Running title: Real-World Size Representations in Natural Images

Human EEG and artificial neural networks reveal disentangled representations of object real-world size in natural images

Zitong Lu & Julie D. Golomb
Department of Psychology, The Ohio State University, Columbus, OH 43212 USA

Please address correspondence to:
Zitong Lu
Department of Psychology
The Ohio State University
Columbus, OH, 43210
Email: lu.2637@osu.edu

Abstract

Remarkably, human brains have the ability to accurately perceive and process the real-world size of objects, despite vast differences in distance and perspective. While previous studies have delved into this phenomenon, distinguishing this ability from other visual perceptions, like depth, has been challenging. Using the THINGS EEG2 dataset with high time-resolution human brain recordings and more ecologically valid naturalistic stimuli, our study uses an innovative approach to disentangle neural representations of object real-world size from retinal size and perceived real-world depth in a way that was not previously possible. Leveraging this state-of-the-art dataset, our EEG representational similarity results reveal a pure representation of object real-world size in human brains. We report a representational timeline of visual object processing: object real-world depth appeared first, then retinal size, and finally, real-world size. Additionally, we input both these naturalistic images and object-only images without natural background into artificial neural networks. Consistent with the human EEG findings, we also successfully disentangled representation of object real-world size from retinal size and real-world depth in all three types of artificial neural networks (visual-only ResNet, visual-language CLIP, and language-only Word2Vec). Moreover, our multi-modal representational comparison framework across human EEG and artificial neural networks reveals real-world size as a stable and higher-level dimension in object space incorporating both visual and semantic information. Our research provides a detailed and clear characterization of the object processing process, which offers further advances and insights into our understanding of object space and the construction of more brain-like visual models.

Keywords: real-world size, depth perception, RSA, artificial neural networks, object recognition

Introduction

Imagine you are viewing an apple tree while walking around an orchard: as you change your perspective and distance, the retinal size of the apple you plan to pick varies, but you still perceive the apple as having a constant real-world size. How do our brains extract object real-world size information during object recognition to allow us to understand the complex world? Behavioral studies have demonstrated that perceived real-world size is represented as an object physical property, revealing same-size priming effects (Setti et al., 2008), familiar-size stroop effects (Konkle & Oliva, 2012a; Long & Konkle, 2017), and canonical visual size effects (Chen et al., 2022; Konkle & Oliva, 2011). Human neuroimaging studies have also found evidence of object real-world size representation (Huang et al., 2022; S.-M. Khaligh-Razavi et al., 2018; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b; Luo et al., 2023; Quek et al., 2023; R. Wang et al., 2022). These findings suggest real-world size is a fundamental dimension of object representation.

However, previous studies on object real-world size have faced several challenges. Firstly, the perception of an object's real-world size is closely related to the perception of its real-world distance in depth. For instance, imagine you are looking at photos of an apple and a basketball: if the two photos were zoomed in such that the apple and the basketball filled the same exact retinal (image) size, you could still easily perceive that the apple is the physically smaller real-world object. But, you would simultaneously infer that the apple is thus located closer to you (or the camera) than the basketball. In previous neuroimaging studies of perceived real-world size (Huang et al., 2022; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b), researchers presented images of familiar objects zoomed and cropped such that they occupied the

same retinal size, finding that neural responses in ventral temporal cortex reflected the perceived real-world size (e.g. an apple smaller than a car). However, while they controlled the retinal size of objects, the intrinsic correlation between real-world size and real-world depth in these images meant that the influence of perceived real-world depth could not be entirely isolated when examining the effects of real-world size. This makes it difficult to ascertain whether their results were driven by neural representations of perceived real-world size and/or perceived real-world depth. MEG and EEG studies focused on temporal processing of object size representations (S.-M. Khaligh-Razavi et al., 2018; R. Wang et al., 2022) have been similarly susceptible to this limitation. Indeed, one recent behavioral study (Quek et al., 2023) provided evidence that perceived real-world depth could influence real-world size representations, further illustrating the necessity of investigating pure real-world size representations in the brain. Secondly, the stimuli used in these studies were cropped object stimuli against a plain white or grey background, which are not particularly naturalistic. More and more studies and datasets have highlighted the important role of naturalistic context in object recognition (Allen et al., 2022; Gifford et al., 2022; Grootswagers et al., 2022; Hebart et al., 2019; Stoinski et al., 2023). In ecological contexts, inferring the real-world size/distance of an object likely relies on a combination of bottom-up visual information and top-down knowledge about canonical object sizes for familiar objects. Incorporating naturalistic background context in experimental stimuli may produce more accurate assessments of the relative influences of visual shape representations (Bracci et al., 2017; Bracci & Op de Beeck, 2016; Proklova et al., 2016) and higher-level semantic information (Doerig et al., 2022; Huth et al., 2012; A. Y. Wang et al., 2022). Furthermore, most previous studies have tended to categorize size rather broadly, such as merely differentiating between big and small objects (S.-M. Khaligh-Razavi et al., 2018; Konkle &

Oliva, 2012b; R. Wang et al., 2022) or dividing object size into seven levels from small to big. To more finely investigate the representation of object size in the brain, it may be necessary to obtain a more continuous measure of size for a more detailed characterization.

Certainly, a minority of fMRI studies have attempted to utilize natural images and also engaged in more detailed size measurements to more precisely explore the encoding of object real-world size in different brain areas (Luo et al., 2023; Troiani et al., 2014). However, no study has yet comprehensively overcome all the challenges and unfolded a clear processing timeline for object retinal size, real-world size, and real-world depth in human visual perception.

In the current study, we overcome all these challenges by combining high temporal-resolution EEG, naturalistic images, artificial neural networks, and novel computational methods to distinguish the neural representations of object real-world size, retinal size, and real-world depth. We applied our novel computational approach to an open EEG dataset, THINGS EEG2 (Gifford et al., 2022). Firstly, the visual image stimuli used in this dataset are more naturalistic and include objects that vary in real-world size, depth, and retinal size. This allows us to employ a multi-model representational similarity analysis to investigate pure representations of object real-world size, partialing out – and simultaneously exploring – these confounding features. Secondly, we are able to explore the neural dynamics of object feature processing in a more ecological context based on natural images in human object recognition. Thirdly, instead of dividing object size into several levels, we applied more detailed behavioral measurements from an online size rating task to obtain a more continuous measure to more finely decode the representation of object size in the brain.

We first focus on unfolding the neural dynamics of pure object real-world size

representations. The temporal resolution of EEG allows us the opportunity to investigate the representational timecourse of visual object processing, asking whether processing of perceived object real-world size precedes or follows processing of perceived depth, if these two properties are in fact processed independently.

We then attempt to further explore the underlying mechanisms of how human brains process object size and depth in natural images by integrating artificial neural networks (ANNs). In the domain of cognitive computational neuroscience, ANNs offer a complementary tool to study visual object recognition, and an increasing number of studies support that ANNs exhibit representations similar to human visual systems (Cichy et al., 2016; Güçlü & van Gerven, 2015; Yamins et al., 2014; Yamins & DiCarlo, 2016). Indeed, a recent study found that ANNs also represent real-world size (Huang et al., 2022); however, their use of a fixed retinal size image dataset with the same cropped objects as described above makes it similarly challenging to ascertain whether the results reflected real-world size and/or depth. Additionally, some recent work indicates that artificial neural networks incorporating semantic embedding and multimodal neural components might more accurately reflect human visual representations within visual areas and even the hippocampus, compared to vision-only networks (Choksi, Mozafari, et al., 2022; Choksi, Vanrullen, et al., 2022; Conwell et al., 2022; Doerig et al., 2022; Jozwik et al., 2023; A. Y. Wang et al., 2022). Given that perception of real-world size may incorporate both bottom-up visual and top-down semantic knowledge about familiar objects, these models offer yet another novel opportunity to investigate this question. Utilizing both visual and visual-semantic models, as well as different layers within these models, ANNs provide us the approach to extract various image features, low-level visual information from early layers and higher-level information including both visual and semantic features from late layers.

The novel computational approach by cross-modal representational comparisons we take with the current study allows us to compare how representations of perceived real-world size and depth emerge in both human brains and artificial neural networks. Unraveling the internal representations of object size and depth features in both human brains and ANNs provides us a deeper approach to not only explore whether both biological and artificial systems represent object real-world size, along with retinal size and real-world depth features, but also investigate possible mechanisms of object real-world size representations.

Materials and Methods

Experimental design, stimuli images and EEG data:

We utilized the open dataset from THINGS EEG2 (Gifford et al., 2022), which includes EEG data from 10 healthy human subjects (age=28.5±4, 8 female and 2 male) in a rapid serial visual presentation (RSVP) paradigm with an orthogonal target detection task to ensure participants paid attention to the visual stimuli. For each trial, subjects viewed one image (sized 500 × 500 pixels) for 100ms. Each subject viewed 16740 images of objects on a natural background for 1854 object concepts from THINGS dataset (Hebart et al., 2019). For the current study, we used the 'test' dataset portion, which includes 16000 trials per subject corresponding to 200 images (200 object concepts, one image per concept) with 80 trials per image. Before inputting the images to the ANNs, we reshaped image sizes to 224 x 224 pixels and normalized the pixel values of images to ImageNet statistics.

EEG data were collected using a 64-channel EASYCAP and a BrainVision actiCHamp amplifier. We used already pre-processed data from 17 channels (O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, P2) overlying occipital and parietal cortex. We re-epoched EEG data ranging from 100ms before stimulus onset to 300ms after onset with a sample frequency of 100Hz. Thus the shape of our EEG data matrix for each trial was 17 channels × 40 time points.

ANN models:

We applied two pre-trained ANN models: one visual model (ResNet-101 (He et al., 2016) pretrained on ImageNet), and one multi-modal (visual+semantic) model (CLIP with a ResNet-101 backbone (Radford et al., 2021) pretrained on YFCC-15M). We used THINGSvision (Muttenthaler & Hebart, 2021) to obtain low- and high-level feature vectors of ANN activations from early and late layers (early layer: second convolutional layer; late layer: last visual layer) for the images.

Word2Vec model:

To approximate the non-visual, pure semantic space of objects, we also applied a Word2Vec model, a natural language processing model for word embedding, pretrained on Google News corpus (Mikolov et al., 2013), which contains 300-dimensional vectors for 3 million words and phrases. We input the words for each image's object concept (pre-labeled in THINGS dataset: Hebart et al., 2019), instead of the visual images themselves. We used Gensim (Řehůřek & Sojka, 2010) to obtain Word2Vec feature vectors for the objects in images.

Representational dissimilarity matrices (RDMs):

To conduct RSA across human EEG, artificial models, and our hypotheses corresponding to different visual features, we first computed representational dissimilarity matrices (RDMs) for different modalities (Figure 2). The shape of each RDM was 200 × 200, corresponding to pairwise dissimilarity between the 200 images. We extracted the 19900 cells from the upper half of the diagonal of each RDM for subsequent analyses.

Neural RDMs. From the EEG signal, we constructed timepoint-by-timepoint neural RDMs for each subject with decoding accuracy as the dissimilarity index (Figure 2A). We first conducted timepoint-by-timepoint classification-based decoding for each subject and each pair of images (200 images, 19900 pairs in total). We applied linear Support Vector Machine (SVM) to train and test a two-class classifier, employing a 5-time 5-fold cross-validation method, to obtain an independent decoding accuracy for each image pair and each timepoint. Therefore, we ultimately acquired 40 (1 per timepoint) EEG RDMs for each subject.

Hypothesis-based (HYP) RDMs. We constructed three hypothesis-based RDMs reflecting the different types of visual object properties in the naturalistic images (Figure 2B): Real-World Size RDM, Retinal Size RDM, and Real-World Depth RDM. We constructed these RDMs as follows:

(1) For Real-World Size RDM, we obtained human behavioral real-world size ratings of each object concept from the THINGS+ dataset (Stoinski et al., 2022). In the THINGS+ dataset, 2010 participants (different from the subjects in THINGS EEG2) did an online size rating task and completed a total of 13024 trials corresponding to 1854 object concepts. The range of possible size ratings was from 0 to 519 in their online size rating task, with the actual mean ratings across subjects ranging from 100.03 ('sand') to

423.09 ('subway'). We used these ratings as the perceived real-world size measure of the object concept pre-labeled in THINGS dataset (Hebart et al., 2019) for each image. We then constructed the representational dissimilarity matrix by calculating the absolute difference between perceived real-world size ratings for each pair of images.

(2) For Retinal Size RDM, we applied Adobe Photoshop (Adobe Inc., 2019) to crop objects corresponding to object labels from images manually, obtaining a rectangular region that precisely contains a single object, then measured the diagonal length of the segmented object in pixels as the retinal size measure (Konkle & Oliva, 2011). Due to our calculations being at the object level, if there were more than one same objects in an image, we cropped the most complete one to get more accurate retinal size. We then constructed the RDM by calculating the absolute difference between measured retinal size for each pair of images.

(3) For Real-World Depth RDM, we calculated the perceived depth based on the measured retinal size index and behavioral real-world size ratings, such that real-world depth / visual image depth = real-world size / retinal size. Since visual image depth (viewing distance) is held constant across images in the task, real-world depth is proportional to real-world size / retinal size. We then constructed the RDM by calculating the absolute difference between real-world depth index for each pair of images.

<u>ANN (and Word2Vec) model RDMs</u>. We constructed a total of five model-based RDMs (Figure 2C). Our primary analyses used four ANN RDMs, corresponding to the early and late layers for both ResNet and CLIP (Figure S1). We also calculated a single Word2Vec RDM for the pure semantic analysis (Figure S2). For each RDM, we got the dissimilarities by calculating 1 –

Pearson correlation coefficient between each pair of two vectors of the model features corresponding to two input images.

Representational similarity analyses (RSA) and statistical analyses:

We conducted cross-modal representational similarity analyses between the three types of RDMs (Figure 2). All decoding and RSA analyses were implemented using NeuroRA (Lu & Ku, 2020).

EEG × ANN (or W2V) RSA. To measure the representational similarity between human brains and ANNs and confirm that ANNs have significantly similar representations to human brains, we calculated the Spearman correlation between the 40 timepoint-by-timepoint EEG neural RDMs and the 4 ANN RDMs corresponding to the representations of ResNet early layer, ResNet late layer, CLIP early layer, CLIP late layer, respectively. We also calculated temporal representational similarity between human brains (EEG RDMs) and the Word2Vec model RDM. Cluster-based permutation tests were conducted to determine the time windows of significant representational similarity. First, we performed one-sample t-tests (one-tailed testing) against zero to get the t-value for each timepoint, and extracted significant clusters. We computed the clustering statistic as the sum of t-values in each cluster. Then we conducted 1000 permutations of each subject's timepoint-by-timepoint similarities to calculate a null distribution of maximum clustering statistics. Finally, we assigned cluster-level p-values to each cluster of the actual representational timecourse by comparing its cluster statistic with the null distribution. Time-windows were determined to be significant if the p-value of the corresponding cluster was <0.05.

EEG × HYP RSA. To evaluate how human brains temporally represent different visual features, we calculated the timecourse of representational similarity between the timepoint-by-timepoint EEG neural RDMs and the three hypothesis-based RDMs. To avoid correlations

between hypothesis-based RDMs (Figure 3A) influencing comparison results, we calculated partial correlations with one-tailed test against the alternative hypothesis that the partial correlation was positive (greater than zero). Cluster-based permutation tests were performed as described above to determine the time windows of significant representational similarity. In addition, we conducted peak latency analysis to determine the latency of peak representational similarity for each type of visual information with the EEG signal. We restricted the time-window to the significant (partial) correlation time-window for real-world size, retinal size, and real-world depth, and got the individual peak timepoint corresponding to the highest partial correlation. Paired t-tests (two-tailed) were conducted to assess the statistical differences in peak latencies between different visual features.

ANN (or W2V) × HYP RSA. To evaluate how different visual information is represented in ANNs, we calculated representational similarity between the ANN RDMs and hypothesis-based RDMs. As in the EEG × HYP RSA, we calculated partial correlations to avoid correlations between hypothesis-based RDMs. We also calculated the partial correlations between hypothesis-based RDMs and the Word2Vec RDM. To determine statistical significance, we conducted a bootstrap test. We shuffled the order of the cells above the diagonal in each ANN (or Word2Vec) RDM 1000 times. For each iteration, we calculated partial correlations corresponding to the three hypothesis-based RDMs. This produced a 1000-sample null distribution for each HYP x ANN (or W2V) RSA. We hypothesized that if the real similarity was higher than the 95% confidence interval of the null distribution, it indicated that ANN (or W2V) features validly encoded the corresponding visual feature.

Additionally, to explore how the naturalistic background present in the images might influence object real-world size, retinal size, and real-world depth representations, we conducted

another version of the analysis by inputting cropped object images without background into ANN models to obtain object-only ANN RDMs (Figure S3). Then we performed the same ANN x HYP similarity analysis to calculate partial correlations between the hypothesis-based RDMs and object-only ANN RDM. (We didn't conduct the similarity analysis between timepoint-by-timepoint EEG neural RDMs with subjects viewing natural images and object-only ANN RDMs due to the input differences.)

Data and code accessibility:

EEG data and images from THINGS EEG2 data are publicly available on OSF (https://osf.io/3jk45/). All Python analysis scripts will be available post-publication on GitHub (https://github.com/ZitongLu1996/RWsize).

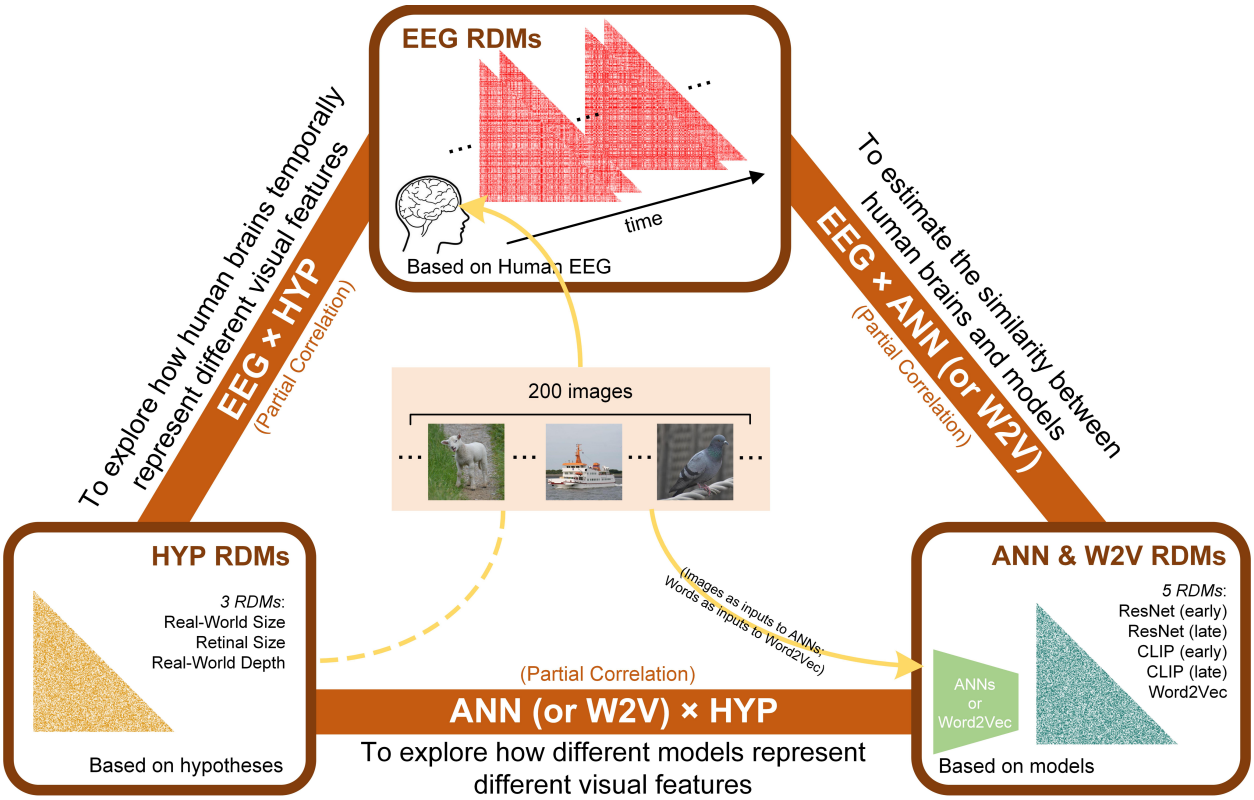

Figure 1 Overview of our analysis pipeline including constructing three types of RDMs and conducting comparisons between them.

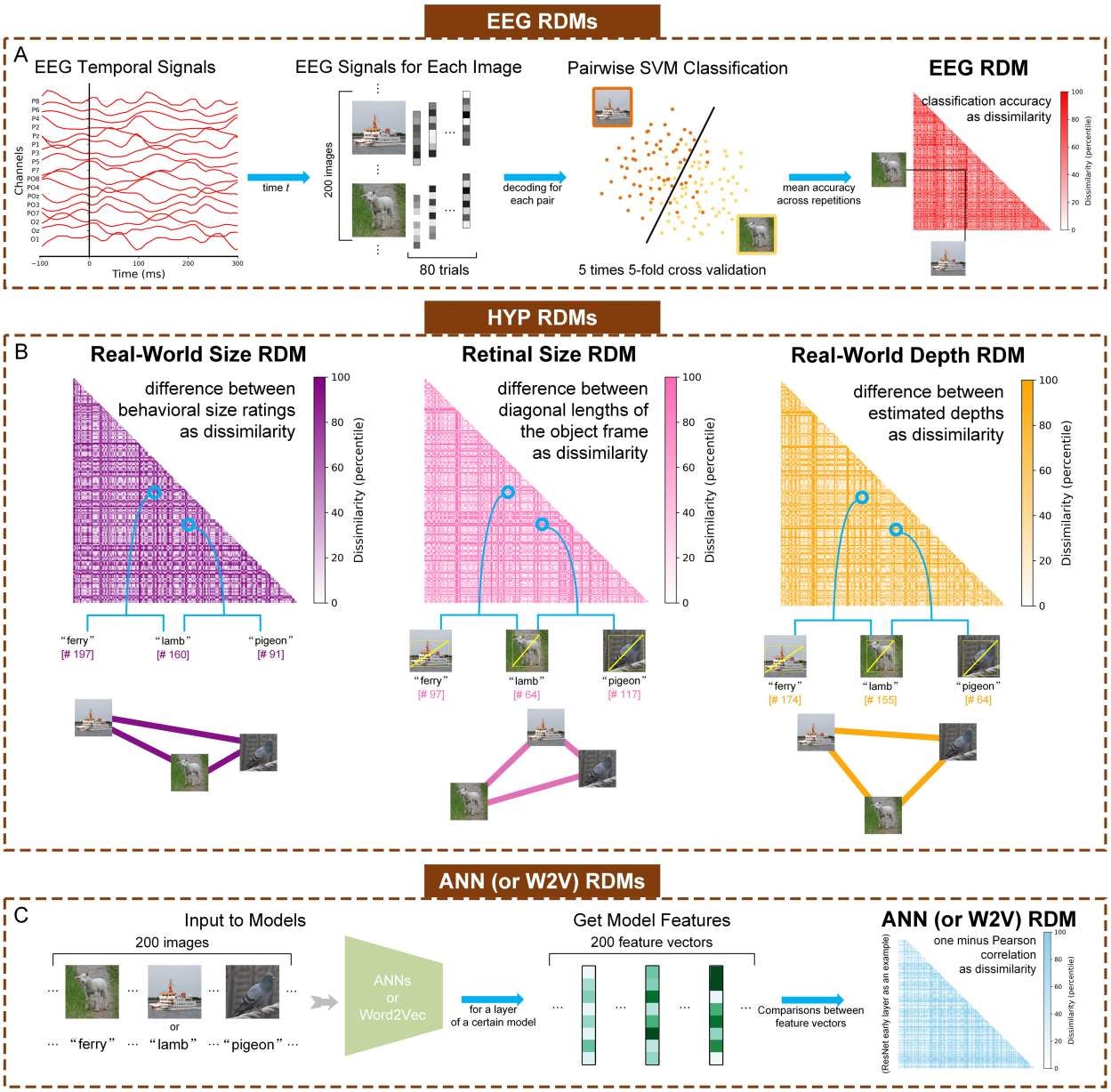

Figure 2 Methods for calculating neural (EEG), hypothesis-based (HYP), and artificial neural network (ANN) & semantic language processing (Word2Vec, W2V) model-based representational dissimilarity matrices (RDMs). (A) Steps of computing the neural RDMs from EEG data. EEG analyses were performed in a time-resolved manner on 17 channels as features. For each time t, we conducted pairwise cross-validated SVM classification. The classification accuracy values across different image pairs resulted in each 200 × 200 RDM for each time point. (B) Calculating the three hypothesis-based RDMs:

Real-World Size RDM, Retinal Size RDM, and Real-World Depth RDM. Real-world size, retinal size, and real-world depth were calculated for the object in each of the 200 stimulus images. The number in the bracket represents the rank (out of 200, in ascending order) based on each feature corresponding to the object in each stimulus image (e.g. "ferry" ranks 197th in real-world size from small to big out of 200 objects). The connection graph to the right of each RDM represents the relative representational distance of three stimuli in the corresponding feature space. (C) Steps of computing the ANN and Word2Vec RDMs. For ANNs, the inputs were the resized images, and for Word2Vec, the inputs were the words of object concepts. For clearer visualization, the shown RDMs were separately histogram-equalized (percentile units).

Results

We conducted a cross-modal representational similarity analysis (Figures 1-2, see Method section for details) comparing the patterns of human brain activation (timepoint-by-timepoint decoding of EEG data) while participants viewed naturalistic object images, the output of different layers of artificial neural networks and semantic language models fed the same stimuli (ANN and Word2Vec models), and hypothetical patterns of representational similarity based on behavioral and mathematical measurements of different visual image properties (perceived real-world object size, displayed retinal object size, and perceived real-world object depth).

Dynamic representations of object size and depth in human brains

To explore if and when human brains contain distinct representations of perceived real-world size, retinal size, and real-world depth, we constructed timepoint-by-timepoint EEG neural

RDMs (Figure 2A), and compared these to three hypothesis-based RDMs corresponding to different visual image properties (Figure 2B). Firstly, we confirmed that the hypothesis-based RDMs were indeed correlated with each other (Figure 3A), and without accounting for the confounding variables, Spearman correlations between the EEG and each hypothesis-based RDM revealed overlapping periods of representational similarity (Figure 3B). In particular, representational similarity with real-world size (from 90 to 120ms and from 170 to 240ms) overlapped with the significant time-windows of other features, including retinal size from 70 to 210ms, and real-world depth from 60 to 130ms and from 180 to 230ms. But critically, with the partial correlations, we isolated their independent representations. The partial correlation results reveal a pure representation of object real-world size in the human brain from 170 to 240ms after stimulus onset, independent from retinal size and real-world depth, which showed significant representational similarity at different time windows (retinal size from 90 to 200ms, and real-world depth from 60 to 130ms and 270 to 300ms) (Figure 3D).

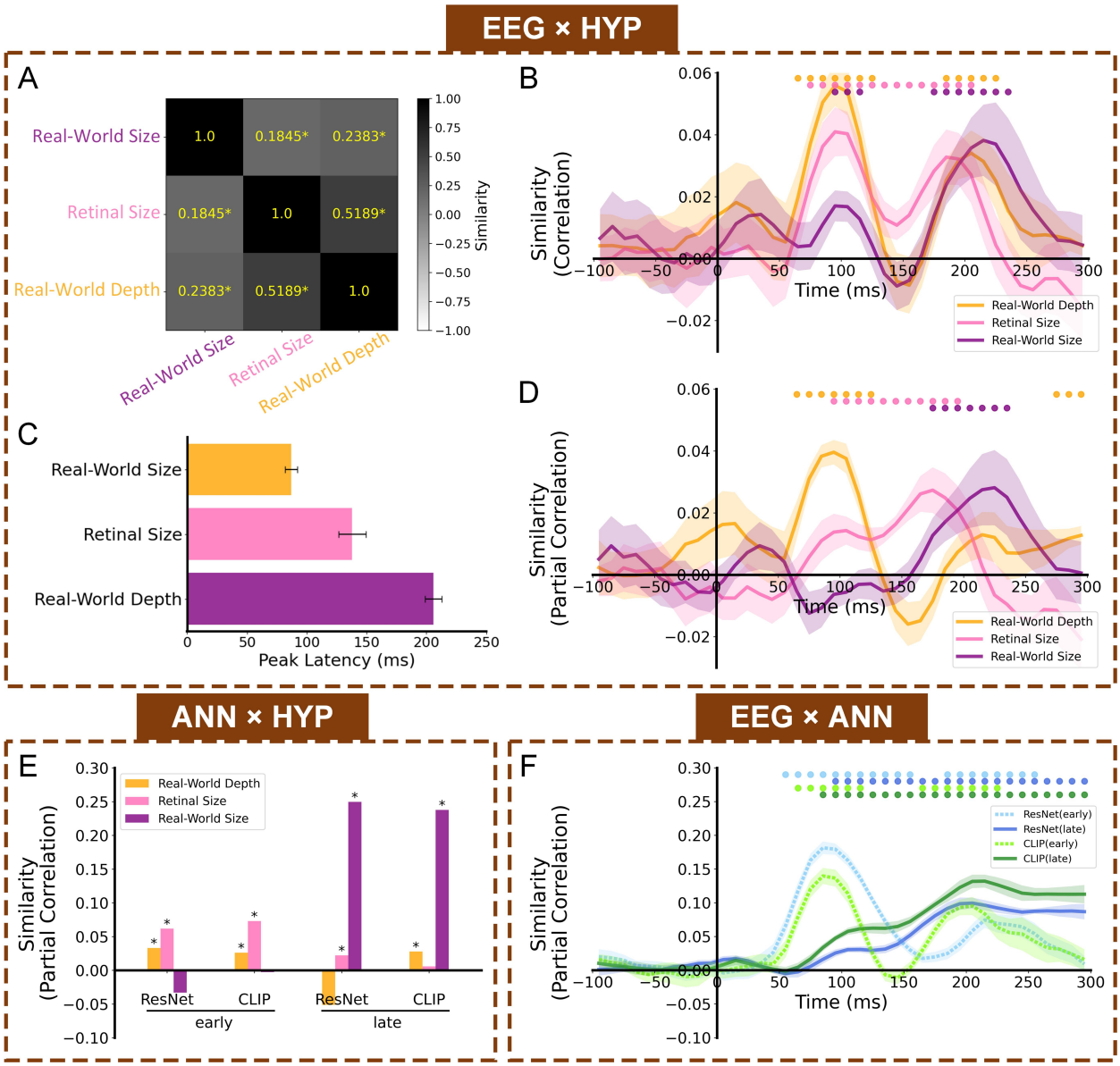

Figure 3 Cross-modal RSA results. (A) Similarities (Spearman correlations) between three hypothesis-based RDMs. Asterisks indicate a significant similarity, p<.05. (B) Representational similarity time courses (full Spearman correlations) between EEG neural RDMs and hypothesis-based RDMs. (C) Temporal latencies for peak similarity (partial Spearman correlations) between EEG and the 3 types of object information. Error bars indicate ±SEM. Asterisks indicate significant differences across conditions (p<.05); (D) Representational similarity time courses (partial Spearman correlations) between EEG neural RDMs and hypothesis-based RDMs. (E) Representational similarities (partial Spearman correlations)

between the four ANN RDMs and hypothesis-based RDMs of real-world depth, retinal size, and real-world size. Asterisks indicate significant partial correlations (bootstrap test, $p<.05$). (F) Representational similarity time courses (Spearman correlations) between EEG neural RDMs and ANN RDMs. Color-coded small dots at the top indicate significant timepoints (cluster-based permutation test, $p<.05$). Shaded area reflects ±SEM.

Peak latency results showed that neural representations of real-world size, retinal size and real-world depth reached their peaks at different latencies after stimulus onset (real-world depth: ~87ms, retinal size: ~138ms, real-world size: ~206ms, Figure 3C). The representation of real-word size had a significantly later peak latency than that of both retinal size ($t=4.2950$, $p=.0020$) and real-world depth ($t=18.5847$, $p<.001$). And retinal size representation had a significantly later peak latency than real-world depth ($t=3.7229$, $p=.0047$). These varying peak latencies imply an encoding order for distinct visual features, transitioning from real-world depth through retinal size, and then to real-world size.

Artificial neural networks also reflect distinct representations of object size and depth

To test how ANNs process these visual properties, we input the same stimulus images into ANN models and got their latent features from early and late layers (Figure 2C), and then conducted comparisons between the ANN RDMs and hypothesis-based RDMs. Parallel to our findings of dissociable representations of real-world size, retinal size, and real-world depth in the human brain signal, we also found dissociable representations of these visual features in ANNs (Figure 3E). Our partial correlation RSA analysis showed that early layers of both ResNet and

CLIP had significant real-world depth and retinal size representations, whereas the late layers of both ANNs were dominated by real-world size representations, though there was also weaker retinal size representation in the late layer of ResNet and real-world depth representation in the late layer of CLIP. The detailed statistical results are shown in Table S1.

Thus, ANNs provide another approach to understand the formation of different visual features, offering convergent results with the EEG representational analysis, where retinal size was reflected most in the early layers of ANNs, while object real-world size representations didn't emerge until late layers of ANNs, consistent with a potential role of higher-level visual information, such as the semantic information of object concepts.

Finally, we directly compared the timepoint-by-timepoint EEG neural RDMs and the ANN RDMs (Figure 3F). The early layer representations of both ResNet and CLIP were significantly correlated with early representations in the human brain (early layer of ResNet: 40-280ms, early layer of CLIP: 50-130ms and 160-260ms), while the late layer representations of two ANNs were significantly correlated with later representations in the human brain (late layer of ResNet: 80-300ms, late layer of CLIP: 70-300ms). This pattern of early-to-late correspondence aligns with previous findings that convolutional neural networks exhibit similar hierarchical representations to those in the brain visual cortex (Cichy et al., 2016; Güçlü & van Gerven, 2015; Kietzmann et al., 2019; Yamins & DiCarlo, 2016): that both the early stage of brain processing and the early layer of the ANN encode lower-level visual information, while the late stage of the brain and the late layer of the ANN encode higher-level visual information. Also, human brain representations showed a higher similarity to the early layer representation of the visual model (ResNet) than to the visual-semantic model (CLIP) at an early stage. Conversely, human brain representations showed a higher similarity to the late layer

representation of the visual-semantic model (CLIP) than the visual model (ResNet) at a late stage. Interestingly, the peaks of significant time windows for the EEG × HYP RSA also correspond with the peaks of the EEG × ANN RSA timecourse (Figure 3D,F).

Real-world size as a stable and higher-level dimension in object space

An important aspect of the current study is the use of naturalistic visual images as stimuli, in which objects were presented in their natural contexts, as opposed to cropped images of objects without backgrounds. In natural images, background can play an important role in object perception. How dependent are the above results on the presence of naturalistic background context? To investigate how image context influences object size and depth representations, we next applied a reverse engineering method, feeding the ANNs with modified versions of the stimulus images containing cropped objects without background, and evaluating the ensuing ANN representations compared to the same original hypothesis-based RDMs. If the background significantly contributes to the formation of certain feature representations, we may see some encoding patterns in ANNs disappear when the input only includes the pure object but no background.

Compared to results based on images with background, the ANNs based on cropped-object modified images showed weaker overall representational similarity for all features (Figure 4). In the early layers of both ANNs, we now only observed significantly preserved retinal size representations (which is a nice validity check, since retinal size measurements were based purely on the physical object dimensions in the image, independent of the background). Real-world depth representations were almost totally eliminated, with only a small effect in the late

layer of ResNet. However, we still observed a preserved pattern of real-world size representations, with significant representational similarity in the late layers of both ResNet and CLIP, and not in the early layers. The detailed statistical results are shown in Table S2. Even though the magnitude of representational similarity for object real-world size decreased when we removed the background, this high-level representation was not entirely eliminated. This finding suggests that background information does indeed influence object processing, but the representation of real-world size seems to be a relatively stable higher-level feature. On the other hand, representational formats of real-world depth changed when the input lacked background information. The deficiency of real-world depth representations in early layers, compared to when using full-background images, might suggest that the human brain typically uses background information to estimate object depth, though the significant effect in the late layer of ResNet in background-absent condition might also suggest that the brain (or at least ANN) has additional ability to integrate size information to infer depth when there is no background.

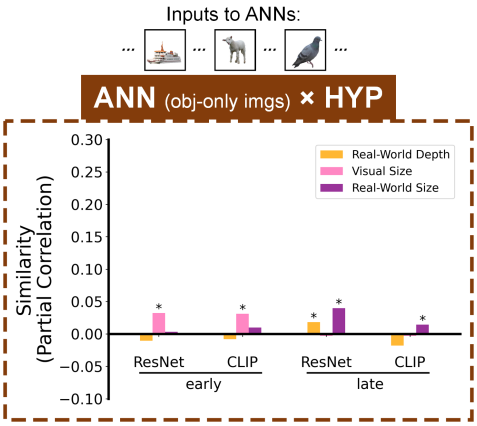

Figure 4 Contribution of image backgrounds to object size and depth representations. Representational similarity results (partial Spearman correlations) between ANNs fed inputs of cropped object images

without backgrounds and the hypothesis-based RDMs. Stars above bars indicate significant partial correlations (bootstrap test, $p<.05$).

The above results reveal that real-world size emerges with later peak neural latencies and in the later layers of ANNs, regardless of image background information. Is this because real-world size is a more conceptual-level dimension in object semantic space? If so, we might expect it to be driven not only by higher-level visual information, but also potentially by purely semantic information about familiar objects. To test this, we extracted object names from each image and input the object names into a Word2Vec model to obtain a Word2Vec RDM (Figure S2), and then conducted a partial correlation RSA comparing the Word2Vec representations with the hypothesis-based RDMs (Figure 5A). The results showed a significant real-world size representation ($r=0.1871$, $p<0.001$) but no representation of retinal size ($r=-0.0064$, $p=0.8148$) or real-world depth ($r=-0.0040$, $p=.7151$) from Word2Vec. Also, the significant time-window (90-300ms) of similarity between Word2Vec RDM and EEG RDMs (Figure 5B) contained the significant time-window of EEG x real-world size representational similarity (Figure 3B).

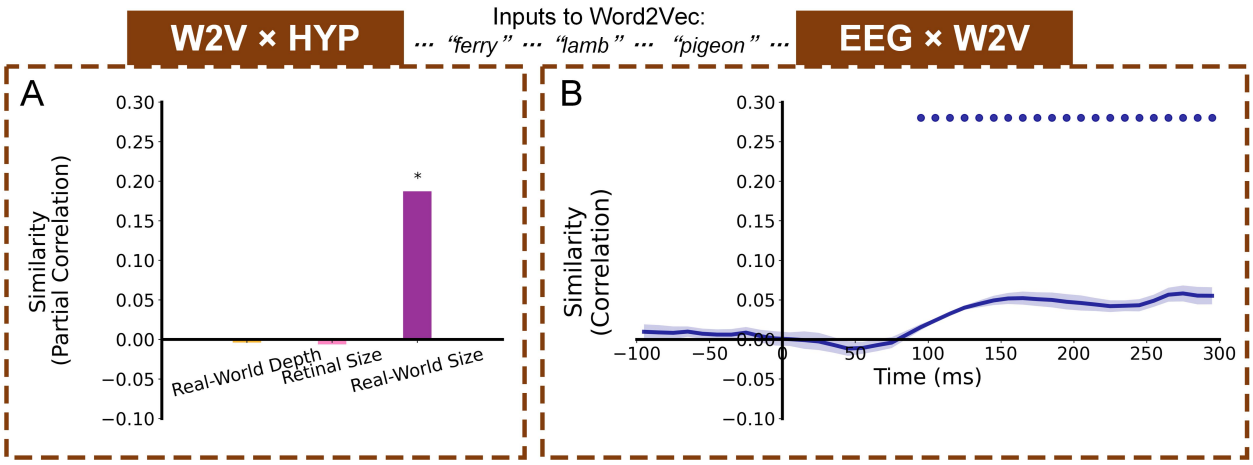

Figure 5 Representation similarity with a non-visual semantic language processing model (Word2Vec) fed word inputs corresponding to the images' object concepts. (A) Representational similarity results (partial Spearman correlations) between Word2Vec RDM and hypothesis-based RDMs. Stars above bars indicate significant partial correlations (bootstrap test, p<.05). (B) Representational similarity time course (Spearman correlations) between EEG RDMs (neural activity while viewing images) and Word2Vec RDM (fed corresponding word inputs). Color-coded small dots at the top indicate significant timepoints (cluster-based permutation test, p<.05). Line width reflects ±SEM.

Both the reverse engineering manipulation and Word2Vec findings corroborate that object real-world size representation, unlike retinal size and real-world depth, emerges in both image- and semantic-level in object space.

Discussion

Our study applied computational methods to distinguish the representations of objects' perceived real-world size, retinal size, and perceived real-world depth features in both human brains and ANNs. Consistent with prior studies reporting real-world size representations (Huang et al., 2022;

S.-M. Khaligh-Razavi et al., 2018; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b; Luo et al., 2023; Quek et al., 2023; R. Wang et al., 2022), we found that both human brains and ANNs contain significant information about real-world size. Critically, compared to the prior studies, our study offers several important theoretical and methodological advances: (a) we eliminated the confounding impact of perceived real-world depth (in addition to retinal size) on the real-world size representation; (b) we conducted analyses based on more ecologically valid naturalistic images; (c) we obtained precise feature values for each object in every image, instead of simply dividing objects into 2 or 7 coarse categories; and (d) we utilized a multi-modal, partial correlation RSA that combines EEG, hypothesis-based models, and ANNs. This novel approach allowed us to investigate representational time courses and reverse engineering manipulations in unparalleled detail. By integrating EEG data with hypothesis-based models and ANNs, this method offers a powerful tool for dissecting the neural underpinnings of object size and depth perception in more ecological contexts, which enriches our comprehension of the brain's representational mechanisms.

Using EEG we uncovered a representational timeline for visual object processing, with object real-world depth information represented first, followed by retinal size, and finally real-world size. While size and depth are highly correlated to each other, our results suggest that the human brain indeed has dissociated time courses and mechanisms to process them. The later representation time-window for object real-world size may suggest that the brain requires more sophisticated, higher-level information to form this representation, perhaps incorporating semantic and/or memory information about familiar objects, which was corroborated by our ANN and Word2Vec analyses. These findings also align with a recent fMRI study (Luo et al., 2023) using natural images to explore the neural selectivity for real-world size, finding that low-level visual

information could hardly account for neural size preferences, although that study did not consider covariables like retinal size and real-world depth.

In contrast to the later-emerging real-world size representations, it makes sense that retinal size representations could be processed more quickly based on more fundamental, lower-level information such as shape and edge discrimination. The intermediate latency for real-world depth processing suggests that this feature may precede real-world size processing. Additionally, there was a secondary, albeit substantially later, significant depth representation time-window, which might indicate that our brains also have the ability to integrate object retinal size and higher-level real-size information to form the final representation of real-world depth. Our comparisons between human brains and artificial models and explorations on ANNs and Word2Vec offer further insights and suggest that although real-world object size and depth are closely related, object real-world size appears to be a more stable and higher-level dimension.

The concept of 'object space' in cognitive neuroscience research is crucial for understanding how various visual features of objects are represented. Historically, various visual features have been considered important dimensions in constructing object space, including animate-inanimate (Kriegeskorte et al., 2008; Naselaris et al., 2012), spikiness (Bao et al., 2020; Coggan & Tong, 2023), and physical appearance (Edelman et al., 1998). In this study, we focus on one particular dimension, real-world size (Huang et al., 2022; Konkle & Caramazza, 2013; Konkle & Oliva, 2012b). How we generate neural distinctions of different object real-world size and where this ability comes from remain uncertain. Some previous studies found that object shape rather than texture information could trigger neural size representations (Huang et al., 2022; Long et al., 2016, 2018; R. Wang et al., 2022). Our results attempt to further advance their findings, that

object real-world size is a stable and higher-level dimension substantially driven by object semantics in object space.

Increasingly, research has begun to use ANNs to study the mechanisms of object recognition (Ayzenberg et al., 2023; Cichy & Kaiser, 2019; Doerig et al., 2023; Kanwisher et al., 2023). We can explore how the human brain processes information at different levels by comparing brain activity with models (Cichy et al., 2016; S. M. Khaligh-Razavi & Kriegeskorte, 2014; Kuzovkin et al., 2018; Xie et al., 2020), and we can also analyze the representation patterns of the models with some specific manipulations and infer potential processing mechanisms in the brain (Golan et al., 2020; Huang et al., 2022; Lu & Ku, 2023; Xu et al., 2021). In current study, our comparisons result between EEG signals and different ANNs showed that the visual model's early layer had a higher similarity to the brain in the early stage, while the visual-semantic model's late layer had a higher similarity to the brain in the late stage. However, for the representation of objects, partial correlation results for different ANNs didn't demonstrate the superiority of the multi-modal model at late layers. This might be due to models like CLIP, which contain semantic information, learning more complex image descriptive information (like the relationship between object and the background in the image). Real-world size might be a semantic dimension of the object itself, and its representation does not require overall semantic descriptive information of the image. In contrast, retinal size and real-world depth could rely on image background information for estimation, thus their representations in the CLIP late layer disappeared when input images had only pure object but no background.

Building on the promising findings of our study, future work may further delve into the detailed processes of object processing and object space. One important problem to solve is how real-world size interacts with other object dimensions in object space. In addition, our approach

could be used with future studies investigating other influences on object processing, such as how different task conditions impact and modulate the processing of various visual features.

Moreover, we must also emphasize that in this study, we were concerned with perceived real-world size and depth reflecting a perceptual estimation of our world, which are slightly different from absolute physical size and depth. The differences in brain encoding between perceived and absolute physical size and depth require more comprehensive measurements of an object's physical attributes for further exploration. Also, we focused on perceiving depth and size from 2D images in this study, which might have some differences in brain mechanism compared to physically exploring the 3D world. Nevertheless, we believe our study offers a valuable contribution to object recognition, especially the encoding process of object real-world size in natural images.

In conclusion, we used computational methods to distinguish the representations of real-world size, retinal size, and real-world depth features of objects in ecologically natural images in both human brains and ANNs. We found an unconfounded representation of object real-world size, which emerged at later time windows in the human EEG signal and at later layers of artificial neural networks compared to real-world depth, and which also appeared to be preserved as a stable dimension in object space. Thus, although size and depth properties are closely correlated, the processing of perceived object size and depth may arise through dissociated time courses and mechanisms. Our research provides a detailed and clear characterization of the object processing process, which offers further advances and insights into our understanding of object space and the construction of more brain-like visual models.

Acknowledgments

This work was supported by research grants from the National Institutes of Health (R01-EY025648) and from the National Science Foundation (NSF 1848939) to JDG. The authors declare no competing financial interests.

References

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

Ayzenberg, V., Blauch, N., & Behrmann, M. (2023). Using deep neural networks to address the how of object recognition. *PsyArXiv*.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108.

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

Bracci, S., & Op de Beeck, H. (2016). Dissociations and associations between shape and

category representations in the two visual pathways. *Journal of Neuroscience*, *36*(2), 432–444.

Chen, Y. C., Deza, A., & Konkle, T. (2022). How big should this object be? Perceptual influences on viewing-size preferences. *Cognition*, *225*, 105114.

Choksi, B., Mozafari, M., VanRullen, R., & Reddy, L. (2022). Multimodal neural networks better explain multivoxel patterns in the hippocampus. *Neural Networks*, *154*, 538–542.

Choksi, B., Vanrullen, R., & Reddy, L. (2022, August 25). Do multimodal neural networks better explain human visual representations than vision-only networks? *Conference on Cognitive Computational Neuroscience 2022*.

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 1–13.

Coggan, D. D., & Tong, F. (2023). Spikiness and animacy as potential organizing principles of human ventral visual cortex. *Cerebral Cortex*, *33*(13), 8194–8217.

Conwell, C., Prince, J. S., Alvarez, G. A., & Konkle, T. (2022). Large-Scale Benchmarking of Diverse Artificial Vision Models in Prediction of 7T Human Neuroimaging Data. *BioRxiv*.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *ArXiv*.

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., Kording, K., Konkle, T., Van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*, 431–450.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309–321.

Gifford, A. T., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, *264*, 119754.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(47), 29330–29337.

Grootswagers, T., Zhou, I., Robinson, A. K., Hebart, M. N., & Carlson, T. A. (2022). Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, *9*(1), 1–7.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., &

Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE*, *14*(10), 1–24.

Huang, T., Song, Y., & Liu, J. (2022). Real-world size of objects serves as an axis of object space. *Communications Biology*, *5*(1), 1–12.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron*, *76*(6), 1210–1224.

Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N., & Mur, M. (2023). Deep Neural Networks and Visuo-Semantic Models Explain Complementary Components of Human Ventral-Stream Representational Dynamics. *Journal of Neuroscience*, *43*(10), 1731–1741.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254.

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.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N.

(2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(43), 21854–21863.

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. (2011). Canonical Visual Size for Real-World Objects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 23–37.

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*(3), 561.

Konkle, T., & Oliva, A. (2012b). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, *74*(6), 1114–1124.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141.

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J. P., Baciu, M., Kahane, P., Rheims, S., Vidal, J. R., & Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications Biology*, *1*(1), 1–12.

Long, B., & Konkle, T. (2017). A familiar-size Stroop effect in the absence of basic-level recognition. *Cognition*, *168*, 234–242.

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*(1), 95–109.

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 of the United States of America*, *115*(38), E9015–E9024.

Lu, Z., & Ku, Y. (2020). NeuroRA: A Python Toolbox of Representational Analysis From Multi-Modal Neural Data. *Frontiers in Neuroinformatics*, *14*, 61.

Lu, Z., & Ku, Y. (2023). Bridging the gap between EEG and DCNNs reveals a fatigue mechanism of facial repetition suppression. *IScience*, *26*(12), 108501.

Luo, A. F., Wehbe, L., Tarr, M. J., & Henderson, M. M. (2023). Neural Selectivity for Real-World Object Size In Natural Images Abbreviated title : Neural Selectivity for Real-World Size. *BioRxiv*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Muttenthaler, L., & Hebart, M. N. (2021). THINGSvision: A Python Toolbox for Streamlining the Extraction of Activations From Deep Neural Networks. *Frontiers in Neuroinformatics*, *15*, 45.

Naselaris, T., Stansbury, D. E., & Gallant, J. L. (2012). Cortical representation of animate and

inanimate objects in complex natural scenes. *Journal of Physiology Paris*, *106*(5–6), 239–249.

Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling Representations of Object Shape and Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of Cognitive Neuroscience*, *28*(5), 680–692.

Quek, G., Theodorou, A., & Peelen, M. V. (2023). Better together : Objects in familiar constellations evoke high-level representations of real-world size. *BioRxiv*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the International Conference on Machine Learning (ICML)*.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Setti, A., Caramelli, N., & Borghi, A. M. (2008). Conceptual information about size of objects in nouns. *European Journal of Cognitive Psychology*, *21*(7), 1022–1044.

Stoinski, L. M., Perkuhn, J., & Hebart, M. N. (2023). THINGSplus: New norms and metadata for the THINGS database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*, 1–21.

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

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2022). Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*.

Wang, R., Janini, D., & Konkle, T. (2022). Mid-level feature differences support early animacy and object size distinctions: Evidence from electroencephalography decoding. *Journal of Cognitive Neuroscience*, *34*(9), 1670–1680.

Xie, S., Kaiser, D., & Cichy, R. M. (2020). Visual Imagery and Perception Share Neural Representations in the Alpha Frequency Band. *Current Biology*, *30*(13), 2621-2627.e5.

Xu, S., Zhang, Y., Zhen, Z., & Liu, J. (2021). The Face Module Emerged in a Deep Convolutional Neural Network Selectively Deprived of Face Experience. *Frontiers in Computational Neuroscience*, *15*, 1–12.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 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. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624.

Supplementary materials

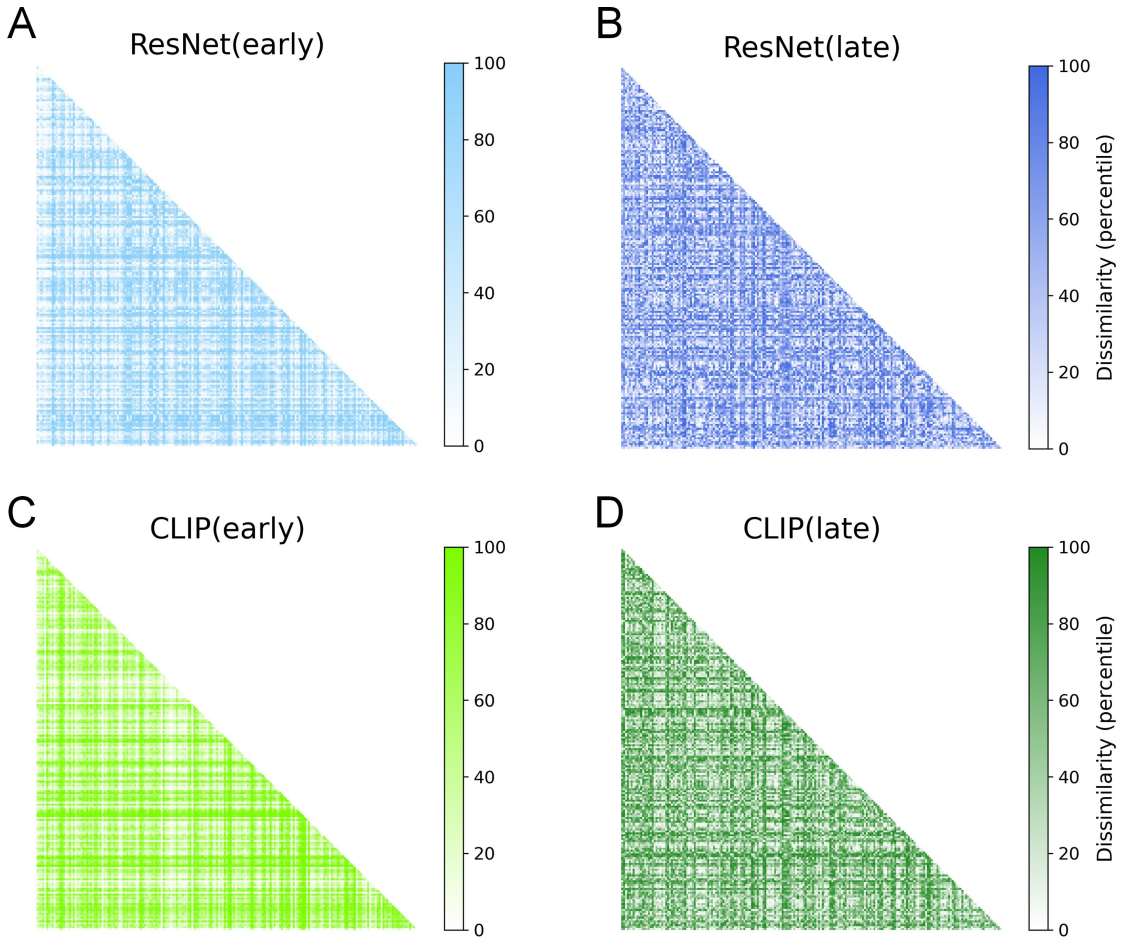

Figure S1. Four ANN RDMs of ResNet early layer, ResNet late layer, CLIP early layer, and CLIP late later.

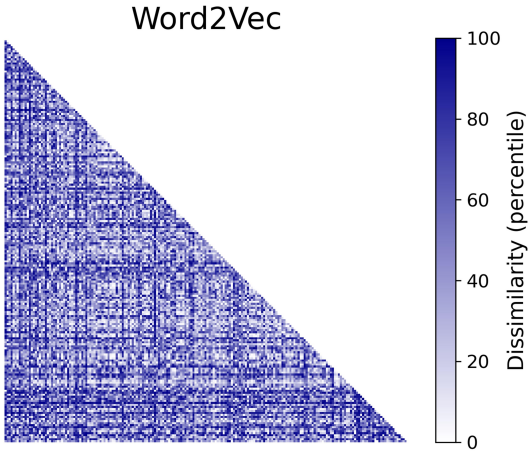

Figure S2. Word2Vec RDMs.

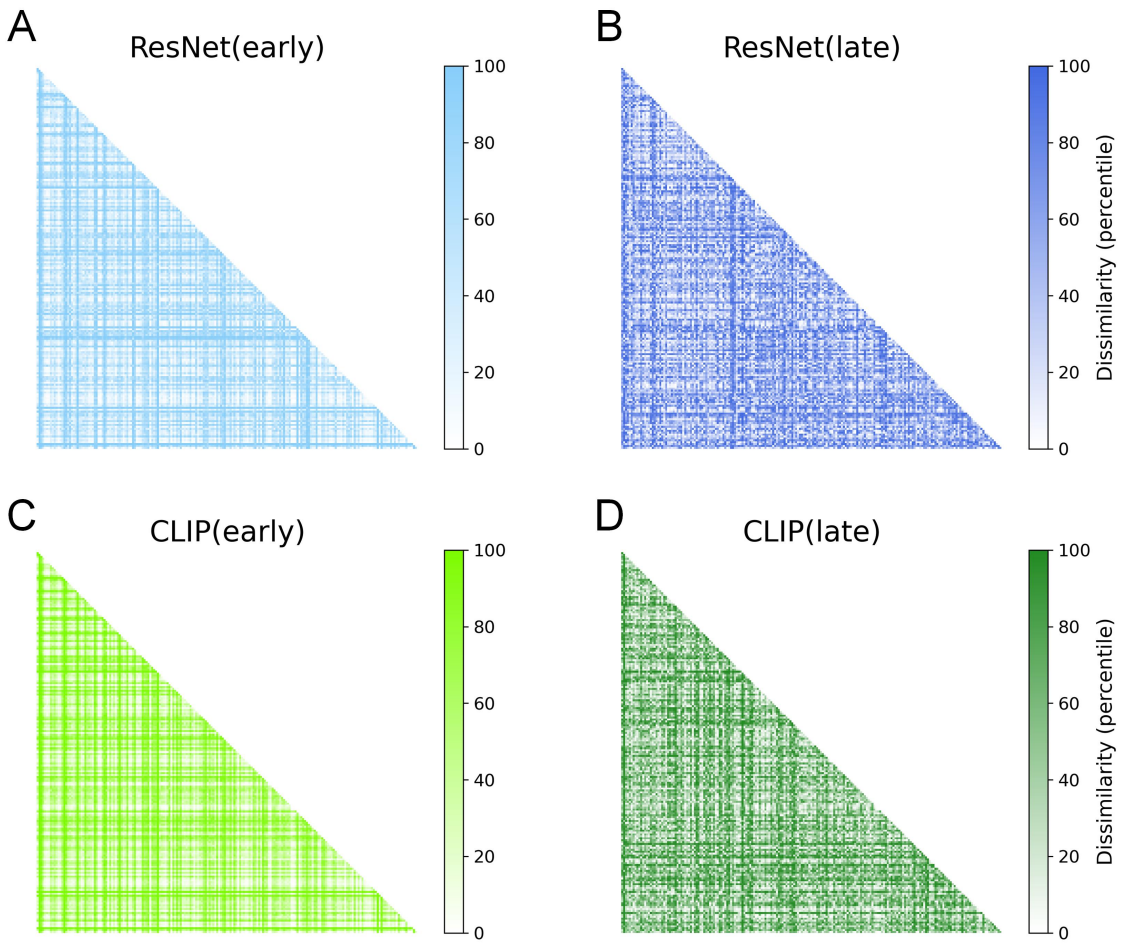

Figure S3. Four ANN RDMs with inputs of cropped object images without background of ResNet early layer, ResNet late layer, CLIP early layer, and CLIP late later.

Table S1 Statistical results of similarities (partial Spearman correlations) between four ANN RDMs and three hypothesis-based RDMs.

ANN × HYP	***ResNet (early)***	***CLIP (early)***	***ResNet (late)***	***CLIP (late)***
Real-World depth	**r=.0330, p<.001**	**r=.0262, p<.001**	r=-.0513, p=1	**r=.0278, p<.001**
Retinal Size	**r=.0618, p<.001**	**r=.0730, p<.001**	**r=.0221, p<.001**	r=.0058, p=.1788
Real-World Size	r=-.0330, p=1	r=-.0027, p=.8710	**r=0.2497, p<.001**	**r=0.2378, p<.001**

Table S2 Statistical results of similarities (partial Spearman correlations) between four ANN RDMs with inputs of cropped object images without background and three hypothesis-based RDMs.

ANN (obj-only imgs) × HYP	***ResNet (early)***	***CLIP (early)***	***ResNet (late)***	***CLIP (late)***
Real-World depth	r=-.0109, p=.9382	r=-.0086, p=.8862	**r=.0183, p=.0049**	r=-.0176, p=.9934
Retinal Size	**r=.0323, p<.001**	**r=.0315, p<.001**	r=-.0032, p=.6725	r=-.0031, p=.6705
Real-World Size	r=.0022, p=.3781	r=.0084, p=.1193	**r=.0402, p<.001**	**r=.0154, p=.0149**