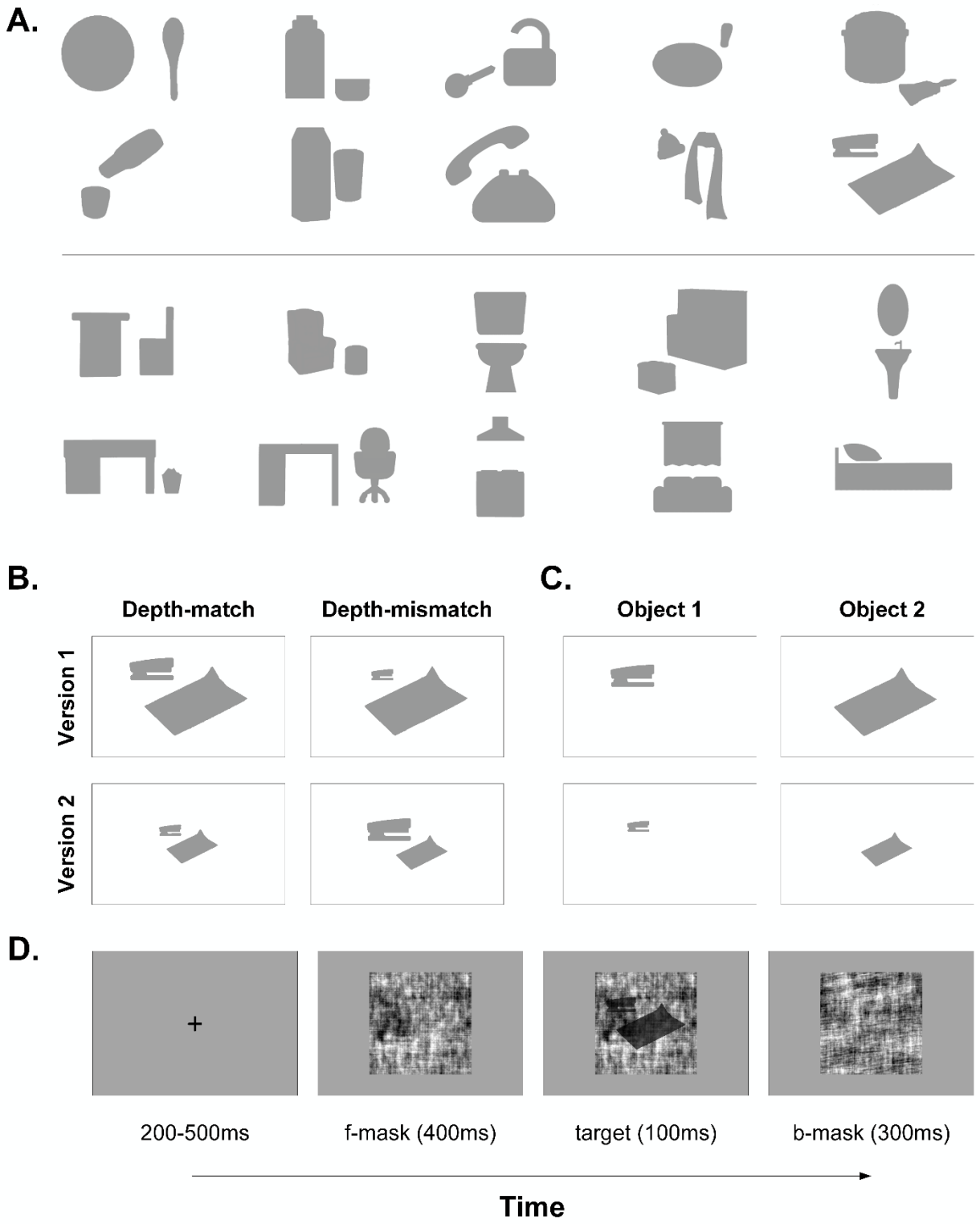

Figure 1. Stimuli used in Experiments 1 and 2. **A)** The 20 base object pairs representing small (top row) and large (bottom row) real-world size. **B)** Each object pair had four versions: A large and small retinal size in which the objects' relative sizing implied they were at the same depth (Depth-match). Recombining the elements of these versions produced two pairs where the objects were implied to be at different depths (Depth-mismatch). **C)** Participants also saw each individual object in isolation. **D)** Basic trial structure for both experiments. Following fixation, the target appeared at 40% contrast embedded in a phase-scrambled noise mask, and was immediately followed by a different phase-scrambled mask. In Experiment 1, the fixation period was fixed at 500 ms.

Procedure

Experiment 1 (behavioural) was programmed in PsychoPy (jsPsych) and hosted online on Pavlovia (Open Science Tools Limited, 2021). Participants completed the experiment in a browser window on either a desktop or laptop computer. Before beginning, we instructed them to minimise any distractions in the environment (e.g., music, TV, other people), and to disconnect any external monitors to ensure the experiment appeared on their main screen. On each trial (see Figure 1D), participants had 1000 ms from target onset to categorise the displayed object(s) as either large or small in real life by pressing either the *f* (small) or *j* (large) key. Response key labels appeared onscreen throughout the trial. We specifically instructed participants to consider the size of the objects in the real world (not the size at which they appeared onscreen), and to maintain fixation on the central cross when it appeared between trials. Prior to commencing the experimental trials, observers saw three example object pairs representing each real-world size category, and completed 16 practice trials in which the target duration was set at 500 ms to facilitate learning. These training trials used an independent stimulus set to the main experiment. In the experiment proper, 80 object pairs and 80 single objects appeared in random, intermingled order for a total of 160 trials. Mask stimuli were sized at 560x560 pixels; we could not fully control the visual angle subtended by the stimuli owing to conducting the study online. Total testing time for Experiment 1 was approximately 10 minutes including a self-paced break at the halfway point.

Experiment 2 (EEG) was conducted in the lab. The trial sequence was programmed in Presentation® (Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com), and displayed on a BenQ XL2420Z computer monitor (120Hz refresh rate; 1920x1080 resolution). Mask stimuli were sized 14x14cm, with viewing distance set at 65cm such that each stimulus subtended 12.3 x 12.3 degrees of visual angle. Experiment 2 saw a 10-fold increase in the number of trials: each of the 160 unique stimuli appeared once per block, with 10 blocks in the full experiment. In contrast to Experiment 1, here participants were engaged in a 1-back task, monitoring the target from trial to trial and pressing the spacebar whenever they saw an exact image repeat (10 instances per block). Prior to the main experiment, observers performed 25 practice trials with a slightly longer target duration (300 ms); this included 5 repeat instances to enable them to learn the task. The full experimental session including EEG set up lasted around 2.5 hours.

We recorded scalp EEG in Experiment 2 using a 64-channel BrainProducts actiCAP active electrode system with a sampling rate of 500Hz (Brain Products GmbH, Gilching, Germany). We used customized electrode positions adapted from the actiCAP 64Ch Standard-2 system (ground placed at AFz; TP10 placed on right mastoid). Impedances for individual scalp channels were held <40 kOhm, with data referenced online to the left mastoid and filtered between 0.016 and 125 Hz using BrainVision Recorder (Brain Products GmbH, Gilching, Germany). We monitored participants' eye movements via external passive electrodes situated at the outer canthus of each eye, as well as immediately above and below the right eye. The ground electrode for passive ocular channels was placed on the tip of the nose. The experimenter visually monitored the EEG trace throughout the experiment, initiating each experimental block manually after ensuring there were no high amplitude deflections resulting from ocular/muscular artefacts. Stimulus onsets were marked in the EEG file using numeric triggers.

Analysis

Experiment 1

For the behavioural real-world size categorisation task, we removed those participants who failed to provide a response on more than 30% of trials (n = 5), and subsequently any participant whose mean accuracy or response time (RT) on correct trials fell more than 2 standard deviations below the group mean (n = 6). We further discarded any trials on which the participant responded sooner than 200 ms following target onset (0.62% of trials), under the rationale that these early responses were unlikely to be informative. We then inspected *i)* response accuracy (treating trials with no response as incorrect responses) and *ii)* RT on correct trials. For the latter, we bounded each participant's RT distribution at 2 standard deviations above and below the mean by replacing values outside this range with the top and bottom RTs respectively (4.6% of trials). For each metric, we ran a 2 x 2 repeated measures ANOVA with the factors *Stimulus Type* (Pair, Single) and *Retinal Size* (Large, Small). Note that for the depth-mismatch pair stimuli, the 'large' and 'small' retinal size labels were simply arbitrary labels, since each depth-mismatch pair necessarily comprised one large retinal size object and one small retinal size object (see Figure 1B). A subsequent analysis inspected the effect of *Relative Depth* (match, mismatch) for the Object Pairs data alone (since there was no relative depth manipulation for single objects).

Experiment 2

EEG preprocessing

We preprocessed the EEG data in MATLAB 2016b using FieldTrip functions (Oostenveld, Fries, Maris & Schoffelen, 2011; http://fieldtriptoolbox.org). This included a bandpass filter between 0.05-100Hz, a line noise filter at 50Hz, 100Hz, & 150Hz, and downsampling the data to 250Hz for easier storage and handling. A maximum of three artefact ridden channels were identified by eye and interpolated using the weighted average of neighbouring electrodes. To combat artefacts introduced by eyeblinks, we relied on an independent components analysis approach (runica decomposition), visually inspecting the components and associated topographies for each participant to identify and remove the eye blink component. No trial or epoch rejection was applied. We then re-referenced each participant's cleaned EEG trace to the average of all channels before segmenting an epoch from -200 ms to 500 ms around the onset of each target stimulus (excluding 1-back repeat instances). The resulting 800 single object epochs and 800 object pair epochs were then baseline-corrected using the 200 ms prior to stimulus onset.

Representational Similarity Analysis (RSA)

RSA using individual versions of each object pair

For the object pair neural data, we used RSA (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008) to build a pairwise representational dissimilarity matrix (RDM) at each timepoint (4 ms), for each participant separately. This involved pairwise decoding of every version of every object pair exemplar (i.e., 80 items) using Linear Discriminant Analysis (LDA) with 10 fold cross-validation as implemented in the CoSMoMVPA toolbox (Oosterhof, Connolly, & Haxby, 2016; https://www.cosmomvpa.org/). We used multiple regression to relate the resulting time-varying RDM to four categorical predictors corresponding to the binary models given by our experimental design (see Figure 3B). These were Pair Identity (i.e., the specific real-world context the objects were drawn from), Retinal Size (i.e., large or small version), Real-World Size (i.e., small- or large-scale objects), and Relative Depth (i.e., implied depth of the objects either matched or mismatched within the pair). Here and in all following analyses, we inspected the utility of each model as a predictor of the neural RDM at each timepoint between 0 and 500 ms via a Bayesian one sample *t*-test that examined whether the participant mean betas were greater than zero.

RSA using object pairs collapsed across version

Next, we ran a simplified RSA that pooled the neural data for the four versions of each object pair (see Figure 1B) for pairwise decoding. This effectively collapsed across all design factors other than Real-World Size, which remained the primary model of interest. Here we aimed to inspect the predictive power of the real-world size model while controlling for the inherent low-level differences between large and small object pairs. To this end, we built four low-level visual control RDMs (see details below), each sized 80x80 (i.e., every version of every object pair treated individually). We then collapsed these into the reduced 20×20 RDMs by averaging values over the four versions of each object pair, e.g., collapsed dissimilarity value for Pair 1 vs. Pair 2 was the average of all RDM intersections of [1A, 1B, 1C, 1D] and [2A, 2B, 2C, 2D]. The visual control models we considered were: *1)* An image similarity model reflecting 1 minus the pixel-wise correlation between each of the 80 object pair images; *2)* A silhouette overlap model (Jaccard, 1901) comprising 1 minus the Jaccard similarity (ranges between 0 and 1) for each pairwise combination of the 80 object pair images; 3) A computational model of low-level visual processes, obtained by applying HMAX (Riesenhuber & Poggio, 1999; Serre et al., 2007; https://maxlab.neuro.georgetown.edu/hmax.html) to each of the 80 object pair images embedded in an example noise mask (just as they appeared in both experiments). RDM values were 1 minus the Spearman correlation between the vectorised responses on the final HMAX C2 layer; and 4) A mid-level feature model comprising the pairwise Euclidean distances between curvature indices for object pairs obtained using a computational model developed by Li and Bonner (2020; https://github.com/shipui2005/Curvature-Model). In this and the following multiple regressions, we inspected the Spearman rank-order correlations between all model RDMs, applying a significance criterion of p <.01, FDR corrected. Since the Pixel Correlation and Jaccard Similarity models were highly correlated (rho = 0.9844, p <.0001) and displayed large Variance Inflation Factors (pix = 31.09, jac = 39.91), we elected to drop the Jaccard model from our multiple regression, leaving 1 categorical predictor (Real-world Size) and 3 continuous control predictors (pixel correlation, HMAX C2, and curvature, see Figure 4)[1]. Since the intercorrelations between these predictors were not especially high (only Pixel Correlation and Curvature were significantly correlated, rho = 0.2587, p = .002), and since VIFs were all close to 1 (RwS = 1.0046, pix =

[1] We confirmed the individual utility of each control RDM as a predictor of the neural RDM series for Object Pairs by running a separate, exclusive regression for each low-level model (see Figure S2 in supplemental material).

1.0645, hmx = 1.0262, cur = 1.0884), we felt justified in including these models as independent predictors in the same multiple regression.

Comparisons with Single Objects
To enable a comparison with real-world size representations evoked by our object silhouettes viewed in isolation (i.e., one at a time, rather than in pairs creating sparse scenes), we duplicated the above collapsed RSA pipeline for the Single Object neural data, producing corresponding real-world size and low-level visual control models for the single object stimuli. For the singles objects, we found significant correlations between the real-world size and curvature models (rho = 0.2516, p = .0028) and the curvature and pixel correlation models (rho = 0.2276, p = .0049). Importantly, variance inflation factors for the predictors were all close to 1 (RwS = 1.0884, pix = 1.0885, hmx = 1.0251, cur = 1.1608), such that we felt confident interpreting the beta estimates from a multiple regression that included these models as independent variables.

Comparisons with a Deep Convolutional Neural Network
To further rule out the possibility that real-world size representations evoked by the object pair stimuli arose in response to low-level featural differences between large and small objects, we examined representations of these sparse scenes in the layers of a deep convolutional neural network trained to label objects using 1.2 million naturalistic images (AlexNet, Krizhevsky, Sutskever, & Hinton, 2012). We obtained output activations corresponding to the 80 object pair images embedded in an example noise mask (i.e., as they appeared to human observers in the two experiments) on each of the 8 AlexNet layers (input layer, five convolutional layers, two fully connected layers), each of which was used to build an RDM (see Figure S4 in Supplemental Materials). Just as for the low-level visual control models, we collapsed the resulting 80×80 RDMs down to 20×20 by averaging values over the four versions of each object pair. Using the same significance criterion as detailed above, we inspected the intercorrelation matrix for these eight AlexNet RDMs and the model of Real-world Size. If featural differences between our large/small stimuli account for the predictive power of real-world size, then we would expect strong correlations between this model and the earlier layers of AlexNet.

Lastly, given that the upper layers of convolutional neural networks capture some important organising principles of object vision (e.g., animacy, Khaligh-Razavi & Kriegeskorte, 2014), we also asked whether the information about real-world size in our neural data could be accounted for by the representational structure evident in these layers of AlexNet (i.e., FC6 & FC7). To this end, we subjected the (reduced) neural RDM series for object pairs to separate multiple regressions that included Real-world Size as a categorical predictor alongside i) the FC6 RDM, and 2) the FC7 RDM (treated as continuous predictors). Although the Real-world Size model was significantly correlated with both AlexNet layer models (FC6: rho = 0.20, p = 0.0079; FC7: rho = 0.24, p = 0.0016, FDR-corrected), Variance Inflation Factors (VIF) for these regressors were close to zero in both cases (RwS & FC6 = 1.0378; RwS & FC7 = 1.0391). As such, we felt justified in treating them as independent predictors in these multiple regressions.

Results

Experiment 1: Behavioural judgements of real-world object size

A 2x2 repeated measures ANOVA for conditional mean accuracies revealed that participants perceived the real-world size of the object silhouettes significantly better when the items appeared in contextually-related pairs that formed sparse scenes (M = 77.3% correct) rather than in isolation (M = 65.32% correct; main effect of *Stimulus Type*, $F(1,63) = 194.72$, p <.00001, ges = 0.512, see Figure 2A). This was qualified by a significant interaction with *Retinal Size*, $F(1,63) = 17.40$, p <.0001, ges = .084, wherein the magnitude of the Pair > Single difference was larger for the small retinal size versions (M_{diff} = .155, SD = .096) than the large retinal size versions (M_{diff} = .084, SD = .098). The main effect of *Retinal Size* was not significant, $F(1,63)$ = 0.25, p = .619, ges = .001. For RT (Figure 2B), we observed a similar main effect of *Stimulus Type*, $F(1,63)$ = 8.16, p <.01, ges = .041. On average, participants were slower to categorise the real-world size of single objects (M = 595 ms, SD = 139 ms) compared to object pairs (M = 590 ms, SD = 136 ms). Neither the main effect of *Retinal Size,* $F(1,63)$ = 3.49, p = .467, ges = .003, nor the two way interaction, $F(1,63)$ = 0.53, p = .066, ges = .016, reached significance in the case of RT.

Contrary to our predictions, there was no evidence in accuracy or RT that participants' real-world size judgements were affected by the relative size of objects within a pair (Figure 2). Paired t-tests of both

accuracy, $t(63) = 0.413$, $p = .681$, and RT, $t(63) = -0.62$, $p = .539$, revealed no difference between responses to depth-match and depth-mismatch object pairs. Note that there was no such comparison for single objects, as relative depth does not apply to isolated objects.

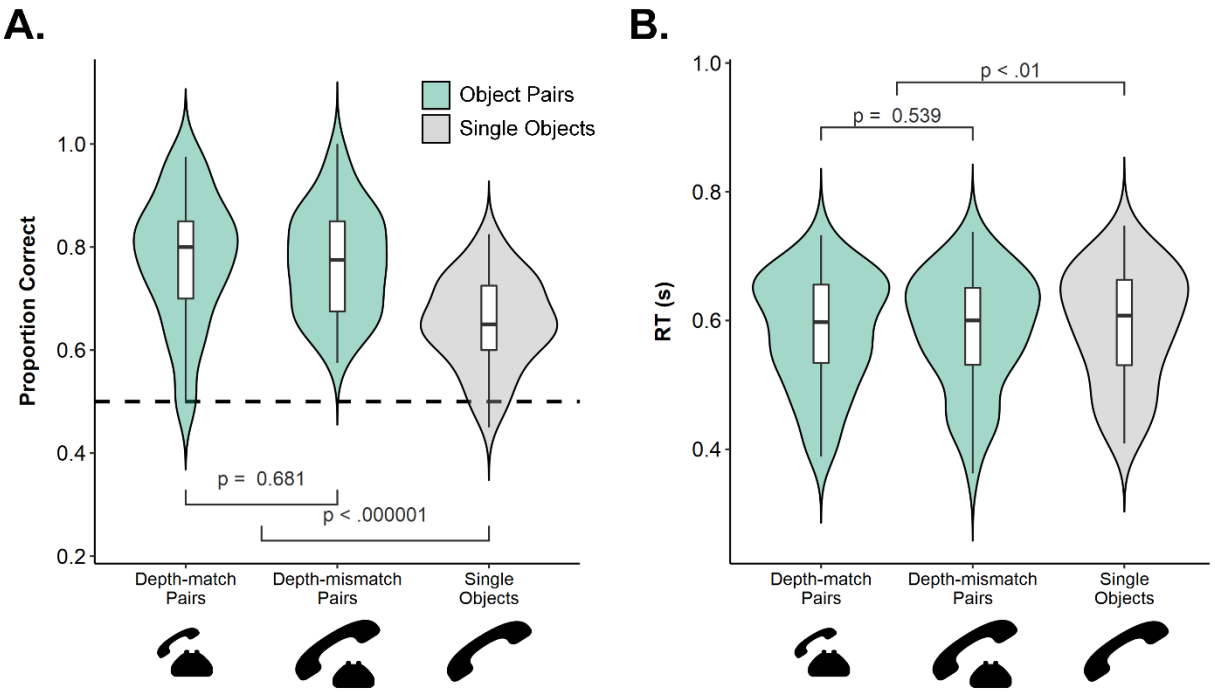

Figure 2. Behavioural judgements of real-world size in Experiment 1. Overlaid violin and box-and-whiskers plots of A) accuracy and B) RTs. Single Object data (in grey) appear at right, with Object Pair data (in green) split into depth-match and depth-mismatch versions (there was no significant difference between these conditions in either metric). Upper and lower box edges correspond to the first and third quartiles (capturing the interquartile range, IQF); whiskers encompass values less than 1.5*IQF from the box edges. Example stimuli for each condition appear below the x-axis labels.

Experiment 2: Neural representations of real-world object size

RSA for the individual versions of each object pair

In Experiment 2, we used RSA applied to EEG data to examine the temporal dynamics of real-world size representations evoked by our silhouetted object pairs. We began by inspecting how well the binary models based on our design factors (Figure 3B) could predict the neural RDM series for object pairs (Figure 3A). In this analysis, we treated the four versions of each object pair (e.g., stapler + paper) independently, building an 80×80 RDM for each timepoint (see Methods for details). Multiple regression beta estimates over time for each of the four predictor models are given in Figure 3C, together with Bayesian one-sample t-tests evaluating beta > 0.

Two of these models predicted the neural RDM from a relatively early stage of stimulus processing: Retinal Size (onset = 84 ms, peak = 152 ms) and Pair Identity (onset = 88 ms, peak = 176 ms). This was expected, since both models capture visual features inherent to the stimuli such as pixel coverage and global shape. Notably, information about both the object identities and their display size was also present in the neural response at much later timepoints, suggesting that the visual system maintains representations of both of these dimensions for some 500 ms after the objects come into view.

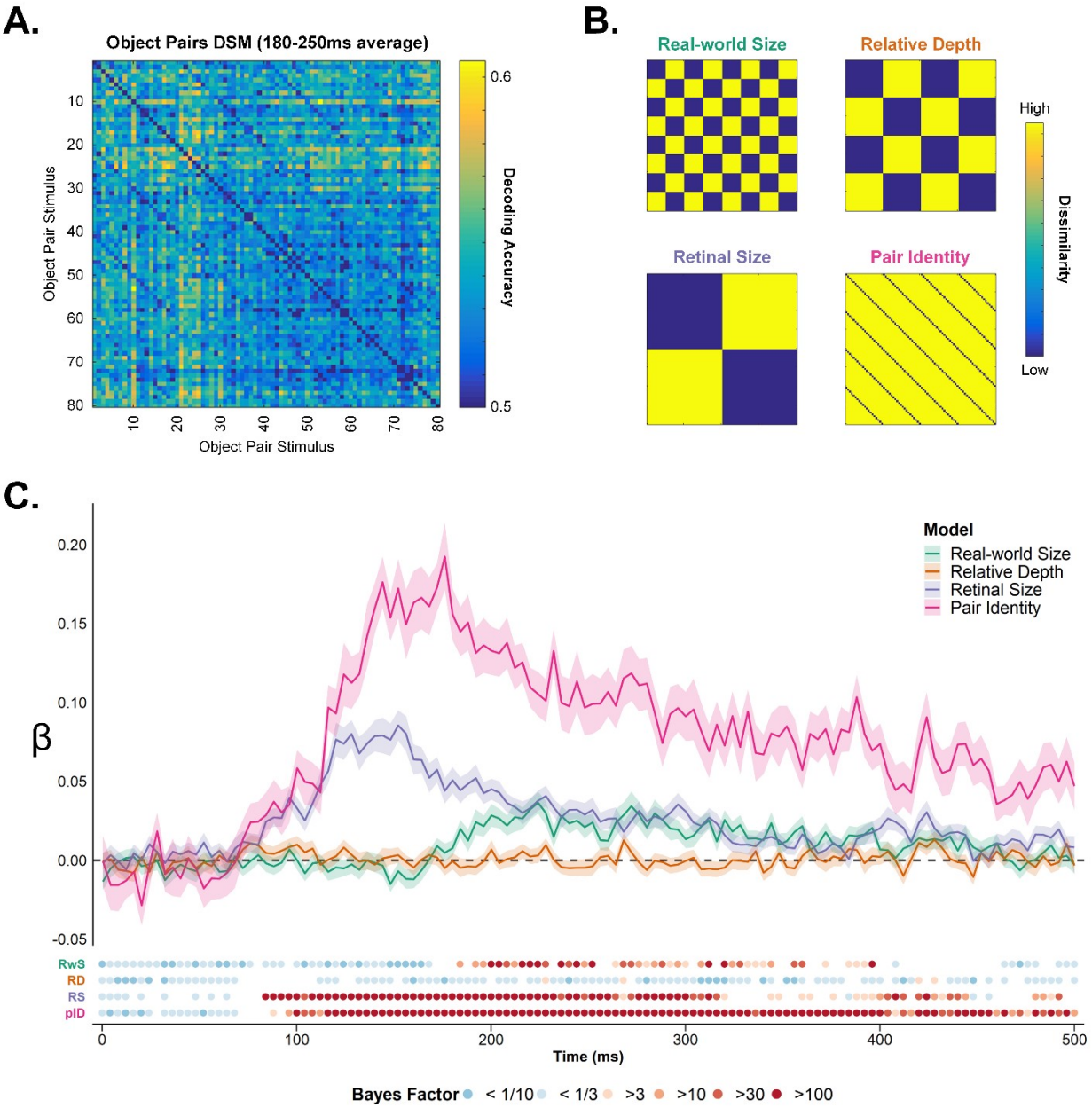

Figure 3. RSA for 80 individual object pair stimuli in Experiment 2 (i.e., treating each of the 4 versions for the 20 object pairs separately). A) The neural RDM averaged over the epoch between 180-250 ms. B) Binary RDMs for the stimulus design factors treated as categorical predictors in a multiple regression of the neural RDM series. C) Beta estimates for the four predictors included in the multiple regression, plotted as a function of time from stimulus onset. Shaded regions are within-subjects standard error. Lower panel shows colour coded Bayes Factors for Bayesian one-sample *t*-tests comparing beta against zero at each timepoint, for each predictor. See Figure S1 in Supplementary Materials for each model's beta-timecourse when used as an isolated predictor of the neural RDM series.

In contrast to the Pair Identity and Retinal Size models, the Real-world Size model requires similar neural responses across stimuli with both distinct object identities and visual features (i.e., the response to *hat+scarf* has to resemble that of *stapler+paper*, while also generalising across large and small versions of those pairs). In this way, we aimed to track real-world size representations reflective of abstract information about the objects' spatial scale. This 'abstract' size model emerged as a significant predictor of the neural response at a comparatively later stage of stimulus processing (onset = 184 ms, peak = 224 ms), and remained a relevant predictor until as late as 400 ms.

The other high-level model yielded by our design was Relative Depth, which captured whether the objects within a pair were implied to be at the same or different depth from the viewer. As is clear in Figure 3C, this model was not a significant predictor of the neural data at any timepoint, suggesting that the objects' relative sizing in each display did not modulate the neural response. Relative Depth was also not a relevant predictor of the object pair neural data when considered as an isolated categorical predictor in a separate regression analysis (see Figure S1 in Supplementary Materials). This was consistent with behavioural observations from Experiment 1, where we found no evidence that the relative depth of objects within a pair influenced participants' real-world size judgements.

RSA for object pairs collapsed across version

Since the relative depth of objects within a pair modulated neither the neural response to our object stimuli nor participants' overt categorisations of real-world size (Experiment 1), we next ran a simplified RSA in which the depth-match and depth-mismatch versions of every object pair were treated as a group. Collapsing across version in this way had the effect of eliminating the Relative Depth, Retinal Size, and Pair Identity models (the latter becomes the RDM diagonal), leaving only the Real-world Size model. We evaluated the utility of this categorical predictor in a multiple regression analysis alongside three low-level visual control models included as continuous predictors (see Methods). Our rationale here was to isolate real-world size representations that could not be explained by inherent differences in visual features between the object-constellations representing small and large spatial scales.

Figure 4 shows the results of this multiple regression analysis collapsing over the four versions of each object pair. Real-world Size did not correlate with any of the three low-level visual control models (**pix:** rho = -0.04, *p* = .608; **hmx:** rho = -0.05, p = .608; **cur:** rho = 0.05, p = .608, FDR-corrected). Bayesian one-sample *t*-tests indicated that the model which predicted the neural response earliest in time was Pixel Correlation (onset = 104 ms, peak = 156 ms), followed by the HMAX (C2) (onset = 148 ms, peak = 176 ms). These early onsets of predictive power were consistent with these models capturing perceptual features of the images, which should be relevant to the feed-forward sweep of visual processing. Interestingly, the Curvature model had no unique predictive power in this analysis, despite being a meaningful predictor when treated as the sole factor in a regression of the neural RDM series analysis (see Figure S2 in Supplementary Materials).

Our primary interest in this analysis was to determine whether the Real-world Size model would offer any unique explanation of the neural response when evaluated in the context of various low-level visual control models. This was indeed the case, with Real-world Size remaining a significant predictor even when simultaneously regressing out three visual control models (Figure 4A). The time course was similar to that observed in the previous analysis (onset = 172 ms, peak = 200 ms), and remained intermittently relevant to the neural response throughout the period examined here (i.e., up to 500 ms).

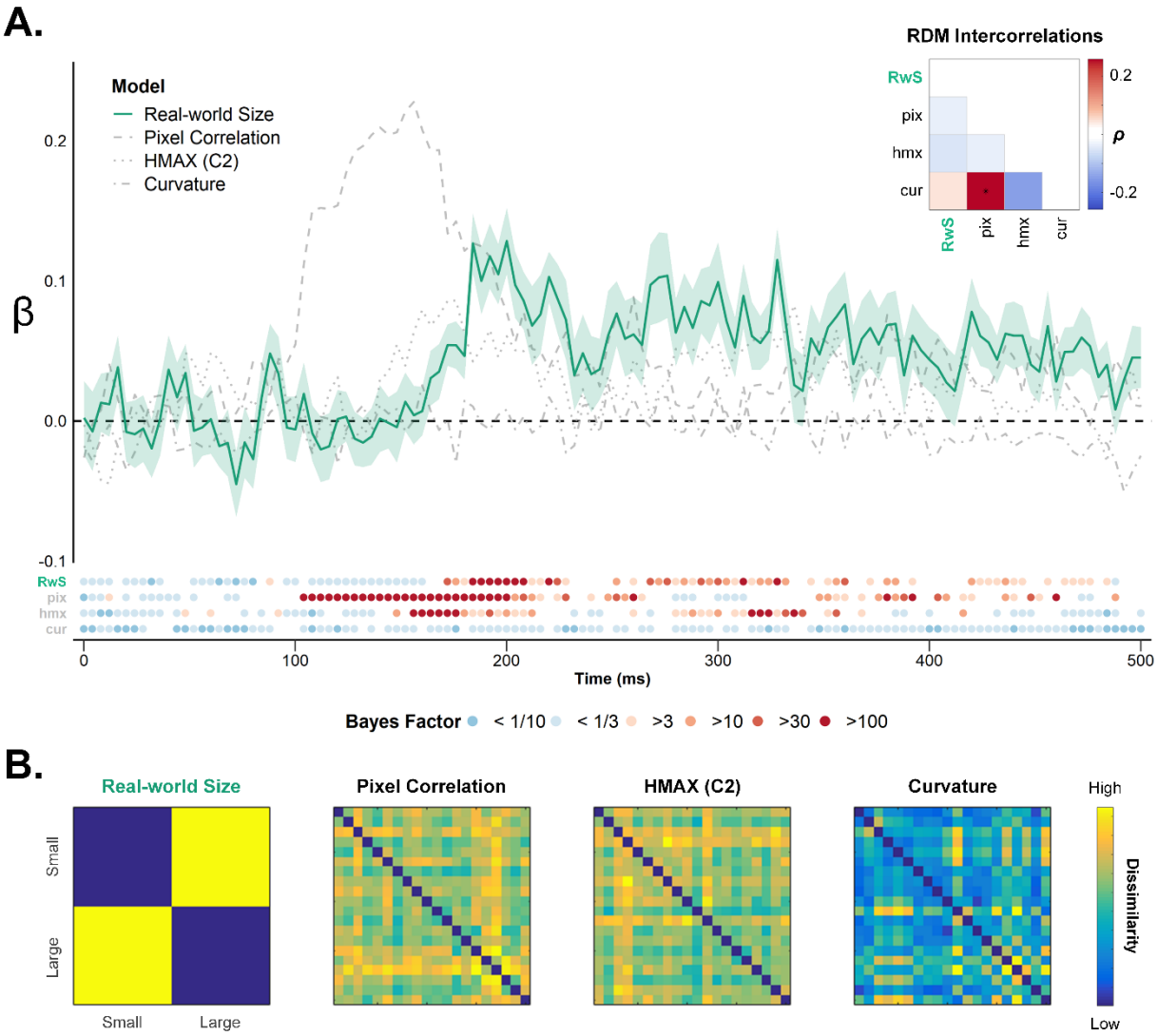

Figure 4. RSA for the 20 object pairs pooled across the four versions (see Figure 1B). A) Multiple regression betas for Real-world Size (in green) and low-level visual control predictors (in grey) of the neural RDM series. Shaded region for the RWS model is within-subjects standard error. Lower panel shows colour coded Bayes Factors for one-tailed t-tests (mean beta > 0) at each timepoint, for each predictor. Inset: Model intercorrelation matrix; asterisks indicate Spearman correlations significant at p <.01 (FDR corrected). B) Models RDMs. See Figure S2 in Supplementary Materials for each model's beta-timecourse when treated as an isolated predictor of the neural RDM series.

Comparison with single objects

We know from Experiment 1 that observers are less able to perceive the size of the objects when viewed one at a time, as compared to when they appeared in related pairs. To test whether the neural representation of real-world size was specific to the object pairs, we next performed the same analyses for single objects. Specifically, we ran an identical multiple regression analysis that included Real-world Size as a categorical predictor alongside three low-level visual control predictors derived from the single object stimuli (i.e., pixel correlation, HMAX (C2), & Curvature). Figure 5 shows the resulting beta-timecourse for the Real-world Size model of Single Object data, with the corresponding beta-timecourse for object pair neural data overlaid in green (to see the full set of beta estimates for the Single Object data, analogous to Figure 4, see Figure S3 in Supplemental Materials). In summary, there was no evidence that real-world size was a relevant predictor for the single object neural response. Moreover, isolated object stimuli did not appear to evoke any kind of size representation at all, in that Real-world Size was also not a relevant predictor when included as the sole categorical predictor in a separate regression of the single object neural RDM series (see Figure S1A in Supplementary Materials). Finally, Real-world Size was a significantly better predictor for

object pairs than for single objects around 200 ms after stimulus onset (Figure 5), providing a possible neural correlate for the behavioural effect observed in Experiment 1.

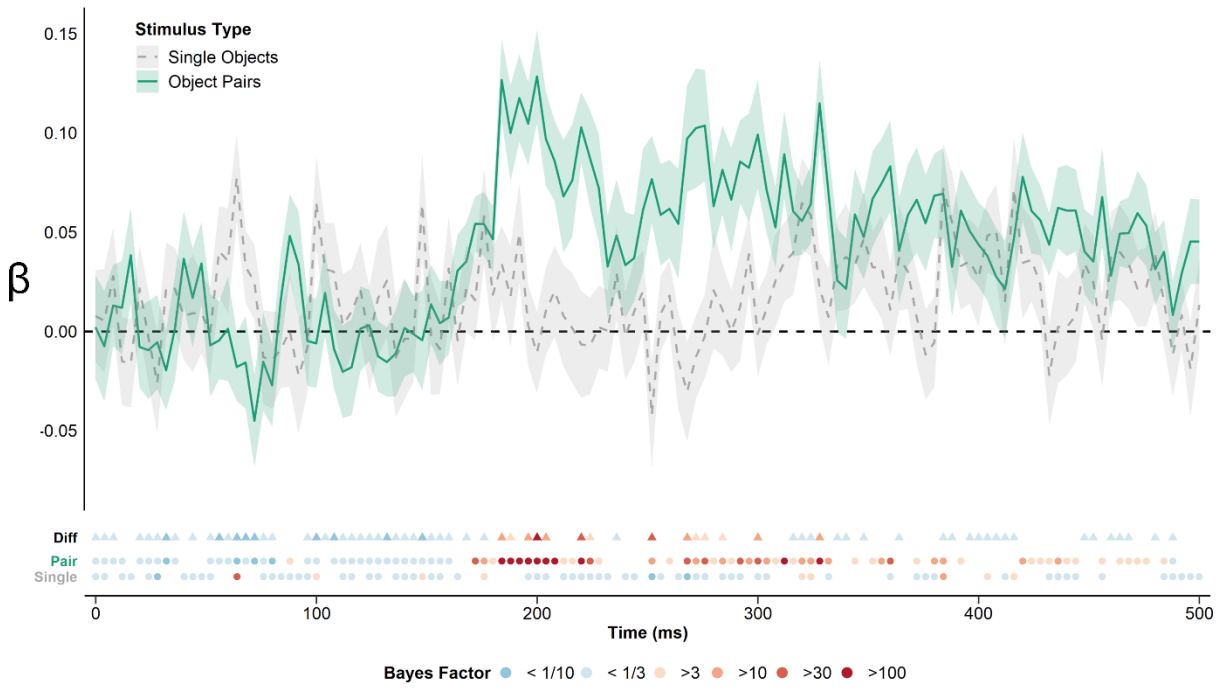

Figure 5. Beta values for Real-world Size included as a categorical predictor in (separate) multiple regression analyses for the single object (in grey) and object pair (in green) neural RDM series. Each regression also included three low-level visual control models as continuous predictors (see Figure 4). Note that the object pair betas here are identical to those in Figure 4 (green lines); shaded region for the RWS model is within-subjects standard error. Lower panel shows colour coded Bayes Factors for each timepoint: circles are one-sample Bayesian *t*-tests evaluating $\beta > 0$; triangles are paired Bayesian *t*-tests evaluating $\beta_{pair} > \beta_{single}$.

Comparisons with a Deep Convolutional Neural Network

Finally, we considered how real-world size information about our object pair stimuli emerged in the layers of a deep convolutional neural network (AlexNet, Krizhevsky et al., 2012). Here we inspected the correlations between the real-world size model and RDMs corresponding to the 5 convolutional and 2 fully-connected layers of AlexNet. As can be seen in Figure 6B, real-world size was significantly correlated with the upper, fully-connected layers (FC6: rho = 0.20, $p = 0.0079$; FC7: rho = 0.24, $p = 0.0016$, FDR-corrected), but had little representational overlap with the early layers of AlexNet. Since low-level features differences between the large and small stimuli should be well-captured by the lowest DNN layers, the fact that these layers did not correlate with Real-world Size once again suggests that this model did not simply reflect visual differences between the large and small stimuli.

To further examine the overlap of the real-world size model with AlexNet, we next ran multiple regressions of the object pair neural RDM series using Real-world Size as a categorical predictor alongside the later RDMs for FC6 and FC7 of AlexNet (treated as continuous predictors, see Figure 6A). Figure 6C and 6D show that in both cases, Real-world Size contributes unique explanatory power for the neural response evoked by object pairs, indicating that representations of spatial scale evoked by Object Pairs are not accounted for by the high-level representations that the DNN gleaned from these stimuli.

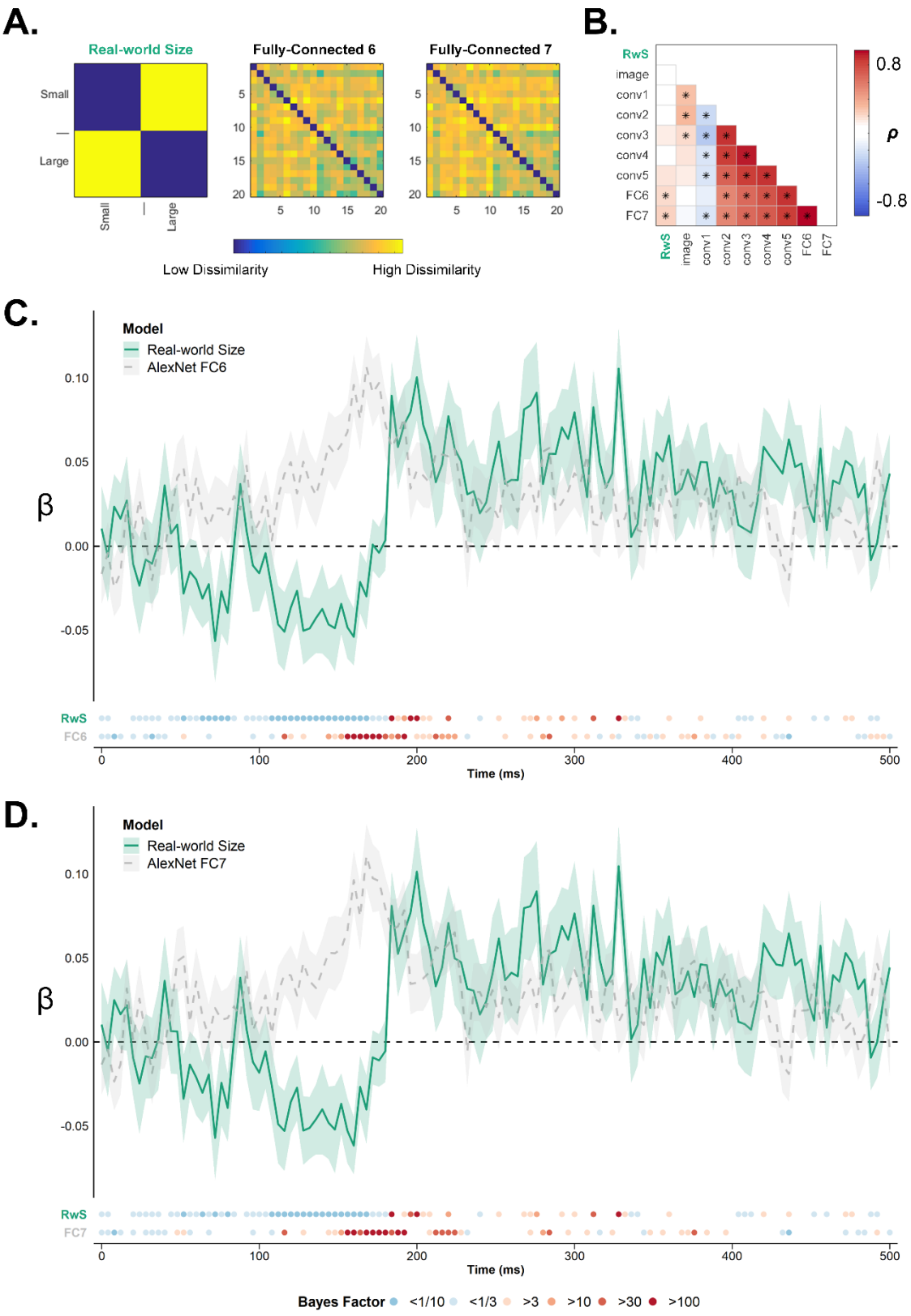

Figure 6. Predicting the neural response to Object Pairs using Real-world Size and DNN activations. **A)** Model RDMs based on Real-world Size and activations on the final two fully-connected layers of AlexNet (RDMs for all layers can be seen in Figure S4 in Supplementary Material). **B)** Intercorrelation matrix for Real-world Size and AlexNet RDMs for all layers. Asterisks indicate correlations significant at p <.01, FDR adjusted. Where Real-world Size was significantly correlated with both fully-connected layers, this model bore little resemblance to the lower layers of AlexNet. **C&D)** Beta estimates from (separate) multiple regressions of the neural RDM series for object pairs that included Real-world Size as a categorical predictor alongside continuous predictors based on AlexNet layers FC6 (C) and FC7 (D). Shaded region is within-subjects standard error. To see a corresponding analysis for the Single Object data, see Figure S5 in Supplemental Materials.

Discussion

To apprehend a familiar object is to rapidly and automatically appreciate its inherent, or 'real-world', size (Hagen et al., 2023; Konkle & Oliva, 2012a). Investigations of both isolated objects and entire scenes have revealed real-world size to be a key organising principle of the visual system, with the spatial scale of objects influencing both their neural representation and the way observers recognise and respond to them (Coutanche & Koch, 2018; He et al., 2013; Khaligh-Razavi et al., 2018; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b; Long et al., 2018; Wang et al., 2022). Our goal in the present study was to exploit observers' tendency to integrate concurrent object information, and examine real-world size representations evoked by groups of contextually-associated objects (i.e., object constellations).

We found that observers recognised the real-world size of ambiguous (silhouetted) objects both faster and more accurately when they appeared in contextually-associated pairs (e.g., cup + saucer) than when they appeared in isolation. This type of inter-object contextual facilitation replicates the findings of Bar and Ullman (1996), who observed that participants were worse to recognise isolated ambiguous objects (e.g., a shoe) compared to when those features appeared in context with another object or object part (e.g., a hat). In the neural domain, we found that representations of real-world size evoked by viewing silhouetted pairs of related objects arose from around 170 ms after stimulus onset and peaked around 200 ms after stimulus onset. These representations endured when controlling for differences in large and small objects' visual features. Importantly, real-world size representations were only evident when the object appeared as part of a contextually-associated pair, and not when the individual object elements appeared in isolation. Furthermore, real-world size representations around 200 ms after stimulus onset were significantly stronger for object pairs than for isolated objects.

We interpret these results as reflecting inter-object facilitation of object recognition (Bar & Ullman, 1996), with improved recognition and neural representation of real-world size following from the improved recognition of contextually-associated objects, made possible by integrating information across related objects. On this interpretation, the facilitatory interaction between the processing of individual objects must have occurred shortly before 200 ms (when real-world size was represented). This is substantially earlier than facilitatory interactions between associated scenes and objects, which are observed some 300 ms after stimulus onset (Brandman & Peelen, 2017, 2023). An explanation for this difference could relate to the finding that scenes and objects are processed in parallel ventral stream pathways (Levy, Hasson, Avidan, Hendler, & Malach, 2001; Park, Brady, Greene, & Oliva, 2011), such that interactions between scenes and objects require interactions across pathways. By contrast, interactions between objects could occur within the object processing pathway itself.

An alternative interpretation of the current results is that they reflect the accumulation of evidence (about real-world size) across objects without facilitatory inter-object interactions. For example, each object may provide weak evidence for either large or small real-world size, with this evidence accumulating for multiple objects presented together. The behavioural and EEG results suggest that such evidence accumulation is non-linear; for example, evidence may have been more likely to reach a critical real-world size recognition threshold (and associated neural representation) for two versus one object. Future studies could distinguish between this account and the inter-object facilitation account by testing a condition in which objects are recombined (or "shuffled"; Kaiser, Stein, & Peelen, 2014) across pairs within each size condition. In such a condition, both objects would still provide individual evidence for large or small real-world size but would not be expected to facilitate each other's recognition.

An additional manipulation in the current experiment concerned the objects' relative depth, with the proportionate/disproportional retinal sizing of the two objects within the pair suggesting they were either at the same or different depth. We hypothesized that observers would have more difficulty integrating the objects to extract the pair's real-world size when the two objects appeared to be at different depths from the viewer, similar to how disrupting objects' relative spatial position impairs recognition (Bar & Ullman, 1996) and reduces EEG measures of integration (Quek & Peelen, 2020). However, our results showed that this manipulation had no influence on either overt recognition of real-world size or on neural representations of real-world size. This suggests that the visual system is tolerant to the relative size of objects when integrating information across them. This may reflect the variability of relative sizes of objects in daily life, where contextually associated objects can be positioned at different distances (and thus appear in different retinal sizes) relative to the observer. For example, a stapler can be observed next to a piece of paper (Figure 1B), but is also frequently observed closer or farther away from this object. On the other hand, given that texture

is a strong cue for distance (Gibson, 1950; Todd & Akerstrom, 1987), the silhouette stimuli we used here may have impaired or reduced the degree to which observers perceived the depth of the objects. Relative-depth effects for contextually-associated objects could well be evident when texture and other internal object cues are preserved, as in photograph stimuli.

In addition to addressing questions about inter-object facilitation, our study also speaks to the literature on the neural representation of real-world size. Real-world size is a dimension that pertains to all stimuli along the object-scene continuum. At the level of individual objects, real-world size is an organising principle within ventral occipitotemporal cortex (Coutanche & Koch, 2018; He et al., 2013; Khaligh-Razavi et al., 2018; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b; Long et al., 2018; Wang et al., 2022). When retinal size is fixed, lateral ventral temporal regions respond more strongly to objects whose real-world size is small (e.g., a tennis ball) compared to large (e.g., a desk), while medial temporal regions exhibit the opposite activation pattern (Konkle & Oliva, 2012b). Such effects appear to reflect regional sensitivity to different featural and functional characteristics that vary systematically across small and large objects (Julian, Ryan, & Epstein, 2016; Konkle & Oliva, 2012b; Long et al., 2018; Mullally & Maguire, 2011; Troiani, Stigliani, Smith, & Epstein, 2014; Wang et al., 2022). Studies investigating the time course of real-world size using single objects have shown that real-world size is reflected in the neural response already 120 ms after stimulus onset (e.g., Khaligh-Razavi et al., 2018; Wang et al., 2022). However, this early size-decoding likely reflects visual feature differences between large and small objects, for example the presence of more rectilinear features in large than small objects (Long et al., 2018; Wang et al., 2022).

To our knowledge, the current study is the first to measure the time course of real-world size representations for object constellations, showing that real-world size is represented at around 200 ms after stimulus onset. Unlike studies reporting earlier size representations for single objects, reviewed above, our results are unlikely to reflect low-or mid-level visual feature differences between large and small constellations, for several reasons. First, we selected our stimuli to avoid obvious differences between large and small conditions and used silhouettes to remove texture and colour features. Second, we included the output of multiple computational models in the EEG analyses to show that visual feature differences could not account for the real-world size representation at 200 ms. Finally, the absence of a size representation for the single objects, consisting of the same shapes, further argues against a feature driven effect.

If not the representation of size-covarying visual features, what could drive the different neural activity patterns for large and small constellations? One possibility is that recognizing the constellations activated corresponding scene representations, such that the activity patterns reflect scene size (e.g., desktop vs living room). Neural responses to scene images capture information about spatial scale as early as 100 ms after stimulus onset (Cichy et al., 2017), though as for objects, such early representations are likely driven by differences in low-level features across small and large scenes (Stansbury, Naselaris, & Gallant, 2013). More abstract signals of scene size (i.e., representations that generalise across image-level changes in clutter, luminance, and contrast) appear to arise later, from around 250 ms (Cichy et al., 2017), and are evident even for auditory stimuli that capture the spatial extent of real-world spaces (Teng, Sommer, Pantazis, & Oliva, 2017).

More generally, our results raise the question of how representations of objects and scenes are related. By moving from single objects to object constellations, it becomes clear that the distinction between objects and scenes is not so clear-cut. Indeed, even the processing of single objects may involve processing their real-world context – i.e., surrounding scene. For example, seeing a large object implies a large space and activates scene-selective cortex (Mullally & Maguire, 2011). Furthermore, object representations in scene-selective regions reflect the context within which objects are usually observed, including their co-occurrences with other objects (Aminoff, Gronau, & Bar, 2007; Bonner & Epstein, 2021; Fu et al., 2022). Our results align with these findings, showing that multiple objects viewed together activate the objects' shared context within 200 ms.

To conclude, our results reveal the time course of real-world size processing of object constellations, showing that neural activity patterns distinguish between large and small constellations at around 200 ms after stimulus onset. We interpret these findings as reflecting the facilitatory influence between contextually associated objects. Our study raises several new questions, including about the relationship between representations along the continuum from objects to scenes.

References

Aminoff, E., Gronau, N., & Bar, M. (2007). The Parahippocampal Cortex Mediates Spatial and Nonspatial Associations. *Cerebral Cortex, 17*(7), 1493-1503. doi:10.1093/cercor/bhl078

Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review, 14*(2), 332-337.

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2017). Human-Object Interactions Are More than the Sum of Their Parts. *Cerebral Cortex, 27*(3), 2276-2288. doi:10.1093/cercor/bhw077

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617.

Bar, M., & Ullman, S. (1996). Spatial Context in Recognition. *Perception, 25*(3), 343-352. doi:10.1068/p250343

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*(4043), 77-80. doi:10.1126/science.177.4043.77

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, 97*(1), 22-27. doi:10.1037/h0033776

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143-177.

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature communications, 12*(1), 4081.

Brandman, T., & Peelen, M. V. (2017). Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. *Journal of Neuroscience, 37*(32), 7700-7710. doi:10.1523/jneurosci.0582-17.2017

Brandman, T., & Peelen, M. V. (2023). Objects sharpen visual scene representations: evidence from MEG decoding. *bioRxiv*, 2023.2004.2006.535903. doi:10.1101/2023.04.06.535903

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. doi:https://doi.org/10.1016/j.neuroimage.2016.03.063

Coutanche, M. N., & Koch, G. E. (2018). Creatures great and small: Real-world size of animals predicts visual cortex representations beyond taxonomic category. *Neuroimage, 183*, 627-634. doi:10.1016/j.neuroimage.2018.08.066

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition, 35*(3), 393-401. doi:10.3758/bf03193280

Epstein, R. A., & Baker, C. I. (2019). Scene Perception in the Human Brain. *Annu Rev Vis Sci, 5*, 373-397. doi:10.1146/annurev-vision-091718-014809

Epstein, R. A., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601. doi:10.1038/33402

Fu, Z., Wang, X., Wang, X., Yang, H., Wang, J., Wei, T., . . . Bi, Y. (2022). Different computational relations in language are captured by distinct brain systems. *Cerebral Cortex, 33*(4), 997-1013. doi:10.1093/cercor/bhac117

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Brain Research: Cognitive Brain Research, 16*(2), 123-144. doi:10.1016/s0926-6410(02)00244-6

Gibson, J. J. (1950). The Perception of Visual Surfaces. *The American Journal of Psychology, 63*(3), 367-384. doi:10.2307/1418003

Green, C., & Hummel, J. E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1107.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58*(2), 137-176. doi:https://doi.org/10.1016/j.cogpsych.2008.06.001

Greene, M. R., & Oliva, A. (2010). High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 1430-1442. doi:10.1037/a0019058

Hagen, S., Zhao, Y., Moonen, L., Ulken, N., & Peelen, M. (2023). What drives the automatic retrieval of real-world object size knowledge?

He, C., Peelen, M. V., Han, Z., Lin, N., Caramazza, A., & Bi, Y. (2013). Selectivity for large nonmanipulable objects in scene-selective visual cortex does not require visual experience. *Neuroimage, 79*, 1-9.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles, 37*, 547-579.

Julian, J. B., Ryan, J., & Epstein, R. A. (2016). Coding of Object Size and Object Category in Human Visual Cortex. *Cerebral Cortex, 27*(6), 3095-3109. doi:10.1093/cercor/bhw150

Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *Neuroimage, 169*, 334-341. doi:https://doi.org/10.1016/j.neuroimage.2017.12.065

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object Vision in a Structured World. *Trends in Cognitive Sciences, 23*(8), 672-685. doi:10.1016/j.tics.2019.04.013

Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America, 111*(30), 11217-11222.

Kaiser, D., Stein, T., & Peelen, M. V. (2015). Real-world spatial regularities affect visual working memory for objects. *Psychonomic Bulletin & Review, 22*(6), 1784-1790. doi:10.3758/s13423-015-0833-4

Khaligh-Razavi, S. M., Cichy, R. M., Pantazis, D., & Oliva, A. (2018). Tracking the Spatiotemporal Neural Dynamics of Real-world Object Size and Animacy in the Human Brain. *Journal of Cognitive Neuroscience, 30*(11), 1559-1576. doi:10.1162/jocn_a_01290

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology, 10*(11), e1003915. doi:10.1371/journal.pcbi.1003915

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience, 33*(25), 10235-10242.

Konkle, T., & Oliva, A. (2012a). A familiar-size Stroop effect: Real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 561-569. doi:10.1037/a0028294

Konkle, T., & Oliva, A. (2012b). A real-world size organization of object responses in occipito-temporal cortex. *Neuron, 74*(6), 1114-1124. doi:10.1016/j.neuron.2012.04.036

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401-412. doi:https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience, 2*, 4. doi:10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Advances in Neural Information Processing Systems.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology, 62*, 621-647.

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center-periphery organization of human object areas. *Nature Neuroscience, 4*(5), 533-539. doi:10.1038/87490

Li, S. P. D., & Bonner, M. (2020). *Curvature as an Organizing Principle of Mid-level Visual Representation: A Semantic-preference Mapping Approach.* Paper presented at the NeurIPS 2020 Workshop SVRHM.

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

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., . . . Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex.

Proceedings of the National Academy of Sciences of the United States of America, 92(18), 8135-8139. doi:10.1073/pnas.92.18.8135

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia, 48*(2), 507-517.

Mullally, S. L., & Maguire, E. A. (2011). A new role for the parahippocampal cortex in representing space. *Journal of Neuroscience, 31*(20), 7441-7449.

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics, 10*. doi:10.3389/fninf.2016.00027

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

Peelen, M., Berlot, E., & de Lange, F. (2023). Predictive processing of scenes and objects. doi:10.31234/osf.io/nuf59

Quek, G. L., & Peelen, M. V. (2020). Contextual and Spatial Associations Between Objects Interactively Modulate Visual Processing. *Cerebral Cortex, 30*(12), 6391-6404. doi:10.1093/cercor/bhaa197

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience, 2*(11), 1019-1025.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in brain research, 165*, 33-56.

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron, 79*(5), 1025-1034. doi:10.1016/j.neuron.2013.06.034

Teng, S., Sommer, V. R., Pantazis, D., & Oliva, A. (2017). Hearing Scenes: A Neuromagnetic Signature of Auditory Source and Reverberant Space Separation. *eNeuro, 4*(1). doi:10.1523/eneuro.0007-17.2017

Todd, J. T., & Akerstrom, R. A. (1987). Perception of three-dimensional form from patterns of optical texture. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 242-255. doi:10.1037/0096-1523.13.2.242

Troiani, V., Stigliani, A., Smith, M. E., & Epstein, R. A. (2014). Multiple object properties drive scene-selective regions. *Cerebral cortex, 24*(4), 883-897.

Võ, M. L. H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology, 29*, 205-210. doi:https://doi.org/10.1016/j.copsyc.2019.03.009

Võ, M. L. H., & Wolfe, J. M. (2013). Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science, 24*(9), 1816-1823. doi:10.1177/0956797613476955

Wang, R., Janini, D., & Konkle, T. (2022). Mid-level Feature Differences Support Early Animacy and Object Size Distinctions: Evidence from EEG Decoding. *Journal of Cognitive Neuroscience*, 1-11. doi:10.1162/jocn_a_01883

Wischnewski, M., & Peelen, M. V. (2021). Causal neural mechanisms of context-based object recognition. *eLife, 10*, e69736. doi:10.7554/eLife.69736