

OPEN

DATA DESCRIPTOR

# BOLD5000, a public fMRI dataset while viewing 5000 visual images

Nadine Chang<sup>1</sup>, John A. Pyles<sup>1</sup>, Austin Marcus<sup>1</sup>, Abhinav Gupta<sup>1</sup>, Michael J. Tarr<sup>1</sup> & Elissa M. Aminoff<sup>1,2</sup> 

Received: 13 September 2018

Accepted: 18 March 2019

Published online: 06 May 2019

Vision science, particularly machine vision, has been revolutionized by introducing large-scale image datasets and statistical learning approaches. Yet, human neuroimaging studies of visual perception still rely on small numbers of images (around 100) due to time-constrained experimental procedures. To apply statistical learning approaches that include neuroscience, the number of images used in neuroimaging must be significantly increased. We present BOLD5000, a human functional MRI (fMRI) study that includes almost 5,000 distinct images depicting real-world scenes. Beyond dramatically increasing image dataset size relative to prior fMRI studies, BOLD5000 also accounts for image diversity, overlapping with standard computer vision datasets by incorporating images from the Scene UNDERstanding (SUN), Common Objects in Context (COCO), and ImageNet datasets. The scale and diversity of these image datasets, combined with a slow event-related fMRI design, enables fine-grained exploration into the neural representation of a wide range of visual features, categories, and semantics. Concurrently, BOLD5000 brings us closer to realizing Marr's dream of a singular vision science—the intertwined study of biological and computer vision.

## Background & Summary

Both human and computer vision share the goal of analyzing visual inputs to accomplish high-level tasks such as object and scene recognition<sup>1</sup>. Dramatic advances in computer vision have been driven over the past few years by massive-scale image datasets such as ImageNet<sup>2</sup> and COCO<sup>3</sup>. In contrast, the number of images used in most studies of biological vision remains extremely small. Under the view that future progress in vision science will rely on the integration of computational models and the neuroscience of vision, this difference in the number of stimuli used in each domain presents a challenge. How do we scale neural datasets to be commensurate with state-of-the-art computer vision, encompassing the wide range of visual inputs we encounter in our day-to-day lives?

The motivation to scale up neural datasets is to leverage and combine the wide variety of technological advances that have enabled significant, parallel progress in both biological and machine vision. Within biological vision, innovations in neuroimaging have advanced both the measurement and analysis of brain activity, enabling finer-scale study of how the brain responds to, processes, and represents visual inputs. Despite these advances, the relationship between the content of visual input from our environment and specific brain responses remains an open question. In particular, mechanistic descriptions of high-level visual processes are difficult to interpret from complex patterns of neural activity.

In an attempt to understand complex neural activity arising from advanced neuroimaging techniques, high-performing computer vision systems have been touted as effective potential models of neural computation<sup>1</sup>. This is primarily for three reasons: (1) the origin of these models is linked to the architecture of the primate visual system<sup>4</sup>; (2) these models learn from millions of real-world images; (3) these models achieve high-performance in diverse tasks such as scene recognition, object recognition, segmentation, detection, and action recognition—tasks defined and grounded in *human* judgments of correctness. In this context, a variety of labs have attempted to understand neural data by focusing on the feed-forward hierarchical structure of deep “convolutional” neural networks (CNNs) trained on millions of visual images<sup>4</sup>. Given a network trained on a dataset for a specific task (e.g., object categorization), one can, for a particular brain area: (1) compare the representational properties of different network layers to the representational properties of that neural region; or, (2) use network layers weights to predict neural responses at the voxel (fMRI) or neuron (neurophysiology) level within that neural region. Demonstrating the efficacy of this approach, recent studies have found that higher-level layers of CNNs tend to

<sup>1</sup>Carnegie Mellon University, Pittsburgh, USA. <sup>2</sup>Fordham University, Bronx, USA. Correspondence and requests for materials should be addressed to E.M.A. (email: [eaminoff@fordham.edu](mailto:eaminoff@fordham.edu))

predict the neural responses of higher-level object and scene regions<sup>5,6</sup>. Similarly, CNNs have been found to better model human dynamics underlying scene representation<sup>7</sup> as compared to more traditional models of scene and object perception (e.g., GIST<sup>8</sup> or HMAX<sup>9,10</sup>).

More broadly, there is growing acceptance that computer vision models such as CNNs are useful for understanding biological vision systems and phenomena<sup>11</sup>. However, computer vision models themselves are still far from approaching human performance and robustness, and there is a growing belief that successful understanding of biological vision will lead to improved computer vision models. As such, integration in both directions will lead to an intertwined approach to biological and computer vision that dates back to Marr<sup>12</sup>. Indeed, as both models and measurement techniques progress, we come closer to this ideal. However, one of the most significant outstanding challenges for integrating across fields is data<sup>13</sup>.

We address this data challenge with the BOLD5000 dataset, a large-scale, slow event-related human fMRI study incorporating 5,000 real-world images as stimuli. BOLD5000 is an order of magnitude larger than any extant slow event-related fMRI dataset, with ~20 hours of MRI scanning per each of four participants. By scaling the size of the image dataset used in fMRI, we hope to facilitate greater integration between the fields of human and computer vision. To that end, BOLD5000 also uniquely uses images drawn from the three most commonly-used computer vision datasets: SUN, COCO, and ImageNet. Beyond standard fMRI analysis techniques, we use both representational similarity analysis<sup>14</sup> and, uniquely, t-distributed stochastic neighbor embedding visualizations<sup>15</sup>, to validate the quality of our data. In sum, we hope that BOLD5000 engenders greater collaboration between the two fields of vision science, fulfilling Marr's dream.

## Methods

**Participant Selection.** Participants were recruited from the pool of students enrolled in graduate school at Carnegie Mellon University. Due to the multi-session nature of the study, we specifically recruited participants who were familiar with MRI procedures and believed themselves capable of completing all scanning sessions with a minimum effect on data quality (e.g., low movement, remaining awake, etc.). This requirement necessarily limited the number of individuals willing and capable of participating in this study. In this context, our study included four participants: three participants (CSI1, CSI2, CSI3) with a full set of data (all 16 sessions), and a fourth participant (CSI4) who completed 10 sessions due to discomfort in the MRI. Participant demographics are as follows: CSI1 - male, age 27, right handed; CSI2 - female, age 26, right handed; CSI3 - female, age 24, right handed; CSI4 - female, age 25, right handed. Participants all reported no history of psychiatric or neurological disorders and no current use of any psychoactive medications. Participants all provided written informed consent and were financially compensated for their participation. All procedures followed the principles in the Declaration of Helsinki and were approved by the Institutional Review Board of Carnegie Mellon University.

**Stimuli.** *Stimulus Selection.* There are two perspectives to consider for data sharing across biological and computer vision. First, for computer vision, what types of neural data will provide insight or improvement in computer vision systems? Second, for biological vision, what types of images will elicit the best neural data for modeling and understanding neural responses? In this larger context, we suggest that three critical data (i.e., image set) considerations are necessary for success.

The first data consideration is *size*. The general success of modern artificial neural networks can be largely attributed to large-scale datasets<sup>2</sup>. High-performing models are trained and evaluated on a variety of standard large-scale image datasets. In contrast, although models trained on large-scale datasets have been applied to neural data, the set of images used in these neural studies is significantly smaller—at best, on the order of a hundred or so distinct images due to time-constrained experimental procedures.

The second data consideration is *diversity*. Stimulus sets of constrained size translate into limited diversity for the set of images: the images used in almost every neural study encompass only a small subset of the entire natural image space. For example, despite many studies focusing on object recognition<sup>16</sup>, few experiments exceed more than 100 categories across all stimuli. In contrast, image datasets used to train and evaluate artificial neural networks contain thousands of categories, covering a wide range of natural images.

The third data consideration is *image overlap*. While multiple recent studies have applied artificial neural networks to the understanding of neural data, the stimulus images used across the two domains have rarely been commensurate with one another. Most significantly, many neural studies use stimuli depicting single objects centered against white backgrounds. In contrast, the images comprising most computer vision datasets contain non-centered objects embedded in realistic, complex, noisy scenes with semantically-meaningful backgrounds. In the instances where studies of biological vision have included more complex images, such as natural scenes, the images were rarely drawn from the same image datasets used in training and testing computer vision models. This lack of overlap across stimuli handicaps the ability to (1) compare neural data and model representations of visual inputs and (2) utilize neural data in network training or design.

BOLD5000 tackles these three concerns in an unprecedented slow event-related fMRI study that includes almost 5,000 distinct images. BOLD5000 addresses the issue of size by dramatically increasing the image dataset size relative to nearly all extant human fMRI studies—scaling up by over an order of magnitude. Of note, although several earlier studies did collect fMRI data while participants viewed real-world movie stimuli<sup>17,18</sup>, effectively showing thousands of images to participants, these images were associated with events that were necessarily overlapping in time, so no slow event-related analyses were possible. As such, analyzing this kind of “large-scale” fMRI data is challenging with respect to disentangling which stimuli gave rise to which neural responses.

BOLD5000 addresses the issues of data diversity and image overlap by including stimulus images drawn from three standard computer vision datasets: scene images from categories largely inspired by Scene UNderstanding (SUN)<sup>19</sup>; images directly from Common Objects in Context (COCO)<sup>3</sup>; and, images directly from ImageNet<sup>2</sup>. SUN, COCO, and ImageNet, respectively, cover the following image domains: real-world indoor and outdoor

scenes; objects interacting in complex, real-world scenes; and objects centered in real-world scenes. These three image datasets encompass an extremely broad variety of image types and categories, thereby enabling fine-grained exploration into the neural representation of visual inputs across a wide range of visual features, categories, and semantics.

By including images from computer vision datasets, we also address two issues in the computer vision field. Firstly, the datasets we used as well as other computer vision datasets are almost all created by automatically “scraping” the web for images, which are then de-noised by humans. As a consequence, these “standard images” for computer vision have not actually been validated as natural or representative of the visual world around us—instead, the generality of CNNs and other modern computer vision models rests on a massive amount of training data, typically covering much of what a model is likely to be tested on in the future. In other words, there is little human behavior or neural data on how such images are processed, perceived, or interpreted. BOLD5000 addresses this by providing some metric—in this case, neural—of how specific images from these datasets are processed relative to one another. Secondly, the scale of BOLD5000 combined with its overlap with common computer vision datasets enables the possibility of machine learning models *training* directly on the neural data associated with images in the training set.

More generally, the scale, diversity, and slow event-related fMRI design of BOLD5000 enable, for the first time, much richer, joint artificial and biological vision models that are critically high-performing in both domains. Beyond the fact that large-scale neural (or behavioral) datasets are necessary for integrating across these two approaches to studying vision, it is also important to note that large-scale neural datasets are equally necessary, in and of themselves, for understanding how complex, real-world visual inputs are processed and represented in the primate brain. In this spirit, BOLD5000 is a publicly-available dataset available through [Figshare](#) (also via Carnegie Mellon’s Figshare portal - [Kilthub](#))<sup>20</sup>, and [OpenNeuro](#)<sup>21</sup>. More details can be found at [BOLD5000.org](#).

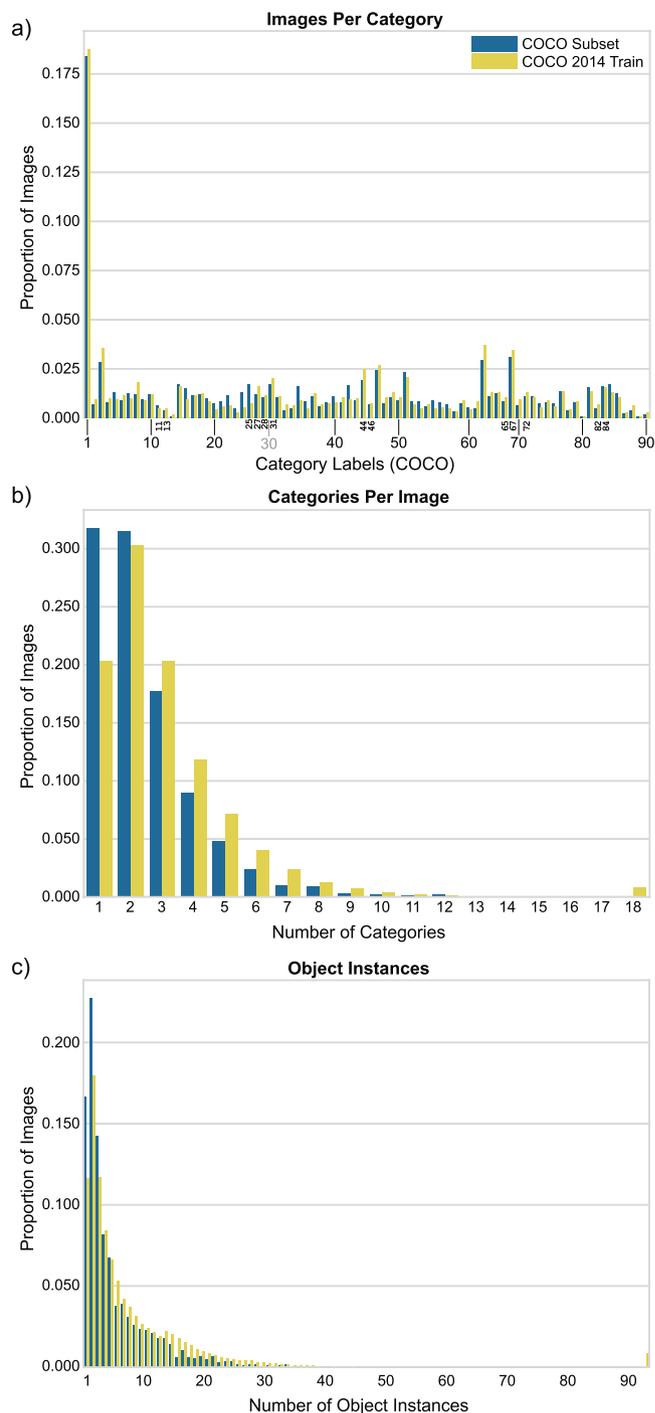
In detail, a total of 5,254 images, of which 4,916 images were unique, were used as the experimental stimuli in BOLD5000. Images were drawn from three particular computer vision datasets because of their prevalence in computer vision and the large image diversity they represent across image categories:

1. *Scenes*. 1,000 hand-curated indoor and outdoor scene images covering 250 categories. These particular 250 scene categories were selected so as to correspond to categories included in the SUN dataset—a standard image dataset for scene categorization tasks<sup>19</sup>. This set of 1,000 images tended to be more scenic, with less of a focus on any particular object, action, or person. We selected images depicting both outdoor (e.g., mountains) and indoor (e.g., restaurant) scenes.
2. *COCO*. 2,000 images of multiple objects from the COCO dataset—a standard benchmark for object detection tasks<sup>3</sup>. Due to the complexity of this task, COCO contains images that are similarly complicated and have multiple annotations. Objects tend to be embedded in a realistic context and are frequently shown as interacting with other objects—both inanimate and animate. COCO is unique because it includes images depicting basic human social interactions.
3. *ImageNet*. 1,916 images of mostly singular objects from the ImageNet dataset—a standard benchmark for object categorization tasks<sup>22</sup> that is also popular for pre-training CNNs and “deep” artificial neural networks<sup>23</sup>. Consistent with this task, ImageNet contains images that tend to depict a single object as the focus of the picture. Additionally, the object is often centered and clearly distinguishable from the image background.

**Stimulus Pre-Processing.** To improve the quality of our neural data, we emphasized image quality for our stimulus images by imposing several selection criteria. Basic image quality checks included image resolution, image size, image blurring, and a hard constraint requiring color images only. Additionally, to ensure that sequentially viewing images would not produce neural changes due to image size variation, all images were constrained to be square and of equal size.

To select the 1,000 scene images, we defined 250 unique scene categories (mostly from the SUN dataset) and set a goal of four exemplars per category. We opted not to use the SUN images directly due to a desire to increase the quality of the pictures chosen (i.e., with regard to resolution, watermarks present, etc.). Thus, for each scene category, we queried Google Search with the scene category name and selected images from among the top results according to the above criteria. We only selected images with sufficient size and resolution, then inspected each image to ensure it was clear and free of watermarks. All images were then downsampled to  $375 \times 375$  pixels, the final size for all stimulus images used in this study.

To select from the COCO images, we sampled 2,000 images from the complete COCO 2014 training dataset. Our goal was to ensure that our chosen images were an accurate representation of the original COCO dataset. Thus, our sampling was structured such that the procedure considered the various annotations that accompany each COCO image. COCO annotations contain 80 object class labels, number of object instances, bounding boxes, and segmentation polygons. Our final 2,000 images adhered to the following criteria: i) the number of images per category is proportional to that of the training set as shown in Fig. 1a; ii) the number of categories in our selected images is proportional to that of the training set as shown in Fig. 1b; iii) the number of instances per image is proportional to that of the training set as shown in Fig. 1c; iv) the final cropped images contain at least 70% of the original bounding boxes, where the boxes are counted if there is an intersection over union (area overlap) of at least 50% between the boxes and the cropped image; v) each image is larger than  $375 \times 375$  pixels. We went through several rounds of sampling, where in each round we randomly sampled according to the above-mentioned criteria before taking a  $375 \times 375$  center crop of the image. Due to the complex, realistic scenes depicted in COCO images, center crops often failed to contain the main image content. Thus, every center cropped image also underwent a manual inspection: if the center crop contained the relevant image content, the



**Fig. 1** Distributional statistics for BOLD5000 stimulus images selected from COCO 2014 training set as compared to the complete COCO 2014 training set. **(a)** The proportion of images per category for all categories. Note that the COCO category labels span from 1 to 90, but are not continuous. Discontinuities in the numerical COCO labels are indicated by explicit labeling of those categories (e.g., 11 13 denotes that the graph shows the proportion of images for categories 11 and 13, but that category 12 is not part of the set). **(b)** The proportion of images containing a specific number of categories. **(c)** The proportion of images containing a specific number of objects.

crop was retained; if the center crop did not contain the relevant image content, we selected a new region of the image from which to crop. If there was no reasonable crop region, the image was rejected. We repeated this process until 2,000 images had been selected.

To select from the ImageNet images, we used the standard 1,000 class categories in ImageNet for our image selection. However, due to the violent or strong negative nature of some images—possibly evoking emotional responses—we removed 42 categories. For each remaining ImageNet category, we randomly selected two exemplars per category from the training set that satisfied our image size and resolution criteria. With 958 categories

Participants	4
fMRI Sessions per Participants	16
Total fMRI Scene Runs Per Participant	142
Total Functional Localizer Runs Per Participant	8
Total Scene Trials Per Participant	5,254
Unique Scene Stimuli Per Participant	4,916

**Table 1.** A general design of the data.

and two exemplars per category, we obtained a total of 1,916 ImageNet images. In light of ImageNet images having varying sizes and resolutions, we only considered images that were larger than  $375 \times 375$  pixels before taking a  $375 \times 375$  center crop. For all randomly sampled center crops, we manually inspected each image to ensure that the crop did not exclude a large portion of the image content and that the image resolution was sufficiently high. We repeated this process until two exemplars per category had been selected.

Finally, we considered the RGB and luminance distribution across all of the selected images. Because visual brain responses are influenced by image luminance, we attempted to render our images invariant to this factor. To this end, for each image we calculated its hue, saturation, and value (HSV). This value represents the brightness of the image. The average brightness per image was determined, and we calculated the difference between average brightness and gray brightness. All values were then multiplied in the image by this new scale—a process known as gray world normalization. This process ensured that luminance was as uniform as possible across all images.

In order to examine the effect of image repetition, we randomly selected 112 of the 4,916 distinct images to be shown four times and one image to be shown three times to each participant. These 113 images were selected such that the image dataset breakdown was proportionally to that of the 4,916 distinct images. Specifically, 1/5 of the images were Scene images, 2/5 of the images were COCO images, 2/5 of the images were ImageNet images. When these image repetitions are considered, we have a total of 5,254 image presentations shown to each participant (4,803 distinct images +  $4 \times 112$  repeated images +  $3 \times 1$  repeated image). For CSI3 and CSI4, a small number of repetitions varied from 2–5 times.

## Experimental Design

**General Procedure.** As seen in Table 1, fMRI data were collected from a total of four participants (referred to as CSI1, CSI2, CSI3, CSI4), with a full dataset collected for three of the four participants (see Participant Selection). A full dataset was collected over 16 MRI scanning sessions, where 15 were functional sessions acquiring task-relevant data, and the remaining session comprised of collecting high resolution anatomical and diffusion data. For CSI4, there were 9 functional sessions, with an additional anatomical session. Participants were scanned using custom headcases (CaseForge Inc., Berkeley, CA) to reduce head movement and maintain consistent head placement and alignment across sessions.

For all participants, each of the functional scanning sessions was 1.5 hours long: 8 sessions had 9 image runs and 7 sessions had 10 image runs. In the sessions with only 9 image runs, we included an additional functional localizer run at the end of the session, providing a total of 8 localizer runs across 15 sessions. The functional localizer runs were used to independently define regions of interest for subsequent analyses. As summarized in Table 1 over the course of the 15 functional sessions, all 5,254 image trials (3,108 for CSI4) were presented. During each functional session 5 functional scans were acquired with less than a 1 minute break in between each scan. Field map scans were then collected over the course of approximately 4 minutes, during which the participants were allowed to close their eyes or view a movie. Finally, the last sequential 5 functional scans (or 4 if a localizer was run at the end of the session) were run. After all scans, the participant filled out a questionnaire (Daily Intake) about their daily routine, including: current status regarding food and beverage intake, sleep, exercise, ibuprofen, and comfort in the scanner.

**Experimental Paradigm for Scene Images.** The following run and session details apply for each participant. Each run contained 37 stimuli. In order for the images in a given run to accurately reflect the entire image dataset, the stimuli in each run were proportionally the same as the overall dataset: roughly 1/5<sup>th</sup> Scene images, 2/5<sup>th</sup> COCO images, and 2/5<sup>th</sup> ImageNet images. Of the 37 images in each run, roughly 2 were from the set of repeated images. For the 35 unique images per run, 7 were Scene images, 14 were COCO images, and 14 were ImageNet images. However, because the total number of images does not divide evenly into 7's, some sessions contained a slightly unbalanced number of image categories by a factor of 1 image. Finally, the presentation order of the stimulus images was randomly and uniquely determined for each participant. These presentations orders were fixed before the start of any experimental sessions for all four participants. All stimuli (both scenes and localizer images) were presented using Psychophysics Toolbox Version 3 (“Psychtoolbox”)<sup>24</sup> running under Matlab (Mathworks, Natick, MA) on a Systems76 computer running Ubuntu presented via 24-inch MR compatible LCD display (BOLDscreen, Cambridge Research Systems LTD., UK) mounted at the head end of the scanner bore. Participants viewed the display through a mirror mounted on the head coil. All stimulus images were  $375 \times 375$  pixels and subtended approximately 4.6 degrees of visual angle.

The following image presentation details apply for each run, each session, and each participant. A slow event-related design was implemented for stimulus presentation in order to isolate the blood oxygen level dependent (BOLD) signal for each individual image trial. At the beginning and end of each run, centered on a blank, black screen, a fixation cross was shown for 6 sec and 12 sec, respectively. Following the initial fixation cross, all 37

Task EPI	5,254 Scenes
	Scene Localizer
Other EPI	Opposite Phase Encoded
Fieldmap	Spin-Echo
Anatomical	T1w_MPRAGE
	T2w_SPACE
Diffusion	Diffusion Spectrum Imaging Scan (Phase Encoding A-P)
	Diffusion Spectrum Imaging Scan (Phase Encoding P-A)
Physiological	Heart Rate
	Respiration
Behavioral	Daily Intake Form
	Localizer Task
	Scene Valence Judgment

**Table 2.** A list of data collected for BOLD5000 dataset.

stimuli were shown sequentially. Each image was presented for 1 sec followed by a 9 sec fixation cross. Given that each run contains 37 stimuli, there was a total of 370 sec of stimulus presentation plus fixation. Including the pre- and post-stimulus fixations, there were a total of 388 sec (6 min 28 sec) of data acquired in each run.

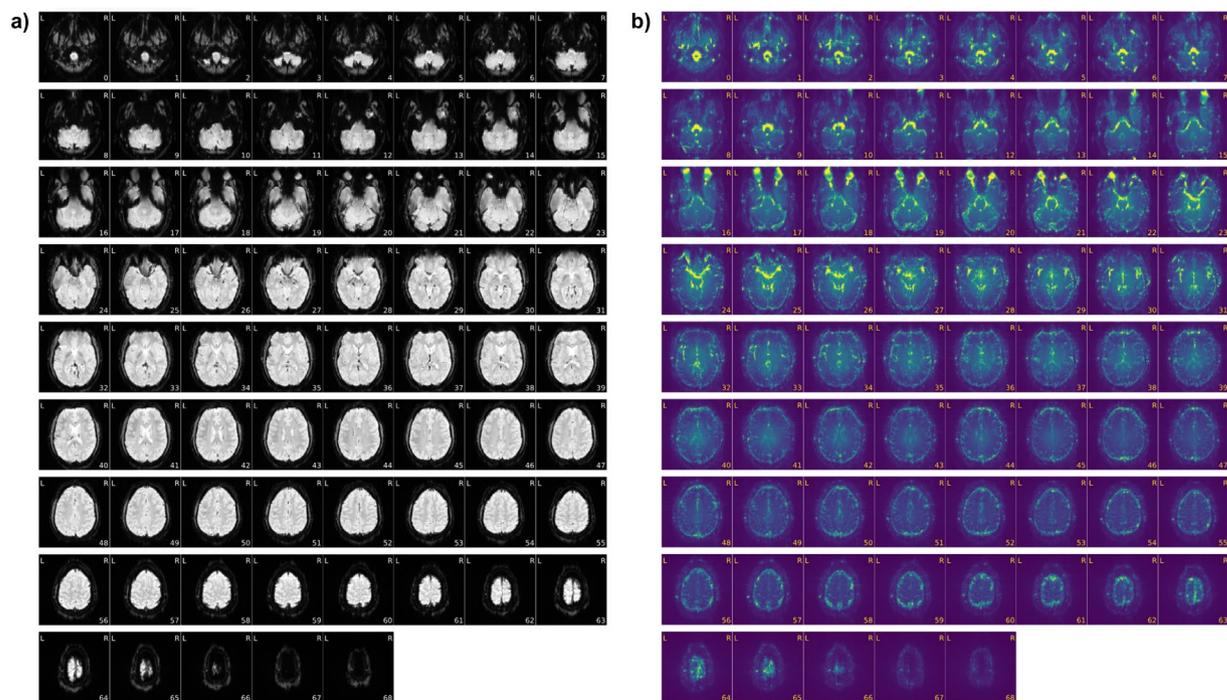
For each stimulus image shown, each participant performed a valence judgment task, responding with how much they liked the image using the metric: “like”, “neutral”, “dislike”. Responses were collected during the 9 sec interval comprising the interstimulus fixation, that is, subsequent to the stimulus presentation, and were made by pressing buttons attached to an MRI-compatible response glove on their dominant hand (right for all participants).

**Experimental Paradigm for the Functional Localizer.** As mentioned above, 8 functional localizers (6 for CSf4) were acquired throughout the 15 functional sessions. These localizers were run at the end of the session for that day. The functional localizer included three conditions: scenes, objects, and scrambled images. The scene condition included images depicting indoor and outdoor environments and were non-overlapping with the 4,916 scenes used in the main experiment. The object condition included objects with no strong contextual association (i.e. weak contextual objects<sup>25</sup>). Scrambled images were generated by scrambling the Fourier transform of each scene image. There were 60 color images in each condition. Images were presented at a visual angle of 5.5 degrees.

Each run was a block design format. Each block had 16 trials, with a stimulus duration of 800 ms, and a 200 ms ISI. Of the 16 trials, there were 14 unique images, with 2 images repeating. Each participant’s task was to press a button if an image immediately repeated (i.e., a one-back task). Between task blocks there were 6 seconds of fixation. Each run started and ended with 12 seconds of fixation. There were 12 blocks per run, with 4 blocks per condition.

**MRI Acquisition Parameters.** MRI data were acquired on a 3T Siemens Verio MR scanner at the Carnegie Mellon University campus using a 32-channel phased array head coil. No identifiable participant information is included in the DICOM files. We do include participant birth year (in the format 01-01-YYYY), gender, weight, and the unique identifier code for each session used by the imaging center (stored in the “PatientID” and “PatientsName” fields) in the DICOM header.

**Functional Images and Field Maps.** Functional images were collected using a T2\*-weighted gradient recalled echoplanar imaging multi-band pulse sequence (cmrr\_mbep2d\_bold) from the University of Minnesota Center for Magnetic Resonance Research (CMRR)<sup>26,27</sup>. Parameters: 69 slices co-planar with the AC/PC; in-plane resolution = 2 × 2 mm; 106 × 106 matrix size; 2 mm slice thickness, no gap; interleaved acquisition; field of view = 212 mm; phase partial Fourier scheme of 6/8; TR = 2000 ms; TE = 30 ms; flip angle = 79 degrees; bandwidth = 1814 Hz/Px; echo spacing = 0.72 ms; excite pulse duration = 8200 microseconds; multi-band factor = 3; phase encoding direction = PA; fat saturation on; advanced shim mode on. Data were saved both with and without pre-scan normalization filter applied. During functional scans, physiological data was also collected using wireless sensors: heart rate was acquired using a Siemens photoplethysmograph attached to the participant’s left index finger; respiration was acquired using a respiratory cushion held to the participant’s chest with a belt and connected via pressure hose to a Siemens respiratory sensor. In the middle of each scanning session, three sets of scans were acquired for use for geometric distortion correction. The three options were provided to allow researchers to use the scans that work best in their processing pipeline. Options: (1) A three volume run with opposite phase encoding (PA) from the functional scan (opposite phase encoding achieved using the “Invert RO/PE polarity” option). (2) Two pairs of three volume spin-echo runs with opposite phase encoding, one with partial Fourier 6/8, one without, both using the cmrr\_mb3p2d\_se pulse sequence. Non partial Fourier spin-echo parameters: geometry and resolution matched to functional parameters; TR = 9240 ms; TE = 81.6 ms; multi-band factor = 1. (3) Partial Fourier spin-echo parameters: geometry and resolution matched to functional parameters; TR = 6708 ms; TE = 45 ms; phase partial Fourier scheme of 6/8; multi-band factor = 1. Opposite phase encoding achieved using “Invert RO/PE polarity” option.



**Fig. 2** A sample of the functional data provided in the dataset. A single participant's (CS1) (a) average BOLD signal and (b) standard deviation of BOLD signal for a single run (Session 1, Run 1).

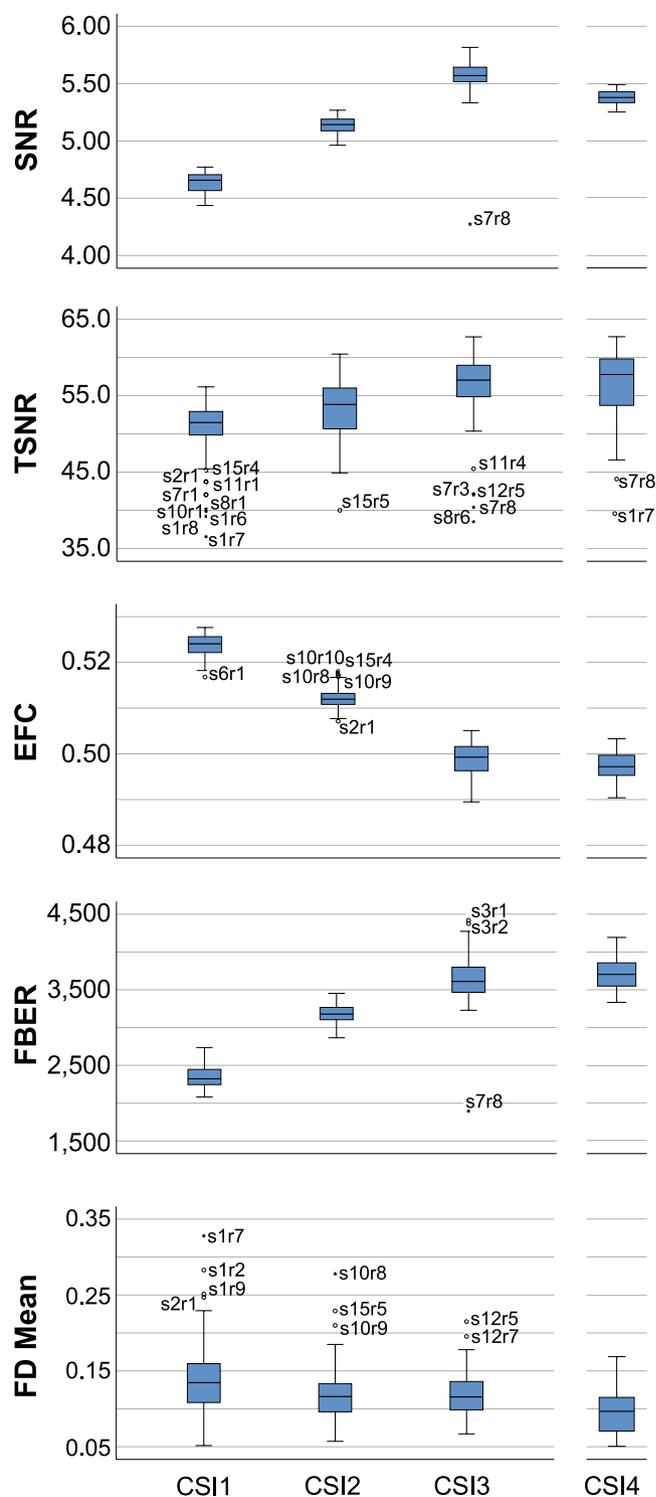
**Anatomical and Diffusion Images.** A T1 weighted MPRAGE scan, and a T2 weighted SPACE scan using Siemens pulse sequences were collected for each participant. MPRAGE parameters: 176 sagittal slices; 1 mm isovoxel resolution; field of view = 256 mm; TR = 2300 ms; TE = 1.97 ms; TI = 900 ms; flip angle = 9 degrees; GRAPPA acceleration factor = 2; bandwidth = 240 Hz/Px. SPACE parameters: 176 sagittal slices; 1 mm isovoxel resolution; field of view = 256 mm; TR = 3000 ms; TE = 422 ms; GRAPPA acceleration factor = 2; bandwidth = 751 Hz/Px; echo spacing = 3.42 ms. In order to protect our participants' privacy and anonymity, faces were removed from all anatomical images, MPRAGE, and SPACE scan using Pydeface (<https://github.com/poldracklab/pydeface>). Two diffusion spectrum imaging (DSI) scans were acquired for each participant using the cmrr\_mbep2d\_diff sequence from CMRR. Parameters: geometry and resolution matched to functional parameters; diffusion spectrum imaging sampling scheme; 230 directions; phase partial Fourier scheme of 6/8; TR = 3981 ms; TE = 121 ms; maximum b-value = 3980 s/mm<sup>2</sup>; bipolar acquisition scheme; AP phase encoding direction. A second scan matching all parameters but with PA phase encoding (achieved using the "Invert RO/PE Polarity" option) was also acquired.

**Data Analyses. fMRI Data Analysis.** All fMRI data were converted from DICOM format into the Brain Imaging Data Structure (BIDS; <http://bids.neuroimaging.io/>) using a modified version of Dcm2bids (<https://github.com/jooh/Dcm2Bids>). Data with the pre-scan normalization filter applied were used for all analyses. Data quality was assessed and image quality metrics were extracted using the default pipeline of MRIQC<sup>28</sup>.

Results included in this manuscript come from preprocessing performed using fMRIPrep 1.1.4<sup>28</sup> (<https://github.com/poldracklab/fmriprep>), a Nipype (0.13.1; <https://github.com/nipy/nipype>)<sup>29</sup> based tool. Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using N4BiasFieldCorrection v2.1.0<sup>30</sup> and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1<sup>31</sup>, and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle<sup>32</sup>. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c<sup>33</sup> was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0<sup>34</sup>, using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast<sup>35</sup> (FSL v5.0.9).

Functional data were motion corrected using mcflirt (FSL v5.0.9<sup>36</sup>). Distortion correction was performed using an implementation of the Phase Encoding POLARity (PEPOLAR) technique using 3dQwarp (AFNI v16.2.07<sup>37</sup>). This was followed by co-registration to the corresponding T1w using boundary-based registration<sup>38</sup> with 9 degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, field distortion correcting warp, and BOLD-to-T1w transformation were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation.

Many internal operations of FM RIPREP use Nilearn<sup>39</sup>, principally within the BOLD-processing workflow. For more details of the pipeline see <https://fmriprep.readthedocs.io/en/latest/workflows.html>. All reports of the fMRIPrep analysis are publicly available, see Data Usage.



**Fig. 3** A sample of the data quality metrics from the MRIQC analysis. Boxplots of representative IQMs for each run of each participant ( $N = 150$  per participant). SNR - signal to noise ratio; higher is better. TSNR - temporal signal to noise ratio; higher is better. EFC - entropy focus criterion - indicating ghosting and blurring induced by head motion; lower is better. FBER - foreground to background energy ratio; higher is better. FD Mean - mean frame displacement of each run given in mm; lower is better. Data for CSI4 is separate due to the different number of datapoints ( $N = 90$ ). Outliers are specified by Session Number (s) and then by Run Number (r).

After preprocessing, the data for the 5,254 images were analyzed and extracted using the following steps. Data from each session, (i.e., including 9 or 10 runs, depending on session) were entered into general linear model (GLM), where nuisance variables were regressed out of the data. There were a total of nine nuisance variables including 6 motion parameters estimates resulting from motion correction, the average signal inside the

IQM	Mean	Std. Dev.	IQM	Mean	Std. Dev.
aor	0.002	0.001	summary bg k	24.238	5.744
aqi	0.011	0.002	summary bg mad	6.722	1.118
dvars_nstd	33.425	3.149	summary bg mean	26.847	3.007
dvars_std	1.274	0.048	summary bg median	9.641	0.906
dvars_vstd	1.013	0.009	summary bg n	512323.309	5825.738
efc	0.509	0.011	summary bg p05	3.905	0.267
fber	3165.148	572.047	summary bg stdv	87.562	24.788
fd_mean	0.120	0.027	summary bg p95	68.888	35.678
fd_num	26.037	16.245	summary fg k	3.720	0.795
fd_perc	13.576	8.517	summary fg mad	81.105	8.961
fwhm_avg	2.572	0.025	summary fg mean	549.945	32.425
fwhm_x	2.316	0.029	summary fg median	553.423	32.733
fwhm_y	3.004	0.044	summary fg n	165740.252	5341.012
fwhm_z	2.394	0.033	summary fg p05	371.228	30.720
gcor	0.015	0.004	summary fg p95	704.422	42.390
gsr_x	-0.006	0.007	summary fg stdv	107.775	8.837
gsr_y	0.027	0.007	tsnr	53.939	3.085
snr	5.157	0.374			

**Table 3.** A list of the image quality metrics (IQMs) produced as a result of the MRIQC analysis. Each mean and standard deviation is calculated across all runs of all participants (CSI1,2,3 N = 150; CSI4 N = 90). For information about each IQM please refer to <http://mriqc.org>.

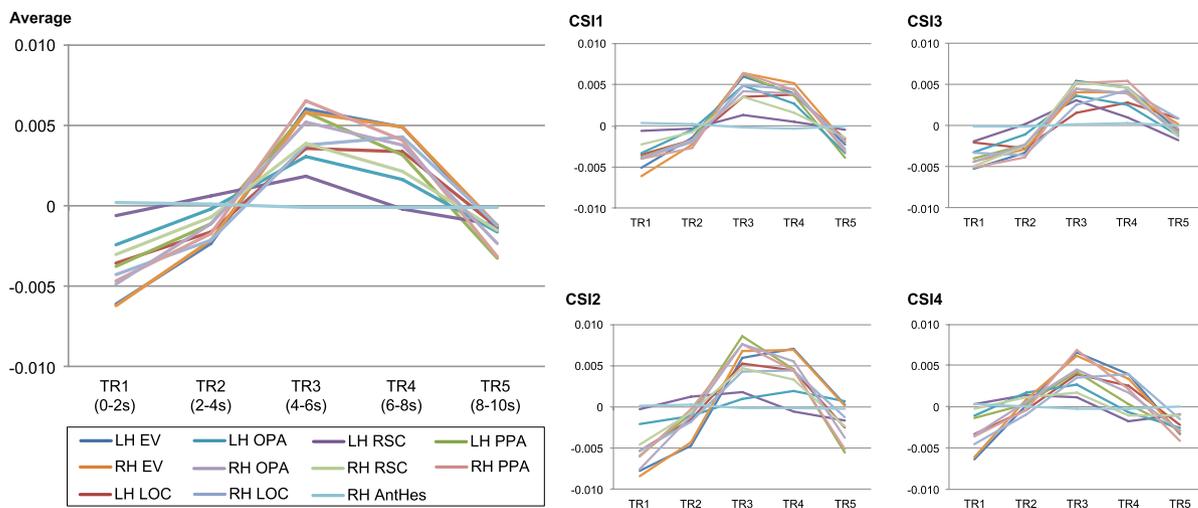
cerebral spinal fluid mask and separately for the white matter mask across time, as well as global signal within the whole-brain mask. All of the nuisance variables were confounds extracted in the fMRIPREP analysis stream. A regressor for each run of the session was used in the GLM. In addition, a high pass filter of 128 s was applied to the data. The residual time series of each voxel within a region of interest (see details below) were extracted, demeaned across all image presentations, and used for all subsequent analyses.

**Functional Localizer Analysis.** Participants CSI1, CSI2, and CSI3 all had eight functional localizer runs scattered throughout the 15 functional sessions, with a maximum of one localizer run per session. CSI4 had six functional localizer runs scattered throughout the 9 functional sessions. All localizer data post fMRIPREP preprocessing were analyzed within the same general linear model using a canonical hemodynamic response function implemented using SPM12<sup>40</sup>. All 9 nuisance regressors mentioned above, plus regressors for each run, and a high pass filter were implemented in the model. Three conditions were modeled: scenes, objects, and scrambled images.

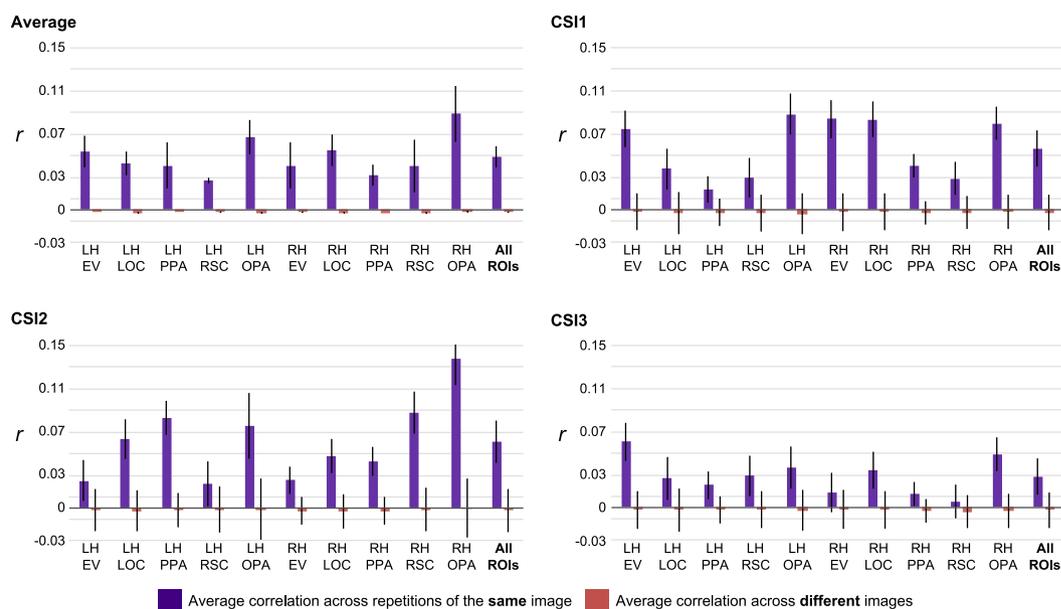
**Region of Interest (ROI) Analyses.** All ROI analyses were performed at the individual level using the MarsBaR toolbox<sup>41</sup> with SPM12 and analyzed within native space on the volume. Scene selective regions of interest (the parahippocampal place area, PPA; the retrosplenial complex, RSC, and the occipital place area, OPA) were all defined using the contrast of scenes compared with objects and scrambled. Although the functional localizer was not specifically designed to locate object selective areas, a lateral occipital complex (LOC) was successfully defined by comparing objects to scrambled images. Finally, an early visual (EV) ROI was defined by comparing the scrambled images to baseline, and the cluster most confined to the calcarine sulcus was used. However, since retinotopy was not used to define the EV ROI, and an anatomical boundary was not applied, this region sometimes extends beyond what is classically considered V1 or V2. A threshold of using a family-wise error correction of  $p < 0.0001$  (or smaller), and  $k = 30$  was used for all ROIs. In cases where including all localizer runs resulted in clusters of activity too big for the intended ROI (e.g., the PPA ROI connected to the RSC ROI), a reduced number of runs were included in the contrast. For the scene selective and object selective regions, this included using just the first, middle, and last run. Across all participants, the EV ROI was only defined using the first run due to massive amount of activity produced from this contrast. Using these ROIs defined from the localizer, the data of the 5,254 scenes were then extracted across the time course of the trial presentation (5TRs; 10 sec). We analyzed the data for each timepoint (TR1 = 0–2 s, TR2 = 2–4 s, TR3 = 4–6 s, TR4 = 6–8 s, TR5 = 8–10 s).

To demonstrate the time course of the data, the time series was extracted for each ROI and averaged across all image presentations (e.g., N = 5,254), see Technical Validation. For CSI1, CSI2, and CSI3 the time course showed that activity peaked across TR3 and TR4 (4–8 s). Due to this pattern of results, all subsequent data analysis used an average of the data from TR3 and TR4. For CSI4, the time course demonstrated a peak of activity most confined to TR3. Thus, only data from TR3 for CSI4 was used for subsequent analyses.

In addition, we used the FreeSurfer parcellation to define a control ROI: the right hemisphere anterior Heschl's gyrus. This control ROI was used to examine the time course of BOLD activity in a non-stimulus responding region.



**Fig. 4** Mean time course of the BOLD signal across all stimulus presentations for each region of interest (ROI). Each graph illustrates the BOLD signal averaged across all voxels within each ROI, averaged across all stimulus trials. The BOLD signal here is measured as the residual values from the model that regressed out the nuisance variables (e.g., motion parameters). The RH anterior Heschl's gyrus ROI is presented as a comparison: being a control region, no response to the stimulus was expected.

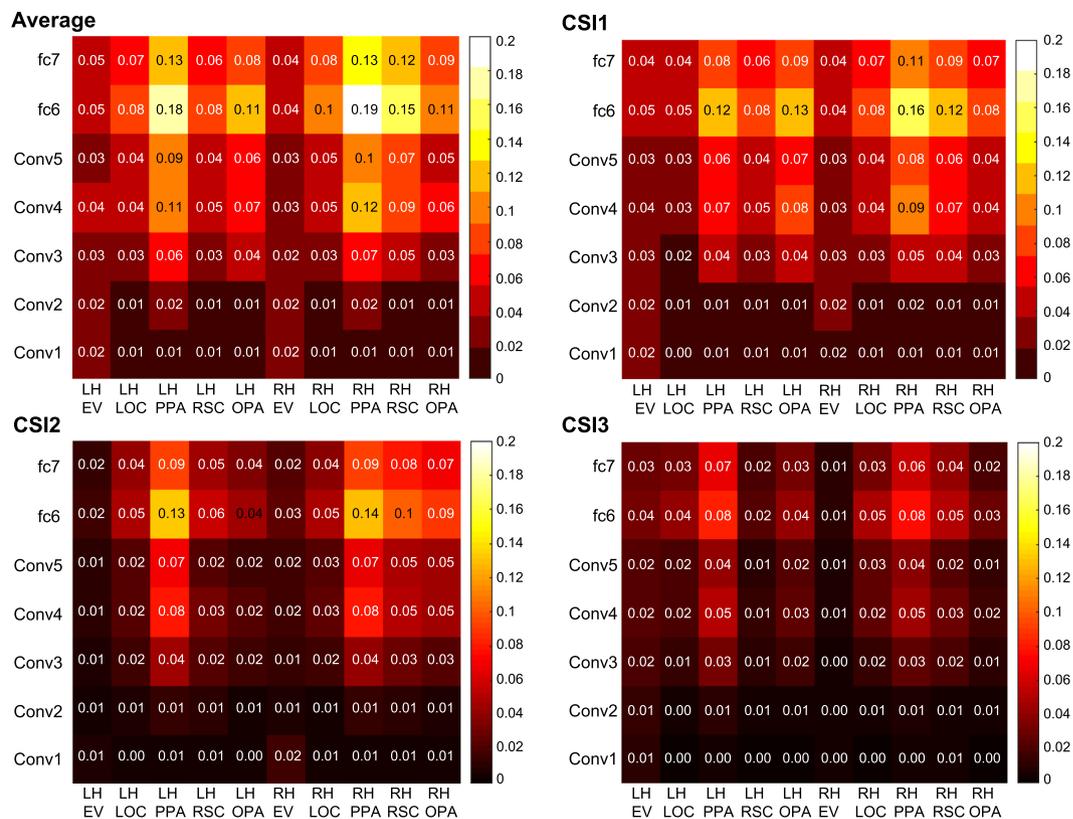


**Fig. 5** Correlations between the BOLD signal across voxels within each ROI across the same image repetitions as compared to correlations across different images. The final column of each graph shows the averages across all ROIs. Error bars represent one standard error.

Finally, data from extracted ROIs were demeaned for normalization purposes. Data were demeaned by each voxel by taking the mean of that voxel across all image presentations and subtracting it from each sample data point.

### Data Records

We have publicly released BOLD5000 online. As seen in Table 2, we have provided a comprehensive list of collected data. Additionally we have made various stages of analyzed data available. The data is available at three locations, with all relevant links and information on [BOLD5000.org](https://www.bold5000.org). Please refer to respective links ([BOLD5000.org](https://www.bold5000.org), [Figshare.com](https://www.figshare.com), and [OpenNeuro.org](https://www.openneuro.org)) for terms of use and relevant license information for the materials provided. Lastly, please note that unless specified by a \*, the data in each three locations are not identical (e.g. Figshare contains unprocessed DICOM data and OpenNeuro contains processed BIDS data).



**Fig. 6** Representational similarity analysis heatmaps comparing the similarity of the 4,916 scenes in voxel space to the similarity of the 4,916 scenes in AlexNet features space. Comparisons were made across each ROI (columns) and each layer of AlexNet (rows).

### 1. [BOLD5000.org](https://BOLD5000.org)

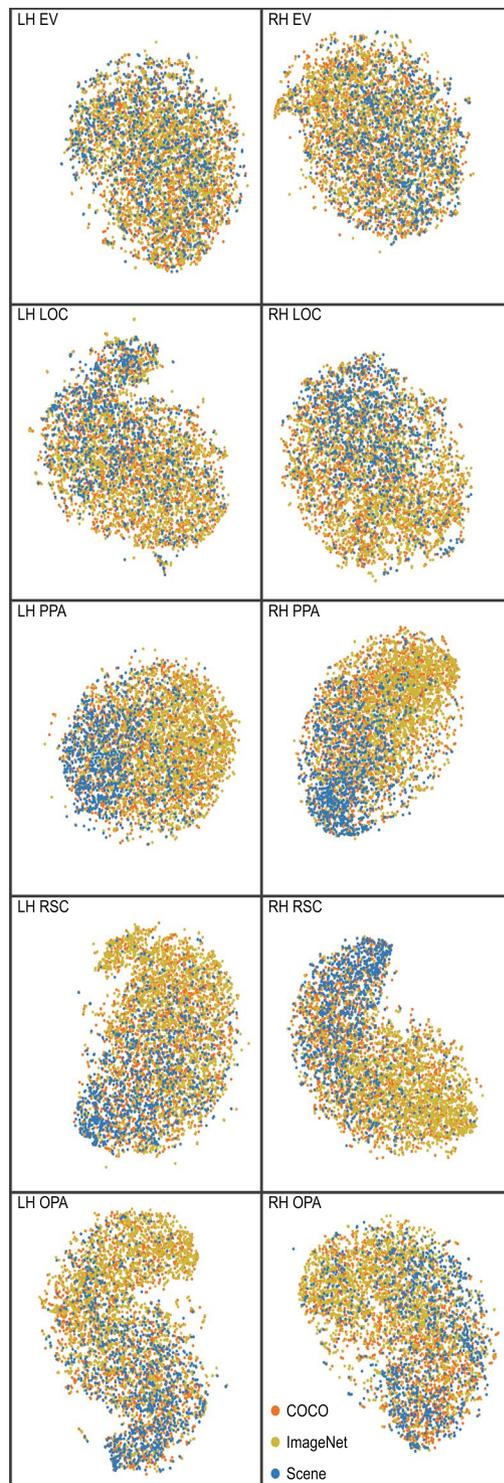
- All relevant links for publicly available data (Figshare, OpenNeuro)
- 5,254 Images and Names\*
- Image labels\*
- News & updates
- t-SNE supplemental results

### 2. [Figshare.com](https://Figshare.com)<sup>20</sup>

- Raw MRI Data (DICOM; all data except defaced anatomicals, which are provided in NIFTI)
- Pre-scan Normalization Filtered fMRI Data (DICOM)
- Physiological Data
- All Behavioral Data
- Data extracted from ROIs
- Scanning Protocols (PDF)
- 5,254 Image Names\*
- Image labels\*
- Experiment Presentation Code (Matlab)

### 3. [OpenNeuro.org](https://OpenNeuro.org)<sup>21</sup>

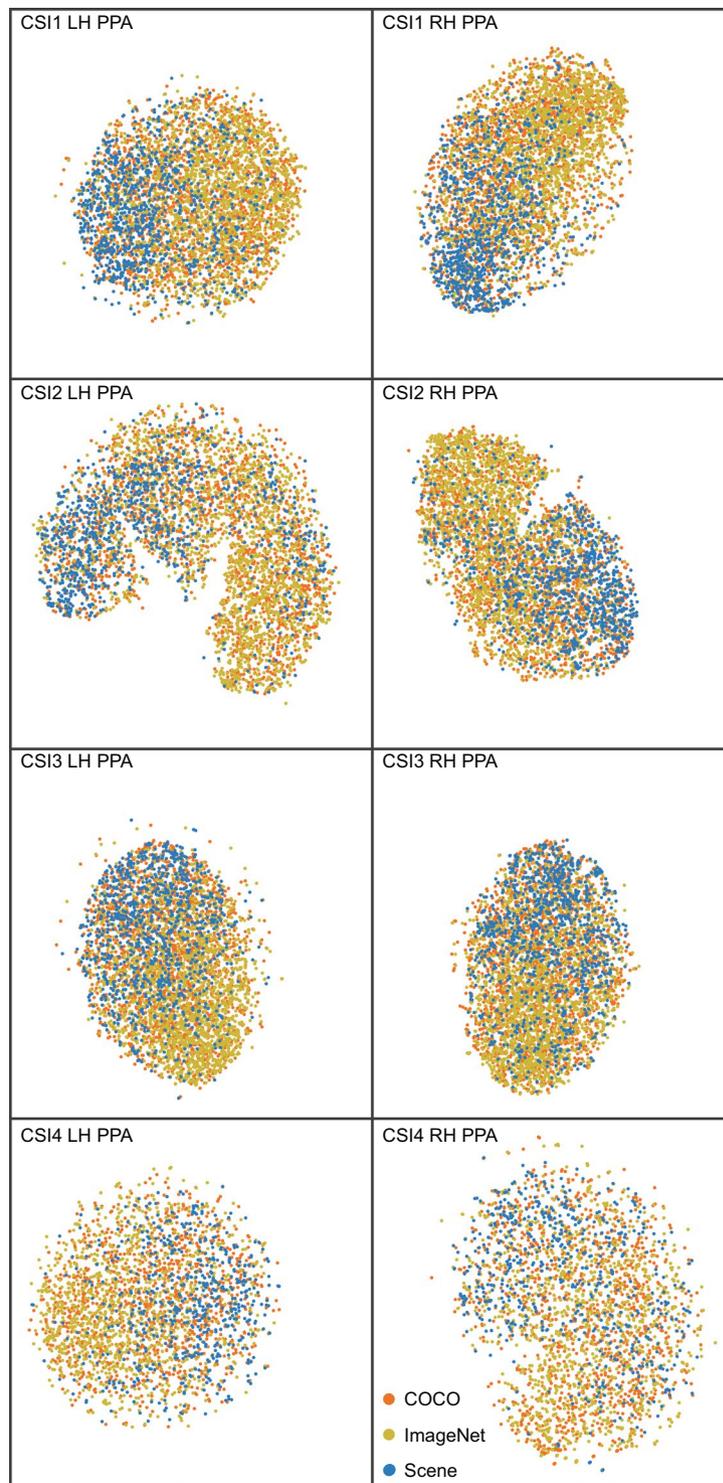
- Pre-scan Normalization Filtered MRI Data (BIDS)
- MRIQC reports
- fMRIPREP preprocessed data & reports
- ROI Masks
- 5,254 Image Names\*



**Fig. 7** t-SNE analysis on the 4,916 scenes in voxel space for CSII. Image datasets (COCO, ImageNet, Scene) are used as labels.

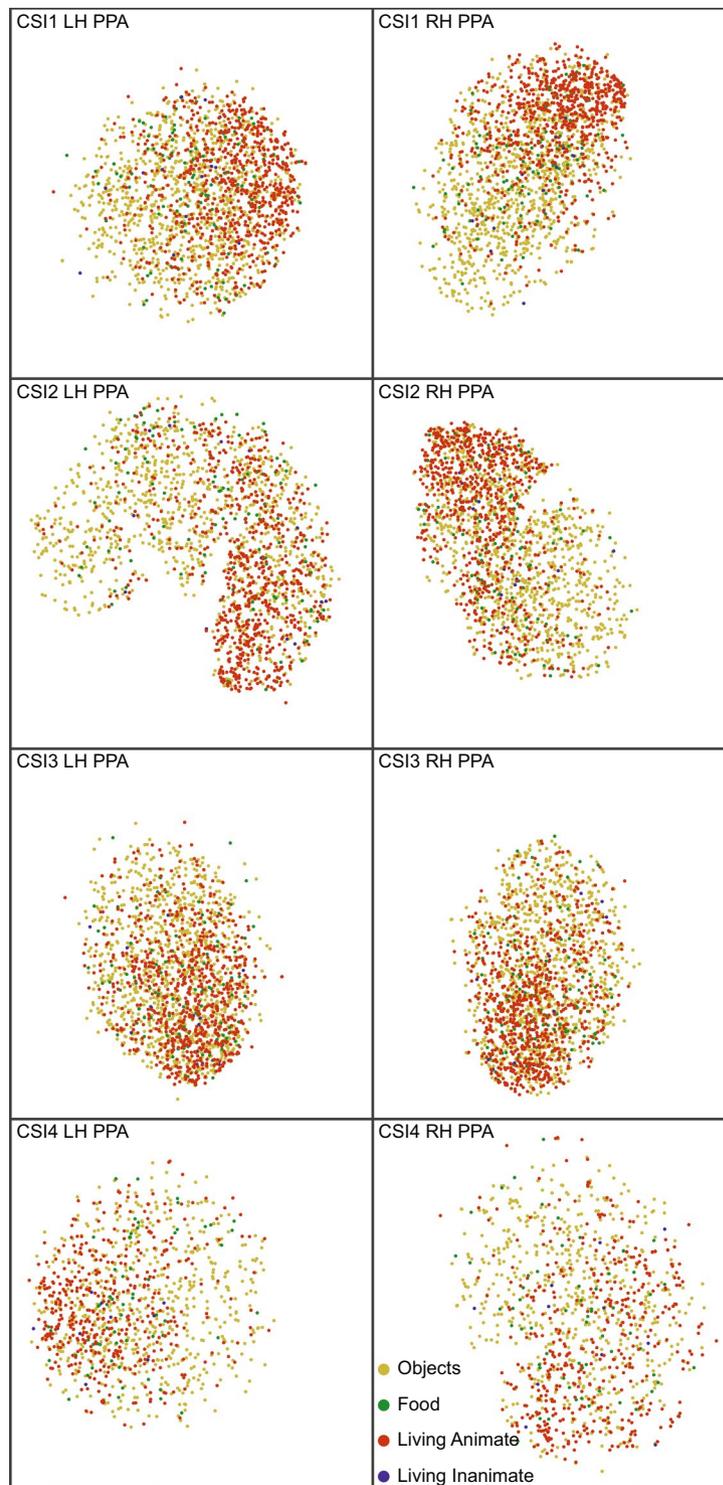
### Technical Validation

**Data Quality.** To provide measures describing the quality of this dataset, the data were analyzed using MRIQC<sup>28</sup>. MRIQC is an open-source analysis program that provides image quality metrics (IQMs) in an effort to provide interoperability, uniform standards, and to assess reliability of a dataset, ultimately with the goal to increase reproducibility. MRIQC analyzes each run (in this case, 150 per participant) and provides IQMs for each run, as well as a figure of the average BOLD signal in the run as well as the standard deviation for each run, see Figs 2 for an example from CSII - Session 1, Run 1. For a full report of each run for each participant, please find



**Fig. 8** t-SNE analysis on the 4,916 scenes in voxel space for all participants, in the PPA. Image datasets (COCO, ImageNet, Scene) are used as labels. Each row corresponds to an individual participant.

the complete analysis available on OpenNeuro.org (see Data Records). Figure 3 also demonstrates the stability of the data for each participant across all 150 runs by showing the variance across five representative IQMs across all runs. Note that CS14 is plotted separately because the magnitudes of the ratios for CS14 are not comparable to the other three participants from a statistical standpoint. Table 3 provides the averages and standard deviations across all participants for all measures from the MRIQC analysis. These measures are on par, if not better, than other studies that have provided IQMs (e.g., see <http://mriqc.org>).



**Fig. 9** t-SNE analysis on the 4,916 scenes (in voxel space for ImageNet images, for all participants, and in the PPA. ImageNet super categories (Objects, Food, Living Inanimate, Living Animate) are used as labels. Each row corresponds to an individual participant.

**Design Validation.** In this study, our goal was to make the data accessible without requiring advanced fMRI analysis tools to understand the data. Therefore, we designed each trial to have minimal bleed over from neighboring trials. In this respect, we chose a slow event-related design in which the stimulus was presented for one second, and a fixation cross was presented for nine seconds between stimulus presentations. As a result of using a slow event-related design, the extracted time course shows the hemodynamic response peaked around 6 seconds post stimulus onset, and returned to baseline before the next stimulus was presented. Using this design,

there was minimal bleed over from neighboring trials and minimizing the need to deconvolve the signal. Figure 4 shows the time course averaged across all stimulus trials for each region of interest. As can be seen in the figure, the timing of the design allowed the BOLD signal to peak and return to baseline.

### Data Validation

**Repeated Stimulus Images.** As noted above, a majority of our images, 4,803 were only presented to the participant on a single trial during the entire 15 functional sessions. However, 113 images had repeated presentations (3+) over the 15 functional sessions. We implemented this design to assess the signal to noise ratio in the data. Note that repeating a stimulus means that we have 4 unique neural representations for the same stimulus. Since the stimulus is the same, the neural representations are expected to be the same, with the exclusion of noise and session to session variance. Thus, we leverage the extra neural representations to our advantage. To demonstrate this, the reliability of the BOLD pattern across voxels within a given ROI for each repetition of a given image was examined. In this analysis, we hypothesized that the correlation of the pattern of BOLD activity across the repetitions of the same image should be considerably higher than the correlation of the patterns of activity across presentations of different images. Figure 5 shows the average correlation across repetitions of the same image compared to the average correlation across images. To do this analysis, the data from the 113 images were extracted from each ROI, which yielded an extracted dataset of 451 images ( $112 \times 4$  repetitions, and  $1 \times 3$  repetitions). Participant CSI4 was not included in this analysis due to the relative insufficient amount of data because of early termination. A Pearson correlation was then calculated for each comparison across repetitions (e.g., correlate repetition 1 with repetition 2, correlation repetition 1 with repetition 3, etc.). The correlations across each pairwise repetition were then averaged for each image to provide a measure of similarity within image (i.e., across repetitions). The pairwise correlation for each image, for each repetition, was also made across all other images (from the pool of 451) and then averaged to give a measure of similarity across different images. The difference between these average correlations provides a measure of the reliability of the signal on a per trial basis.

If there is reliability in the signal, the correlation across repetitions should be considerably higher than the correlation across images. This is indeed what we found.

**Representational Similarity Analysis.** One type of analysis used in the literature thus far to compare computational models to brain activation patterns is representational similarity analysis (RSA). A popular comparison<sup>7,11</sup>, and one relevant to our mission, is comparing BOLD activity to the unit activity in a convolutional neural network (CNN) aimed at recognizing objects - AlexNet<sup>23</sup>. We have done the same analysis as a reference point to put in context with other related studies. We implemented this analysis by comparing the similarity space derived from the pattern of BOLD activity across each voxel in a given region of interest to the similarity spaced derived from feature space of AlexNet for the set of scenes. We used 4,916 scenes in this analysis, where each scene is presented on a single trial, i.e., the trials in which a scene was repeated were not included in this analysis. For voxel space, the BOLD signal was extracted from each voxel of a given ROI such that we had a matrix of images by voxels (see Region of Interest Analysis). To measure similarity, the cosine distance of each pairwise comparison across images were made. This was performed for each ROI ( $N = 10$ , PPA, RSC, OPA, LOC, EV, in each hemisphere, see Region of Interest Analyses). In model space, we extracted the weights from each layer of AlexNet. All images were passed through an ImageNet pre-trained AlexNet, and all layer weights were extracted. Similarities were then computed for all weights in their original shape using the cosine distance metric across each pairwise comparison of images. This was then computed for each layer of AlexNet, such that 7 (5 convolutional, and 2 fully connected) feature spaces are derived from AlexNet. All cosine distances were Fisher Z transformed before a Pearson correlation was used to compare the voxel similarity space of a given ROI (i.e., cosine distances) to the feature similarity space of a given AlexNet layer.

Figure 6 shows the RSA analysis for the average of three participants, as well as the results for each individual participant (the fourth was not included due to an incomplete set of data). The average was calculated by first averaging the cosine distance measurements of the fMRI data across participants and then correlating that with AlexNet features. The results of this analysis demonstrate the typical pattern: high-level visual regions (e.g., PPA) correlated more strongly with higher layers of AlexNet (e.g., convolutional layer 5; presumably where high-level information, like semantics, is represented). Lower-level visual regions (e.g., EV) correlated stronger with lower layers (e.g., convolutional layer 2) compared to higher regions. Although we do not show a strong correlation with early visual regions and low levels of AlexNet, as commonly shown, we believe that this may result from our region of interest being across early visual regions as a whole and not confined to a retinotopically defined V1 or V2. However, the pattern of results in lower layers to brain regions (low level regions compared with high level regions) still holds. In addition, the correlation is at times lower than what has been reported in the literature. However, this may be a consequence of stimuli numbers, with low numbers inflating the correlation effect. Important to note from this analysis is the consistency of the results across hemispheres and across individual participants.

**t-distributed Stochastic Neighbor Embedding Analysis.** A common type of analysis used in machine learning is t-distributed Stochastic Neighbor Embedding (t-SNE)<sup>15</sup>. The purpose of t-SNE is to embed high-dimensional data into a 2D space, such that high-dimensional data can be visualized. Importantly, t-SNE preserves similarity relations among high-dimensional data by modeling similar data with close points and divergent data with distant points. Similar to our RSA analysis, we perform t-SNE on the BOLD signal from each of the unique 4,916 (2,900 for CSI4) scenes trials. The BOLD signal was extracted from each voxel for each ROI for each participant. We visualize our t-SNE results with different categorical labels. Specifically, each t-SNE figure contains the *same* data points/coordinates, and different labels are purely for visualization purposes.

First, we examine the similarity space across the different image datasets. Specifically, we will be exploiting the implicit image attributes of these datasets: Scene contains whole scenes, ImageNet is focused on a single object, and COCO is in between with images of multiple objects in an interactive scene. Given the ROIs tend to process visual input with specific properties (e.g., category selectivity) we would expect to see a separation in t-SNE space of the different image datasets, especially with regard to Scenes vs. ImageNet. In Fig. 7 the t-SNE results are visualized for CSII and each ROI. Here, the data points are labeled by the dataset the image belongs to (e.g. Scene, COCO, ImageNet). In ROIs commonly associated with scene processing (e.g. PPA, RSC, OPA), there were a clustering of the Scene images, a clustering of ImageNet images, and a more uniform scattering of COCO images. The strongest clustering of Scene images was observed in the PPA, regardless of participant (Fig. 8). However, note that in lower-level visual regions (e.g. EV), we observed a uniform scatter for *all* images, regardless of their datasets. This clustering and uniform scattering contrast reaffirms that higher-level scene regions have a stronger selectivity for processing categorical information. This pattern holds for all participants, and their t-SNE results for all ROIs can be seen on [BOLD5000.org](https://bold5000.org).

Second, the t-SNE results are visualized with only ImageNet images, and our labels and categories are broken down further. Specifically, the images are labeled with ImageNet super categories. These super categories were created by using the WordNet hierarchy<sup>42</sup>. For each image synset, which describes the image's object of interest, we found all of its WordNet ancestors. From 61 final WordNet categories, we labeled each one as "Living Animate", "Living Inanimate", "Objects", or "Food". Images labeled as "Geography" were removed because only 20 images applied. An example of our label mapping is "Dog" to "Living Animate" and "Vehicle" to "Objects". The t-SNE results are shown in Fig. 9. Only the PPA region is presented to conserve space. The observations stated below also apply to other higher order regions, and all participants' ROIs for this t-SNE result are available on [BOLD5000.org](https://bold5000.org). In Fig. 9, the data demonstrated that "Living Animate" object based images clustered in a small area for all participants. However, "Objects" based images are uniformly scattered in space. From the "Living Animate" clustering, the results demonstrate that the higher order ROIs process images with living animate objects differently from objects, such as "cup". From these results, it is evident that "Living Animate" is a specific object category within the general object domain. Further, observing Figs 8 and 9, one can see that the "Living Animate" clustering occupies a separate spatial location than the Scene images clustering. Finally, it is important to emphasize that the results are consistent across hemispheres and across individual participants.

Using t-SNE, category selectivity in high-level visual regions was demonstrated using a data-driven method. The clustering observed demonstrates a small glimpse of the rich information inherent in the neural representations of each scene.

### Usage Notes

Marr's<sup>12</sup> nearly four-decades-old dream of a singular vision science—the intertwined study of biological and computer vision—is, at long last, being realized. Fueled by dramatic progress in neuroimaging methods and high-performing computer vision systems, new synergies are arising almost daily. However, connecting these two domains remains challenging. Here, we tackle one of the biggest obstacles for integrating across these two fields: data. In particular, neural datasets studying biological vision are typically lacking in: (1) *size*; (2) *diversity*; and (3) *stimulus overlap*; relative to extant computer vision datasets. We address these three concerns in BOLD5000 in which we have collected a large-scale, diverse fMRI dataset across 5,254 stimulus images. Critically, the human neuroimaging data available in BOLD5000 is: (1) significantly larger than prior slow event-related fMRI datasets by an order of magnitude; (2) extremely diverse in stimuli; (3) overlaps considerably with standard computer vision datasets. At the same time, BOLD5000 represents a significant dataset for the study of human vision in and of itself. As mentioned above, it is, by far, the largest slow event-related fMRI dataset using real-world images as stimuli. Moreover, given the diversity of content within these images and the fact that our fMRI data covers the whole brain, BOLD5000 may be sufficient to cover a wide range of high-level vision experiments (in the context of automatic visual processing during non-task-related, free viewing).

While we believe the scale of the BOLD5000 dataset is a major step forward in the study of vision across biological and computer vision, we should acknowledge its limitations. First and somewhat ironically, one salient limitation is the total number of stimulus images. Although 5,000 is significantly more than the number of images included in previous human neuroimaging studies, it is still relatively small as compared to either human visual experience across one's lifespan or the million of images used to train modern artificial vision systems. Given the practicalities of running the same individuals across multiple neuroimaging sessions, scaling up future datasets will necessitate collecting partially-overlapping data across participants and then applying methods for "stitching" data together<sup>43,44</sup>. Second, another obvious limitation is that our dataset includes only four participants. Again, the practicalities of human experimental science come into play, necessarily limiting how many suitable participants we were able to recruit and run. However, we argue that the sort of in-depth, detailed functional data we have collected at the individual level are as valuable as small amounts of data across many participants. Indeed, there have been recent arguments for exactly the sort of "small-N" design we have employed here<sup>45</sup>. Of course, we also encourage others in the field to add to the BOLD5000 dataset by running one or more participants using our exact experimental protocol and stimuli—see the Code Availability section below. Ultimately, we see a straightforward solution to these two limitations: expansion of our the number of stimulus images by another order of magnitude, but with subsets of the stimuli run across many more participants. Such an undertaking is not for the experimentally faint-of-heart—we suggest that the best approach for realizing this scale of human neuroscience would involve a large number of labs collaborating to run subsets of the overall study with coordinated data storage, distribution, and analysis.

## Code Availability

The complete set of Psychtoolbox Matlab scripts for running this study are available for download at [BOLD5000.org](http://BOLD5000.org). The Psychtoolbox Version 3 Matlab toolbox and documentation are both available for download at <http://psychtoolbox.org>. As per convention in the field, the complete set of images used as stimuli are available for download at [BOLD5000.org](http://BOLD5000.org).

## References

1. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
2. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 248–255 (2009).
3. Lin, T.-Y. *et al.* Microsoft COCO: Common Objects in Context. *Comput. Vis. ECCV* 740–755 (2014).
4. LeCun, Y., Bengio, Y. & Hinton, G. E. Deep learning. *Nat* **521**, 436–444 (2015).
5. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
6. Guclu, U. & van Gerven, M. A. J. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
7. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep* **6**, 27755 (2016).
8. Oliva, A. & Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. of Comp. Vis* **42**, 145–175 (2001).
9. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–25 (1999).
10. Serre, T., Wolf, L. & Poggio, T. Object recognition with features inspired by visual cortex. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 994–1000 (2005).
11. Groen, I. I. *et al.* Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife* **7**, e32962 (2018).
12. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Freeman, San Francisco, 1982).
13. Tarr, M. & Aminoff, E. Can big data help us understand human vision? In Jones, M. (ed.) *Big Data in Cognitive Science*, chap. 15, 343–363 (Psychology Press, 2016).
14. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
15. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
16. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
17. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nat* **532**, 453–458 (2016).
18. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. Intersubject synchronization of cortical activity during natural vision. *Sci* **303**, 1634–1640 (2004).
19. Xiao, J., Hays, J., Ehinger, K. A., Oliva, A. & Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 3485–3492 (2010).
20. Chang, N. *et al.* BOLD5000. *figshare*. <https://doi.org/10.1184/R1/6459449.v4> (2019).
21. Chang, N. *et al.* BOLD5000. *OpenNeuro*. <https://doi.org/10.18112/openneuro.ds001499.v1.3.0> (2019).
22. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis* **115**, 211–252 (2015).
23. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012).
24. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis* **10**, 437–442 (1997).
25. Bar, M. & Aminoff, E. Cortical analysis of visual context. *Neuron* **38**(2), 347–358 (2003).
26. Moeller, S. *et al.* Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magn. Reson. Med.* **63**(5), 1144–1153 (2010).
27. Feinberg, D. A. *et al.* Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging. *PLoS ONE* **5**, e15710 (2010).
28. Esteban, O. *et al.* MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS ONE* **12**, e0184661 (2017).
29. Gorgolewski, K. *et al.* Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinformatics* **5**, 13 (2011).
30. Tustison, N. J. *et al.* N4itk: Improved n3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
31. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. *NeuroImage* **9**, 179–194 (1999).
32. Klein, A. *et al.* Mindboggling morphometry of human brains. *PLoS Comput. Biol.* **13**, e1005350 (2017).
33. Fonov, V., Evans, A., McKinstry, R., Almlri, C. & Collins, D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009).
34. Avants, B., Epstein, C., Grossman, M. & Gee, J. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).
35. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
36. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17**, 825–841 (2002).
37. Cox, R. W. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comp. Biomed. Res.* **29**, 162–173 (1996).
38. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).
39. Abraham, A. *et al.* Machine learning for neuroimaging with scikit-learn. *Front. Neuroinformatics* **8** (2014).
40. Penny, W., Friston, K., Ashburner, J., Kiebel, S. & Nichols, T. *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (Academic Press, 2006).
41. Brett, M., Anton, J. L., Valabregue, R. & Poline, J. B. Region of interest analysis using an SPM toolbox. *NeuroImage* **16**, 497 (2002).
42. Miller, G. A. WordNet: a lexical database for English. *Comm. ACM* **38**, 39–41 (1995).
43. Bishop, W. E. & Yu, B. M. Deterministic symmetric positive semidefinite matrix completion. *Adv. Neural Inf. Process. Syst.* **27**, 2762–2770 (2014).
44. Bishop, W. E. *et al.* Leveraging low-dimensional structure in neural population activity to combine neural recordings. In *Cosyne Abs.* 1–69 (2018).
45. Smith, P. L. & Little, D. R. Small is beautiful: In defense of the small-N design. *Psychon. Bull. Rev.* **25**(6), 2081–2101 (2018).

## Acknowledgements

We thank Scott Kurdilla for his help and patience as our MRI technologist throughout all data collection. We would also like to thank Jayanth Koushik for his assistance in AlexNet feature extractions, and Ana Van Gulick for her assistance with public data distribution and open science issues. This dataset was collected with the support of NSF Award BCS-1439237 to Elissa M. Aminoff and Michael J. Tarr, ONR MURI N000141612007 and Sloan, Okawa Fellowship to Abhinav Gupta, and NSF Award BSC-1640681 to Michael J. Tarr. Finally, we thank our participants for their participation and patience, without them this dataset would not have been possible.

## Author Contributions

N.C. participated in stimulus selection, stimulus pre-processing, stimulus analysis, experimental design, collecting fMRI data, t-SNE data analysis, writing the manuscript, public distribution of the data, creating the website, and consulted on the remaining sections of the project. J.A.P. participated in developing and testing the MRI protocols, in experimental design, collecting fMRI data, the MRI processing pipeline, writing the manuscript, public distribution of the data, and consulted for the remaining sections of the project. A.M. participated in the MRI processing pipeline and public distribution of the data. A.G. participated in conceiving the original project and writing the manuscript. A.G. consulted regarding stimulus selection, stimulus analysis, experimental design, and t-SNE data analysis. M.J.T. participated in conceiving the original project and writing the manuscript. M.J.T. consulted regarding stimulus selection, experimental design, data analysis, and public distribution of the data. E.M.A. helped conceive the original project and participated in stimulus selection, experimental design, fMRI pre-processing data, general fMRI data analysis (GLM, ROIs), specific subsequent data analysis (design validation, data validation, representational similarity analysis), writing the manuscript, public distribution of the data and consulted on the remaining sections of the project.

## Additional Information

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019