A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence

Emily J. Allen[1,2], Ghislain St-Yves[3,$], Yihan Wu[4],
Jesse L. Breedlove[3,%], Jacob S. Prince[5,^], Logan T. Dowdle[6,7], Matthias Nau[8], Brad Caron[9,10],
Franco Pestilli[11,12,13], Ian Charest[14,15], J. Benjamin Hutchinson[16], Thomas Naselaris[3,$,*], Kendrick Kay[1,*,#]

[1]Center for Magnetic Resonance Research (CMRR), Department of Radiology, University of Minnesota, Minneapolis, Minnesota, USA
[2]Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA
[3]Department of Neuroscience, Medical University of South Carolina, Charleston, South Carolina, USA
[4]Graduate Program in Cognitive Science, University of Minnesota, Minneapolis, Minnesota, USA
[5]Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[6]Department of Neuroscience, Center for Magnetic Resonance Research (CMRR), University of Minnesota, Minneapolis, Minnesota, USA
[7]Department of Neurosurgery, Center for Magnetic Resonance Research (CMRR), University of Minnesota, Minneapolis, Minnesota, USA
[8]National Institute of Mental Health (NIMH), Bethesda, USA
[9]Program in Neuroscience, Indiana University, Indiana, USA
[10]Program in Vision Science, Indiana University, Indiana, USA
[11]Department of Psychology, University of Texas at Austin, Texas, USA
[12]Center for Perceptual Systems, University of Texas at Austin, Texas, USA
[13]Institute for Neuroscience, University of Texas at Austin, Texas, USA
[14]Center for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK
[15]cerebrUM, Département de Psychologie, Université de Montréal, Montréal, Canada
[16]Department of Psychology, University of Oregon, Eugene, Oregon, USA

*Co-senior author
[#]Corresponding author (kay@umn.edu)
[$]Current address: Department of Neuroscience, University of Minnesota, Minneapolis, Minnesota, USA
[%]Current address: Department of Psychology, University of Minnesota, Minneapolis, Minnesota, USA
[^]Current address: Department of Psychology, Harvard University, Cambridge, Massachusetts, USA

Keywords: neuroimaging, big data, high-resolution fMRI, natural stimuli, denoising methods, recognition memory

Abstract

Extensive sampling of neural activity during rich cognitive phenomena is critical for robust understanding of brain function. We present the Natural Scenes Dataset (NSD), in which high-resolution fMRI responses to tens of thousands of richly annotated natural scenes are measured while participants perform a continuous recognition task. To optimize data quality, we develop and apply novel estimation and denoising techniques. Simple visual inspections of the NSD data reveal clear representational transformations along the ventral visual pathway. Further exemplifying the inferential power of the dataset, we use NSD to build and train deep neural network models that predict brain activity more accurately than state-of-the-art models from computer vision. NSD also includes substantial resting-state and diffusion data, enabling network neuroscience perspectives to constrain and enhance models of perception and memory. Given its unprecedented scale, quality, and breadth, NSD opens new avenues of inquiry in cognitive neuroscience and artificial intelligence.

Introduction

Neuroscience has an insatiable appetite for data. Many ongoing efforts to extensively sample brain activity[1–3] and structure[4–6] are motivated, in part, by the availability of new computational methods that make analysis of massive datasets feasible. Equally as important is the growing desire to understand how the brain coordinates complex sensory and motor behaviors and the realization that the neural networks supporting such behaviors span multiple scales, from single neurons to local circuits to whole systems. Understanding massive, complex networks will inevitably require commensurately massive amounts of data.

The need for massive data is especially acute in visual neuroscience, a model system for understanding brain function. The network that mediates our ability to flexibly and efficiently perceive the visual world occupies approximately one-third of human cerebral cortex[7] and interconnects brain areas with profoundly different functional properties[8]. This network both encodes visual stimuli and interfaces visual representations into a cognitive context, including information about what one has already seen[9], might see[10], or is selectively attending[11]. Understanding vision thus means interrogating a high-dimensional, context-dependent neural network.

Given these considerations, it is clear that extensive experimental data providing access to whole-brain responses to complex stimuli are critical in the quest to understand the human visual system. The ideal dataset should include naturalistic stimuli: the visual system is distributed widely across the brain, and natural scenes, in addition to being ecologically relevant, are effective activators of the entire system[12]. Moreover, the ideal dataset should be large: in order to take full advantage of powerful data analysis and machine learning (ML) techniques that have recently become available, we need considerably more data than is currently available. How much? Modern ML methods used in computer vision to process natural scenes (e.g. deep convolutional neural networks) require tens to hundreds of thousands of image samples for training[13,14]. A dataset that sampled brain activity at these scales would raise the exciting possibility of exploiting these methods to develop better models of how the brain processes natural scenes[15–20], and would accelerate efforts to bridge cognitive neuroscience and artificial intelligence[21].

In this paper, we present a dataset that achieves sampling at this ambitious scale. The Natural Scenes Dataset (NSD) consists of high-resolution (1.8 mm) whole-brain 7T fMRI of 8 carefully screened human participants who each viewed 9,000–10,000 color natural scenes (22,000–30,000 trials) during 30–40 scan sessions distributed over the course of a year. Aggregated across participants, NSD includes responses to 70,566 distinct natural scene images—this is more than an order of magnitude larger than comparable datasets involving fMRI sampling of many images[22–24]. Moreover, as we show, the high quality of the NSD dataset makes it possible to leverage the full power of modern ML methods for developing better models of visual representation. Achieving high data quality was afforded, in part, by the use of ultra-high magnetic field strength (7T) to improve signal-to-noise ratio over what is attained at lower field strengths[25].

NSD incorporates several innovations in addition to its unprecedented scale and quality. To reconcile extensive sampling with a practical time commitment, we used an aggressive rapid event-related design. This drove the development of new analysis techniques that accurately compensate for the overlap of hemodynamic responses across successive trials. To ensure participant engagement and control cognitive state, we incorporated a continuous recognition task[26] in which participants were instructed to indicate whether they have seen each presented image at any point in the past. In addition to making the experiment tolerable (and even somewhat interesting) for participants, inclusion of this task makes NSD, to our knowledge, the longest-term continuous recognition memory fMRI study in history and, thus, a likely source of new insights into long-term memory formation and the cognitive context of vision. Finally, to ensure the broad reach of the NSD dataset, we incorporated design input from a large network of collaborators with diverse scientific interests (e.g., low-level vision, high-level vision, memory, connectivity, neuroanatomy) and technical expertise (e.g., mapping, multivariate pattern analysis, encoding models, representational similarity analysis, neural network modeling). This input helped precipitate a carefully curated dataset with extensive auxiliary measures.

The goal of this paper is to provide a comprehensive description of the design, acquisition, and preparation of the NSD dataset. In particular, we detail the state-of-the-art acquisition and analysis methods that we developed for the dataset, and perform comprehensive assessments that evidence the high quality of the data. We also perform initial analyses of the NSD dataset demonstrating the feasibility of using data-driven analyses to reveal insights into vision and memory. We expect that NSD will serve as a valuable resource with widespread application in neuroscience and its intersection with artificial intelligence.

Results

Sampling thousands of images during continuous recognition

We obtained 73,000 color natural scenes from the richly annotated Microsoft Common Objects in Context (COCO) image dataset[14], a dataset that is heavily used in the computer vision and machine learning communities. Our experimental design specified that each of 8 subjects would view 10,000 distinct images and a special set of 1,000 images would be shared across subjects (8 subjects × 9,000 unique images + 1,000 shared images = 73,000 images). This sampling strategy was chosen to maximize the number of distinct images in NSD, while also facilitating investigations of similarities and differences in brain representations across individuals[27]. Each image would be presented 3 times to a given subject. While this is a low number, we reasoned that 3 trials would be sufficient to produce robust responses given our use of ultra-high field (7T) fMRI. Furthermore, images would be presented using a rapid event-related design consisting of 4-s trials (**Figure 1A**). This was done to maximize statistical power and to create an engaging experience for the subjects. In addition, the continuous nature of task engagement—in contrast to slow event-related designs and block designs where engagement is likely to fluctuate—helps avoid unwanted respiratory variations[28] and arousal-related confounds[29].

The NSD experiment was split across 40 scan sessions for each subject (**Figure 1B**). To control cognitive state and encourage deep processing of the images, subjects were instructed to perform a continuous recognition task in which they reported whether the current image had been presented at any previous point in the experiment. We controlled the distributions of image presentations such that both short-term and long-term repetitions were probed (**Extended Data Figure 1A**). Parameters were selected such that even in the first scan session, images were not always new, and even in the last scan session, images were not always old (**Extended Data Figure 1B**).

Neuroimaging data collection on carefully selected subjects

All fMRI data in NSD were collected at 7T using a whole-brain 1.8-mm 1.6-s gradient-echo EPI pulse sequence. After verbally screening a number of potential participants with respect to basic eligibility criteria, we recruited 14 subjects to participate in an initial 7T fMRI screening session which involved population receptive field (pRF)[30] and category localizer (fLoc)[31] experiments. Based on data from this scan session, we ranked the 14 subjects with respect to data quality—specifically, we quantified BOLD variance explained in the pRF and fLoc experiments, behavioral performance in the pRF and fLoc experiments, and two metrics of head motion, normalized these six measures, and then averaged the measures (for details, see 'Rankings from the 7T fMRI screening session' in the Methods). We then invited the top 8 subjects to participate in the full NSD experiment (all subjects accepted). This selection process was conducted to ensure the best possible data quality for NSD. Analyses conducted after completion of the NSD experiment confirm that the ranking procedure successfully identified subjects that yield high-quality data and that data quality would have suffered substantially had we omitted the selection process (**Figure 2C**).

Data were collected from the 8 NSD subjects over the course of a year (**Figure 1C**). Subjects consistently engaged with the task: the average response rate across scan sessions was above 99% for all subjects and the response rate never dropped below 96% in any single scan session. Moreover, all subjects exhibited successful recognition performance (**Figure 1D**), issuing 'old' responses at a higher rate for previously presented images (blue and orange lines) than for novel images (yellow lines). The full NSD dataset includes a variety of anatomical neuroimaging measures (including T_1, T_2, diffusion, venogram, and angiogram), functional neuroimaging measures (including the pRF and fLoc experiments, the NSD experiment, resting-state data, and two additional experiments involving synthetic stimuli and visual imagery), and behavioral measures (**Figure 2A–B**). In some fMRI sessions, physiological data (10 sessions per subject) and eyetracking data (2–4 sessions per subject) were also collected. Analysis of the eyetracking data indicates that subjects were able to successfully maintain central fixation most of the time, with some variability in fixation performance across subjects (**Extended Data Figure 4**). With regards to the core NSD experiment, we completed the full set of 40 NSD scan sessions for four of the subjects, but due to unforeseen summer absences and scheduled decommissioning of the 7T scanner, we completed 30–32 NSD scan sessions for each of the other subjects. A full breakdown of data collection and analysis procedures is provided in **Extended Data Figures 2–3**.

Stable high-resolution imaging across scan sessions

In our experience, although visual inspection is non-quantitative and somewhat subjective, it is still the most effective way to assess many common aspects of fMRI pre-processing[32]. Accordingly, we generated a comprehensive set of visualizations that detail the excellent quality of the raw and pre-processed NSD data. These include detailed inspections of raw time-series data to confirm the presence of stimulus-evoked signals (**Supplementary Figure 3**); movies that assess the co-registration of the different imaging modalities (e.g. T_1, T_2, EPI; **Supplementary Video 1**); movies that assess the manually-edited cortical surface reconstructions generated using FreeSurfer (**Supplementary Video 2**); movies that assess the registration of the NSD subjects to the fsaverage (**Supplementary Video 3**) and MNI (**Supplementary Video 4**) group spaces; movies that inspect raw and pre-processed EPI volumes (**Supplementary Video 5**); and movies that provide volume and surface visualizations of the stability of mean EPI intensity across sessions (**Supplementary Videos 6 and 7; Supplementary Figure 4**) and the stability of BOLD responses across sessions (**Supplementary Videos 8 and 9**). All movies are readily viewable online (https://osf.io/zyb3t/). The visualizations—in particular, **Supplementary Video 9**—indicate that the quality of the NSD data enables precision functional mapping[33]: activity patterns are fine-scale and highly reliable within individual subjects and these patterns are distinct across subjects.

In addition to visual inspection, quantitative data quality metrics were computed for each NSD scan session. This was in fact done on a rolling basis as the data were acquired, allowing us to monitor data quality and provide performance bonuses to the subjects. Inspecting the metrics, we see that temporal signal-to-noise ratio (tSNR) is stable across scan sessions for each subject (**Figure 2D, left**). One subject, subject 8, exhibits low tSNR compared to the other subjects; this can be attributed to higher levels of head motion for this subject (**Figure 2D, middle**). We also observe that BOLD responses (quantified as median variance explained across voxels and runs by a simple ON-OFF GLM) are stable across scan sessions for each subject, though there is substantial variation in the strength of BOLD responses across subjects (**Figure 2D, right**).

One feature we implemented in the pre-processing of the fMRI data was to interpolate the data on a fine temporal grid and a fine spatial grid in the same steps used to correct for slice timing differences and spatial displacements (e.g. head motion). This upsampling strategy preserves fine-scale detail that is present in the raw fMRI data due to the temporal jitter of the acquired fMRI volumes relative to the experimental paradigm and the spatial jitter of the acquired fMRI volumes relative to the brain's anatomy[32,34]. An illustration of the benefits of upsampling is provided in **Extended Data Figure 5**. This example highlights the existence of fine-scale detail in fMRI image intensities (**Extended Data Figure 5B, top row**) as well as in BOLD responses extracted from the fMRI data (**Extended Data Figure 5B, bottom row** and **Extended Data Figure 5C**). Importantly, this fine-scale detail is replicable across different scan sessions (**Extended Data Figure 5C, bottom** and **Extended Data Figure 5D**), indicating that the upsampled preparation reveals meaningful detail that is lost under a non-upsampled approach.

Extensive auxiliary measures to complement the NSD data

To enrich the fMRI data from the NSD experiment, we collected and prepared a large set of auxiliary measures. These measures include substantial amounts of resting-state data (minimum 100 minutes per subject), external physiological measures during the resting-state scan sessions, diffusion data and associated derivatives (white-matter tracts, structural connectivity matrices), and an extensive collection of manually defined regions of interest (ROIs) including retinotopic and category-selective areas as well as subregions of the thalamus and medial temporal lobe. Results and discussion of these resources can be found in **Supplementary Note 1**, **Extended Data Figures 6–7**, and **Supplementary Figure 5**.

Accurate estimation of single-trial fMRI response amplitudes

We performed a general linear model (GLM) analysis of the data from the NSD experiment in order to help streamline subsequent analyses of the data. The goal of the GLM was to obtain single-trial betas, i.e., estimates of the fMRI response amplitude of each voxel to each trial conducted. Given the low signal-to-noise ratio of fMRI and the overlap of the hemodynamic response from trial to trial, estimating accurate betas is a challenging endeavor. We thus developed a novel GLM approach consisting of three components. First, we used a library of hemodynamic response functions (HRFs) derived from an initial analysis of the dataset as an efficient and well-regularized method for estimating voxel-specific HRFs (**Figure 3A–C**). Second, we adapted the GLMdenoise technique[35] to the single-trial GLM framework, thereby enabling the use of data-driven nuisance regressors (**Figure 3D**). Third, to address the challenge posed by highly correlated single-trial regressors, we developed an efficient implementation of ridge regression[36] and used this to regularize and improve the accuracy of the betas (**Figure 3E**). To assess the efficacy of these various GLM techniques, we generated three versions of the betas, reflecting increasing sophistication (**Extended Data Figure 8A–C**). Beta version 1 (b1) is the result of simply using a canonical HRF for all voxels. Beta version (b2) is the result of fitting an HRF to each voxel using the library-of-HRFs approach. Beta version (b3) uses the library-of-HRFs approach like b2 but also adds the use of GLMdenoise and ridge regression in an attempt to improve the accuracy of the betas.

We quantified the quality of the different beta versions (b1, b2, b3) by calculating noise ceilings for individual voxels. The noise ceiling is a measure of trial-to-trial reliability, quantifying the percentage of variance in a voxel's responses that can be attributed to the stimulus and not to measurement noise (see Methods). Surface maps of noise ceiling results reveal locations of reliable responses to the NSD stimuli: high noise ceilings are present in occipital cortex and extend into temporal and parietal cortex (**Figure 3F** and **Supplementary Video 10**). Importantly, the maps reveal very large increases in noise ceilings from b1 to b2 to b3, indicating that the additional GLM techniques incorporated into b2 and b3 improve reliability of responses. Detailed quantifications show that these improvements are highly consistent across voxels and subjects (**Figure 3G** and **Supplementary Figure 6A**) and that noise ceiling estimates are highly reliable (**Supplementary Figure 6B**). For b3, the noise ceiling levels in visual cortex are, on average, 36% (calculated by computing the median across the nsdgeneral ROI and then averaging across subjects). This means that a typical visual cortex voxel in the NSD dataset has associated with it a set of 10,000 responses (30,000 trials divided by 3 trials per image = 10,000 images) and a large percentage, 36%, of the variance in these 10,000 values is a signal that is, in theory, predictable. Expressed in terms of Pearson's correlation (r), this is equivalent to a prediction accuracy of $r = 0.60$. Complementing the noise ceiling analysis, we also performed simple univariate analyses of the NSD betas (**Extended Data Figure 8D–E**); these analyses demonstrate that the NSD dataset contains high response reliability across trials within a subject as well as high response reliability across subjects.

A massive increase in equivalent trials

To put the quality of the NSD data into perspective, we propose the concept of 'equivalent trials' which allows comparison of different datasets that vary in signal-to-noise ratio and trial distribution (see Methods for details). The next largest data collection effort that is similar in nature to NSD is BOLD5000[22]. Using the same GLM analysis methods on both NSD and BOLD5000, we find that the signal-to-noise ratio per trial is approximately 0.260 for NSD and 0.187 for BOLD5000. Combining these values with the number of trials conducted in each dataset, we estimate that the total size of the NSD dataset is 213,000 trials × $(0.260)^2$ = 14,399 equivalent trials, whereas the total size of BOLD5000 is 18,870 trials × $(0.187)^2$ = 660 equivalent trials. Thus, using the metric of equivalent trials, NSD can be viewed as 14,399/660 = ~22 times as large as the BOLD5000 dataset. This is a massive increase in statistical power. Note that even if we do not take into account the higher SNR per trial in the NSD dataset, NSD still has substantially more subjects (8 vs. 4), trials per subject (26,625 vs. 4,718, on average), and hours of fMRI per subject (35.5 vs. 13.7, on average) than BOLD5000.

Successful recovery of retinotopy

Having demonstrated the quality of the NSD data, we now turn to example analyses that illustrate the rich scientific insights that can be derived from the data. As a simple starting example, we fit a voxelwise pRF model that uses local contrast in the NSD images to account for the NSD betas. This simple model is expected to recover spatial tuning in early visual cortex where responses co-vary with stimulus energy[37]. Indeed, in all eight subjects, high-quality maps of angle and eccentricity estimates are obtained in early visual cortex, and these estimates extend all the way to the fovea (**Extended Data Figure 9** and **Supplementary Modeling Note 1**). These results provide a check of the validity of the NSD betas. They also demonstrate that subjects were able to maintain central fixation reliably enough to support detailed mapping of visual space. This finding is consistent with our analysis of the eyetracking data (see **Extended Data Figure 4**).

Reliable and long-term recognition memory effects

The use of a continuous recognition task establishes NSD as one of the largest datasets relevant to human memory. Despite the challenging nature of the task, we find that subjects were able to successfully discriminate old images from new images (average *d*' across subjects: 1.28, maximum: 1.47, minimum: 0.94). Further, recognition memory remained above chance even at long timescales between repetitions (**Figure 4A**). Specifically, for each session, we calculated a measure of recognition accuracy accounting for guessing (adjusted hit rate: hit rate minus false alarm rate) and binned this measure by the time since last exposure (considering only those trials involving a previously shown image). At the group level, subjects exhibit performance levels greater than chance (adjusted hit rate > 0) in all measured intervals, ranging from one second to one year. At the level of individuals, all subjects show a positive adjusted hit rate in the longest time bin for which data are available for every subject (when binning on a log scale; 7 out of 8 subjects when binning on a linear scale). These results indicate that from its behavioral component alone, NSD is powered to address questions concerning human memory spanning short (seconds) to relatively long (months) timescales.

But what about neural effects? To assess whether recognition effects are present in the fMRI data, we performed two-sample *t*-tests contrasting NSD betas observed for hits with NSD betas observed for correct rejections (the so-called 'old/new effect'[38]). We find highly consistent old/new effects at the level of individual scan sessions (**Figure 4B, top;** see also **Supplementary Figure 7**). Moreover, these effects occur in expected frontal and parietal regions[39], and persist at the group level (**Figure 4B, bottom**). The scale and statistical power afforded by the NSD dataset also provides additional insight. Whereas old/new effects are typically studied using group-level analyses, the quality of the NSD dataset reveals highly statistically significant results at the level of individual subjects. Indeed, when pooling trials across all NSD scan sessions, several subjects exhibit statistically significant activity differentiating hits and correct rejections in nearly the entire cerebral cortex (see results for a representative subject in **Figure 4B, top**). Reminiscent of past datasets employing extensive sampling of individuals[40], the current results suggest that the extent of cortex engaged by basic memory processes is much more widespread than previously appreciated, though a careful consideration of effect sizes would be important for a full understanding of the effect.

Rich stimulus sampling for probing brain representations

NSD samples a huge variety of natural scenes. To gain insight into the breadth of stimulus sampling available, we constructed representational dissimilarity matrices (RDMs) from the NSD betas and performed *t*-distributed stochastic neighbor embedding[41] (t-SNE) to visualize the underlying representations. We computed t-SNE embeddings in different regions along the ventral visual pathway for an example subject (**Figure 5A**). These embeddings reflect arrangements of stimuli that are driven by the overall similarity of multivoxel activity patterns in the brain, independent of their anatomical organization within a given ROI. Visualizing the data in this way reveals intriguing patterns of semantic representation that are clearly visible by eye. For example, by color-coding the resulting embeddings according to animacy attributes (**Figure 5B**), we find that in posterior ventral temporal cortex (pVTC), there is a clear large-scale pattern progressing from images containing people (gray dots; lower left), images containing animals (red dots; middle), and images containing inanimate objects (blue dots; upper right), whereas the pattern is not present in early visual areas V1, V2, and V3. This aspect of semantic representation is consistent with previous studies[42,43].

Other intriguing patterns are also visible. In anterior ventral temporal cortex (aVTC), the animacy progression is present to some extent, but a different, more clustered representation emerges that presumably reflects more complex categorical and semantic clusters. Indeed, zooming in on small sections of the t-SNE embedding for aVTC reveals that these clusters contain images with relatively homogeneous semantic content (**Figure 5C**): the blue cluster is dominated by images of round edible objects, while the gray cluster is dominated by images of people interacting with objects. Note that the clustering of semantically related images does not necessarily mean that these representations are truly semantic in the sense of being invariant or independent of visual features; the clustering could be driven by certain visual features that are diagnostic of object categories[44]. To tease apart these possibilities, further detailed analyses would be necessary. Overall, these findings show how simple visual inspections of the NSD dataset can be used to generate hypotheses about visual representations in the human brain.

To further characterize brain representations using a quantitative analysis, we calculated how well brain RDMs are captured by a model RDM constructed from category labels in the COCO image dataset. Consistent with the clustering observed in the t-SNE embeddings, we find that categorical structure is pronounced in VTC compared to early visual areas (**Figure 5D**). Finally, to assess the utility of NSD for investigating similarities of brain representations across subjects, we isolated images that were common across subjects and created a second-order RDM that quantifies the similarity of brain RDMs across ROIs and subjects (**Figure 5E**). In this second-order RDM, we observe high levels of consistency in each ROI's representation across subjects (red outlines). We also observe distinct representations across ROIs, with the largest distinctions occurring between early visual areas and VTC. One noticeable finding is the existence of strong off-diagonal elements (white arrows); these elements indicate spatial noise correlations that are typical in fMRI and other neural measurement techniques. To counteract these noise correlations, one simple approach is to compare representations across ROIs using data from distinct trials[45]. To further summarize the second-order RDM, we computed the average correlation of brain RDMs across all ROI pairs, restricting this calculation to distinct subjects in order to avoid the effects of spatial noise correlations (**Figure 5F**). We observe that correlations are highest for brain RDMs from the same ROI (e.g. a given subject's V1 RDM is more correlated with other subjects' V1 RDMs compared to other ROIs), confirming consistencies in brain representations across subjects (for a complementary univariate analysis of across-subject consistency, see **Extended Data Figure 8D–E**).

A brain-optimized neural network model of the visual system

One of the main motivations for NSD was to amass sufficient sampling of brain activity to be able to drive data-hungry machine learning techniques. As an intriguing test case, we specifically investigated whether we could successfully use the scale of NSD to train, from scratch, a deep convolutional neural network (CNN) to accurately predict brain activity[17]. Adopting the framework of encoding models[46], we took NSD betas from visual areas V1–hV4, divided these data into a training set (used for parameter tuning) and validation set (used to assess prediction performance), and evaluated how accurately different computational models predict brain responses in the validation set based on the presented image. The primary encoding model of interest is based on a new network we refer to as 'GNet', a *brain-optimized* CNN whose parameters are trained using image-response pairings observed in the training set. For comparison, we also evaluated an encoding model based on AlexNet[47], a *task-optimized* CNN whose parameters are pre-trained using explicit labels of objects taken from an image database. AlexNet has been previously shown to provide state-of-the-art performance in modeling visual responses[15,19]. Finally, we included a simple V1-like control model based on oriented Gabor filters[24]. Details of modeling procedures are provided in **Supplementary Modeling Note 2** and **Extended Data Figure 10**.

Varying the amount of training data provided to the models, we find that the performance of the GNet-based encoding model is relatively poor when only small amounts of training data are available (**Figure 6A, orange arrows**). This is expected since the feature extractors in GNet are not pre-trained and thus require data for tuning. However, when large amounts of training data are available, the GNet model exhibits an impressive increase in performance, achieving approximate parity with the AlexNet-based encoding model (**Figure 6A, blue arrows**). Interestingly, when we trained a single GNet model using brain activity from multiple subjects, we find that the model is able to outperform the AlexNet model (two-tailed paired *t*-test across subjects, $p = 0.013$), albeit modestly (**Figure 6A, red arrows**). Noticeably, the simple Gabor model accounts for substantial variance in the responses; nonetheless, the more complex CNN-based models provide additional predictive power, consistent with previous observations[48]. For additional insight into model performance, we compared voxel-wise performance levels of the GNet model to noise ceiling estimates (**Figure 6B**). Across voxels, prediction accuracy is tightly correlated with the noise ceiling, suggesting that voxel-wise differences in prediction accuracy simply reflect differences in signal-to-noise ratio. In addition, performance levels are close to, but do not reach, the noise ceiling. Finally, cortical surface maps indicate voxel-wise performance levels vary across foveal and peripheral representations (**Figure 6C**).

The demonstration that an encoding model based on a brain-optimized CNN (GNet) outperforms an encoding model based on a task-optimized CNN (AlexNet) is significant for two reasons. First, it indicates NSD is large enough to successfully train a complex neural network architecture. Had the NSD dataset been smaller in scale or lower in quality, qualitatively different patterns of model performance would have been obtained (in **Figure 6A**, compare orange arrows reflecting a few thousand trials to red arrows reflecting tens of thousands of trials). Second, the successful training of a brain-optimized CNN opens the possibility of new avenues of investigation into the nature of the features used in CNNs. It is an interesting open question whether the features learned by task-optimized networks like AlexNet are similar to, or diverge from, the features present in brain-optimized networks like GNet. In general, brain-optimized networks[17] are a useful alternative to task-optimized networks[16,20], as the narrowly defined tasks that task-optimized networks are typically trained to solve do not necessarily respect the diversity of functions supported by the human visual system[49] nor necessarily match properties found in biological visual systems[50].

Discussion

In the last several years, there have been a number of large-scale neuroimaging datasets that have been made publicly available for re-use (e.g., refs [5,33,51–53]). Several distinguishing aspects of the present work set NSD apart from past datasets. One is the *unprecedented scale of the dataset*. NSD shares the motivation of recent 'deep' (or 'precision') neuroimaging efforts[33,54–57] seeking to amass large amounts of data from individual subjects, as opposed to modest amounts of data on a large number of subjects. In this context of deep neuroimaging, NSD is, to our knowledge, the most extensive fMRI data collection effort that has been performed to date. This can be gauged not only in terms of the number of hours of fMRI data acquisition per subject (30–40 hours of data for each of 8 subjects on the core NSD experiment) and the high spatial resolution of the acquired data (1.8 mm), but also the wealth of additional measures beyond the core experiment, including substantial amounts of resting-state and diffusion data, physiological data, and functional localizers. The availability of extensive measures provides the opportunity to build complete models of how individual brains support vision and memory[58]. Of course, the emphasis on depth in individuals comes at the cost of sampling fewer individuals; datasets emphasizing large numbers of participants, such as the Human Connectome Project[5], are better suited for studying variability in the general population and how psychological traits broadly relate to brain structure and function.

A second aspect is the *unusually high quality of the data*. Although quality of neuroimaging data is more complex to assess than quantity, assessment of data quality is essential since MRI data have relatively low sensitivity and are prone to errors and artifacts. In particular, when acquiring massive datasets, there is a risk of accumulating unknown sources of noise and artifact. The work presented in this paper (and in the accompanying files in the data release) guards against this possibility by crafting a customized and highly optimized approach to pre-processing the NSD data and providing comprehensive documentation of the high data quality (see also **Supplementary Note 2**). Several factors likely contributed to the high data quality: these include the use of ultra-high magnetic field strength (7T) which enhances BOLD contrast-to-noise ratio; the screening, training, and incentivization of participants; the detailed inspection and supervision of data processing; and the large network of collaborators who helped guide the design and trajectory of the dataset.

A third aspect of the present work lies in the *novel analysis techniques* developed for improved GLM analysis of fMRI time-series data. These include (i) an efficient and robust method to estimate voxel-specific HRFs, (ii) adaptation of the GLMdenoise technique[35] to a single-trial GLM framework, and (iii) development of ridge regression as an effective method for regularizing single-trial response estimates. These three techniques have been integrated into a toolbox that can be applied to other neuroimaging datasets, and are the subject of a forthcoming paper. An important lesson stemming from our results is that well-executed data collection is important but not the only factor to consider: data preparation methods exert a major influence on the quality of a dataset and hence its scientific value. One can view improvements in data quality as equivalent to increases in data quantity, in the sense that analysis methods that reduce unwanted variability (noise) can be interpreted as increasing the effective amount of data collected[35]. Thus, by improving data quality, the methods introduced with NSD are contributing to the massive scale of the dataset.

The NSD dataset has many potential applications. Given its extensive sampling of natural scenes (70,566 distinct images aggregated across 8 subjects) and high signal-to-noise ratio, the dataset will be useful for investigating a variety of phenomena in low-, mid-, and high-level vision. In addition, the memory component of the NSD experiment provides a unique opportunity to study the neural mechanisms of both short- and long-term memory (ranging from seconds to many months), as well as potential interactions between vision and memory. From a methodological perspective, the repeated scanning of individuals using a consistent experimental manipulation (up to 40 scan sessions of the NSD experiment per subject) provides a unique opportunity for development and evaluation of neuroimaging pipelines. Finally, perhaps the most exciting use of NSD is as a common dataset to bridge the disciplines of cognitive science, neuroscience, and artificial intelligence[21]. As we have shown in the context of deep neural network modeling (see **Figure 6**), there are sufficient data in NSD to successfully drive the training of neural network models with thousands of free parameters. This demonstration exemplifies how NSD—with its large amounts of carefully curated fMRI data collected during a rich cognitive paradigm—enables data-driven approaches towards understanding the complexities of information processing in the brain.

Acknowledgements

We thank the NSD subjects for their time and endurance; E. Aminoff, J. Pyles, M. Tarr, M. Hebart, and C. Baker for advice on experimental design and data collection; J. Power and A. Schapiro for consultation on resting-state and physiological data; V. Carr and R. Olsen for consultation on hippocampal subfield scanning protocols; A. Grant for assistance with scanner peripherals; F. Gosselin and J. Tardif for contrast sensitivity analysis; B. Klimes-Dougan and K. Cullen for designing the valence/arousal assessment; W. Guo for segmentations of the medial temporal lobe; M. Arcaro, A. Bratch, D. Finzi, A. White, and J. Winawer for assistance with ROI definition; C. Gorgolewski and R. Poldrack for discussion of BIDS and data sharing; R. Cichy, E. Yacoub, K. Grill-Spector, K. Jamison, A. Rokem, A. Huth, S. Anzellotti, N. Kriegeskorte, and J. Winawer for general discussions; and K. Ugurbil for overall project advice. We also thank our NSD collaborators for shaping the trajectory of the project. This work was supported by NSF CRCNS grants IIS-1822683 (K.K.) and IIS-1822929 (T.N.); NIH grants P41 EB015894, P30 NS076408, S10 RR026783, S10 OD017974-01, and the W. M. Keck Foundation; NIMH Intramural Research Program ZIAMH002909 (M.N.); and NSF BCS-1734853, NIH NIBIB R01EB030896, NIH NIBIB R01EB029272, and NIH IIS-1912270 (F.P.).

Author Contributions

E.J.A. collected the neuroimaging data, coordinated the data collection effort, and performed manual brain segmentations. G.S.-Y. performed neural network analyses. Y.W. performed subject recruitment, assisted with scanning, and prepared eyetracking videos. J.L.B. assisted in data analysis. J.S.P. performed the equivalent-trials analysis on NSD and BOLD5000. L.T.D. organized and prepared data in BIDS format. M.N. analyzed the eyetracking data. B.C. and F.P. analyzed the diffusion data. I.C. performed representational similarity analyses. J.B.H. analyzed the behavioral data. K.K. and T.N. conceived of the project and designed the main experiment. J.B.H. and I.C. designed the nsdmeadows and nsdmemory behavioral assessments. K.K. developed analysis methods, analyzed the neuroimaging data, and directed the overall project. K.K., T.N., E.J.A., M.N., B.C., F.P., I.C., and J.B.H. wrote the paper. All authors discussed and edited the manuscript.

Competing Interests

The authors declare no competing interests.

Figure Captions

Figure 1. Design of the NSD experiment. *A*, Trial design. While maintaining central fixation, participants viewed sequences of color natural scenes and judged whether each image had been previously shown at any point in the past. The scenes, taken from Microsoft's COCO[14], are richly annotated with object information (as depicted). *B*, Run and session design. Each run lasted 5 minutes and consisted of 62 or 63 stimulus trials with occasional interspersed blank trials. Each scan session consisted of 12 runs (750 stimulus trials). *C*, Timeline of 7T fMRI scan sessions. Each subject participated in an initial screening session (prffloc), 30–40 NSD core sessions, and two final sessions (nsdsynthetic, nsdimagery). The first NSD core session corresponds to day 0. *D*, Behavioral performance. For each of three trial types, we quantify the percentage of trials on which the subject indicated an 'old' response.

Figure 2. Overview of acquired data. *A*, Auxiliary fMRI experiments. Data from the pRF and fLoc experiments were used to define retinotopic visual areas and category-selective regions,

respectively. Resting-state data were collected before and after the NSD runs in a subset of the NSD core sessions (totaling 100 or 180 minutes per subject). *B*, Available measures. Examples of the actual data are depicted. *C*, Participant selection. Data quality from the initial screening session was used to rank a set of 14 participants. On the right is an illustration of one measure contributing to the ranking, specifically, variance explained in the fLoc experiment (one slice per participant; identical color range). The inset compares the participant ranking against the b3 noise ceiling calculated on the full NSD dataset (see **Figure 3**). A line fit to the 8 NSD subjects (green dots) is extrapolated to predict noise ceilings for the subjects who were not selected for participation in NSD (red circles). *D*, Metrics of data quality (for details, please see 'Data quality metrics' in the Methods). Results for individual subjects (thin colored lines) and the median across subjects (thick black line) are shown. The insets show detail on tSNR and head motion for one sample run (see **Supplementary Figures 1–2** for more information).

Figure 3. Improving signal-to-noise ratio through novel response estimation and denoising methods. *A*–*C*, Library of HRFs. Hemodynamic response functions (HRFs) were estimated within a subspace spanned by 3 principal components (PCs). Distributions of voxel-specific HRFs are shown for individual subjects (panel A) and the group average (panel B). These distributions reside on the unit sphere with coordinate axes corresponding to 3 PC timecourses (see panel B, inset). We defined a series of points on the unit sphere (cyan dots), and the timecourses associated with these points are used as the HRF library (panel C). *D*, GLMdenoise. Horizontal lines indicate the average number of GLMdenoise regressors identified in a scan session (1.8-mm preparation; error bars indicate bootstrapped 68% confidence intervals). *E*, Ridge regression. Optimal ridge regression fractions are shown for an example scan session (subject 5, nsd10, 1-mm preparation). *F*, Noise ceilings for the case where responses are averaged across 3 trials. Results from individual subjects (nativesurface preparation) were mapped to fsaverage and then averaged. Right inset shows results thresholded at 15% on the inflated left hemisphere (see also **Supplementary Video 10**). *G*, Performance summary. Each bar indicates the median noise ceiling across vertices in the nsdgeneral ROI.

Figure 4. Reliable and long-term recognition memory effects. *A,* Behavioral recognition effects. Adjusted hit rate indicates recognition accuracy accounting for guessing (hit rate minus false alarm rate), and is binned by time between repetitions on a linear (left) or log scale (right). Dashed line indicates chance performance. Each dot in each bin summarizes relevant trials from one scan session. Black line indicates the mean across subjects, with ribbon indicating ± 1 SEM. *B*, Neural recognition effects. We performed two-sample t-tests on NSD betas contrasting 'hits' > 'correct rejections'. All results are shown on a flattened left hemisphere fsaverage surface and thresholded at $|t| > 3$ (inset shows inflated surface). Tests were performed for trials taken from individual NSD scan sessions (columns 1 through 4) as well as for trials pooled across all NSD scan sessions (column 5). In addition, we perform a control in which trial labels in the pooled analysis are shuffled (column 6). Results for subject 1 (top row) and a simple average of results across subjects (bottom row) are shown.

Figure 5. Representational similarity analysis (RSA) reveals transformations of representations along the ventral visual stream. *A*, Illustration of fsaverage ROIs used for the RSA analysis. *B*, t-SNE embedding for each ROI in an example subject (subject 1). Each dot represents a distinct image (total 10,000). Using category labels from the COCO image dataset, we color each dot according to whether the associated image contains particular combinations of people, animals, and inanimates. *C*, t-SNE embedding for anterior ventral temporal cortex with actual images depicted. Insets highlight an inanimate cluster (blue inset) and a cluster of people with inanimate objects (gray inset). *D*, Categorical brain repesentations. We plot the correlation between brain RDMs and a model RDM constructed from category labels in the COCO dataset.

Color shaded regions indicate within-subject error (mean and standard error across distinct groups of images), while the gray shaded region indicates across-subject error (mean and standard error across subjects). *E*, Similarities of brain representations across ROIs and subjects. Depicted are correlations across brain RDMs obtained for different ROIs and subjects. Thin white lines separate groups of 8 subjects. *F*, Quantitative summary. We summarize the results of panel E by averaging the upper triangle of each group of 8 × 8 subjects, reflecting the correlation of RDMs from different subjects. Shaded regions indicate standard errors estimated by bootstrapping subjects with replacement.

Figure 6. Prediction of brain activity using a brain-optimized neural network. We used encoding models[46] to predict voxel activity in V1–hV4. NSD betas were divided into a training set (consisting of up to 9,000 images × 3 trials = 27,000 training samples per subject) and validation set (consisting of up to 1,000 images × 3 trials = 3,000 validation samples per subject), and the accuracy of different encoding models was quantified as the voxel-wise correlation between model predictions and responses observed in the validation set. *A*, Performance as a function of amount of training data used. Models include an encoding model based on AlexNet which is a task-optimized neural network (blue); encoding models based on GNet which is a brain-optimized neural network trained using data from single subjects (orange) or data from multiple subjects (red); and a V1-like control model based on Gabor filters (purple). Plotted lines and error bars indicate mean and standard deviation across results obtained from different bootstrap samples of the data. *B*, Detailed view of the performance of the multi-subject GNet model for a representative subject. *C*, Surface maps depicting spatial distribution of validation accuracy for the multi-subject GNet model.

References

1. de Vries, S. E. J. et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. Nat Neurosci 23, 138–151 (2020).
2. Siegle, J. H. et al. Survey of spiking in the mouse visual system reveals functional hierarchy. Nature (2021) doi:10.1038/s41586-020-03171-x.
3. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. Nature 571, 361–365 (2019).
4. Markram, H. et al. Reconstruction and Simulation of Neocortical Microcircuitry. Cell 163, 456–492 (2015).
5. Van Essen, D. C. et al. The WU-Minn Human Connectome Project: an overview. NeuroImage 80, 62–79 (2013).
6. Zheng, Z. et al. A Complete Electron Microscopy Volume of the Brain of Adult Drosophila melanogaster. Cell 174, 730-743.e22 (2018).
7. Van Essen, D. C. et al. Mapping visual cortex in monkeys and humans using surface-based atlases. Vision Res 41, 1359–1378 (2001).
8. Grill-Spector, K. & Malach, R. The human visual cortex. Annual review of neuroscience 27, 649–677 (2004).
9. Wheeler, M. E., Petersen, S. E. & Buckner, R. L. Memory's echo: Vivid remembering reactivates sensory-specific cortex. PNAS 97, 11125–11129 (2000).
10. Breedlove, J. L., St-Yves, G., Olman, C. A. & Naselaris, T. Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images. Curr Biol 30, 2211-2224.e6 (2020).
11. Kay, K. N., Weiner, K. S. & Grill-Spector, K. Attention reduces spatial uncertainty in human ventral temporal cortex. Curr Biol 25, 595–600 (2015).
12. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224 (2012).
13. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. University of Toronto (2009).
14. Lin, T.-Y. et al. Microsoft COCO: Common Objects in Context. in Computer Vision – ECCV 2014 vol. 8693 740–755 (Springer, Cham, 2014).
15. Güçlü, 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).
16. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology 10, e1003915 (2014).
17. Seeliger, K. et al. End-to-end neural system identification with neural information flow. PLOS Computational Biology 17, e1008558 (2021).
18. Stansbury, D. E., Naselaris, T. & Gallant, J. L. Natural scene statistics account for the representation of scene categories in human visual cortex. Neuron 79, 1025–1034 (2013).
19. St-Yves, G. & Naselaris, T. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. NeuroImage (2017) doi:10.1016/j.neuroimage.2017.06.035.
20. Yamins, D. L. K. et al. 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, 8619–8624 (2014).
21. Naselaris, T. et al. Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline. Trends Cogn Sci 22, 365–367 (2018).
22. Chang, N. et al. BOLD5000, a public fMRI dataset while viewing 5000 visual images. Sci Data 6, 49 (2019).
23. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. Nature Communications 8, 15037 (2017).
24. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. Nature 452, 352–355 (2008).
25. Triantafyllou, C. et al. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. Neuroimage 26, 243–250 (2005).
26. Brady, T. F., Konkle, T., Alvarez, G. A. & Oliva, A. Visual long-term memory has a massive storage capacity for object details. PNAS 105, 14325–14329 (2008).
27. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. eLife 9, e56601 (2020).
28. Power, J. D., Lynch, C. J., Adeyemo, B. & Petersen, S. E. A Critical, Event-Related Appraisal of Denoising in Resting-State fMRI Studies. Cereb Cortex 30, 5544–5559 (2020).
29. Roth, Z. N., Ryoo, M. & Merriam, E. P. Task-related activity in human visual cortex. PLoS Biol 18, e3000921 (2020).
30. Benson, N. C. et al. The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. J Vis 18, (2018).
31. Stigliani, A., Weiner, K. S. & Grill-Spector, K. Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. J. Neurosci. 35, 12412–12424 (2015).
32. Kay, K. et al. A critical assessment of data quality and venous effects in sub-millimeter fMRI. NeuroImage 189, 847–869 (2019).
33. Gordon, E. M. et al. Precision Functional Mapping of Individual Human Brains. Neuron 95, 791-807.e7 (2017).
34. Kang, X., Yund, E. W., Herron, T. J. & Woods, D. L. Improving the resolution of functional brain imaging: analyzing functional data in anatomical space. Magnetic resonance imaging 25, 1070–1078 (2007).

35. Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F. & Wandell, B. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front Neurosci 7, 247 (2013).

36. Rokem, A. & Kay, K. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. GigaScience 9, (2020).

37. Albrecht, D. G. & Hamilton, D. B. Striate cortex of monkey and cat: contrast response function. Journal of neurophysiology 48, 217–237 (1982).

38. Wagner, A. D., Shannon, B. J., Kahn, I. & Buckner, R. L. Parietal lobe contributions to episodic memory retrieval. Trends Cogn Sci 9, 445–453 (2005).

39. Spaniol, J. et al. Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. Neuropsychologia 47, 1765–1779 (2009).

40. Gonzalez-Castillo, J. et al. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. PNAS 109, 5487–5492 (2012).

41. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008).

42. Connolly, A. C. et al. The Representation of Biological Classes in the Human Brain. J. Neurosci. 32, 2608–2618 (2012).

43. Naselaris, T., Stansbury, D. E. & Gallant, J. L. Cortical representation of animate and inanimate objects in complex natural scenes. Journal of physiology, Paris 106, 239–249 (2012).

44. Long, B., Yu, C.-P. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. PNAS 115, E9015–E9024 (2018).

45. Henriksson, L., Khaligh-Razavi, S.-M., Kay, K. & Kriegeskorte, N. Visual representations are dominated by intrinsic fluctuations correlated between areas. NeuroImage 114, 275–286 (2015).

46. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. NeuroImage 56, 400–410 (2011).

47. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. 1097–1105 (2012).

48. Cadena, S. A. et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLOS Computational Biology 15, e1006897 (2019).

49. Wang, A. Y., Wehbe, L. & Tarr, M. J. Neural Taskonomy: Inferring the Similarity of Task-Derived Representations from Brain Activity. bioRxiv 708016 (2019) doi:10.1101/708016.

50. Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M. & Tolias, A. S. Engineering a Less Artificial Intelligence. Neuron 103, 967–979 (2019).

51. Aliko, S., Huang, J., Gheorghiu, F., Meliss, S. & Skipper, J. I. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. Sci Data 7, 347 (2020).

52. Nastase, S. A., Liu, Y.-F., Hillman, H., Norman, K. A. & Hasson, U. Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. Neuroimage 217, 116865 (2020).

53. Taylor, J. R. et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage 144, 262–269 (2017).

54. Bellec, P. & Boyle, J. A. Bridging the gap between perception and action: the case for neuroimaging, AI and video games. https://osf.io/3epws (2019) doi:10.31234/osf.io/3epws.

55. Pinho, A. L. et al. Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. Sci Data 5, 180105 (2018).

56. Poldrack, R. A. et al. Long-term neural and physiological phenotyping of a single human. Nat Commun 6, 8885 (2015).

57. Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E. & Gerven, M. A. J. van. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. bioRxiv 687681 (2019) doi:10.1101/687681.

58. Naselaris, T., Allen, E. & Kay, K. Extensive sampling for complete models of individual brains. Current Opinion in Behavioral Sciences 40, 45–51 (2021).

Methods

Subject recruitment

The NSD study was advertised to the University of Minnesota community. We sought to recruit right-handed individuals (18–65 years old) with no known cognitive deficits nor color blindness and with normal or corrected-to-normal vision. Those who were interested in participating were contacted for a phone interview to explain the nature of the study and to screen them for eligibility. We discussed the long-term nature of the study, the time commitment it would involve, and the feasibility of traveling to the scanner on a regular basis. We paid attention to the communicativeness of potential participants and their general attitude towards study participation. Selecting participants whom we were confident would provide high-quality data was more important to us than obtaining a random sample of the general population. Based on the phone interviews, we invited 14 people who we felt were strong candidates to participate in an initial 7T fMRI screening session. Of these, 8 were selected to participate in the full NSD experiment.

Subjects

Eight subjects (two males, six females; age range 19–32) participated in the NSD dataset (subj01–subj08). There were six additional subjects (four males, two females; age range 20–53) who participated in the initial 7T fMRI screening session but not in the remainder of data collection. No statistical methods were used to pre-determine the sample size; rather, our experimental approach[58] emphasizes collecting extensive data from each subject, which enables the demonstration and replication of effects in individual subjects. Subjects were naïve to the design of the NSD dataset. All subjects had normal or corrected-to-normal visual acuity. Informed written consent was obtained from all subjects, and the experimental protocol was approved by the University of Minnesota Institutional Review Board. Subjects were compensated at a rate of $30 per hour, plus performance bonuses. Additional subject information including height, weight, handedness, and visual acuity was logged and is available online.

Subjects participated in a number of neuroimaging and behavioral data collection sessions (a full breakdown is provided in **Extended Data Figure 2**). Neuroimaging included 3T structural scan sessions and 7T functional scan sessions. The 7T functional scan sessions included an initial screening session termed ‘prffloc’, referring to the population receptive field (pRF) and functional localizer (fLoc) experiments conducted in that session. The 7T sessions also included, for each subject, 30–40 sessions in which the main NSD experiment was conducted (‘nsd01–nsd40’). These sessions are collectively termed the ‘NSD core’. In some of these sessions, resting-state data were acquired before and after the NSD experiment. Finally, the 7T sessions also included two sessions conducted after completion of the NSD core; these sessions, termed ‘nsdsynthetic’ and ‘nsdimagery’, involved measuring responses to synthetic stimuli and cognitive task manipulations (including mental imagery), respectively. The total number of 7T fMRI scan sessions were 43, 43, 35, 33, 43, 35, 43, and 33 for subj01–subj08, respectively. The average number of hours of resting-state fMRI conducted for each subject was 2.0 hours, and the average number of hours of task-based fMRI conducted for each subject was 38.5 hours. Each subject also participated in several behavioral assessments after scanning was complete. These included a variety of behavioral measures (‘nsdpostbehavior’), a final memory test (‘nsdmemory’), and an image-similarity assessment (‘nsdmeadows’).

MRI data acquisition

MRI data were collected at the Center for Magnetic Resonance Research at the University of Minnesota. Some data were collected using a combination of a 3T Siemens Prisma scanner and a standard Siemens 32-channel RF head coil. Most data were collected using a combination of a 7T Siemens Magnetom passively-shielded scanner and a single-channel-transmit, 32-channel-receive RF head coil (Nova Medical, Wilmington, MA). Illustrations of the different types of MRI data acquired are provided in **Figure 2B**. Below we summarize the scanning protocols (full protocol printouts are available online).

At 3T, we collected a number of anatomical measures (T_1, T_2, diffusion, angiogram). The motivation for collecting data at 3T was to ensure acquisition of T_1 volumes with good gray/white-matter contrast and homogeneity, which is difficult to achieve at ultra-high field[59]. To increase contrast-to-noise ratio and enable the ability to assess reliability, we acquired several repetitions of T_1- and T_2-weighted volumes. For each subject, we collected between 6–10 scans of a whole-brain T_1-weighted MPRAGE sequence (0.8-mm isotropic resolution, *TR* 2400 ms, *TE* 2.22 ms, *TI* 1000 ms, flip angle 8°, bandwidth 220 Hz/pixel, no partial Fourier, in-plane acceleration factor (iPAT) 2, TA 6.6 min/scan) and 2–3 scans of a whole-brain T_2-weighted SPACE sequence (0.8-mm isotropic resolution, *TR* 3200 ms, *TE* 563 ms, bandwidth 744 Hz/pixel, no partial Fourier, in-plane acceleration factor (iPAT) 2, TA 6.0 min/scan). In addition to T_1 and T_2 data, we also acquired 4 high angular resolution diffusion-weighted spin-echo EPI scans, using protocols from the Lifespan Human Connectome Project Development effort[60]. These protocols involved varying the number of diffusion directions and the phase-encode direction (1.5-mm isotropic resolution, *TR* 3230 ms, *TE* 89.20 ms, flip angle 78°, refocusing flip angle 160°, bandwidth 1700 Hz/pixel, echo spacing 0.69 ms, partial Fourier 6/8, no in-plane acceleration, multiband slice acceleration factor 4, TA 5.6 min/scan for 99 directions, TA 5.7 min/scan for 100 directions). The 4 scans included 99 directions AP (anterior-to-posterior phase-encode direction), 99 directions PA (posterior-to-anterior phase-encode direction), 100 directions AP, and 100 directions PA. Diffusion volumes were acquired at *b*-values of 0, 1,500, or 3,000 s/mm^2. We also acquired an angiogram using a time-of-flight (TOF) multi-slab 3D sequence (0.39 mm × 0.39 mm × 0.5 mm resolution, *TR* 19.0 ms, *TE* 2.91 ms, flip angle 18°, bandwidth 186 Hz/pixel, phase partial Fourier 6/8, slice partial Fourier 6/8, in-plane acceleration factor (iPAT) 2, *TA* 5.5 min).

At 7T, we collected functional data and associated fieldmaps and a few additional anatomical measures (venogram, high-resolution T_2). Functional data were collected using gradient-echo EPI at 1.8-mm isotropic resolution with whole-brain (including cerebellum) coverage (84 axial slices, slice thickness 1.8 mm, slice gap 0 mm, field-of-view 216 mm (FE) × 216 mm (PE), phase-encode direction anterior-to-posterior, matrix size 120 × 120, *TR* 1600 ms, *TE* 22.0 ms, flip angle 62°, echo spacing 0.66 ms, bandwidth 1736 Hz/pixel, partial Fourier 7/8, in-plane acceleration factor (iPAT) 2, multiband slice acceleration factor 3). The use of moderate spatial resolution capitalizes on the signal-to-noise ratio benefits provided by ultra-high magnetic field strength. At the beginning of each 7T session, we acquired a short test EPI scan and adjusted the gain factor (FFT scale factor) accordingly to ensure large dynamic range while avoiding clipping. Empirical measurements indicate that the acoustic noise caused by the EPI sequence is 112 dBA; assuming a conservative noise reduction estimate of 26 dB for the earplugs that we used, the resulting noise level is 86 dBA, which can be safely endured for approximately 8–16 continuous hours according to guidelines from the National Institute for Occupational Safety and Health (NIOSH) 1998 and Occupational Safety and Health Administration (OSHA) 2009.

In addition to the EPI scans, the 7T sessions also included dual-echo fieldmaps for post-hoc correction of EPI spatial distortion (same overall slice slab as the EPI data, 2.2 mm × 2.2 mm × 3.6 mm resolution, *TR* 510 ms, TE_1 8.16 ms, TE_2 9.18 ms, flip angle 40°, bandwidth 301 Hz/pixel, partial Fourier 6/8, *TA* 1.3 min/scan). Fieldmaps were periodically acquired over the course of each scan session to track changes in the magnetic field (details provided below). In one of the 7T sessions held for each subject, we acquired a venogram using a susceptibility-weighted imaging (SWI) 3D sequence (0.5625 mm × 0.5625 mm × 0.6 mm resolution, *TR* 28 ms, *TE* 21 ms, flip angle 17°, bandwidth 120 Hz/pixel, phase partial Fourier 6/8, slice partial Fourier 6/8, in-plane acceleration factor (iPAT) 3, *TA* 10.1 min). This venogram could be useful for investigating the impact of vasculature on fMRI signals[32]. In addition, for the purposes of hippocampal segmentation, we acquired in one of the 7T sessions a high-resolution T_2-weighted TSE scan (0.357 mm × 0.357 mm × 1.5 mm resolution, 56 oblique slices oriented perpendicular to the long axis of the hippocampus, field-of-view 160 mm (FE) × 156.4 mm (PE), *TR* 16000 ms, *TE* 53 ms, bandwidth 100 Hz/pixel, no partial Fourier, in-plane acceleration factor (iPAT) 2, turbo factor 15, *TA* 4.5 min).

In the prffloc 7T fMRI session, the acquisition structure was [F BWLL F BWLL F BWLL F], where F indicates a fieldmap, B indicates a multibar run of the pRF experiment (188 TRs), W indicates a wedgering run of the pRF experiment (188 TRs), and L indicates a run of the fLoc experiment (195 TRs). In the NSD 7T fMRI sessions, the acquisition structure was either [F NNNN F NNNN F NNNN F] or [F RNNNN F NNNN F NNNNR F], where F indicates a fieldmap, N indicates a run of the NSD experiment (188 TRs), and R indicates a resting-state run (188 TRs).

Stimulus display and scanner peripherals

Ear plugs were used to reduce acoustic noise experienced by the subjects. To minimize head motion, we acquired a headcase[61] for each of the 8 NSD subjects (Caseforge, Berkeley, CA; http://caseforge.co) and deployed the headcases starting from the second NSD core scan session (nsd02). To ensure maximal subject comfort, only the posterior half of the headcases were used (omitting the anterior half). Standard foam padding was used to mitigate head motion prior to that point (prffloc, nsd01).

Stimuli were presented using a Cambridge Research Systems BOLDscreen 32 LCD monitor positioned at the head of the 7T scanner bed, placed flush against the scanner bore. We chose to use an LCD monitor because it delivers a sharp, high-quality image, in contrast to typical scanner setups involving projectors and backprojection screens. The monitor operated at a resolution of 1920 pixels × 1080 pixels at 120 Hz. The size of the full monitor image was 69.84 cm (width) × 39.29 cm (height). Subjects viewed the monitor via a mirror mounted on the RF coil. The viewing distance was 5 cm from the subjects' eyes to the mirror + 171.5 cm from the mirror to the monitor image = 176.5 cm total. Measurements of the display spectral power density were obtained using a PR-655 spectroradiometer (Photo Research). The BOLDscreen is designed by the manufacturer to behave as a linear display device, and our measurements confirmed this to be the case.

We determined the maximum square extent visible in both eyes given the constraints of the RF coil to be 8.4° × 8.4° (714 pixels × 714 pixels). Thus, stimuli from the various experiments (e.g., pRF, fLoc, NSD) were adjusted to fill 8.4° of visual angle (details provided below). At the beginning of each scan session, we made an effort to position the monitor in the same location relative to the scanner and to position the subject's head and RF coil in the same location relative to the scanner. We also used a calibration square (8.4° in size) to determine any incidental horizontal or vertical offsets needed in that session in order for the participant to see the entire square in each eye, unobstructed. Given these efforts, we believe that consistent and high-quality visual stimulation was achieved across scan sessions. Nonetheless, we caution that due to limitations in positioning and/or potential drift over the course of a scan session, some slight occlusion of the corners of the 8.4° × 8.4° square extent may have occurred some of the time.

A Mac Pro computer controlled stimulus presentation using code based on Psychophysics Toolbox 3.0.14[62,63]. Behavioral responses were recorded using a button box (Current Designs, Philadelphia, PA). In some scan sessions (nsd21–nsd30, the same sessions in which the primary set of resting-state data were acquired), physiological data were collected using a pulse oximeter and a respiratory belt (stock Siemens equipment). Care was taken to secure the oximeter with tape to the left index finger of the subject and to secure the respiratory belt snugly to the subject's torso. Physiological data were carefully synchronized with the fMRI data and cropped, but are not further analyzed in this paper.

In several scan sessions (see **Extended Data Figure 2** for details), eyetracking was performed using an EyeLink 1000 system (SR Research, Mississauga, Ontario, Canada) combined with a custom infrared illuminator mounted on the RF coil. Eyetracking was performed for the left eye, and eyetracking data were obtained at 2000 Hz using the Pupil-CR centroid mode. We caution that the eyetracking data are variable in quality, as achieving sufficient pupil contrast was often difficult given the constraints of the scanner setup. For information complementary to the eyetracking data, we also captured video recordings of the eyetracker computer display (see **Figure 2B**) using a cell phone secured to a mount. These video recordings are useful for checking the accuracy of the eyetracker, and provide information in scan sessions where pupil tracking and data acquisition failed completely. Details of pre-processing and analysis of eyetracking data are provided in **Supplementary Note 3**.

Day-to-day acquisition procedures

Participants were scanned roughly once a week, with attempts to keep a regular weekly scan time. At the beginning of each session (starting at approximately nsd07), participants were asked to rate on a five-point scale how well they slept the night before, their mood, how hungry they were, and their stress level. We also asked whether they had had caffeine in the past three hours. At the end of each scan session, participants were asked to rate how comfortable they were during the session and to provide any general feedback they had about the session. These various measures, as well as any technical issues that arose during the session, were logged onto a spreadsheet (available online).

In the first several scan sessions, we emphasized the importance of fixation and performed simple tests prior to scanning in which we watched the subject’s eyes while they attempted to fixate and while they deliberately broke fixation. This was done to help subjects understand what good fixation feels like. In every scan session, we reminded subjects about the importance of fixation and about the correct mapping between buttons and responses.

During data collection, we monitored aspects of data quality including overall image quality, head motion, quality of physiological data, and behavioral performance. Between functional runs, we checked in with the subject to assess their energy level, enthusiasm, and compliance. If we noticed any substantial drops in response rate, we politely notified the subject and offered short breaks before continuing.

To promote subject engagement and retention, participants were given the opportunity to earn monetary bonuses that gradually increased in size over the course of the NSD study. These bonuses were contingent on achieving certain performance levels on data quality metrics such as head motion and response rate (details available online). Information regarding performance was supplied to participants in the form of a continually updated “leaderboard” figure. We found that this figure greatly helped to motivate participants.

The NSD experiment

Basic design

In the NSD experiment, participants performed a long-term continuous recognition task while viewing a large number of color natural scenes. We chose this recognition task because it engages and challenges the observer and is unbiased with respect to the specific content of the images (unlike other tasks such as animacy judgment). In addition, it infuses the experiment with a rich memory dimension that is likely of interest to memory researchers. A total of 73,000 distinct images were prepared. We intended that the 8 NSD subjects would each view 10,000 distinct images presented 3 times each over the course of 40 scan sessions. We designated a special set of 1,000 images (chosen randomly from the full set of prepared images) as shared images that would be seen by all subjects (referred to as the ‘shared1000’); all other images would be mutually exclusive across subjects. The distribution of the 3 presentations of each image was tightly controlled, and subjects were naïve as to both the number and distribution of the presentations. Note that because some NSD subjects completed only 30 of the 40 prescribed scan sessions, there are ultimately 515 images, out of the shared 1,000 images, that are viewed all 3 times by all 8 subjects (referred to as the ‘shared515’).

Images were presented using a 3-s ON / 1-s OFF trial structure (**Figure 1A**). In informal piloting, we found that this pacing made the recognition task feasible and not overly taxing. In addition, we reasoned that the relatively long stimulus duration would increase neural activity and that the rapidity of the design would allow more trials to be collected and thereby increase overall experimental power. Finally, we speculated that the 3/1 trial structure would yield a pleasant experience for participants, at least compared to slow event-related designs where most experimental time is spent viewing a blank screen.

Image preparation

The NSD stimuli are prepared as a single brick of RGB images with dimensionality 425 pixels × 425 pixels × 3 RGB channels × 73,000 images and unsigned 8-bit integer format.

Images were taken from Microsoft’s Common Objects in Context (COCO) image database[14]. COCO images are photographs harvested from online repositories; each image is supplemented by a rich set of annotations (e.g., boundary polygons around objects, natural language captions, body-pose estimates). Out of the 90 original COCO categories, there are a total of 80 COCO categories that exist in the 73,000 NSD images. We used COCO images in the 2017 train/val split[14], and restricted selection to the subset of images for which pixel-level annotations of “stuff”[64] (e.g., sky, land, wall, road) in addition to “things” (e.g., car, skateboard, hat) were available.

We selected only images whose smaller dimension (height or width) was at least 425 pixels. Where necessary, we squared image dimensions by cropping out pixels along the largest dimension. For example, if the original image was 425 × 585, we cropped away 160 pixels from the larger dimension, resulting in an image that is 425 × 425. The median number of pixels cropped per image was 160. After cropping, images were downsampled, if needed, to 425 × 425.

Cropping an image can change the way the viewer interprets it. We refer to this effect of cropping as “semantic loss”. In order to be able to take full advantage of the rich annotations available for the COCO images, we attempted to minimize semantic loss when cropping images. For landscape-oriented images, we selected between a center, left, or right crop. For portrait-oriented images, we selected between a center, top, or bottom crop (finer grids of cropping options had little effect on results). Selection of crops were carefully performed based on quantitative analysis and visual inspection (details provided in the NSD Data Manual).

In addition to screening to minimize semantic loss, we implemented a screening procedure to remove duplicate images. Some of the COCO images are extremely similar to each other, differing only by a post-processing operation (i.e., grayscaling or sharpening) or by a few frames in a motion-capture sequence. To remove these near-duplicates, we downsampled all images to 40 × 40 and then computed the correlation of grayscale pixel intensities between all image pairs. We manually inspected the image pairs with the 500 highest correlation values. Of these, 38 image pairs were observed to be near-duplicates. We randomly selected another image from the COCO dataset to replace one image in each near-duplicate pair. Finally, we screened captions for all images for indications of violent or salacious content. No images were deemed too offensive to include in the experiment.

The distribution of “thing” categories across the final images selected for NSD was nearly identical to distribution in the full COCO dataset. As a result, the “person” category was over-represented; however, with a few exceptions, all 80 COCO object categories are displayed in at least 100 images to each subject. Note that images tend to depict more than one category, so that a given object category frequently appeared in the same image with other categories. For each subject’s images, at least 90% of the images contain 2 or more of the 80 COCO categories.

Distribution of image presentations

We determined the ordering of the 10,000 images × 3 trials = 30,000 trials in advance and kept the ordering fixed across subjects. The idea is that these 10,000 images are actually treated as slots into which different NSD images are inserted. We designated the first 1,000 slots as corresponding to the special shared1000 images; the remaining 9,000 slots were filled with unique images for each subject. Note that because the trial ordering and repetition structure are identical across subjects, the difficulty of the recognition task is comparable across subjects (up to the fact that some images might be more difficult to remember than others).

We controlled the distribution of image presentations in order to prevent the recognition task from becoming too difficult (and risking loss of subject morale). In the procedure, we conceptualized the task of determining the trial ordering as equivalent to placing image presentations on a circle that would eventually be cut and unraveled. The rationale for this circular design is to minimize the extent to which certain points in the experiment differ from others; of course, since the circle eventually becomes a line, there is some imperfection (see discussion below regarding “burn-in” and “dead” time). To determine presentation times, we created a circular probability distribution by mixing a von Mises distribution and a uniform distribution (**Extended Data Figure 1A**). Using random draws from the resulting distribution (positioning the distribution at a random location on the circle for each image), we determined 3 presentation times for each of the 10,000 images. After completing the placement of all 30,000 trials, we then cut the circle, unraveled it into a linear sequence of image presentations, and divided this sequence into 40 consecutive segments corresponding to the 40 NSD scan sessions (750 trials per session).

To determine presentation times, we created a circular probability distribution by mixing a von Mises distribution and a uniform distribution (**Extended Data Figure 1A**). For each image, we positioned the peak of the von Mises distribution at a random position on the circle (i.e., we randomly sampled the mean parameter from –180 to 180 degrees) and then randomly sampled presentation times for each of the three image repetitions from the mixture distribution. We chose specific parameters for the probability distribution: we used a von Mises distribution with concentration parameter of 3^6 and a mixing ratio of 60% and 40% for the von Mises and uniform distributions, respectively. This choice of parameters yields appealing properties. First, the distribution is relatively narrow (see **Extended Data Figure 1A**) and therefore ensures that there will be many trials involving an image that has been presented in the recent past (thus, making the trials easy), while still allowing the probing of more distant memory events. Second, there is minimal “burn-in” time at the beginning of the experiment: even in the first scan session, there is still a substantial number of trials involving old images (see **Extended Data Figure 1B, blue line**). Third, there is minimal “dead” time at the end of the experiment: even in the last scan session, there is still a substantial number of trials involving new images (see **Extended Data Figure 1B, blue line**).

To provide a sense of the overall experimental design, we computed basic statistics on each NSD scan session. For a typical session, the total number of distinct images shown once, twice, and all three times within that session is 437, 106, and 34, respectively (these numbers reflect the mean across scan sessions, rounding to the nearest integer).

Trial and run design

Each trial lasted 4 s, and consisted of the presentation of an image for 3 s, followed by a 1-s gap. A total of 75 trials were conducted in a run; thus, each run lasted 300 s. The first 3 trials (12 s) and last 4 trials (16 s) were blank trials. The remaining 68 trials were divided into 63 stimulus trials and 5 blank trials. The blank trials were randomly positioned in each run such that the minimum and maximum continuous number of stimulus trials was 9 trials (36 s) and 14 trials (56 s), respectively (see **Figure 1B**). For even-numbered runs, the 63rd stimulus trial was designated to be a blank trial. A total of 12 NSD runs were collected in one NSD session, yielding a total of (63 + 62) × 6 = 750 stimulus trials. Moreover, this design was repeated for all 40 NSD sessions: 750 stimulus trials × 40 sessions = 30,000 stimulus trials. The temporal ordering of stimulus and blank trials was generated once and kept fixed across subjects.

Note that the experimental design involves minimal trial jittering: for the most part, the time interval separating consecutive stimulus images is fixed at 1 s, though occasionally, due to blank trials, the time interval is 5 s. This design was intended to maximize statistical power, and differs from conventional fMRI practice where intervals are often chosen randomly from a fixed range.

Stimulus presentation and task

Since the BOLDscreen is calibrated to behave as a linear display device, we used a squaring luminance response when presenting the NSD experiment in order to simulate the typical viewing of digital images. At time of presentation, the prepared NSD images were resized using linear interpolation from their native resolution of 425 pixels × 425 pixels to 714 pixels × 714 pixels in order to occupy 8.4° × 8.4° on the display. Throughout each run (including blank trials), a small semi-transparent red fixation dot with a black border (0.2° × 0.2°; 50% opacity) was present at the center of the stimuli (**Figure 1A**). Stimuli were shown against a gray background with RGB value (127,127,127).

Subjects were instructed to fixate the central dot and to press button 1 using the index finger of their right hand if the presented image was new, i.e. the image had never been presented before, or button 2 using the middle finger of their right hand if the presented image was old, i.e. the image is identical to one that had been presented before, either in the current scan session or any previous scan session. Subjects were additionally instructed to continue to fixate and wait for the next image in the event of blank trials.

Before the start of the NSD experiment, we showed the subjects a version of the experiment involving cartoon images in order for them to become familiarized with the feel and timing of the task. During the NSD experiment, minimal feedback was provided to the subjects regarding their performance on the recognition task. Participants were blind to the precise details of the NSD experiment (e.g., total number of images, total number of presentations per image). Participants were informed only about their response rate (fraction of trials on which they successfully made a response) and a vague “performance metric” which, unbeknownst to them, quantified their percent correct for easy trials (trials that involved the presentation of an image that had occurred earlier in the same scan session). We revealed the nature of the design in a debriefing session after the completion of the NSD experiment (details below).

Details on experiment timing

Stimulus presentation was locked to the refresh rate of the BOLDscreen monitor. Empirical measurements confirmed that the monitor refresh rate was nearly exactly 120 Hz: duration of runs were highly reliable, ranging from 299.95–299.98 s. To compensate for the slight offset from 300 s, the fMRI data were pre-processed to achieve a sampling rate of 0.999878 s (high-resolution preparation) or 0.999878 s × (4/3) = 1.333171 s (standard-resolution preparation). For brevity, we refer to these numbers as 1.000 s and 1.333 s. Experimental runs were started by a trigger issued by the MR scanner. Due to input polling and monitor refresh, there was slight variability in the delay between trigger detection and the presentation of the first stimulus frame, ranging from 3–22 ms. We did not attempt to compensate for this delay.

Acquisition

Due to constraints on subject availability (including unplanned out-of-town absences in the summer of 2019) and due to constraints on scanner availability (the 7T scanner was decommissioned in November 2019), we did not complete the full NSD experiment for every participant. Fortunately, we were able to collect a sizable amount of data: 40, 40, 32, 30, 40, 32, 40, and 30 NSD sessions for subj01–subj08, respectively. In these collected data, each subject viewed 9,209–10,000 distinct images and participated in 22,500–30,000 trials. Aggregated across subjects, the total number of distinct images shown was 70,566, and the total number of trials was 213,000.

Debriefing

After completion of the final memory test (details below), participants filled out a post-NSD questionnaire. This questionnaire probed topics such as strategies used for performing the NSD task and estimates for the number of images viewed and the number of image repetitions. After filling out this questionnaire, the design of the NSD experiment was then revealed to the participants.

Other experiments

Population receptive field (pRF) experiment

We adapted the experiment used in the Human Connectome Project (HCP) 7T Retinotopy Dataset[30]. Stimuli consisted of slowly moving apertures filled with a dynamic colorful texture (see **Figure 2A**). Apertures and textures were updated at a rate of 15 Hz. Two run types were used. The first, termed 'multibar', involves bars sweeping in multiple directions (same as RETBAR in the HCP 7T Retinotopy Dataset). The second, termed 'wedgering', involves a combination of rotating wedges and expanding and contracting rings. Both run types included blank periods.

For consistency with the NSD experiment, stimuli were resized to fill a circular region with diameter 8.4°. Each run lasted 300 s (exact empirical timings were highly accurate and ranged between 299.95–300.00 s). Throughout stimulus presentation, a small semi-transparent dot (0.2° × 0.2°) was present at the center of the stimuli. The color of the central dot switched randomly to one of three colors (black, white, or red) every 1–5 s. Subjects were instructed to maintain fixation on the dot and to press a button whenever the color of the dot changed. To further aid fixation, a semi-transparent fixation grid was superimposed on the stimuli and was present throughout the experiment[65]. A total of 6 runs (3 multibar, 3 wedgering) were collected in the first 7T fMRI session (prffloc).

Functional localizer (fLoc) experiment

This experiment was developed by the Grill-Spector lab[31] (stimuli and presentation code available at http://vpnl.stanford.edu/fLoc/). The experiment consisted of the presentation of grayscale images depicting different stimulus categories (see **Figure 2A**). There were 10 categories, grouped into 5 stimulus domains: characters (word, number), bodies (body, limb), faces (adult, child), places (corridor, house), and objects (car, instrument). Stimuli were presented on a scrambled background (different backgrounds for different stimuli). Stimuli were presented in 4-s trials. In a trial, 8 images from a given category were sequentially presented (image duration 0.5 s). Each run included 6 presentations of each of the 10 categories as well as blank trials (also of 4-s duration).

For consistency with the NSD experiment, stimuli were resized to fill a square region filling 8.4° × 8.4° of visual extent. Each run lasted 300 s (exact empirical timings were highly accurate and ranged between 300.000–300.002 s). Throughout stimulus presentation, a small red fixation dot was present at the center of the stimuli. Subjects were instructed to maintain fixation on the dot and to press a button whenever they noticed an image in which only the background was present ("oddball" task). A total of 6 runs were collected in the first 7T fMRI session (prffloc).

Resting-state experiment

Stimuli consisted of a white fixation cross (0.5° × 0.5°) on a gray background (see **Figure 2A**). Each resting-state run lasted 300 s. In the second resting-state run held within a given scan session, the fixation cross turned red after 12 s had elapsed and remained red for 4 s before returning to white.

Resting-state data were acquired in several NSD core scan sessions: nsd21–nsd38 for subj01 and subj05, and nsd21–nsd30 for all other subjects. Thus, a total of 100 or 180 minutes of resting-state data were acquired for each subject. In each session, one resting-state run was acquired at the beginning of the session (prior to the NSD runs) and another resting-state run was acquired at the end of the session (after the NSD runs).

In the first resting-state run, subjects were instructed to stay awake and fixate the cross but otherwise rest. In the second resting-state run, subjects were additionally instructed to inhale deeply when the fixation cross turned red. This instructed breath was designed to aid analysis of the physiological data collected concomitantly with the resting-state data. Prior to each resting-state run, subjects were asked to report their current sleepiness level using the Stanford Sleepiness Scale[66] (1–7 where 1 is most active and 7 is most sleepy). After each resting-state run, subjects were asked to report their sleepiness level during the run that had just completed.

After the last scan session involving resting-state data, participants filled out a post-resting-state questionnaire. This questionnaire queried what the participants were doing during the resting-state runs and whether they thought about the images from the NSD experiment.

Synthetic stimuli experiment (nsdsynthetic)

After completion of the NSD experiment, we conducted an additional 7T fMRI scan session in which responses were measured to a variety of carefully controlled synthetic (non-naturalistic) stimuli while the subject performed either a fixation task or a one-back task. These data will be described and released in a forthcoming manuscript.

Visual imagery experiment (nsdimagery)

After completion of the nsdsynthetic experiment, we conducted an additional 7T fMRI scan session in which responses were measured while participants engaged in visual imagery and other cognitive tasks. These data will be described and released in a forthcoming manuscript.

Additional behavioral measures (nsdpostbehavior, nsdmemory, nsdmeadows)

A number of behavioral assessments were conducted after completion of the NSD experiment. Several of these were relatively brief, and included the following (nsdpostbehavior): open-ended questions regarding language ability; the Vividness of Visual Imagery Questionnaire[67]; the Test of Word Reading Efficiency[68], including both Sight Word Efficiency and Phonemic Decoding Efficiency; the Cambridge Memory Test for Faces[69]; ultra-fast measurement of contrast sensitivity[70]; and an assessment of chromatic sensitivity (participants adjusted intensities of red, green, and blue channels on the BOLDscreen display until minimal luminance flicker was perceived).

We also conducted a final memory test in which we collected various memory-related measures regarding the images shown to the subjects during the NSD experiment (nsdmemory). These data will be described and released in a forthcoming manuscript.

Finally, using the web-based Meadows platform (http://meadows-research.com), we conducted an assessment of how the NSD subjects perceive and interpret the NSD images (nsdmeadows). First, we selected a small set of images that maximally span semantic space. This was done by isolating the shared515 images; computing shifted inverse frequency sentence embeddings for the sentence captions provided by the COCO dataset[71]; and using a greedy approach to determine the subset of 100 images that maximize the average distance between each image's embedding and its closest neighbor. We then asked participants to perform a Multiple Arrangements Task[72] in which they arrange using a drag-and-drop interface the 100 images within a white circular arena according to the similarity of their content. Using an adaptive procedure, subsequent arrangements were conducted using subsets of the images in order to maximize information gain. This was done until 45 minutes had elapsed. Using a similar interface on Meadows, participants then provided valence and arousal ratings for the 100 images as well as 3 additional images pulled from the shared515 images. Ratings were performed separately for valence and arousal, and were accomplished by freely arranging using a drag-and-drop interface the images (delivered in small batches) along a one-dimensional axis ranging from low to high. This assessment took about 15 minutes.

Overview of data analysis

We designed custom analysis strategies to maximize the quality of derived measures from the NSD data. A number of methods are based on recent work[32,73] where further details can be found. Data analysis and visualization were performed using custom code in MATLAB and Python as well as tools from various packages such as FreeSurfer, SPM, FSL, ANTs[74], and ITK-SNAP[75]. An archive of code used is provided online (https://github.com/cvnlab/nsddatapaper/), and specific code files are referenced in the text below.

A comprehensive schematic outlining the data analysis performed in this paper is provided in **Extended Data Figure 3**. The analysis of the NSD data can be divided into three components: (i) pre-processing of the anatomical, diffusion, and functional data, (ii) time-series analysis of the fMRI data to estimate trial-wise betas, and (iii) further analyses of the trial-wise betas to answer specific scientific questions. The first two components produce the so-called 'prepared data' that are generally useful to the community, whereas the third component refers to analyses performed for the purposes of this paper (estimation of pRFs from the NSD data, univariate memory analysis, representational similarity analysis, brain-optimized neural network training). Data collection and analysis were not performed blind to the conditions of the experiments. No data were excluded from analyses, with the exception of a few T_1 volumes (2 of 52 volumes = 4%) and certain portions of the eyetracking data that were corrupted by noise (11 of 160 eyetracking runs = 7%).

The pre-processing approach that we designed for the NSD dataset prioritizes accuracy and preservation of information (e.g. avoiding spatial smoothing). We avoid "baking in" unnecessary assumptions (e.g. aggressively removing signal fluctuations without careful assessment of validity) and we avoid assuming the accuracy of automated methods; care is taken to manually inspect each pre-processing step to ensure satisfactory results. While we believe our pre-processing is general and likely suitable for most downstream uses of the data, the raw data are also available for those who wish to explore other pre-processing approaches such as fmriprep[76]. We note several aspects of the NSD dataset that may render the dataset challenging from a pre-processing standpoint: the relatively high spatial resolution of the fMRI data (1.8 mm) places higher demands on spatial accuracy, the ultra-high field strength (7T) used for the fMRI data yields higher levels of EPI spatial distortion compared to lower field strengths, and the emphasis on many repeated scans of individuals heightens the importance of achieving consistent imaging results across scan sessions.

Pre-processing of MRI data

Details of the pre-processing of anatomical, functional, and diffusion data are provided in **Supplementary Notes 4–5**. Functional data were pre-processed using one temporal resampling to correct for slice time differences and one spatial resampling to correct for head motion within and across scan sessions, EPI distortion, and gradient nonlinearities. Two versions of the functional data were prepared: a 1.8-mm standard-resolution preparation (temporal resolution 1.333 s) and an upsampled 1.0-mm high-resolution preparation (temporal resolution 1.000 s). Analyses of the pRF and fLoc localizer experiments were used to define retinotopic and category-selective regions of interest (ROIs), respectively. Other ROIs were also defined, including an ‘nsdgeneral’ ROI indicating occipital regions generally responsive in the NSD experiment and a ‘corticalsulc’ ROI collection indicating major cortical sulci and gyri. Annotations for several of the corticalsulc ROIs are shown in **Figure 3F** and **Figure 4B**. Abbreviations: CGS = cingulate sulcus, PrCS = precentral sulcus, CS = central sulcus, PoCS = postcentral sulcus, SFRS = superior frontal sulcus, IFRS = inferior frontal sulcus, LS = lateral sulcus, Calc = calcarine sulcus, OTS = occipitotemporal sulcus, CoS = collateral sulcus, STS = superior temporal sulcus, IPS = intraparietal sulcus.

Data quality metrics

Several data quality metrics were calculated (*export_runmetrics.m*) and summarized in **Figures 1D** and **2D**. Temporal signal-to-noise ratio (tSNR) was computed from raw fMRI volumes (no pre-processing) by first detrending the time-series data from each voxel (quadratic polynomial fit) and then dividing the mean signal intensity by the standard deviation of signal intensity values (*autoqc_fmri.m*). We calculated the median tSNR across voxels within a simple brain mask (mean volume thresholded at 1/10th of the 99th percentile of values) and then computed the median across runs. Head motion was quantified by calculating framewise displacement (FD)[77] based on motion parameter estimates (1.8-mm preparation). We calculated the mean FD across volumes in a run and then computed the median across runs. BOLD response was quantified by calculating the percentage of variance explained by a simple ON-OFF GLM model (1.8-mm preparation). We calculated the median variance explained across voxels within the nsdgeneral ROI and then computed the median across runs. (Further details on the ON-OFF GLM can be found in the ‘GLMsingle algorithm’ section.) Response rate was quantified by calculating the percentage of trials for which the subject pressed a button and then computing the mean across runs. Behavioral performance was quantified by dividing trials into easy trials (trials for which the presented image had been previously presented in the same scan session), hard trials (trials for which the presented image had been previously presented but in a previous scan session), and novel trials (trials for which the presented image had never been previously presented) and then calculating, for each trial type, the percentage of trials on which the subject indicated an ‘old’ response.

To identify EPI signal dropout regions (*export_signaldropout.m*), we divided the T_2 volume (resampled to match the EPI data) by the mean EPI volume (1-mm preparation). The resulting volume is useful as it indicates which voxels have high signal intensity in the T_2 but are corrupted by signal dropout in the EPI. We mapped the volume to the cortical surface (cubic interpolation; mean across depth), transformed the result to fsaverage, and then used a data-driven threshold to mark atypically high values. Vertices marked in at least four of the eight subjects are indicated in **Figure 3F**. To visualize surface imperfections, we took the voxels that were marked in the 0.8-mm anatomical space (during the manual inspection of FreeSurfer surface imperfections), smoothed this binary volume with a 3D Gaussian with full-width-half-max of 2 mm, mapped the result to the cortical surface (cubic interpolation; max across depth), and then transformed the result to fsaverage. Vertices exceeding 0.01 in at least one of the eight subjects are indicated in **Figure 3F***.*

Rankings from the 7T fMRI screening session

Six quality measures (pRF BOLD, fLoc BOLD, pRF behavior, fLoc behavior, raw motion, detrended motion) were computed for each of the 14 subjects who participated in the screening session. BOLD quality was quantified as the percentage of voxels for which variance explained by modeling the fMRI time-series data (either pRF model fitting or GLM model fitting) exceeded 20%. Behavior quality was quantified as described above. Motion was quantified by calculating the median voxel displacement relative to the reference volume used for motion correction, computing the median of this quantity across volumes, and then computing the mean across runs. This motion quantification was performed using raw motion parameter estimates (thereby providing a measure of global head displacement over the course of the session) as well as using motion parameter estimates that are linearly detrended within each run (thereby providing a measure of within-run head instability). Each of the six measures was linearly scaled to span the range 1–5 where 1 corresponds to the worst performance and 5 corresponds to the best performance observed across subjects. Finally, the normalized measures were averaged to produce an overall ranking for each subject, as depicted in **Figure 2C**.

Analysis of behavioral data from the NSD experiment

The behavioral data from the NSD experiment were lightly reformatted for the convenience of subsequent analyses (*analyzebehavior_nsd.m*). We first checked whether the subject had accidentally positioned their fingers on incorrect buttons on the button box, and compensated for this if necessary. (In a few instances, we deliberately instructed subjects to use alternative buttons due to hardware malfunction of the button box.) We then recorded, for each stimulus trial, several quantities including time of image presentation, whether the image presented was new or old, whether the response was correct or incorrect, and the reaction time. Button responses were extracted from a time window extending 250–4250 ms after image onset. In the case of multiple buttons pressed during a trial, we scored the final button pressed, excluding any redundant presses of that button (subjects sometimes repeated button presses for good measure).

GLM analysis of the NSD experiment

Overview of approach

We performed a GLM analysis of the pre-processed time-series data from the NSD experiment. To maximize flexibility for subsequent analyses, the GLM approach was designed to provide estimates of BOLD response amplitudes ('betas') for single trials. Due to low signal-to-noise ratio, single-trial estimation in fMRI is challenging. We therefore developed several analysis components in order to optimize the quality of single-trial betas. These components are packaged into a tool called *GLMsingle*, and is the subject of a forthcoming manuscript where additional details and discussion can be found.

The first analysis component of GLMsingle is the use of a library of hemodynamic response functions (HRFs) whereby the best-fitting HRF from the library is chosen for each voxel. This simple approach for compensating for differences in hemodynamic timecourses across voxels[78] has several appealing features: it is efficient and can be executed with little computational cost (and hence can accommodate the massive scale of NSD); and it invariably provides well-regularized HRF estimates. The second analysis component is an adaptation of GLMdenoise to a single-trial GLM framework. GLMdenoise[35] is a technique in which data-derived nuisance regressors are identified and used to remove noise from—and therefore improve the accuracy of—beta estimates. The third component is an application of ridge regression[79] as a method for dampening the noise inflation caused by correlated single-trial GLM predictors. To determine the optimal level of regularization for each voxel, we make use of a recently developed efficient re-parameterization of ridge regression called 'fractional ridge regression'[36].

Derivation of the library of HRFs

To generate a library of HRFs that accurately capture empirically occurring timecourse variation, we performed an initial analysis of data from the first NSD core session (nsd01). This library was fixed and used for the analysis of all subsequent NSD sessions. The first step was to create a comprehensive summary of observed timecourses (*hrf_derivecanonicalpcs.m*). The time-series data from each subject's nsd01 session was fit using a finite impulse response model (0–30 s) where all of the stimulus trials are treated as instances of a single experimental condition (this simplification is necessary to make estimation feasible). We identified voxels for which model variance explained (R^2) was greater than 10%, and from these voxels randomly drew 20,000 voxels (with replacement). Pooling across subjects, timecourse estimates from the resulting 160,000 voxels were subjected to singular value decomposition to determine the top 3 principal components (shown in **Figure 3B, inset**). To fine-tune timecourse estimates, we re-fit the time-series data from the nsd01 session using these 3 principal components as the basis (as opposed to the finite impulse response basis). Finally, adopting the visualization approach of the Temporal Decomposition Method[73], we projected voxel timecourse estimates onto the unit sphere (using the same voxel selection criterion of R^2 > 10%), and constructed a 2D histogram for each subject (shown in **Figure 3A**).

The second step was to define a set of timecourses that span the observed timecourse variation (*hrf_constructmanifold.m*). To do this, we converted the 2D histograms to units of relative frequency and then averaged the histograms across subjects. Inspecting the group-average histogram (shown in **Figure 3B**), we manually clicked a sequence of points on the unit sphere that follow the data density as closely as possible. We then parameterized the path traced by these points (a simple 1D manifold) by positioning regularly spaced points where successive points are separated by six angular degrees (**Figure 3B, cyan dots**). The timecourses corresponding to the resulting set of 20 points were cubic interpolated to a sampling rate of 0.1 s and normalized to peak at 1 (**Figure 3C**). Finally, we fit each timecourse using a double-gamma function as implemented in SPM's *spm_hrf.m* (*hrf_fitspmhrftomanifold.m*). This yielded a library of 20 canonical HRFs that may be useful for application to other experimental datasets (*getcanonicalhrflibrary.m*). We note that variation in timecourse shape is likely due to the influence of macrovasculature on BOLD temporal dynamics[73].

Cross-validation framework for single-trial GLM

The GLMdenoise and ridge regression analysis components of GLMsingle both require tuning of hyperparameters. To determine the optimal setting of hyperparameters, we use a cross-validation approach in which out-of-sample predictions are made for single-trial beta estimates, as opposed to time-series data. This simplifies and reduces the computational requirements of the cross-validation procedure. Note that because of cross-validation, although GLMsingle produces estimates of responses to single trials, it does require the existence of and information regarding repeated trials, i.e., trials for which the stimulus is the same.

The first step of the cross-validation procedure is to analyze all of the available data using no regularization. In the case of GLMdenoise, this amounts to the inclusion of zero nuisance regressors; in the case of ridge regression, this amounts to the use of a shrinkage fraction of one, indicating ordinary least-squares regression. In both cases, the analysis produces a full set of unregularized single-trial betas (e.g., in one NSD session, there are 750 single-trial betas distributed across 12 runs). The second step of the procedure is to perform a grid search over values of the hyperparameter (e.g., number of nuisance regressors; shrinkage fraction). For each value, we assess how well the resulting beta estimates generalize to left-out runs. For example, in leave-one-run-out cross-validation, one run is held out as the validation run, stimuli that occur in both the training runs and the validation run are identified, and squared errors between the regularized beta estimates from the training runs and the unregularized beta estimates from the validation run are calculated. This procedure is iterated with each run serving as the validation run, and errors are summed across iterations.

GLMsingle algorithm

Having described the essential aspects of the estimation framework above, we now turn to the steps in the GLMsingle algorithm. GLMsingle involves fitting several different GLM variants. Each variant includes polynomial regressors to characterize the baseline signal level: for each run, we include polynomials of degrees 0 through round(L/2) where L is the duration in minutes of the run.

1. *Fit a simple ON-OFF GLM.* In this model, stimulus trials are treated as instances of a single experimental condition, and a canonical HRF is used (*getcanonicalhrf.m*). Thus, there is a single “ON-OFF” predictor that attempts to capture signals driven by the experiment. The utility of this simple model is to provide variance explained (R^2) values that help indicate which voxels carry experiment-driven signals.
2. *Fit a baseline single-trial GLM.* In this model, each stimulus trial is modeled separately using the canonical HRF. This model provides a useful baseline for comparison.
3. *Identify HRF for each voxel.* We fit the data multiple times with a single-trial GLM, each time using a different HRF from the library of HRFs. For each voxel, we identify which HRF provides the best fit to the data (highest variance explained), and inherit the single-trial betas associated with that HRF. Note that the final model for each voxel involves a single chosen HRF from the library (not a weighted sum of HRFs).
4. *Use GLMdenoise to determine nuisance regressors to include in the model.* We define a pool of noise voxels (brain voxels that have low ON-OFF R^2) and then perform principal components analysis on the time-series data associated with these voxels. The top principal components are added one at a time to the GLM until cross-validation performance is maximized on-average across voxels.
5. *Use fractional ridge regression to regularize single-trial betas.* With the nuisance regressors determined, we use fractional ridge regression (fracridge[36]) to estimate the single-trial betas, systematically evaluating different shrinkage fractions. For each voxel, in the context of a GLM that incorporates the specific HRF chosen for that voxel, cross-validation is used to select an optimal shrinkage fraction for that voxel. To mitigate bias on the overall scale of betas, we apply a post-hoc scaling and offset on betas obtained for a given voxel in order to match, in a least-squares sense, the unregularized betas obtained for that voxel.

Application of GLMsingle to the NSD data

We used GLMsingle to analyze the time-series data independently for each NSD scan session (*glm_nsd.m*). Major algorithmic parameters included the following: we evaluated up to 10 nuisance regressors; we evaluated shrinkage fractions from 0.05 to 0.90 in increments of 0.05 and from 0.91 to 1 in increments of 0.01 (representing a finer grain for voxels with the best signal-to-noise ratio); we performed 6-fold cross-validation (consecutive pairs of runs) for Steps 4 and 5; and we used an ON-OFF R^2 threshold of 5% in Step 4.

Three different versions of the single-trial betas were computed and saved. The first beta version (b1, 'betas_assumehrf') is the result of Step 2, and reflects the use of a canonical HRF. The second beta version (b2, 'betas_fithrf') is the result of Step 3, and reflects the result of voxel-wise HRF estimation. The third beta version (b3, 'betas_fithrf_GLMdenoise_RR') is the result of Step 5, and reflects the additional GLMdenoise and ridge regression procedures. Betas were converted to units of percent BOLD signal change by dividing amplitudes by the mean signal intensity observed at each voxel and multiplying by 100. While we provide betas in units of percent signal change, we suggest that users may wish to *z*-score the responses of each voxel within each scan session in order to eliminate potential non-stationarities and to equalize units across voxels.

For user convenience, we created preparations of the single-trial betas in additional spaces other than the native 1.8-mm and 1.0-mm functional spaces. For the 'nativesurface' preparation, we performed cubic interpolation of the 1.0-mm betas onto each of the 3 cortical surface depths and averaged across depths (*analysis_transformfsaverage.m*). The result was then mapped using nearest-neighbor interpolation to fsaverage space to create the 'fsaverage' preparation. For the 'MNI' preparation, we mapped the 1.0-mm betas to MNI space using cubic interpolation (*analysis_transformMNI.m*).

GLM analysis of the resting-state experiment

As an optional resource, we fit the time-series data from the resting-state experiment using methods that parallel those used for the NSD experiment (*glm_nsdresting.m*). For each scan session involving resting-state, we took the two resting-state runs (first and last run acquired) and analyzed the data using the design matrix of the neighboring NSD runs and the same voxel-wise HRFs determined from analyzing the NSD runs in that scan session (this is analogous to beta version b2). Although there is no reason to think that spontaneous resting-state activity conforms to the 4-s trial structure of the NSD experiment, these resting-state betas may be useful as a direct comparison for the NSD betas.

Noise ceiling estimation

To obtain a measure of data quality, noise ceilings were estimated for the NSD betas (*export_noiseceiling.m*). The noise ceiling for a given voxel is defined as the maximum percentage of variance in the voxel's responses that can in theory be explained, given the presence of measurement noise. Our method for estimating the noise ceiling follows the general framework laid out in previous studies[80,81]. Several assumptions are made: (i) the signal contained in the voxel's response is determined solely by the presented image, (ii) the variability of the signal across different images is Gaussian-distributed, (iii) the noise is Gaussian-distributed with zero mean, and (iv) the response to an image is equal to the signal plus noise. Given these assumptions, any observed response is a sample from a sum of Gaussian distributions:

$$RESP \sim \mathcal{N}(\mu_{signal}, \sigma_{signal}) + \mathcal{N}(0, \sigma_{noise})$$

where *RESP* indicates the NSD beta observed on a given trial, μ_{signal} is the mean signal across different images, σ_{signal} is the standard deviation of the signal across different images, and σ_{noise} is the standard deviation of the noise (for illustration of these concepts, see **Extended Data Figure 8C**). Note that the first Gaussian distribution characterizes true signal variability, whereas the second Gaussian characterizes variability due to noise. Also, note that this framework treats response variability unrelated to the stimulus as "noise", but such variability may in fact reflect "signal" from the perspective of functional connectivity[82].

To compute the noise ceiling, we first take the trial-wise NSD betas for each voxel and *z*-score these betas within each scan session. This simple normalization compensates for nonstationarities that may exist across sessions. We then calculate the variance of the betas across the three presentations of each image (using the unbiased estimator that normalizes by *n*–1 where *n* is the sample size), average this variance across images, and then compute the square-root of the result. This produces an estimate of the noise standard deviation:

$$\hat{\sigma}_{noise} = \sqrt{\mathrm{mean}({\beta_{\sigma}}^{2})}$$

where ${\beta_{\sigma}}^{2}$ indicates the variance across the betas obtained for a given image. Next, given that the variance of the *z*-scored betas is 1, we estimate the signal standard deviation as follows:

$$\hat{\sigma}_{signal} = \sqrt{\left|1 - {\hat{\sigma}_{noise}}^{2}\right|_{+}}$$

where $|\,|_{+}$ indicates positive half-wave rectification. Finally, we simplify by calculating a single scalar quantity:

$$ncsnr = \frac{\hat{\sigma}_{signal}}{\hat{\sigma}_{noise}}$$

where *ncsnr* indicates the noise ceiling signal-to-noise ratio.

Given the framework described above, the noise ceiling can be calculated as the amount of variance contributed by the signal expressed as a percentage of the total amount of variance in the data:

$$NC = 100 \times \frac{{\sigma_{signal}}^{2}}{{\sigma_{signal}}^{2} + {\sigma_{noise}}^{2}}$$

where *NC* indicates the noise ceiling. We would like to be able to calculate the noise ceiling based on the single scalar *ncsnr*. Moreover, since a researcher may wish to average across multiple presentations of each image before attempting to explain the NSD betas, we would like a method for flexibly expressing the noise ceiling for different levels of trial-averaging. With some algebra, it can be shown that the noise ceiling can be expressed as follows:

$$NC = 100 \times \frac{ncsnr^{2}}{ncsnr^{2} + \frac{1}{n}}$$

where *n* indicates the number of trials that are averaged together (see the NSD Data Manual for the derivation and additional details). We note that there is a direct relationship between the commonly used metric of split-half reliability and the noise ceiling: if a voxel has two sets of responses that reflect the same image presentations, then the correlation between the two sets of responses multiplied by 100 is equal to the noise ceiling for single-trial responses expressed in percent variance explained.

Using the above methods, we calculated noise ceilings for each of the beta versions and for each of various spatial preparations (1.8-mm, 1-mm, fsaverage, nativesurface). For simplicity, noise ceiling estimates were calculated using betas associated with images with all three presentations available. To assess stability, we also computed split-half noise ceiling estimates. This was achieved by splitting the available images into two mutually exclusive groups and computing noise ceiling estimates independently for each group. The noise ceiling results shown in **Figure 3F–G** and **Supplementary Figure 6** were computed assuming *n* = 3, reflecting the scenario in which trial-wise betas are averaged across three trials for each image. The noise ceiling results shown in **Figure 6A–B** were computed assuming *n* = 1 and are expressed in correlation units (square root of percent variance explained).

A few important notes: Even though NSD consists of only up to three trials for a given image, the estimate of response variability for each voxel (i.e. the noise standard deviation) is averaged across a very large number of images, thus stabilizing the noise ceiling estimate. Also, note that our noise ceiling metric refers to activity levels in individual voxels in individual subjects. It is thus quite different from, for example, noise ceiling metrics computed for group-average representational dissimilarity matrices[83]. The latter are more abstracted away from the data given that they summarize properties observed across a collection of voxels, reflect second-order computations on activity levels and not activity levels themselves, and probe responses at the group level and not at the individual level.

Calculation of equivalent trials

To provide a common basis for comparing different datasets, we define the number of equivalent trials present in a dataset as $N \times ncsnr^2$ where N indicates the number of trials conducted and *ncsnr* is the noise ceiling signal-to-noise ratio (as defined earlier). The assumptions here are that (i) every trial has equal value, irrespective of whether it is used to measure brain responses to an image that has already been shown or a new image (e.g., two trials for one image is equivalent to one trial for two distinct images), and (ii) increases in signal-to-noise ratio are equivalent to the collection of additional trials. For an illustrative example of the second assumption, suppose an experimenter chooses to improve signal-to-noise ratio by averaging the response to a given image across *p* repetitions of that image. This effectively reduces the noise standard deviation by a factor of $\sqrt{p}$ and *ncsnr* will thus increase by a factor of $\sqrt{p}$. Alternatively, the experimenter could choose to not average and instead use the *p* trials as-is. In the former case, the number of equivalent trials is $1 \times (\sqrt{p} \times ncsnr)^2 = p \times ncsnr^2$, whereas in the latter case, the number of equivalent trials is $p \times ncsnr^2$. Thus, the two cases correspond to the same number of equivalent trials.

We conducted an auxiliary analysis that directly compares NSD against the BOLD5000 dataset[22]. The goal of this analysis was to calculate a summary *ncsnr* value for each dataset, so that the number of equivalent trials can be calculated. For fair comparison, both NSD and BOLD5000 were analyzed using the exact same GLM methods described in this paper (beta version b3). We then defined a common brain region on which data quality can be compared. This was done by transforming the nsdgeneral ROI to MNI space and then mapping the resulting MNI mask to each subject in the two datasets. Finally, we computed the median *ncsnr* observed across voxels in the mask in each subject.

The median *ncsnr*, averaged across subjects, was 0.260 for NSD (averaged across the first four NSD subjects), and 0.187 for BOLD5000 (averaged across the four subjects in BOLD5000). This indicates that, despite the longer time duration allocated per trial in BOLD5000 (10 s) compared to NSD (4 s), the quality of a single-trial beta in NSD is higher than that in BOLD5000. Specifically, one NSD trial is approximately equivalent to $(0.260)^2/(0.187)^2 = 1.93$ BOLD5000 trials. This increase in quality is likely due, in part, to the screening of subjects and the ultra-high magnetic field strength (7T) used in NSD. Note that the *ncsnr* metric quantifies the SNR per trial and is expected to be unbiased with respect to the number of repeated trials used to calculate it. Thus, although the exact number of trials per image is different in the NSD and BOLD5000 datasets, the *ncsnr* values can still be directly compared.

Univariate analysis of memory recognition

For this analysis (results shown in **Figure 4B**), we used version 3 of the NSD betas (b3) in the fsaverage preparation. Betas for each surface vertex were kept in percent signal change units. Using the behavioral responses, we identified trials involving hits (subjects responded ‘old’ to a previously presented image) and trials involving correct rejections (subjects responded ‘new’ to a novel image). Then, for each subject, we calculated two-sample *t*-values at each surface vertex. This was done both for trials pooled within individual NSD scan sessions as well as for trials pooled across all sessions.

Representational similarity analysis

For this analysis (results shown in **Figure 5**), we used version 3 of the NSD betas (b3) in the fsaverage preparation. Betas for each surface vertex were *z*-scored within each scan session, concatenated across sessions, and averaged across repeated trials for each distinct image. To support the representational similarity analysis[84], we defined a set of ROIs (V1, V2, V3, pVTC, aVTC) on the fsaverage surface. This was done by mapping the manually-defined V1, V2, and V3 from each subject to fsaverage, averaging across subjects, and using the result to guide the definition of group-level ROIs. We also defined a posterior and anterior division of ventral temporal cortex (pVTC and aVTC, respectively) based on anatomical criteria. For each subject, we extracted betas for vertices within each ROI (concatenating across hemispheres). We then computed Pearson's correlation between beta patterns across all possible pairs of images. This yielded representational dissimilarity matrices (RDMs) with rows and columns indexing distinct images (e.g., the RDMs for subject 1 have dimensionality 10,000 × 10,000 with correlations corresponding to 49,995,000 possible pairs).

To help visualize and interpret these large dissimilarity matrices, we performed *t*-distributed stochastic neighbor embedding[41,85] (t-SNE) using a perplexity level of 100 (**Figure 5B–C**). This projects the high-dimensional representations onto a two-dimensional plane such that the distance of a given pair on the plane reflects that pair's distance in the high-dimensional representation as accurately as possible. To verify the strong categorical structure visible in pVTC and aVTC (see **Figure 5B**), we quantified the similarity of the brain RDMs to a model RDM constructed from the category labels in the COCO dataset. Specifically, we constructed an RDM from a binary matrix indicating the presence or absence of each of the 80 COCO categories (cosine distance metric), and correlated this model RDM with each brain RDM. This process was performed for mutually exclusive groups of 100 images drawn from all images presented 3 times to a given subject (the number of groups was 100, 100, 62, 54, 100, 62, 100, and 54 for the eight subjects, respectively). We calculated the mean and standard error across results obtained for different groups of images (**Figure 5D**). Finally, we investigated similarity of brain representations across ROIs and subjects. This was done by isolating the shared515 images, constructing brain RDMs for these images, and correlating brain RDMs across ROIs and subjects. The resulting second-order RDM is shown in **Figure 5E**, with further quantification of this matrix shown in **Figure 5F**.

Data availability

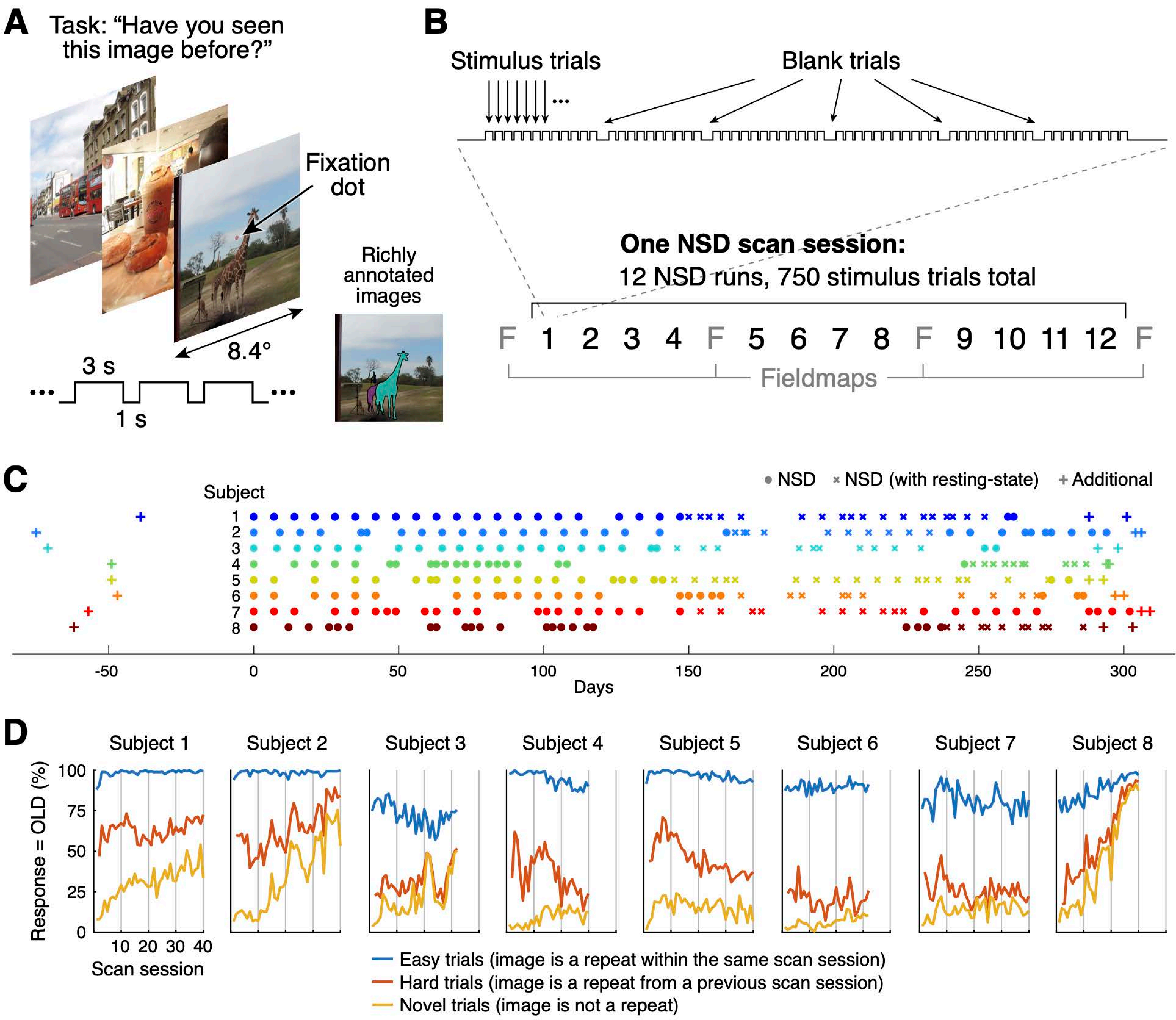
The NSD dataset is freely available at http://naturalscenesdataset.org. The data are hosted in the cloud, allowing researchers to exploit high-performance cloud computing to efficiently analyze the dataset. We provide both raw data in BIDS format[86] and prepared data files, along with extensive technical documentation in the NSD Data Manual. To ensure strict validation for an upcoming Algonauts prediction challenge[87], the initial public release will withhold the last three NSD scan sessions from each participant (about 8.4% of the NSD data). Images used for NSD were taken from the Common Objects in Context database[14] (https://cocodataset.org).

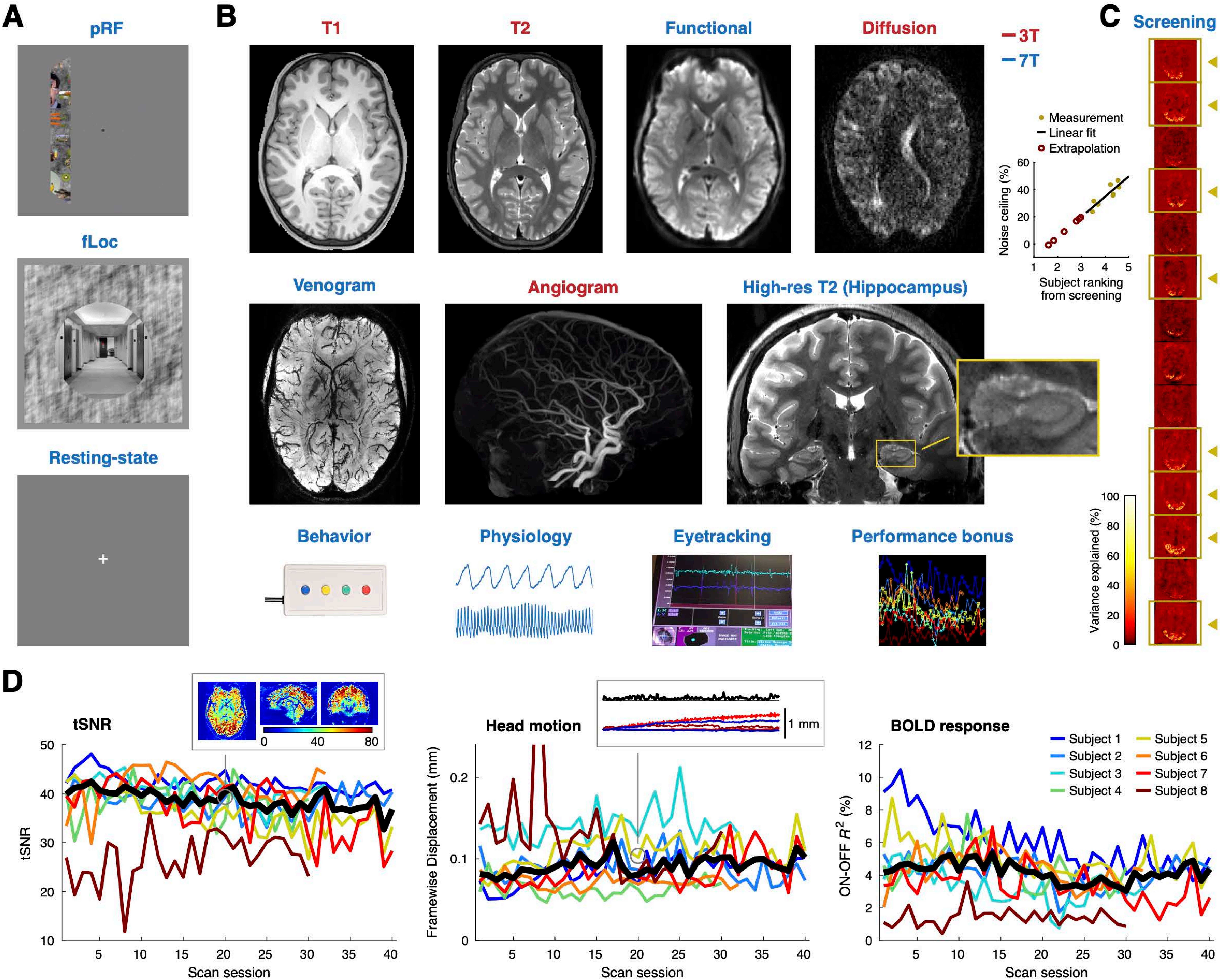
Code availability

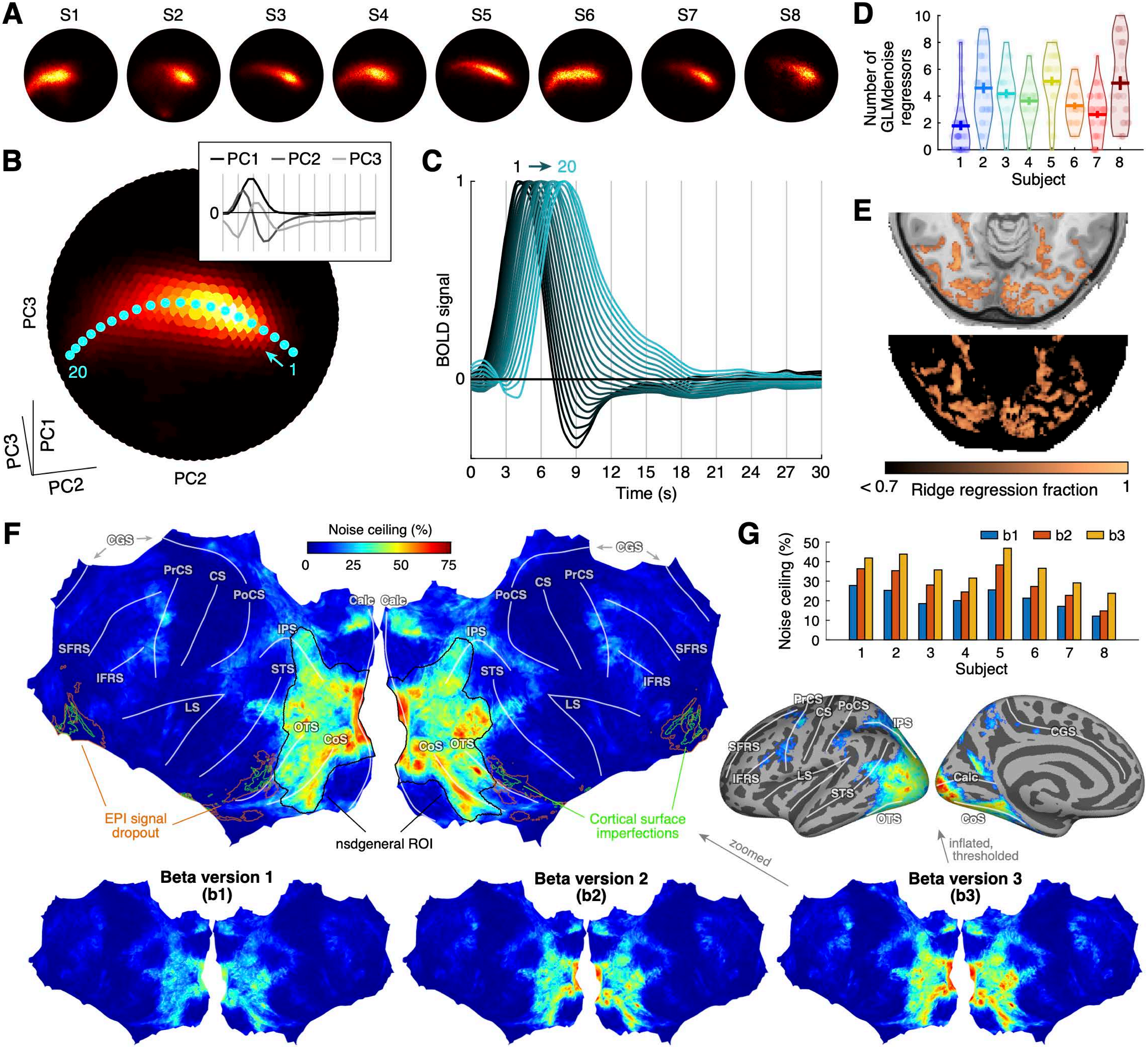
We provide an archive of code used in this paper (https://github.com/cvnlab/nsddatapaper/), as well as utility functions for working with the prepared NSD data (https://github.com/cvnlab/nsdcode/). Custom algorithms developed for this paper include GLMsingle (https://github.com/cvnlab/GLMsingle/) and fracridge (https://github.com/nrdg/fracridge/). Example scripts demonstrating scientific analyses of the NSD data are available (https://github.com/cvnlab/nsdexamples/); these scripts may be useful for teaching purposes.

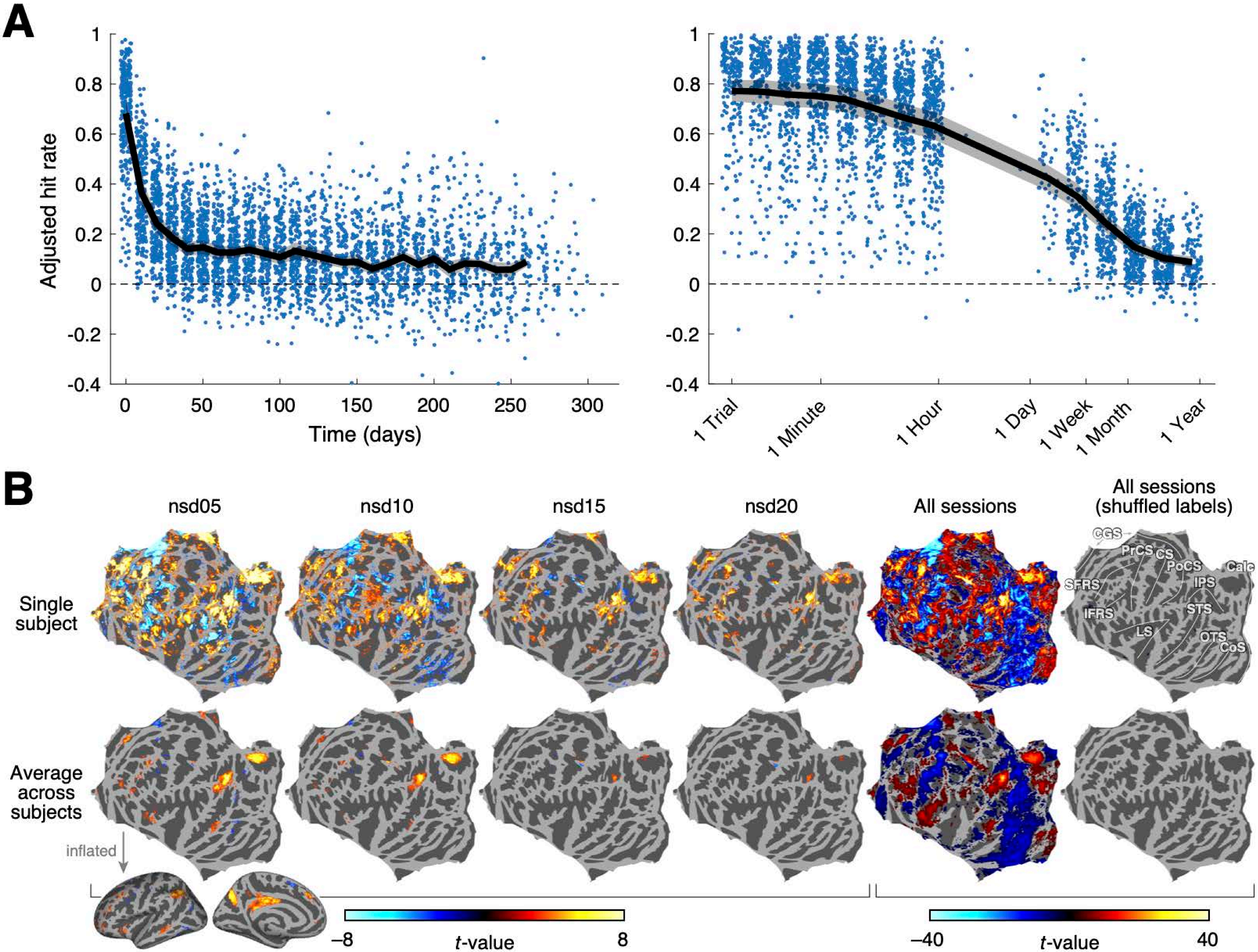
Methods References

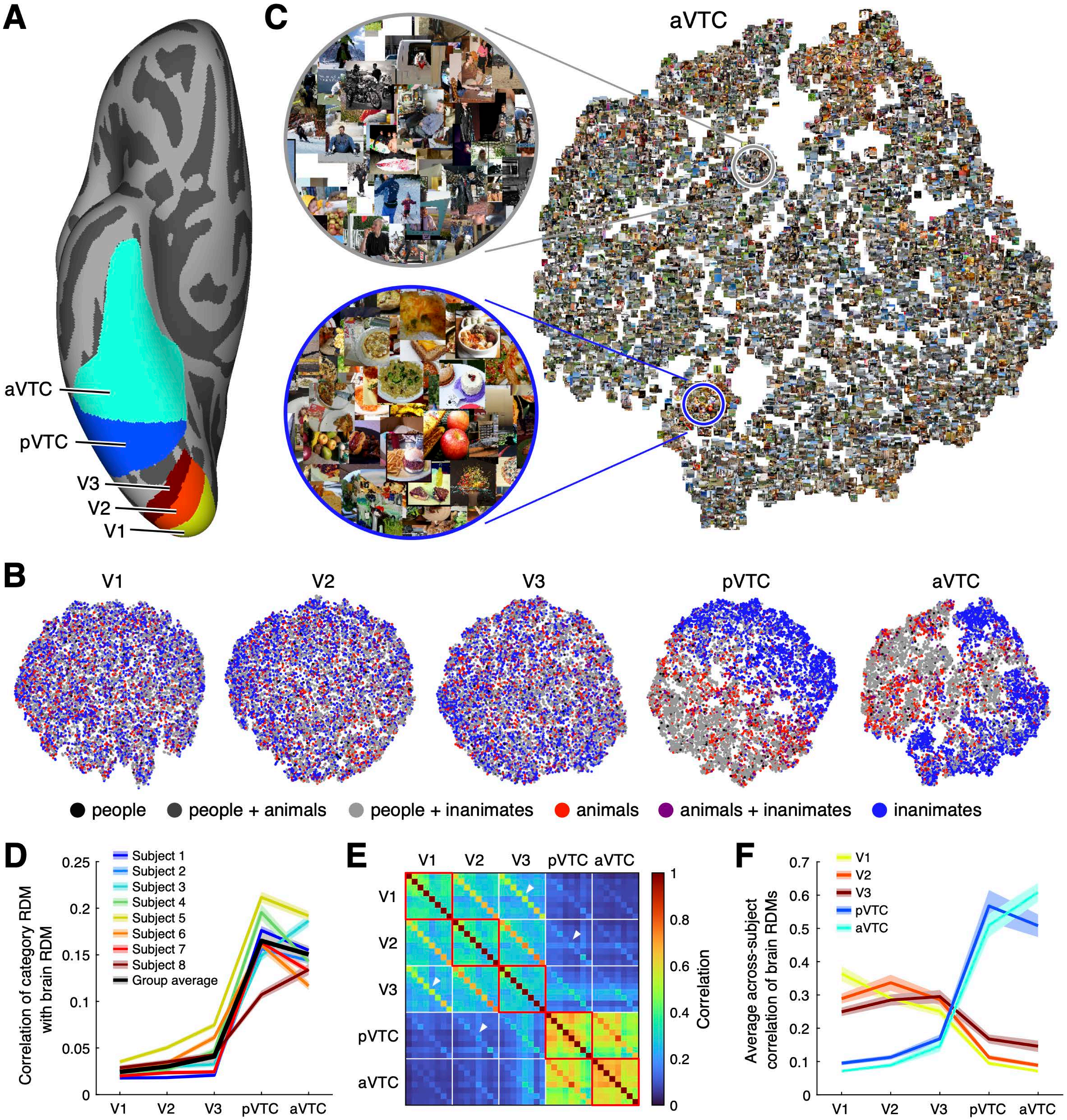
59. Polimeni, J. R., Renvall, V., Zaretskaya, N. & Fischl, B. Analysis strategies for high-resolution UHF-fMRI data. NeuroImage 168, 296–320 (2018).
60. Harms, M. P. et al. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. NeuroImage 183, 972–984 (2018).
61. Power, J. D. et al. Customized head molds reduce motion during resting state fMRI scans. NeuroImage 189, 141–149 (2019).
62. Brainard, D. H. The Psychophysics Toolbox. Spat Vis 10, 433–436 (1997).
63. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10, 437–442 (1997).
64. Caesar, H., Uijlings, J. & Ferrari, V. COCO-Stuff: Thing and Stuff Classes in Context. in 1209–1218 (2018).
65. Schira, M. M., Tyler, C. W., Breakspear, M. & Spehar, B. The foveal confluence in human visual cortex. J. Neurosci. 29, 9050–9058 (2009).
66. Shahid, A., Wilkinson, K., Marcu, S. & Shapiro, C. M. Stanford Sleepiness Scale (SSS). in STOP, THAT and One Hundred Other Sleep Scales (eds. Shahid, A., Wilkinson, K., Marcu, S. & Shapiro, C. M.) 369–370 (Springer, 2012). doi:10.1007/978-1-4419-9893-4_91.
67. Marks, D. F. Visual Imagery Differences in the Recall of Pictures. British Journal of Psychology 64, 17–24 (1973).
68. Torgesen, J. K., Wagner, R. & Rashotte, C. Test of word reading efficiency:(TOWRE-2). (Pearson Clinical Assessment, 2012).
69. Duchaine, B. & Nakayama, K. The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. Neuropsychologia 44, 576–585 (2006).
70. Tardif, J., Watson, M., Giaschi, D. & Gosselin, F. Measuring the Contrast Sensitivity Function in just three clicks. Journal of Vision 16, 966–966 (2016).
71. Arora, S., Liang, Y. & Ma, T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. in (2017).
72. Kriegeskorte, N. & Mur, M. Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. Front Psychol 3, 245 (2012).
73. Kay, K., Jamison, K. W., Zhang, R.-Y. & Uğurbil, K. A temporal decomposition method for identifying venous effects in task-based fMRI. Nat Methods 17, 1033–1039 (2020).
74. Avants, B. B. et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044 (2011).
75. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage 31, 1116–1128 (2006).
76. Esteban, O. et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods 16, 111–116 (2019).
77. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59, 2142–2154 (2012).
78. Handwerker, D. A., Gonzalez-Castillo, J., D'Esposito, M. & Bandettini, P. A. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. NeuroImage 62, 1017–1023 (2012).
79. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. null 12, 55–67 (1970).
80. Kay, K. N., Winawer, J., Mezer, A. & Wandell, B. Compressive spatial summation in human visual cortex. Journal of neurophysiology 110, 481–494 (2013).
81. Lage-Castellanos, A., Valente, G., Formisano, E. & De Martino, F. Methods for computing the maximum performance of computational models of fMRI responses. PLoS Comput Biol 15, e1006397 (2019).
82. Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34, 537–541 (1995).
83. Nili, H. et al. A toolbox for representational similarity analysis. PLoS computational biology 10, e1003553 (2014).
84. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in systems neuroscience 2, 4 (2008).
85. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).
86. Gorgolewski, K. J. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 3, 1–9 (2016).
87. Cichy, R. M., Roig, G. & Oliva, A. The Algonauts Project. Nature Machine Intelligence 1, 613–613 (2019).


A
Task: "Have you seen this image before?"
Fixation dot
8.4°
3 s
1 s
Richly annotated images
B
Stimulus trials
Blank trials
One NSD scan session:
12 NSD runs, 750 stimulus trials total
F 1 2 3 4 F 5 6 7 8 F 9 10 11 12 F
Fieldmaps
C
Subject
● NSD
× NSD (with resting-state)
+ Additional
1
2
3
4
5
6
7
8
-50
0
50
100
150
200
250
300
Days
D
Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Subject 7
Subject 8
Response = OLD (%)
100
75
50
25
0
10 20 30 40
Scan session
Easy trials (image is a repeat within the same scan session)
Hard trials (image is a repeat from a previous scan session)
Novel trials (image is not a repeat)

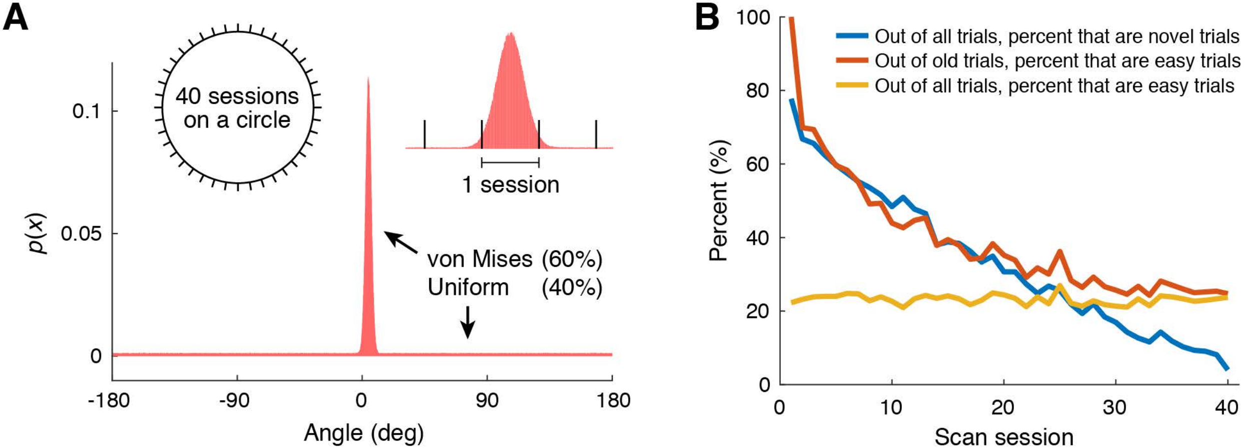

A
pRF
fLoc
Resting-state
B
T1
T2
Functional
Diffusion
3T
7T
Venogram
Angiogram
High-res T2 (Hippocampus)
Behavior
Physiology
Eyetracking
Performance bonus
C
Screening
Measurement
Linear fit
Extrapolation
Noise ceiling (%)
60
40
20
0
1
2
3
4
5
Subject ranking from screening
Variance explained (%)
100
80
60
40
20
0
D
tSNR
0
40
80
50
40
30
20
10
tSNR
5
10
15
20
25
30
35
40
Scan session
Head motion
1 mm
Framewise Displacement (mm)
0.2
0.1
0
5
10
15
20
25
30
35
40
Scan session
BOLD response
Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Subject 7
Subject 8
ON-OFF R^2 (%)
12
10
8
6
4
2
0
5
10
15
20
25
30
35
40
Scan session

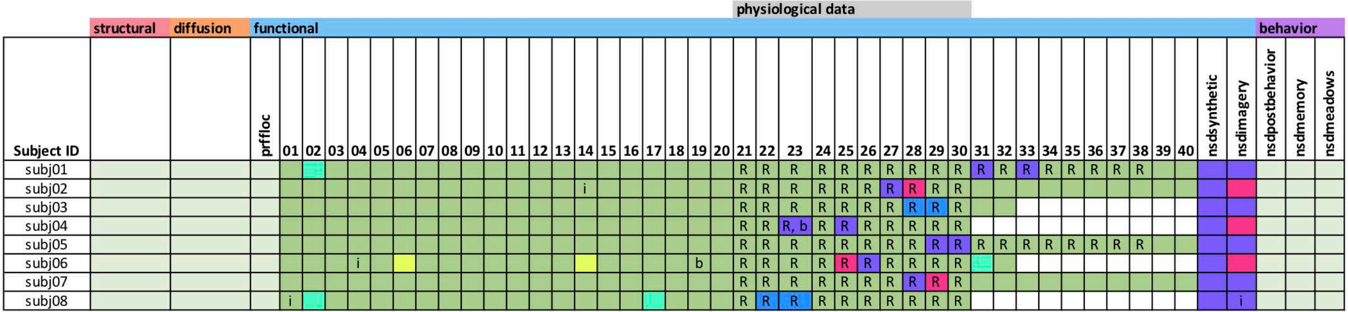

A
S1
S2
S3
S4
S5
S6
S7
S8
B
PC1
PC2
PC3
0
PC3
20
1
PC3
PC1
PC2
PC2
C
1 → 20
1
0
BOLD signal
0
3
6
9
12
15
18
21
24
27
30
Time (s)
D
10
8
6
4
2
0
Number of GLMdenoise regressors
1
2
3
4
5
6
7
8
Subject
E
< 0.7
Ridge regression fraction
1
F
Noise ceiling (%)
0
25
50
75
CGS
PrCS
CS
PoCS
SFRS
IFRS
LS
STS
IPS
Calc
OTS
CoS
EPI signal dropout
nsdgeneral ROI
Cortical surface imperfections
zoomed
inflated, thresholded
G
b1
b2
b3
50
40
30
20
10
0
Noise ceiling (%)
1
2
3
4
5
6
7
8
Subject
Beta version 1 (b1)
Beta version 2 (b2)
Beta version 3 (b3)


A
Adjusted hit rate
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
0
50
100
150
200
250
300
Time (days)
1 Trial
1 Minute
1 Hour
1 Day
1 Week
1 Month
1 Year
B
nsd05
nsd10
nsd15
nsd20
All sessions
All sessions (shuffled labels)
Single subject
Average across subjects
CGS
PrCS
CS
PoCS
Calc
SFRS
IPS
IFRS
STS
LS
OTS
CoS
inflated
–8
t-value
8
–40
t-value
40


A
aVTC
pVTC
V3
V2
V1
C
aVTC
B
V1
V2
V3
pVTC
aVTC
● people
● people + animals
● people + inanimates
● animals
● animals + inanimates
● inanimates
D
Correlation of category RDM with brain RDM
0.25
0.2
0.15
0.1
0.05
0
V1
V2
V3
pVTC
aVTC
Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Subject 7
Subject 8
Group average
E
V1
V2
V3
pVTC
aVTC
V1
V2
V3
pVTC
aVTC
Correlation
1
0.8
0.6
0.4
0.2
0
F
Average across-subject correlation of brain RDMs
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
V1
V2
V3
pVTC
aVTC
V1
V2
V3
pVTC
aVTC

A
S1
S2
S5
S6
Noise ceiling
Median validation accuracy
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
2
3
6
9
18
27
Training samples (x1000)
Gabor
AlexNet
GNet (single-subject)
GNet (multi-subject)
B
S1-V1
S1-V2
S1-V3
S1-hV4
Validation accuracy
0.8
0.6
0.4
0.2
0.0
Noise ceiling
Individual voxels
Noise ceiling
Median across voxels
C
S1
S2
S5
S6
V1
V2
V3
hV4
Validation accuracy
0.50
0.25
0


A
40 sessions
on a circle
1 session
von Mises (60%)
Uniform (40%)
p(x)
0.1
0.05
0
-180
-90
0
90
180
Angle (deg)
B
Out of all trials, percent that are novel trials
Out of old trials, percent that are easy trials
Out of all trials, percent that are easy trials
Percent (%)
100
80
60
40
20
0
0
10
20
30
40
Scan session

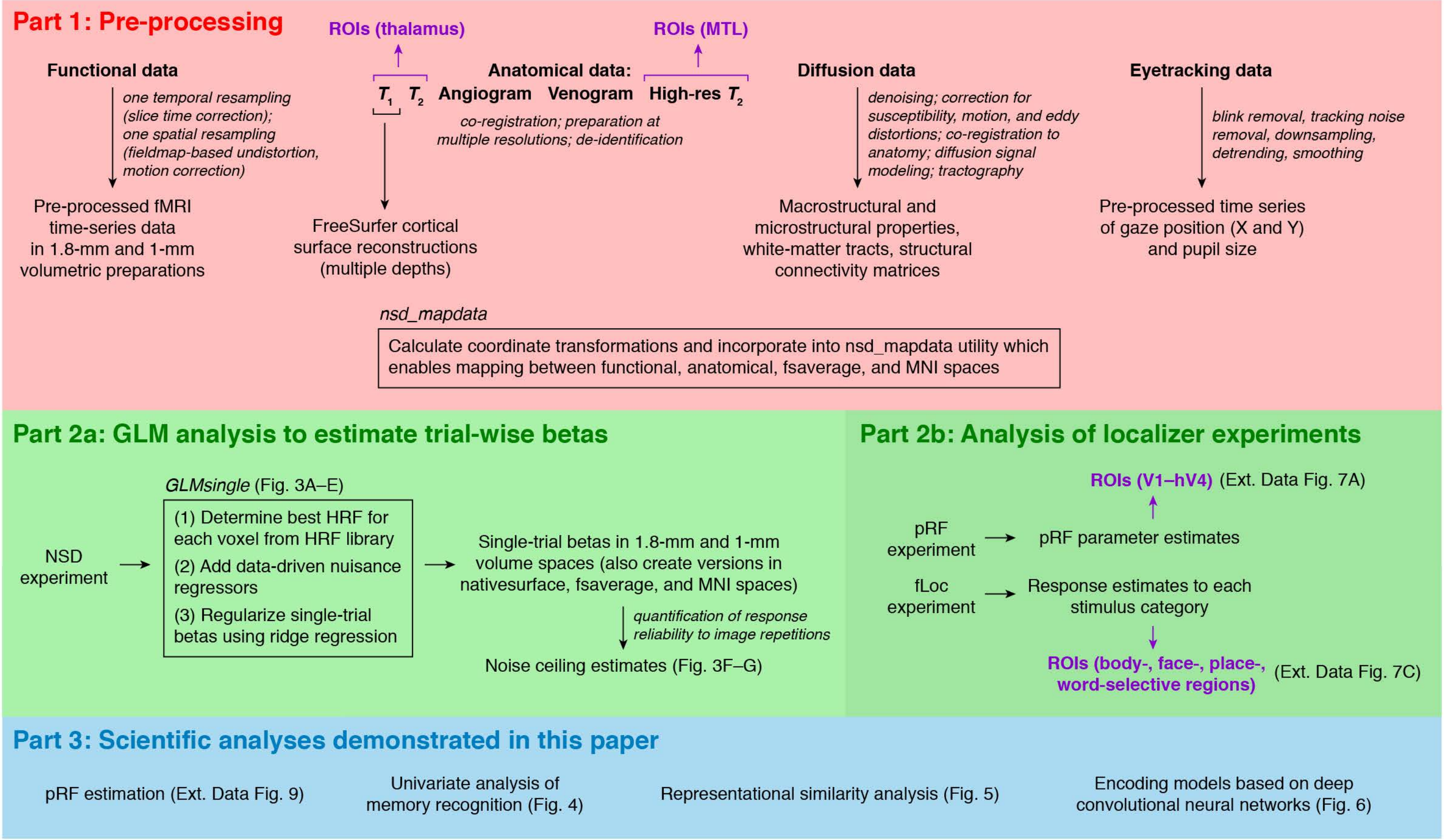


																							physiological data																									
	structural	diffusion	functional																																										behavior		
Subject ID			prffloc	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	nsdsynthetic	nsdimagery	nsdpostbehavior	nsdmemory	nsdmeadows
subj01	░	░	░	▓	◆	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	🟪 R	▓ R	🟪 R	▓ R	▓ R	▓ R	▓ R	▓ R	▓	▓	🟪	🟪	░	░	░
subj02	░	░	░	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓ i	▓	▓	▓	▓	▓	▓	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	🟪 R	🟥 R	▓ R	▓ R	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	🟪	🟥	░	░	░
subj03	░	░	░	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	🟦 R	🟦 R	▓ R	▓	▓									🟪	🟪	░	░	░
subj04	░	░	░	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓ R	▓ R	🟪 R, b	▓ R	🟪 R	▓ R	▓ R	▓ R	▓ R	▓ R											🟪	🟥	░	░	░
subj05	░	░	░	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	🟪 R	🟪 R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓	▓	🟪	🟪	░	░	░
subj06	░	░	░	▓	▓	▓	▓ i	▓	◇	▓	▓	▓	▓	▓	▓	▓	◇	▓	▓	▓	▓	▓ b	▓	▓ R	▓ R	▓ R	▓ R	🟥 R	🟪 R	▓ R	▓ R	▓ R	▓ R	◆	▓									🟪	🟥	░	░	░
subj07	░	░	░	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	🟪 R	🟥 R	▓ R	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	🟪	🟪	░	░	░
subj08	░	░	░	▓ i	◆	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	▓	◆	▓	▓	▓	▓ R	🟦 R	🟦 R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R	▓ R											🟪	🟪 i	░	░	░

░ data acquired
▓ data acquired (NSD core)
◆ split-session (different days)
◇ split-session (same day)

R = resting-state acquired
i = some missing voxels (invalid voxels)
b = behavioral responses missing in one of the runs

🟥 eyetracking data
🟦 eyetracking video feed
🟪 both



Part 1: Pre-processing

ROIs (thalamus)

ROIs (MTL)

Functional data

one temporal resampling (slice time correction); one spatial resampling (fieldmap-based undistortion, motion correction)

Pre-processed fMRI time-series data in 1.8-mm and 1-mm volumetric preparations

Anatomical data:

T_1 T_2 Angiogram Venogram High-res T_2

co-registration; preparation at multiple resolutions; de-identification

FreeSurfer cortical surface reconstructions (multiple depths)

Diffusion data

denoising; correction for susceptibility, motion, and eddy distortions; co-registration to anatomy; diffusion signal modeling; tractography

Macrostructural and microstructural properties, white-matter tracts, structural connectivity matrices

Eyetracking data

blink removal, tracking noise removal, downsampling, detrending, smoothing

Pre-processed time series of gaze position (X and Y) and pupil size

nsd_mapdata

Calculate coordinate transformations and incorporate into nsd_mapdata utility which enables mapping between functional, anatomical, fsaverage, and MNI spaces

Part 2a: GLM analysis to estimate trial-wise betas

NSD experiment

GLMsingle (Fig. 3A–E)

(1) Determine best HRF for each voxel from HRF library

(2) Add data-driven nuisance regressors

(3) Regularize single-trial betas using ridge regression

Single-trial betas in 1.8-mm and 1-mm volume spaces (also create versions in nativesurface, fsaverage, and MNI spaces)

quantification of response reliability to image repetitions

Noise ceiling estimates (Fig. 3F–G)

Part 2b: Analysis of localizer experiments

ROIs (V1–hV4) (Ext. Data Fig. 7A)

pRF experiment

pRF parameter estimates

fLoc experiment

Response estimates to each stimulus category

ROIs (body-, face-, place-, word-selective regions) (Ext. Data Fig. 7C)

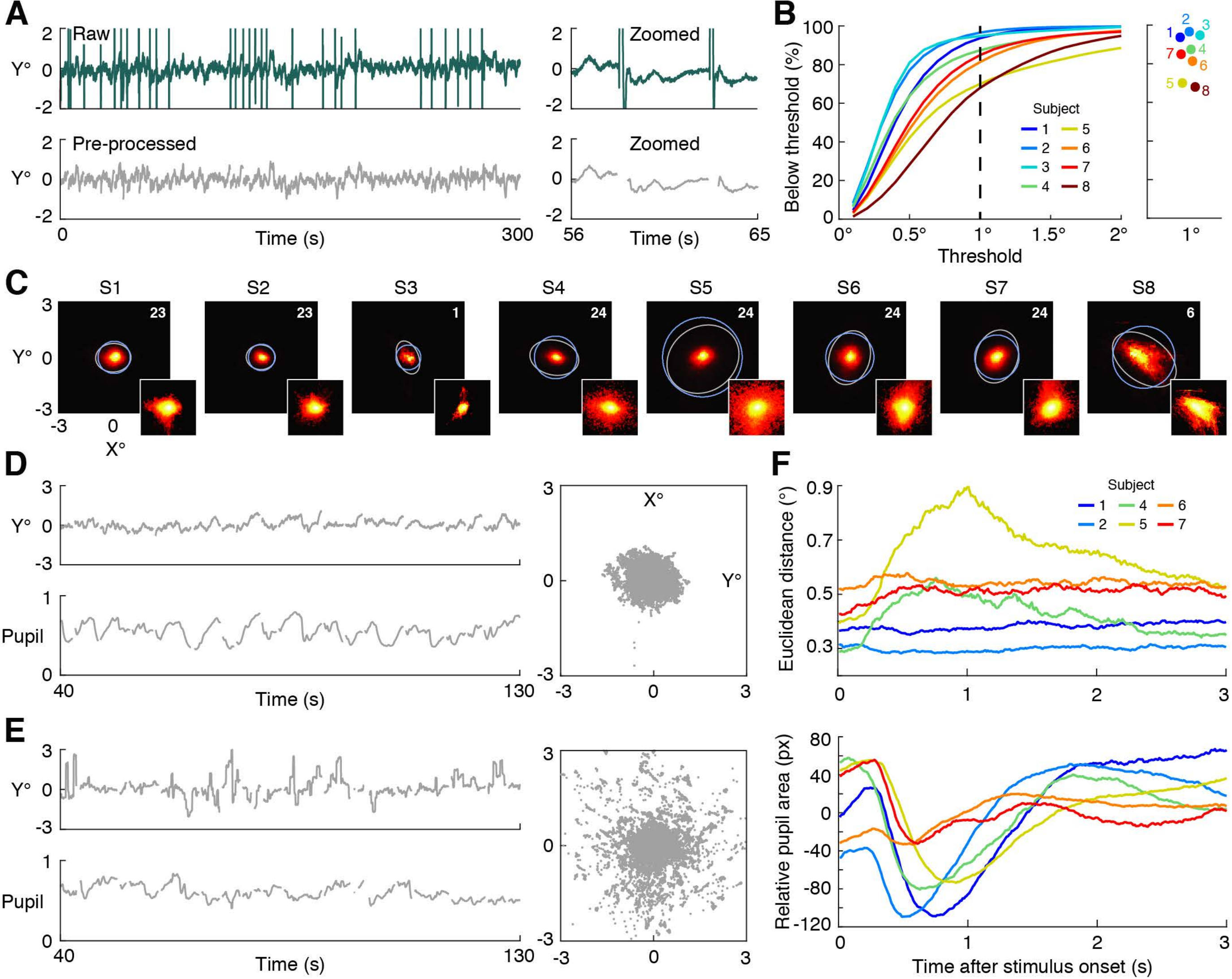
Part 3: Scientific analyses demonstrated in this paper

pRF estimation (Ext. Data Fig. 9)

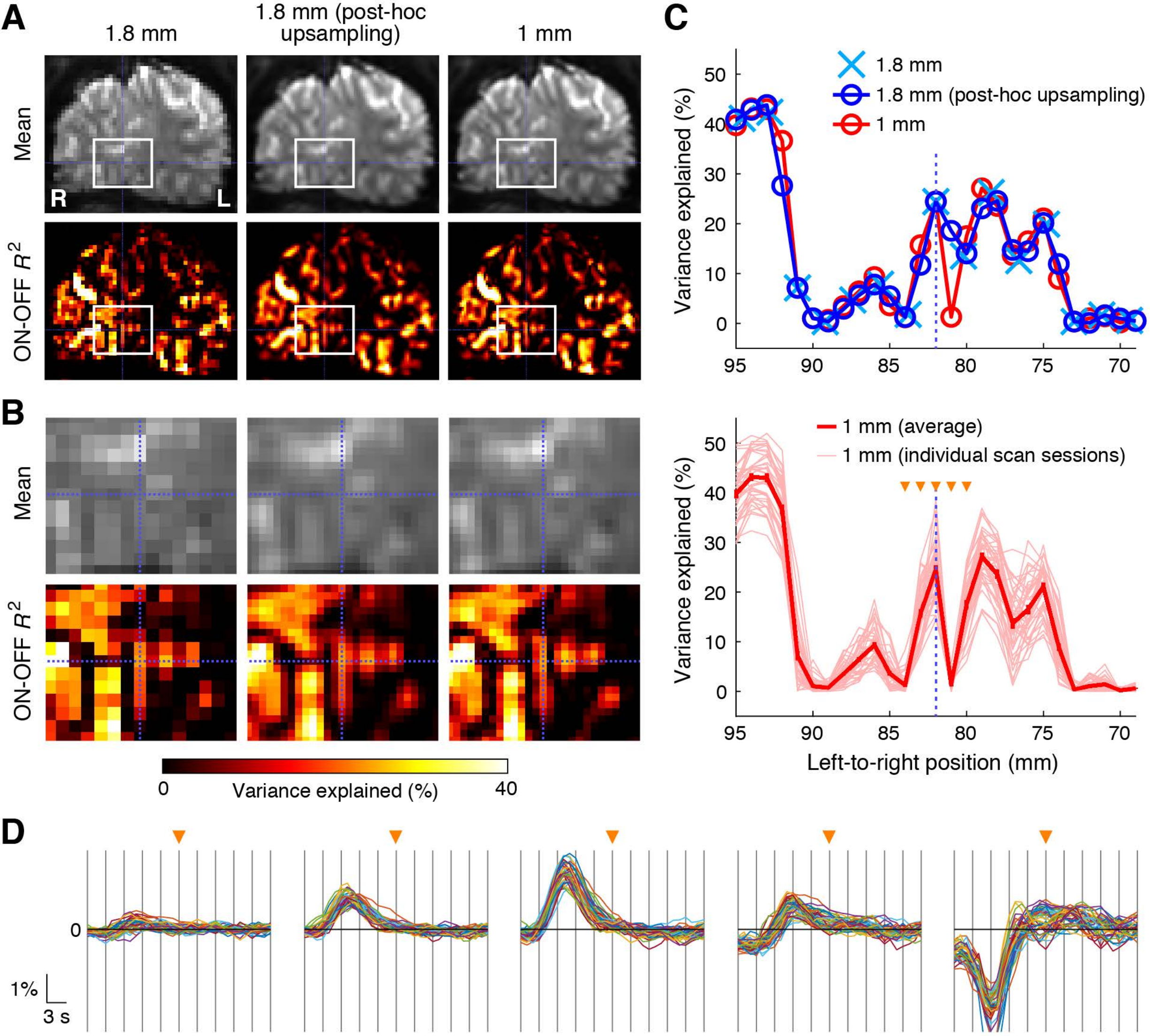
Univariate analysis of memory recognition (Fig. 4)

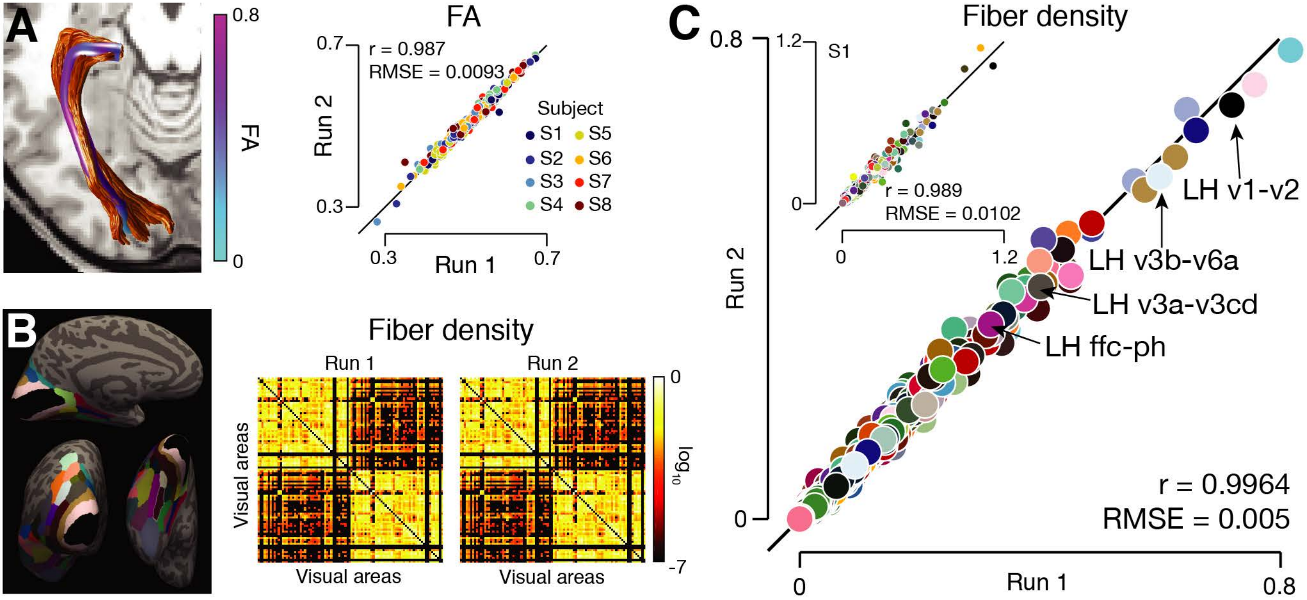
Representational similarity analysis (Fig. 5)

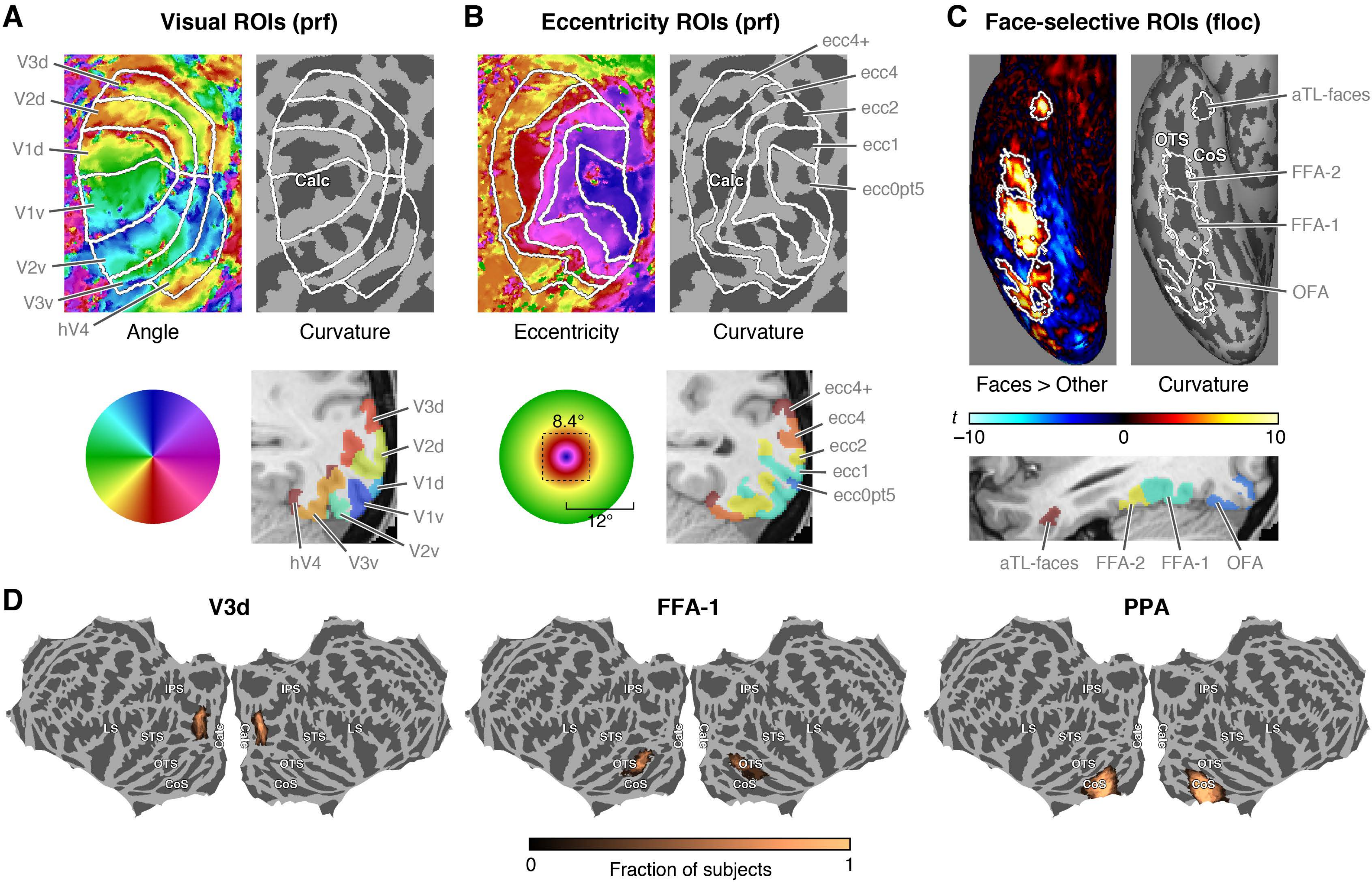
Encoding models based on deep convolutional neural networks (Fig. 6)

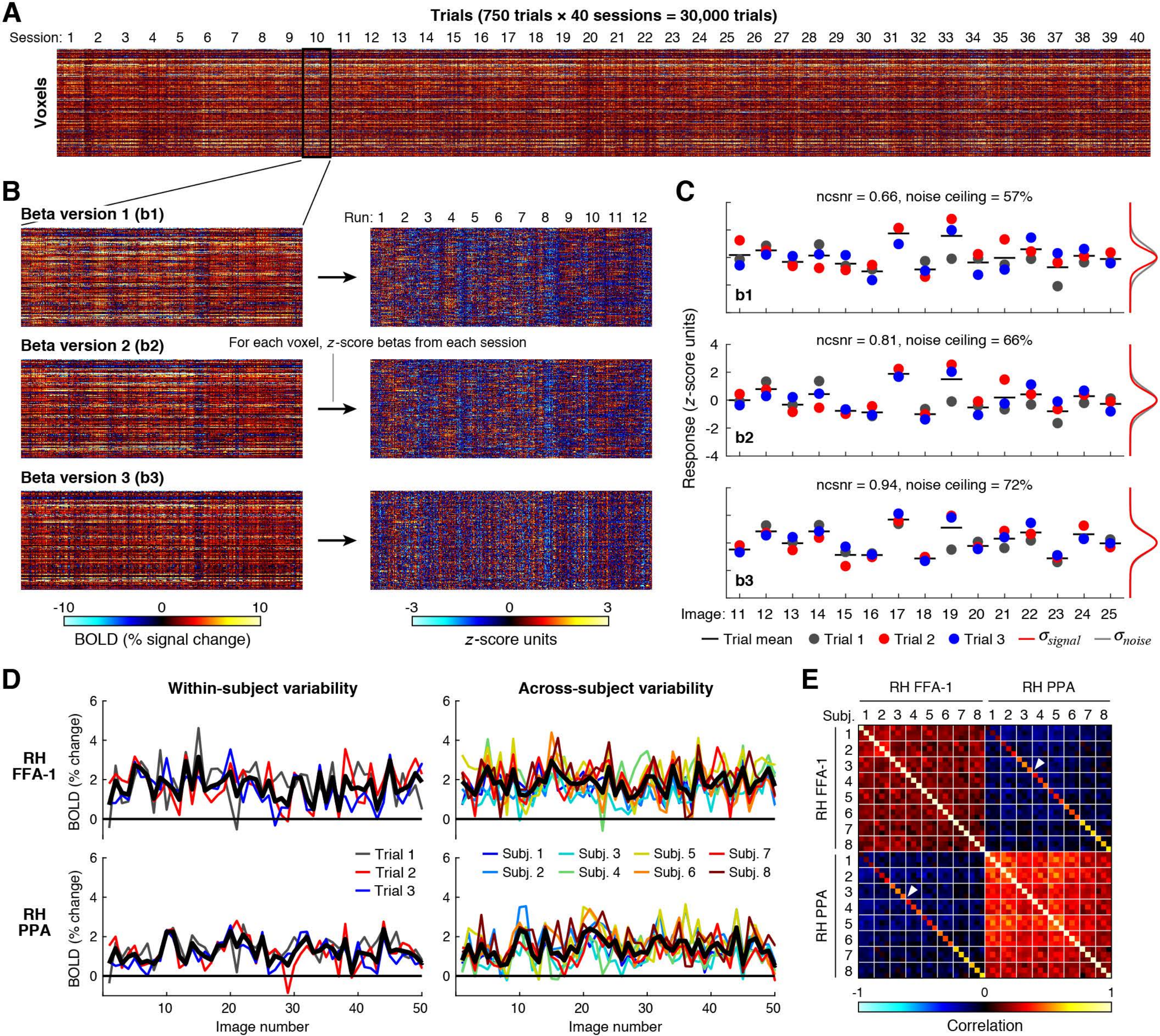


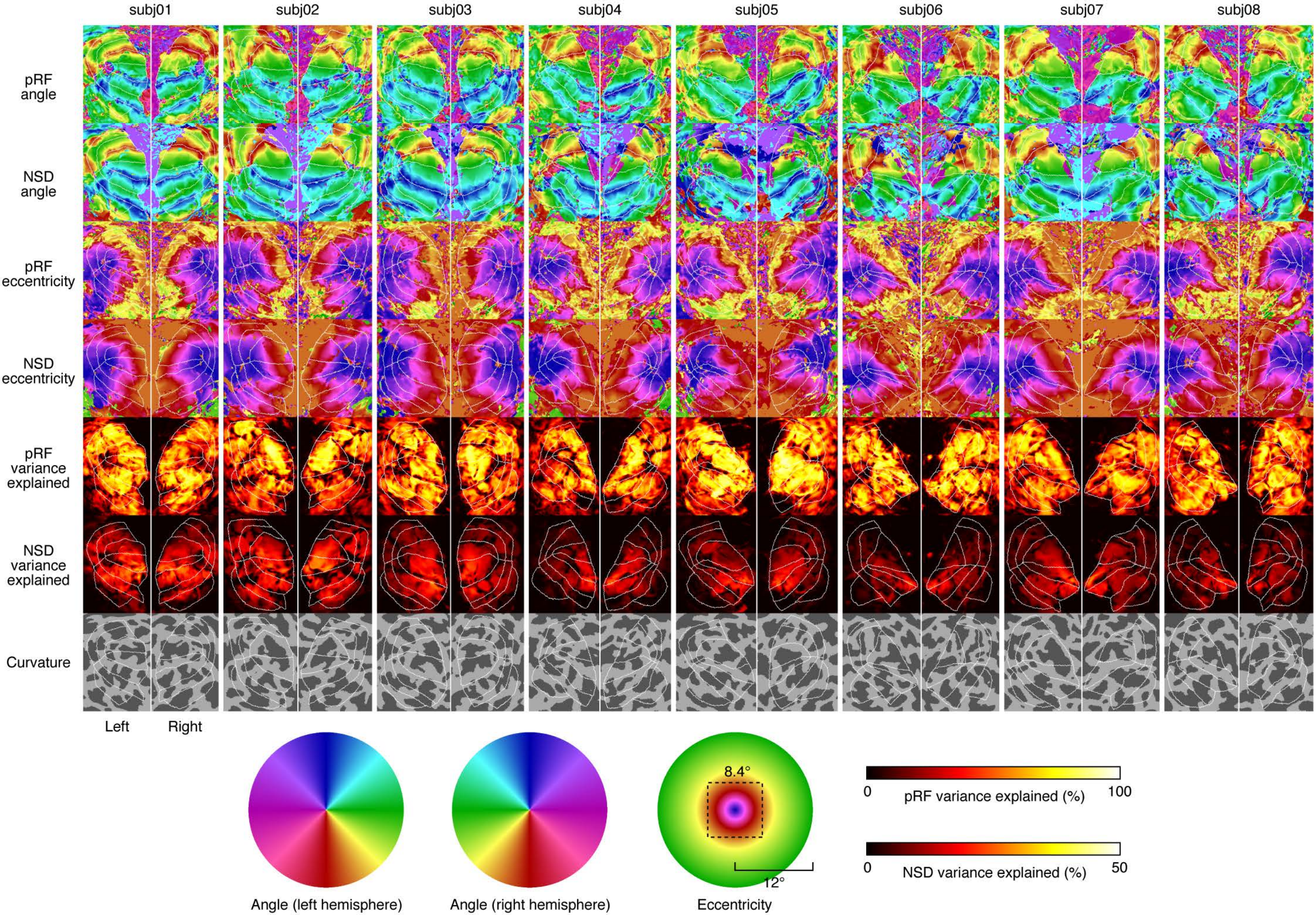
A
Raw
Y°
2
0
-2
Zoomed
Pre-processed
Time (s)
0
300
56
65
B
Below threshold (%)
100
80
60
40
20
0
Threshold
0°
0.5°
1°
1.5°
2°
Subject
1
2
3
4
5
6
7
8
C
S1
S2
S3
S4
S5
S6
S7
S8
23
23
1
24
24
24
24
6
3
-3
X°
D
3
-3
Pupil
1
40
130
X°
Y°
E
F
Euclidean distance (°)
0.9
0.7
0.5
0.3
Relative pupil area (px)
80
40
-40
-80
-120
Time after stimulus onset (s)

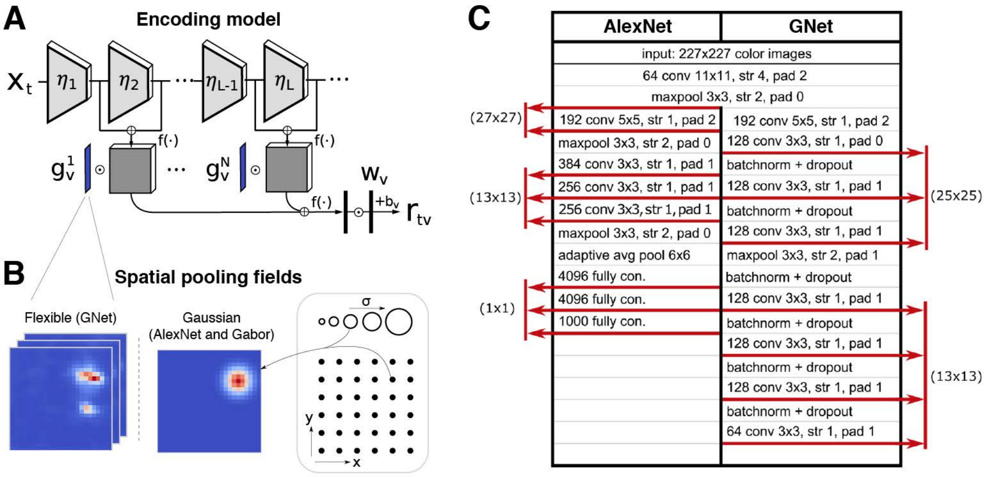

A
1.8 mm
1.8 mm (post-hoc upsampling)
1 mm
Mean
R
L
ON-OFF R^2
B
Mean
ON-OFF R^2
0
Variance explained (%)
40
C
1.8 mm
1.8 mm (post-hoc upsampling)
1 mm
Variance explained (%)
50
40
30
20
10
0
95
90
85
80
75
70
1 mm (average)
1 mm (individual scan sessions)
Left-to-right position (mm)
D
0
1%
3 s


A
0.8
FA
0
FA
0.7
0.3
r = 0.987
RMSE = 0.0093
Run 2
Subject
S1
S2
S3
S4
S5
S6
S7
S8
0.3
Run 1
0.7
B
Fiber density
Run 1
Run 2
Visual areas
Visual areas
Visual areas
0
log10
-7
C
Fiber density
0.8
1.2
S1
0
r = 0.989
RMSE = 0.0102
0
1.2
Run 2
LH v1-v2
LH v3b-v6a
LH v3a-v3cd
LH ffc-ph
0
r = 0.9964
RMSE = 0.005
0
Run 1
0.8


A
Visual ROIs (prf)
V3d
V2d
V1d
V1v
V2v
V3v
hV4
Calc
Angle
Curvature
V3d
V2d
V1d
V1v
V2v
hV4
V3v
B
Eccentricity ROIs (prf)
ecc4+
ecc4
ecc2
ecc1
ecc0pt5
Calc
Eccentricity
Curvature
8.4°
12°
ecc4+
ecc4
ecc2
ecc1
ecc0pt5
C
Face-selective ROIs (floc)
aTL-faces
OTS
CoS
FFA-2
FFA-1
OFA
Faces > Other
Curvature
t
−10
0
10
aTL-faces
FFA-2
FFA-1
OFA
D
V3d
FFA-1
PPA
IPS
LS
STS
OTS
CoS
Calc
0
Fraction of subjects
1


A
Trials (750 trials × 40 sessions = 30,000 trials)
Session: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Voxels
B
Beta version 1 (b1)
Run: 1 2 3 4 5 6 7 8 9 10 11 12
Beta version 2 (b2)
For each voxel, z-score betas from each session
Beta version 3 (b3)
-10 0 10
BOLD (% signal change)
-3 0 3
z-score units
C
ncsnr = 0.66, noise ceiling = 57%
b1
ncsnr = 0.81, noise ceiling = 66%
b2
ncsnr = 0.94, noise ceiling = 72%
b3
Response (z-score units)
4 2 0 -2 -4
Image: 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Trial mean
Trial 1
Trial 2
Trial 3
σ_signal
σ_noise
D
Within-subject variability
Across-subject variability
RH FFA-1
RH PPA
BOLD (% change)
6 4 2 0
Trial 1
Trial 2
Trial 3
Subj. 1
Subj. 2
Subj. 3
Subj. 4
Subj. 5
Subj. 6
Subj. 7
Subj. 8
10 20 30 40 50
Image number
E
RH FFA-1
RH PPA
Subj. 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8
-1 0 1
Correlation


subj01
subj02
subj03
subj04
subj05
subj06
subj07
subj08
pRF angle
NSD angle
pRF eccentricity
NSD eccentricity
pRF variance explained
NSD variance explained
Curvature
Left
Right
Angle (left hemisphere)
Angle (right hemisphere)
8.4°
12°
Eccentricity
0
pRF variance explained (%)
100
0
NSD variance explained (%)
50



AlexNet	GNet
input: 227x227 color images	
64 conv 11x11, str 4, pad 2	
maxpool 3x3, str 2, pad 0	
192 conv 5x5, str 1, pad 2	192 conv 5x5, str 1, pad 2
maxpool 3x3, str 2, pad 0	128 conv 3x3, str 1, pad 0
384 conv 3x3, str 1, pad 1	batchnorm + dropout
256 conv 3x3, str 1, pad 1	128 conv 3x3, str 1, pad 1
256 conv 3x3,str 1,pad 1	batchnorm + dropout
maxpool 3x3, str 2, pad 0	128 conv 3x3, str 1, pad 1
adaptive avg pool 6x6	maxpool 3x3, str 2, pad 1
4096 fully con.	batchnorm + dropout
4096 fully con.	128 conv 3x3, str 1, pad 1
1000 fully con.	batchnorm + dropout
	128 conv 3x3, str 1, pad 1
	batchnorm + dropout
	128 conv 3x3, str 1, pad 1
	batchnorm + dropout
	64 conv 3x3, str 1, pad 1

Supplementary Note 1:
Auxiliary data and resources in the NSD dataset

Substantial amounts of resting-state data with physiological measurements

A minimum of 100 minutes of resting-state data were acquired for each NSD subject. This large amount of data is appealing as it enables stable estimates of network correlations[88]. External physiological measures (pulse oximeter, respiratory belt) were also acquired in most of the scan sessions that included resting-state acquisition. Visual inspections of the physiological data (available online) suggest that the data are of excellent quality and that more than 90% of the pulse data and more than 95% of the respiratory data are usable. These physiological data play a critical role in identifying potential contaminants of resting-state signals[89,90]. Overall, the resting-state component of NSD is valuable not only for enriching interpretation of the core NSD experiment, but also as a standalone resource due to the sizable amount of data and use of 7T imaging.

Highly reliable diffusion data and derivatives

The diffusion data included with the NSD dataset complement the extensive fMRI measurements. We pre-processed the raw diffusion data using the state-of-the-art DESiGNER pipeline methodology[91] as implemented on brainlife.io[92]. As there is no accepted map of white matter in the human brain, the field lacks consensus on how to validate the accuracy of tractography results. We instead focused on reliability as a measure of quality and adopted a statistical approach that evaluates reliability of major tracts and connectivity matrices[93]. We find that the quality of the pre-processed diffusion data for each subject is high, as evidenced by the signal-to-noise ratio (**Supplementary Figure 5B**). We then proceeded to perform diffusion signal modeling[94–98], anatomically-informed tractography[99], and profilometry[100]. White-matter microstructural properties are found to be highly reliable for each subject (**Extended Data Figure 6A**). Structural connectivity matrices[101] derived from tractography results are also highly reliable, both at the group level (**Extended Data Figure 6B–C**) as well as at the single-subject level (**Extended Data Figure 6C, inset**). The ready-to-use diffusion derivatives provided with NSD include a variety of macrostructural and microstructural measures, white-matter tracts, and structural connectivity matrices, as well as intermediary pre-processed outputs (such as denoised and spatially corrected diffusion volumes). These derivatives can be easily integrated into machine learning workflows, and serve as launching points for scientific investigations seeking to apply network neuroscience perspectives[102,103] to understanding brain function in the NSD dataset.

Extensive set of manually defined ROIs

To increase the value of the NSD dataset to the broader community, we performed analysis of the data from the pRF and fLoc experiments and manually defined regions of interest (ROIs) based on the results. The defined ROIs include retinotopic visual areas based on the pRF results (V1, V2, V3, hV4), eccentricity-based regions based on the pRF results (bands between 0°, 0.5°, 1°, 2°, 4°, and beyond), and category-selective regions based on the fLoc results (body-, face-, place-, and word-selective regions). Representative examples illustrating the high quality of the localizer results and associated ROIs are shown in **Extended Data Figure 7**. NSD also includes manual segmentations of the thalamus (LGN, pulvinar, superior colliculus) and the medial temporal lobe (hippocampal subfields and surrounding subregions). These resources reduce overhead and facilitate scientific analyses of the NSD dataset.

Supplementary Note 2:
Limitations of the NSD fMRI data

Our examination of the NSD fMRI data did not reveal any severe problems or artifacts that could not be compensated for in data pre-processing. Nonetheless, there are several known limitations of the data that should be considered. EPI pulse sequences invariably suffer from signal dropout and spatial distortion in locations with magnetic field inhomogeneties. The NSD data exhibit signal dropout in typical locations such as near the ear canals and the frontal sinuses (see **Figure 3F**), and our approach for distortion correction may have some imperfections (see **Supplementary Video 6**). In a few fMRI scan sessions (6 out of 308; 1.9%), the subject exited the scanner and re-entered either on the same day or a different day to complete the session. We compensated for these occurrences in the pre-processing of the data, but they nonetheless contribute some variability. Due to hardware errors, behavioral responses are missing for a few of the NSD runs (2 out of 3,408; 0.06%). A few fMRI scan sessions (4 out of 308; 1.3%) had slightly incomplete brain coverage due to subject motion. Finally, while NSD subjects were instructed to fixate, eye movements are not fully avoidable[104] and are likely present to some degree in the data (see eyetracking results in **Extended Data Figure 4**). A comprehensive summary of data anomalies is available in the online materials.

Supplementary Note 3:
Pre-processing and analysis of eyetracking data

Eyetracking data and video recordings were carefully synchronized and cropped to match the fMRI data acquisition (*nsd_et_crop.m*). The eyetracking data were then pre-processed to reduce noise as detailed below (*nsd_et_preprocessing.m*). Note that the eyetracking data are variable in quality: the eyetracker frequently lost track of the pupil, thereby introducing noise to the data and causing missing samples. We carefully cleaned the data to the best of our ability, taking care to select parameters that work well for the majority of subjects and scan sessions.

We implemented the following pre-processing steps. First, blinks and tracking noise were removed by excluding samples at which the pupil was lost entirely, excising data 100 ms before and 150 ms after each occurrence. Further, because recorded gaze positions tended to jump erratically to the screen edge whenever the pupil was lost, we excluded samples deviating more than 6° from central fixation, excising data 250 ms before and 250 ms after each occurrence. Next, we removed slow signal drift by linear detrending and median-centering of the gaze position time series (X and Y). This step assumes that the median gaze position corresponds to central fixation. Finally, the time-series data for gaze position and pupil size were downsampled to 100 Hz and smoothed using a 50-ms running average. Any eyetracking run containing fewer than 1/3 valid samples after pre-processing was deemed unusable and excluded from further analysis.

Detailed analyses of the eyetracking data obtained for the NSD experiment are shown in **Extended Data Figure 4**. Because two of the subjects (S3 and S8) do not have eyetracking data acquired during the NSD experiment, we considered for these two subjects eyetracking data acquired during the nsdsynthetic experiment (which also required central fixation) as a proxy. This allowed aggregate analyses to be performed for these subjects (panels B and C), but precluded trial-wise time-resolved analyses (panel F), due to the different experimental design used in the nsdsynthetic experiment.

Note that despite our best efforts to reduce noise in the eyetracking data, the data are still noisy. This can be appreciated by inspecting the eyetracking video recordings available online. Thus, not all deviations in the recorded data reflect actual eye movements, and results shown in **Extended Data Figure 4** likely reflect an underestimation of the true fixation accuracy of the subjects. As an alternative approach, it may be possible to infer eye gaze from fMRI signal intensities in and around the eyeballs[105,106]. This could provide robust estimates of fixation accuracy, even for scan sessions where eyetracking was not conducted.

Supplementary Note 4:
Pre-processing of the MRI data

Pre-processing of anatomical data

T_1 and T_2 volumes

T_1- and T_2-weighted volumes were corrected for gradient nonlinearities using a custom Python script (https://github.com/Washington-University/gradunwarp) and the proprietary Siemens gradient coefficient file retrieved from the scanner. The multiple T_1 and T_2 volumes acquired for a given subject were then co-registered (*preprocess_nsd_structuralalignment.m*). This was accomplished by first co-registering the T_1 volumes to each other (rigid-body transformation; correlation metric; the first T_1 volume serving as the target) and then by co-registering the T_2 volumes to the T_1 volumes (rigid-body transformation; mutual information metric; the prepared T_1 data serving as the target). In the estimation of registration parameters, a manually defined 3D ellipse was used to focus the cost metric on brain tissue. Individual volumes were manually inspected and rejected if substantial image artifacts were visible (only the 2nd and 4th T_1 volumes for subject 8 were rejected). The final T_1 and T_2 data were created by performing cubic interpolation at a resolution of 0.5 mm. Results for the multiple acquired volumes were averaged (within modality) to increase contrast-to-noise ratio. Finally, the 0.5-mm volumes were resampled to create alternative 0.8-mm and 1.0-mm versions. These resampled versions are provided for the convenience of users.

SWI volume (venogram)

We co-registered the SWI volume (magnitude component only; corrected for gradient nonlinearities) to the prepared 1.0-mm EPI volume (*preprocess_nsd_SWI.m*). To compensate for the acquisition being performed on different scanners, we used a slightly flexible nonlinear warp as implemented in ANTs 2.1.0 (BSplineSyN with parameters [0.1, 400, 0, 3]). The final SWI volume was prepared in the subject-native anatomical space by performing B-spline interpolation at a resolution of 0.5 mm. The resulting volume was then resampled to create alternative 0.8-mm and 1.0-mm versions.

TOF volume (angiogram)

We co-registered the TOF volume to the prepared 1.0-mm T_1 volume (*preprocess_nsd_TOF.m*). This was accomplished using a slightly flexible nonlinear warp as implemented in ANTs 2.1.0 (BSplineSynN with parameters [0.1, 200, 0, 3]). To aid estimation of registration parameters, a temporary version of the TOF volume was used in which extremely bright pixels were dampened. The final TOF volume was prepared in the subject-native anatomical space by performing B-spline interpolation at a resolution of 0.5 mm. The resulting volume was then resampled to create alternative 0.8-mm and 1.0-mm versions.

High-resolution T_2 volume

We co-registered the high-resolution T_2 volume to the prepared 0.5-mm T_2 volume (*external_mtl.m*). Given that these volumes were acquired on different scanners, we evaluated several strategies for achieving accurate co-registration. We obtained the best results by performing a simple linear co-registration (affine transformation; correlation metric) in combination with a rectangular box that focused the cost metric on regions of interest in the medial temporal lobe. The estimated registration was subsequently used to map labels defined on the high-resolution T_2 volume to the subject-native anatomical space.

De-identification

We mapped a liberal brain mask defined in MNI space to the subject-native anatomical space, and then used the result to mask and thus de-identify the anatomical volumes (*preprocess_nsd_applybrainmask.m*).

FreeSurfer processing

The prepared 0.8-mm T_1 volume was processed using FreeSurfer version 6.0.0 with the *-hires* option (*analysis_freesurfer.m*). Manual edits of tissue segmentation (labeling voxels as gray matter, white matter, or cerebrospinal fluid) were performed for each of the eight subjects to optimize the accuracy of the cortical surface representations generated by FreeSurfer. The prepared 0.8-mm T_2 volume was used to inform manual segmentation decisions, but was not explicitly used in the FreeSurfer processing. We also manually marked surface imperfections that remained even after manual edits; these are labeled in the surface inspections (**Supplementary Video 2**) and are largely confined to a few difficult regions located in the inferior aspects of the temporal and frontal lobes (see **Figure 3F**).

Several additional FreeSurfer processing steps were performed. Using *mris_expand*, we generated cortical surfaces positioned at 25%, 50%, and 75% of the distance between the pial surface and the boundary between gray and white matter. These surfaces are useful for creating surface representations of the fMRI data. Multiple surfaces at different gray-matter depths were created given the relatively high spatial resolution of the fMRI data (1.8-mm acquisition); this may represent a departure from standard fMRI workflows geared towards lower-resolution data. We also generated several flattened surface representations: for each hemisphere in each subject, we created a flattened version of the entire cortical sheet (using manually defined cuts) as well as flattened versions of cortical patches covering ventral temporal cortex and early visual cortex (patches were determined automatically based on a set of cortical patches defined on fsaverage). Finally, in line with the 'surface voxels' visualization technique[32], we sampled 1-, 2-, and 3-mm volumetric test patterns onto surface vertices using nearest-neighbor interpolation (*analysis_surfacevoxels.m*). The test patterns, distributed with the dataset, may be useful to users for understanding the impact of cortical curvature on surface visualizations.

Pre-processing of diffusion data

Code scripts used to analyze the diffusion data are accessible via the hyperlinks indicated below, which refer to brainlife.io apps (http://brainlife.io)[107].

The prepared 0.8-mm T_1-weighted volume for each subject was segmented into different tissue types using MRTrix3[108] (https://doi.org/10.25663/brainlife.app.239). The gray- and white-matter interface mask was subsequently used as a seed mask for white-matter tractography. For network generation, the HCP-MMP cortical parcellation[109] was mapped to subject-native surfaces and then to the volumetric Freesurfer segmentation (ribbon.mgz) for each subject (https://doi.org/10.25663/bl.app.23).

The raw data from the four diffusion scans (99 AP, 99 PA, 100 AP, 100 PA) were corrected for gradient nonlinearities, concatenated, and then pre-processed following a published protocol[91]. Specifically, diffusion volumes were denoised and cleaned with respect to Gibbs ringing using MRTrix3 before being corrected for susceptibility, motion, and eddy distortions using FSL's *topup* and *eddy* functions (https://doi.org/10.25663/brainlife.app.287). We note that the raw acquired diffusion volumes exhibit substantial 'striping' artifacts (in which every other slice appears spatially displaced), possibly reflecting within-volume motion caused by physical vibrations of the RF coil. We attempted to mitigate these effects using the mporder functionality of *eddy*, but we caution that some residual artifact may exist in the pre-processed results. Following these corrections, the diffusion volumes were bias-corrected and had background noise removed using MRTrix3. Finally, the diffusion volumes were co-registered to the 0.8-mm T_1-weighted anatomical volume using FSL's *epi_reg* (rigid-body transformation, boundary-based registration), and then resliced to 0.8-mm isotropic voxels. The diffusion data were organized into two runs: data from the 99 AP and 99 PA scans constitute 'Run 1' and data from the 100 AP and 100 PA scans constitute 'Run 2' (https://doi.org/10.25663/brainlife.app.371). To assess data quality, we calculated signal-to-noise ratio in the corpus callosum using workflow provided by Dipy 1.1[110] (https://doi.org/10.25663/brainlife.app.120).

Following pre-processing, brain masks were generated using Dipy's *median_otsu* (https://doi.org/10.25663/bl.app.70). This mask was used in subsequent model fitting and tractography. Multiple models of myelinated microstructural organization were fit to the diffusion data from each run. This included the diffusion tensor (DTI) model[96], diffusion kurtosis (DKI) model[95], and the neurite orientation dispersion and density imaging[94,98] (NODDI) models (https://doi.org/10.25663/bl.app.9, https://doi.org/10.25663/brainlife.app.365). The NODDI model was fit twice for each run: once for white-matter tract microstructure using an intrinsic free diffusivity parameter ($d_{/\!/}$) of 1.7×10^{-3} mm^2/s, and once for cortical microstructure using $d_{/\!/} = 1.1 \times 10^{-3}$ mm^2/s, following previously described procedures[111]. The constrained spherical deconvolution[112] (CSD) model was fit for 4 spherical harmonic orders ($L_{\max}$ = 2, 4, 6, 8) using MRTrix3 (https://doi.org/10.25663/brainlife.app.238). The fiber orientation distribution functions for $L_{\max}$ = 6 and $L_{\max}$ = 8 were subsequently used to guide anatomically-constrained probabilistic tractography[99] using MRTrix3 (https://doi.org/10.25663/brainlife.app.297). A total of 3 million streamlines across $L_{\max}$ = 6 and $L_{\max}$ = 8 for each run were generated, using a step size of 0.2 mm, minimum length of 25 mm, maximum length of 250 mm, and a maximum angle of curvature of 35°. Finally, structural connectivity matrices representing fiber density were generated using a combination of SIFT2[113] and MRTrix3 (https://doi.org/10.25663/brainlife.app.394). Fiber density was quantified as the number of streamlines connecting two regions divided by the average volume of the two regions.

Following model fitting and tractography, 61 major white-matter tracts were segmented for each run using a customized version of the white matter query language[114] (https://doi.org/10.25663/brainlife.app.188). Then, using Vistasoft (https://github.com/vistalab/vistasoft), outlier streamlines were removed (https://doi.org/10.25663/brainlife.app.195) and tract profiles (each tract sampled with 200 nodes) were generated for DTI, DKI, and NODDI measures (https://doi.org/10.25663/brainlife.app.361). Finally, these measures were mapped to the cortical surface following previously published procedures[111] using FreeSurfer 7.0 and Connectome Workbench 1.4.2 (https://github.com/Washington-University/workbench) (https://doi.org/10.25663/brainlife.app.379).

Pre-processing of functional data

Overall strategy

We implemented a pre-processing approach that aimed to preserve as much spatial and temporal detail as possible. In short, the fMRI data were pre-processed by performing one temporal resampling to correct for slice time differences and one spatial resampling to correct for head motion within and across scan sessions, EPI distortion, and gradient nonlinearities. This produced volumetric fMRI time-series data in subject-native space for each NSD subject. The functional data were pre-processed independently of the anatomical data; this was done intentionally in order to avoid dependence of the pre-processed functional data on choices such as how to co-register the functional and anatomical data. Also, to minimize the risk of inaccurate or unwanted assumptions, we did not include any temporal filtering (e.g. detrending, confound regression, censoring). Pre-processing results were carefully visually inspected to ensure quality control. There were a few anomalous cases, such as acquisition being split across two different scan sessions; special modifications were made to the pre-processing to accurately compensate for these occurrences (see online notes for details).

First-stage pre-processing

Given the fMRI data acquired in a scan session, a series of steps were performed (*preprocess_nsd.m*, *preprocessfmri.m*):

1. *Temporal resampling*. A cubic interpolation of each voxel's time-series data in each run was performed. This interpolation corrected differences in slice acquisition times (as determined from the DICOM header) and also upsampled the data (in the same step) to either 1.333 s (standard-resolution preparation) or 1.000 s (high-resolution preparation). Data were prepared such that the first time-series data point coincides with the acquisition time of the first slice acquired in the first volume of each run. The upsampling exploits the benefits of temporal jitter between the acquisition and the experiment and synchronizes the time-series data to convenient multiples of the experiment trial structure[73]. For example, in the standard-resolution preparation, there are 3 time points for each 4-s trial in the NSD experiment.
2. *Fieldmap preparation.* The multiple fieldmaps acquired in the scan session (3.6-mm slices) were upsampled using nearest-neighbor interpolation to match the slice resolution of the fMRI data (1.8-mm slices). The fieldmaps were then phase-unwrapped using the FSL utility *prelude* and regularized by performing 3D local linear regression using an Epanechnikov kernel with radius 5 mm. Values in the magnitude component of the fieldmaps were used to regularize the fieldmaps and the regression in order to improve robustness of the field estimates. Finally, the fieldmaps were linearly interpolated over time, producing an estimate of the field for each fMRI volume acquired. This time-varying fieldmap strategy is atypical for fMRI workflows, but we have found it to be highly effective[32].
3. *Spatial undistortion*. The temporally resampled volumes from Step 1 were undistorted based on the field estimates from Step 2 using the standard unwarping method[115]. Undistorted volumes were generated using cubic interpolation.
4. *Motion estimation*. The undistorted volumes from Step 3 were used to estimate rigid-body motion parameters using the SPM5 utility *spm_realign* (the first fMRI volume in the scan session served as the reference). A manually defined 3D ellipse was used to focus the cost metric on brain regions unaffected by gross susceptibility effects. Note that the estimated motion parameters reflect temporally upsampled data and should be interpreted accordingly (e.g. when assessing framewise displacement). Also, note that the motion parameters may reflect apparent image motion due to respiration-induced B_0 fluctuations[116]; this was particularly apparent in subject 3.

5. *Spatial resampling.* A single cubic interpolation was performed on the temporally resampled volumes from Step 1 in order to correct for the combined effects of head motion and spatial distortion.

Gradient nonlinearity correction and session registration

Given the results of the first-stage pre-processing, we computed the mean fMRI volume and corrected this volume for gradient nonlinearities (*preprocess_nsd_epigradunwarp.m*). We then co-registered this gradient-corrected volume to the gradient-corrected volume from the first NSD scan session (affine transformation, correlation metric). Thus, the first NSD scan session determined the target space for preparing fMRI data from different scan sessions (*preprocess_nsd_epialignment.m*).

Second-stage pre-processing

We repeated the pre-processing steps (Steps 1–5 above) but with the final spatial resampling step incorporating the effects of the gradient nonlinearity correction and the session registration (*preprocess_nsd_secondstage.m*). In this way, a single cubic interpolation is used to compensate for the effects of head motion, spatial distortion, gradient nonlinearities, and session registration. For this final interpolation step, we used either a 1.8-mm grid (standard-resolution preparation) or a 1.0-mm grid (high-resolution preparation). The latter approach intentionally upsamples the data in order to exploit the benefits of small head displacements and preserve as much spatial detail as possible[32,34]. To minimize storage requirements, the interpolations were performed within a 3D box that was just large enough to cover the brain of each subject.

To facilitate assessment of T_2* effects, we created a bias-corrected version of the mean EPI volume (*analysis_biascorrection.m*). For each preparation, we took the mean EPI volume and fit a 5th-degree 3D polynomial, considering only voxels labeled as cortical or cerebellar gray matter in the FreeSurfer aseg file. The fitted volume (‘coilbias’) was then divided from the mean EPI volume, producing the bias-corrected volume (‘bc’).

Final outputs

The final result of pre-processing was volumetric fMRI time-series data in subject-native space. Two versions were generated: the standard-resolution version was prepared at a spatial resolution of 1.8 mm and a temporal resolution of 1.333 s, whereas the high-resolution version was prepared at a spatial resolution of 1 mm and a temporal resolution of 1.000 s. These two volumetric versions of the fMRI data are the main versions of the data. However, we do create some alternative versions of the data: for example, a surface-based version of the NSD betas (‘nativesurface’) is prepared for the convenience of users. We prioritize the volume-based format as the main version of the data; this is primarily because it is a simple format, amenable for both cortical and sub-cortical analyses, and does not incorporate specific decisions about how to map functional data onto cortical surface representations.

Calculation of coordinate transformations between volumetric and surface-based representations of functional and anatomical images

We performed several analyses related to mapping data between different spaces:

- *Mapping between functional and anatomical spaces.* We co-registered the mean fMRI volume (1-mm preparation; mean of first five NSD sessions) to the prepared 1.0-mm T_2 volume (*preprocess_nsd_functionaltostructuralalignment.m*). To compensate for acquisition on different scanners, we used a slightly flexible nonlinear warp as implemented in ANTs 2.1.0 (BSplineSyN with parameters [0.1, 400, 0, 3]). A small amount of nonlinearity was necessary to achieve accurate co-registration (see inspections provided online).
- *Changing resolutions in anatomical space.* For resampling data to different anatomical resolutions (0.5-, 0.8-, or 1.0-mm), we used an ideal Fourier filter (10th-order low-pass Butterworth filter) followed by cubic interpolation (*changevolumeres.m*).
- *Mapping to and from fsaverage.* We calculated the indexing information that maps subject-native surfaces to and from fsaverage using nearest-neighbor interpolation in the spherical space defined by FreeSurfer. Visual inspections confirm the quality of the folding-based alignment achieved by FreeSurfer (**Supplementary Video 3**).

- *Mapping to and from MNI.* Using FSL's utility *fnirt*, we co-registered the subject-native prepared 1.0-mm T_1 volume to the MNI152 T_1 1-mm template provided by FSL (*preprocess_nsd_MNIandbrainmask.m*). Visual inspections confirm the quality of the registration (**Supplementary Video 4**).
- *Converting surface data to volumetric format.* We implemented a method that, for a given target anatomical volume with resolution *R* mm, allows each surface vertex to contribute a triangular (linear) kernel of size +/– *R* mm and then calculates a weighted average of data values at the position of each voxel in the volume (*cvnmapsurfacetovolume_helper.m*).

Based on the results of the above analyses, we calculated coordinate transformations that indicate how to map between the functional spaces, the anatomical spaces, the subject-native surfaces, fsaverage, and MNI space (*preprocess_nsd_calculatetransformations.m*). Finally, we created a lightweight utility (*nsd_mapdata.{m,py}*) that uses the coordinate transformations to map user-supplied data from one space to another space under a given interpolation scheme (nearest-neighbor, linear, cubic, winner-take-all). For example, data in subject-native functional space can be mapped to MNI space or a subject-native cortical surface using a single interpolation of the functional data. The use of interpolation to map volumetric data onto surface representations (as opposed to incorporating spatial kernels tailored to the cortical surface) helps maximize spatial resolution and avoids making strong assumptions about cortical topology. Nonetheless, the user is free to apply other methods (e.g., FreeSurfer, Connectome Workbench) to perform the mapping. We used the nsd_mapdata utility to perform a number of mundane but useful transformations, such as generating versions of the anatomical volumes that are matched to the functional volumes (*analysis_transforms.m*). Label data (e.g. ROI labels) were transformed by performing a separate interpolation for each label and then a winner-take-all operation.

Supplementary Note 5: Localizers and regions of interest

Analysis of the pRF experiment

The pre-processed fMRI data from the pRF experiment were analyzed using the Compressive Spatial Summation model[80] as implemented in analyzePRF (*analysis_prf.m*). First, the time-series data from the three repetitions of each run type (multibar, wedgering) were averaged. Stimulus apertures indicating the position of the texture were prepared at a resolution of 200 pixels × 200 pixels. We then used analyzePRF (http://cvnlab.net/analyzePRF/) to estimate pRF parameters for each voxel (canonical hemodynamic response function; seedmode 2). Results were mapped to the cortical surface by performing linear interpolation on the volumes (1-mm preparation) and then averaging across cortical depth. To quantify behavioral performance, we calculated, for each run, $(A - B)/C \times 100$, where A indicates the number of successful detections of color changes (button pressed within 1 s of a color change), B indicates the number of extraneous button presses, and C indicates the total number of color changes. Performance averaged across the six runs ranged between 93.5–98.9% for the eight NSD subjects.

Analysis of the fLoc experiment

The pre-processed fMRI data from the fLoc experiment were analyzed using GLMdenoise[35,117], a data-driven denoising method that derives estimates of correlated noise from the data and incorporates these estimates as nuisance regressors in a general linear model (GLM) analysis of the data (*analysis_floc.m*). We coded the 10 stimulus categories using a "condition-split" strategy[32] in which the trials associated with a single category were split into separate conditions in each run. We used six condition-splits, thereby producing six response estimates (betas) for each category. After fitting the GLM, *t*-values were computed from the GLM betas in order to quantify selectivity for different categories and domains (e.g., selectivity for faces was quantified by calculating a *t*-value that contrasts adult and child faces vs. all other categories). Results were mapped to the cortical surface by performing linear interpolation on the volumes (1-mm preparation) and then averaging across cortical depth. To quantify behavioral performance, we calculated the hit rate for each run (button pressed within 1 s of an oddball image). Performance averaged across the six runs ranged between 90.8–97.5% for the eight NSD subjects.

Regions of interest (ROIs)

Volume-based subject-native ROIs

The *thalamus* ROI collection consists of ROIs related to the lateral geniculate nucleus, pulvinar, and superior collicus (*external_subcortical.m*). Manual labeling of these ROIs was performed by an expert (M. Arcaro) based on the T_1 and T_2 anatomical volumes obtained for each subject as well as functional results obtained in prior studies[118], projected from MNI space to the native space of each subject. Labels were defined in 0.5-mm anatomical space and were resampled to create alternative 0.8-mm and 1.0-mm versions. To provide labels in functional space, the 0.8-mm anatomical volume was mapped to the 1.0-mm functional space, and the 1.0-mm anatomical volume was mapped to the 1.8-mm functional space.

The *MTL* ROI collection consists of ROIs related to the hippocampus and surrounding brain regions in the medial temporal lobe (*external_mtl.m*). Manual labeling of these ROIs was performed by an expert (W. Guo) based on the high-resolution T_2 volume obtained for each subject, following a published protocol[119]. Labels were defined on the raw high-resolution volume, and were mapped to 0.5-mm anatomical space using the previously determined affine transformation. Note that the resulting labels have some amount of jaggedness due to the anisotropy of the voxels in the high-resolution T_2 volume. Alternative versions of the labels were created in the same way as described for the *thalamus* labels.

Surface-based subject-native ROIs

Results of the pRF experiment were used to define *prf-visualrois*, a collection of ROIs consisting of the dorsal and ventral subdivisions of V1, V2, and V3, and area hV4 (*analysis_drawrois_prf*.m*). These ROIs were manually drawn on cortical surfaces by experts (K. Kay, J. Winawer) based on pRF angle and eccentricity estimates, following common practices[120]. The ROIs extended to the fovea (0° eccentricity) but were restricted to the extent of cortex stimulated by the pRF experiment.

The pRF results were also used to define *prf-eccrois*, a collection of ROIs consisting of concentric regions with increasing eccentricity coverage (0.5°, 1°, 2°, 4°, >4°). Labeled regions were confined to the same cortical extent labeled in *prf-visualrois*.

Results of the fLoc experiment were used to define several collections of category-selective ROIs, including commonly used ROIs such as extrastriate body area (EBA), fusiform face area (FFA), parahippocampal place area (PPA), and visual word form area (VWFA) (*analysis_drawrois*.m*). These ROIs were manually drawn on cortical surfaces by experts (K. Kay, A. White, A. Bratch) based on a combination of anatomical location (relative to sulci and gyri) and stimulus selectivity *t*-values obtained from the fLoc experiment, following general procedures used in prior studies[31,121,122]. For each ROI collection (*floc-bodies*, *floc-faces*, *floc-places*, *floc-words*), several ROIs exhibiting preference for the associated category were defined (e.g., *floc-faces* was based on *t*-values for the contrast of faces > non-faces). ROIs were defined by drawing a polygon around a given patch of cortex and then restricting the ROI to vertices within the polygon that satisfy $t > 0$. This liberal criterion was used to provide maximum flexibility (the user can easily restrict the ROI further using the provided *t*-value volumes).

Surface-based atlas ROIs

To help summarize results in this paper, we defined *nsdgeneral*, an ROI in occipital cortex reflecting regions generally responsive in the NSD experiment (*analysis_drawnsdgeneral.m*). This ROI was drawn on fsaverage based on group-average results for variance explained by the b3 version of the GLM. The ROI is shown in **Figure 3F**.

To provide anatomical reference, we defined *corticalsulc*, a collection of ROIs consisting of major sulci and gyri, and *streams*, a collection of ROIs reflecting large-scale divisions of visual cortex (early, midventral, midlateral, midparietal, ventral, lateral, parietal). These ROI collections were manually drawn on fsaverage.

For convenience, the NSD dataset also includes a few publicly available atlases. These include *Kastner2015*[123], an atlas of visual topography, and *HCP_MMP1*[109], a whole-brain cortical parcellation based on multimodal measures. Both atlases were prepared in fsaverage and converted as described below.

Conversion of surface-based ROIs

A number of conversions were performed to prepare volumetric versions of surface-based ROIs (*analysis_surfaceroistovolume.m*). ROIs defined on fsaverage were mapped to subject-native surfaces using nearest-neighbor interpolation. ROIs defined on subject-native surfaces were mapped to 0.8-mm anatomical space by assigning labels to the 3 depth-dependent surfaces and then performing weighted linear conversion (as described earlier). The 0.8-mm volume was then mapped to the 1.0-mm and 1.8-mm functional spaces.

Supplementary Modeling Note 1: Estimation of pRFs from the NSD responses

For this analysis (results shown in **Extended Data Figure 9**), we used version 3 of the NSD betas (b3) in the nativesurface preparation. Betas for each surface vertex were *z*-scored within each scan session, concatenated across sessions, averaged across repeated trials for each distinct image, and then re-normalized using a scale and offset such that 0 corresponds to 0% BOLD signal change and the standard deviation of the betas equals 1.

To prepare stimuli for pRF estimation, the NSD images were converted to grayscale, resized to 800 pixels × 800 pixels (cubic interpolation), and squared to mimic the luminance response of the display. The images were then placed against the gray background and divided into a 51 × 51 grid such that the first and last grid elements were centered at the edges of the stimulus (each grid element spanned 0.168° × 0.168°). Finally, to quantify local contrast, we computed the standard deviation of pixel values within each grid element.

Based on the local-contrast preparation of the NSD images, we used analyzePRF (http://cvnlab.net/analyzePRF/) to fit the Compressive Spatial Summation pRF model[80] to the trial-averaged betas obtained for each vertex. The non-shared NSD images were used as training data; the shared NSD images were used as validation data. pRFs were constrained to have non-negative gain. No offset term was included in the model (opt.maxpolydeg = NaN); thus, the model necessarily predicts a response of 0 for an image with zero contrast. For model fitting, an initial gridding of model parameters was performed (opt.seedmode = 2), and parameter optimization started from the best parameter combination (opt.modelmode = 2; opt.algorithm = 'trust-region-reflective'). Model fitting produced, for each vertex, an estimate of pRF angle, eccentricity, size, exponent, and gain, as well as variance explained in the training data and the validation data.

Supplementary Modeling Note 2: Encoding models based on deep convolutional neural networks

For this analysis (results shown in **Figure 6**), we used version 3 of the NSD betas (b3) in the 1.8-mm volume-based preparation. Before modeling, betas for each voxel were z-scored within each scan session and concatenated across sessions. Models were implemented using PyTorch.

Model architecture

We considered several variants of voxel-wise encoding models[46] that attempt to predict the NSD betas. All three models consist of (i) a feature extractor implemented as a convolutional neural network (CNN) and (ii) a network-to-brain coupling model that maps extracted features into predictions of activity observed for individual voxels.

In the first model (AlexNet), the feature extractor is the AlexNet CNN[124], a *task-optimized* network that has been trained to classify object categories in the ImageNet database[125]. In the second model (GNet), the feature extractor is a different CNN—referred to here as 'GNet'—a *brain-optimized* network that is trained to directly predict brain activity in the NSD dataset. The third model is a simple control model in which the feature extractor consists of a single fixed layer of Gabor filters[24]. The specific network architectures for AlexNet and GNet are illustrated in **Extended Data Figure 10**.

To facilitate direct comparison, all models are designed to have comparable coupling models. For GNet, both the feature extractor and coupling model are trained jointly using brain data; for AlexNet and the Gabor models, the feature extractors are fixed and only the coupling model is trained using brain data.

The CNNs in the AlexNet, GNet, and Gabor models consist of hierarchically composed functions of an input image x:

$$e_l(x) = \eta_l \circ e_{l-1}(x)$$

where η_l is a feature extractor that operates at layer l on the output of $e_{l-1}(x)$ (also a composite function). e_l may denote an arbitrary sequence of transformations. The encoding models leverage the intermediate representations $e_l(x)$, which are feature maps with pixels denoted by $[e_l(x)]_{kji}$, where (i,j) is the location of the pixel in the kth feature map. Predicted brain activity for voxel v, $\widehat{r_v}$,
is given by the expression:

$$\widehat{r_v} = b_v + \sum_k w_{vk} f\big(\Phi_k(x)\big)$$

where w_{vk} are feature weights for voxel v and feature k, b_v is a bias term,

$$\Phi_k(x) = \sum_{i,j} f\big([e_1(x)]_{k_1 ji}\big) g^1_{vji} \oplus \dots \sum_{i,j} f\big([e_L(x)]_{k_L ji}\big) g^L_{vji}$$

$f(\cdot)$ is typically a compressive nonlinearity, g^l_{vji} indicates a weight assigned to pixel (i,j) in the lth feature map, and $\oplus$ denotes summation along the feature axis $k = (k_1, \dots k_L)$. Note that this formulation incorporates feature-space separability, which reduces overfitting and generally improves prediction accuracy for brain activity[19].

In the Gabor model, the feature extractor consists of a single fixed set of convolutions involving 12 log-spaced spatial frequency Gabor wavelets between 3 and 72 cycles/stimulus and constructed at 6 evenly spaced orientations between 0 and π[19].

Spatial pooling fields

Constraints were placed on the weights (g^l_{vji})—termed 'spatial pooling fields'—that couple the feature maps to voxel activity. For the AlexNet- and Gabor-based encoding models, the spatial pooling field for each voxel was a 2D isotropic Gaussian that was applied to all feature maps (see **Extended Data Figure 10B, middle**). We find that this constrained model of spatial pooling typically yields better prediction accuracy (relative to other possible variants) in the scenario where feature extraction parameters are fixed[19]. For the GNet-based encoding model, the weights of the spatial pooling fields were independently adjustable; hence, we refer to these as flexible spatial pooling fields (see **Extended Data Figure 10B, left**). Feature maps with the same spatial resolution were grouped together, and a distinct, independently optimized spatial pooling field was applied to each group. Thus, the GNet model for each voxel was specified by multiple, independently optimized spatial pooling fields.

Model training and validation

Given the demanding memory requirements of training large-scale neural networks to jointly predict tens of thousands of voxels, we selected the four NSD subjects with the highest noise ceilings (see **Figure 3G**). For the selected subjects (1, 2, 5, 6), NSD betas were extracted from visual areas V1–hV4. These betas were separated into those evoked by the shared1000 images and those that were not; the former were designated as the validation set, while the latter were designated as the training set. For example, for subject 1, there were 9,000 images × 3 trials = 27,000 samples in the training set, and 1,000 images × 3 trials = 3,000 samples in the validation set. After model training, accuracy was quantified as the voxel-wise correlation between model predictions and observed responses in the validation set.

For the AlexNet-based encoding model, parameters of the feature extractors were pre-trained based on classification of objects in the ImageNet database[47]. For both the AlexNet- and Gabor-based encoding models, feature weights for the coupling model were optimized via ridge regression, with the ridge parameter selected to maximize accuracy on a held-out subset (20%) of the training data. Line search was used to optimize the position and size of the Gaussian spatial pooling field for each voxel (see **Extended Data Figure 10B, right**). In total, the AlexNet-based encoding model consisted of 2,692 free parameters per voxel (2,688 feature weighting parameters, 3 spatial pooling parameters, 1 bias term), and the Gabor-based encoding model consisted of 76 free parameters per voxel (72 feature weighting parameters, 3 spatial pooling parameters, 1 bias term).

For the GNet-based encoding model, parameters of the feature extractors, spatial pooling fields, and feature weights were all optimized via stochastic gradient descent of an L_2-norm weighted loss function:

$$L(r_v, \widehat{r_v}) = \frac{\sum_v \lfloor {\rho_v}^2 \rfloor (r_v - \widehat{r_v})^2}{\sum_v \lfloor {\rho_v}^2 \rfloor}$$

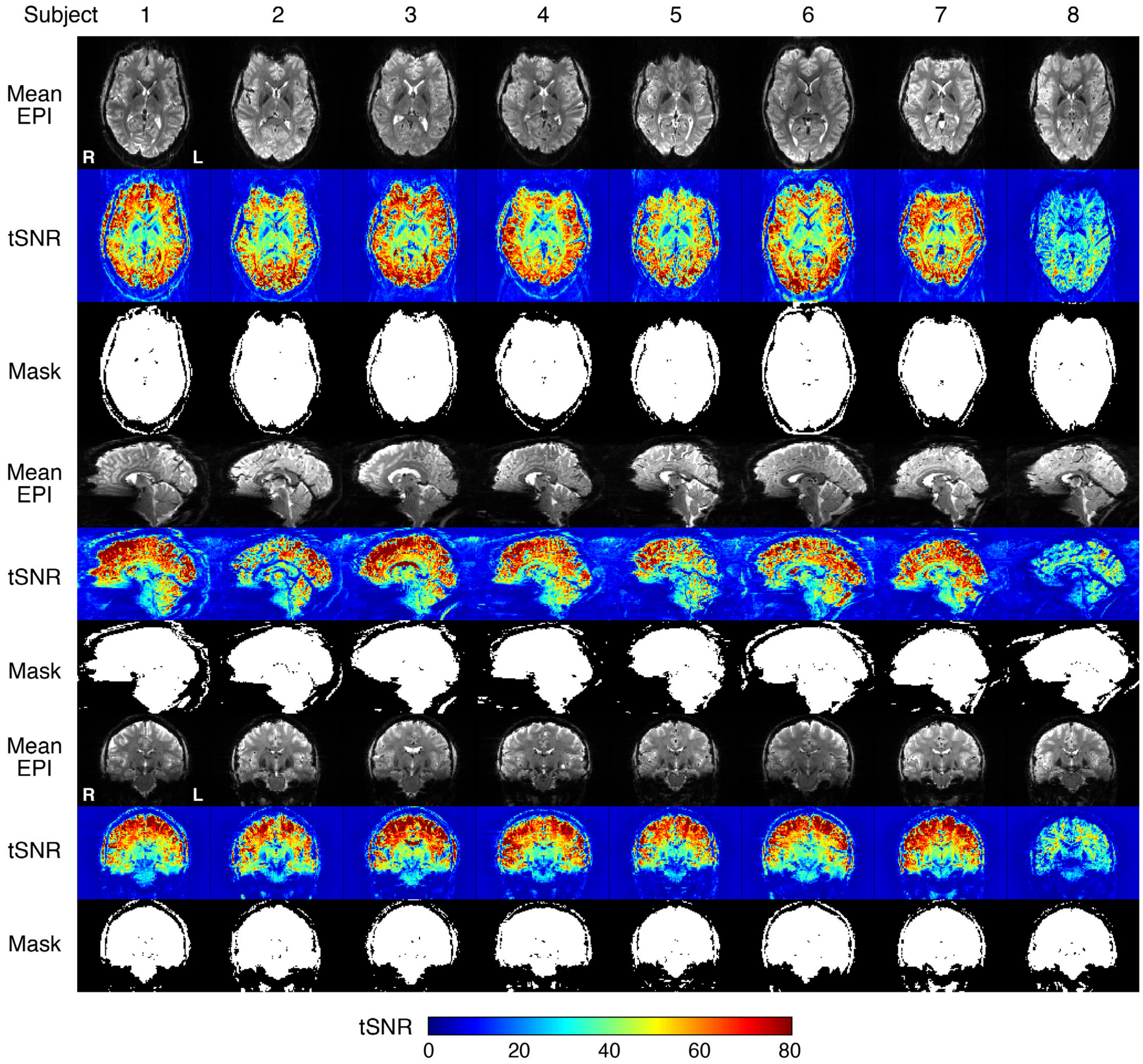
where $\lfloor {\rho_v}^2 \rfloor$ is the batchwise prediction accuracy for a given voxel v with an imposed floor of 0.1 in order to permit contribution of yet-to-be-predicted voxels. In total, the GNet-based encoding model consisted of 1,034,944 free parameters that are shared across voxels, plus 1,307 free parameters per voxel.

Two versions of the GNet model were developed and evaluated. In the single-subject GNet model, different instantiations of GNet were created for different subjects, and only the data from a given subject were used to train the GNet-based encoding model for that subject. In the multiple-subject GNet model, a single instantiation of GNet was created for all four subjects, and data from all subjects were used to train the GNet-based encoding models. In this scheme, all subjects share a common feature extractor, but each subject has independently adjusted coupling models and feature weights.

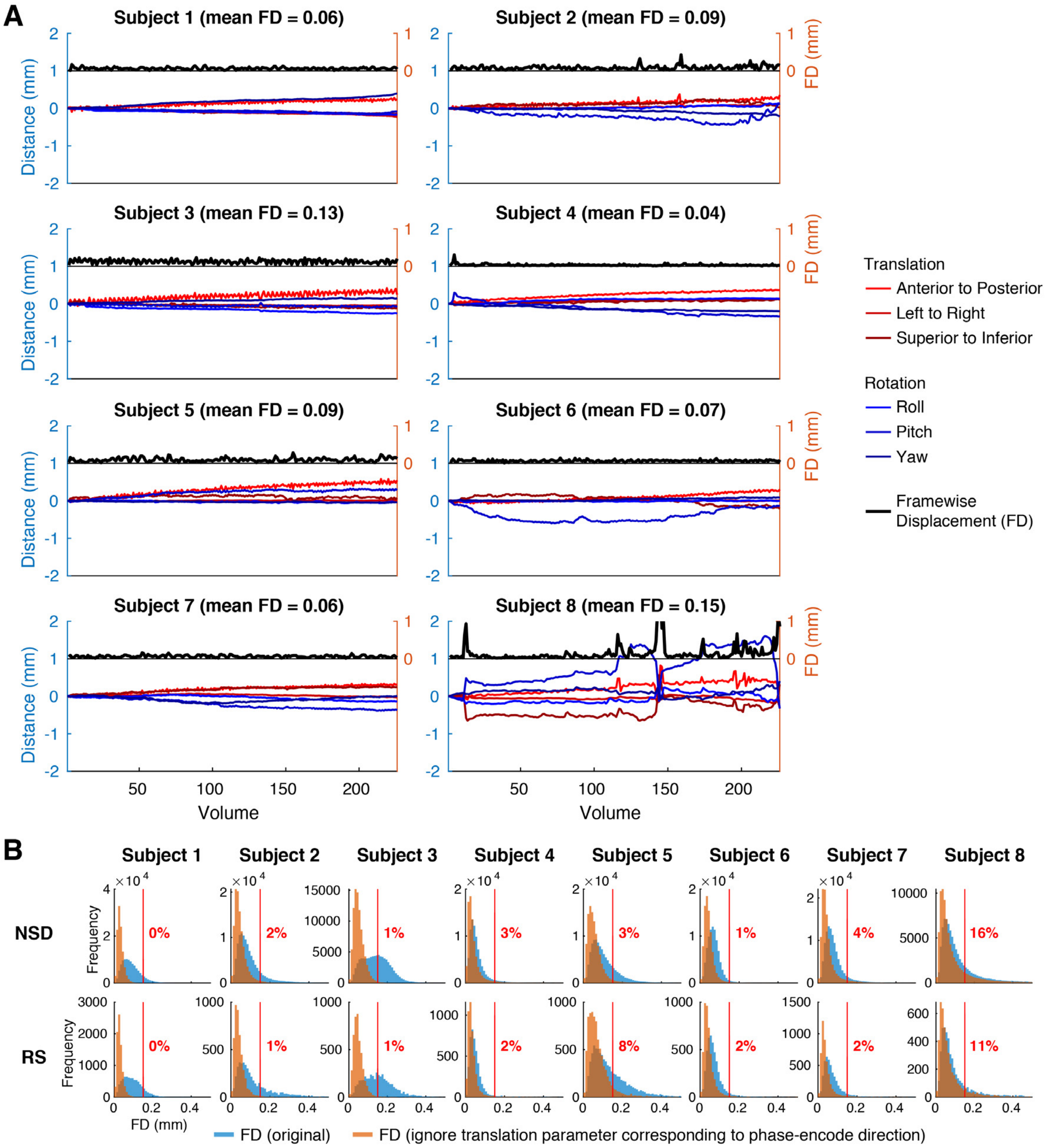
To train the GNet-based encoding model, stochastic gradient descent with early stopping was performed using the ADAM optimizer[126] $(\mathrm{lr} = 10^{-3}, \beta_1 = 0.99, \beta_2 = 0.99)$. Parameter updates for feature extractors, spatial pooling fields, and feature weights were alternated to promote stability of the training procedure.

Note that the NSD subjects view largely non-overlapping sets of images. Thus, when training GNet on data from multiple subjects, we used a modified procedure for selecting batches of training data. For each iteration of training, we first extracted a batch of training samples from one subject's data and calculated the gradient with respect to the loss function. Coupling model parameters for that subject and feature extractor parameters were then updated and the process was repeated until all batches from all subjects were used. This corresponded to one training epoch.

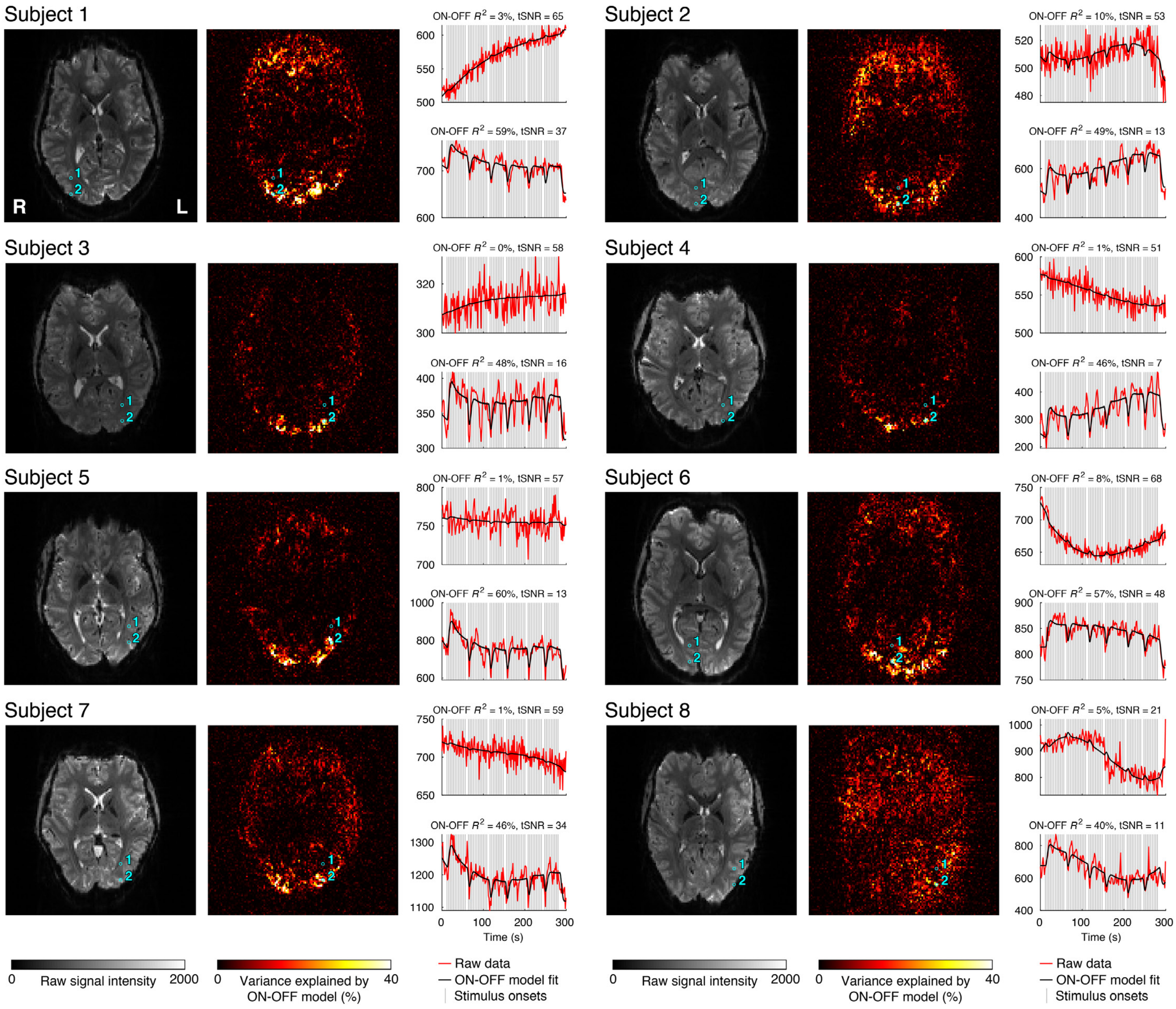
Supplementary Figures



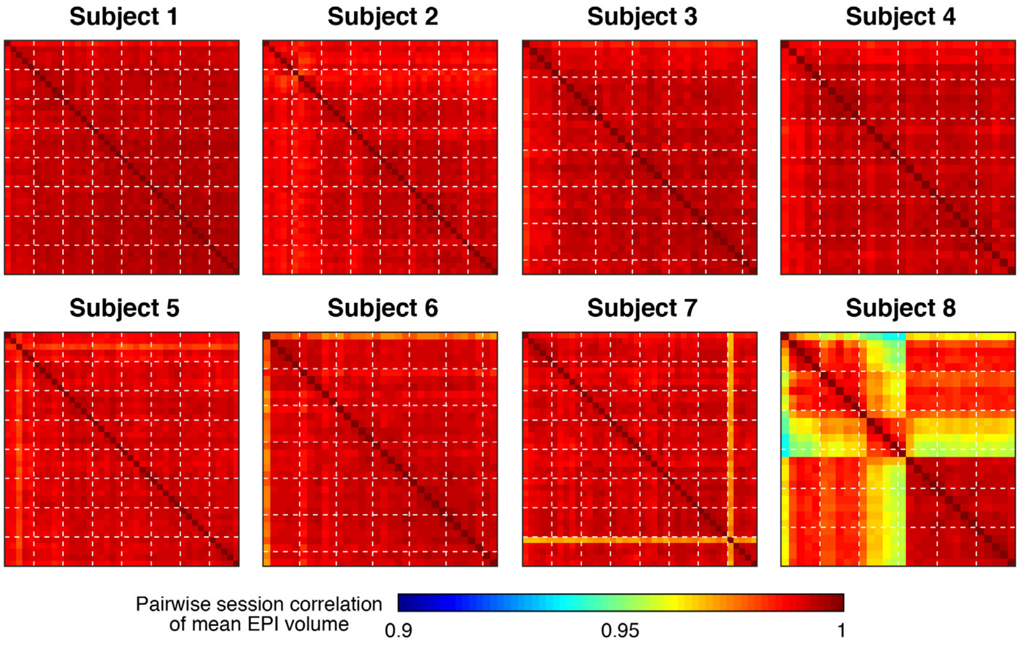
Supplementary Figure 1. Details on the quantification of tSNR. This figure shows example tSNR results (nsd20 scan session, first NSD run). The middle slice in each of three orthogonal views (axial, sagittal, coronal) is displayed. To compute tSNR, the raw fMRI volumes (with no pre-processing) from a given run are obtained, and the mean across volumes is computed (Mean EPI). A brain mask is computed by identifying voxels whose intensity is at least 10% of the 99th percentile of intensities in the mean volume (Mask). tSNR is calculated by quadratically detrending the time-series of each voxel (preserving the mean) and then computing the mean divided by the standard deviation of the time-series values (tSNR). A summary tSNR value is determined by calculating the median tSNR across voxels within the brain mask. This corresponds to the summary metric shown in **Figure 2D, left** (the inset shows results from subject 2).



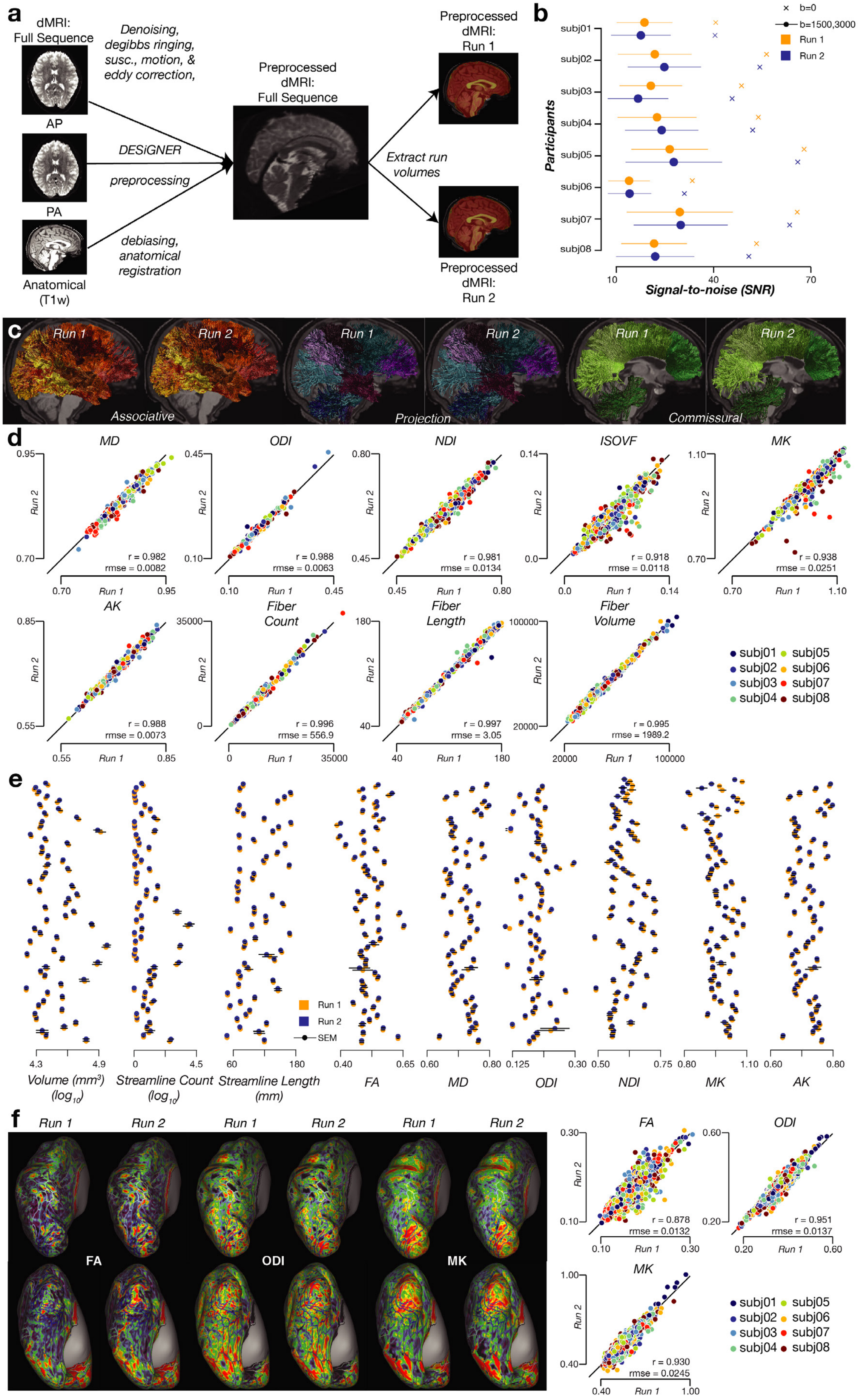
Supplementary Figure 2. Details on the quantification of head motion. *A*, Example motion parameter estimates (nsd20 scan session, first NSD run, 1.8-mm data preparation). The rotation parameters, originally in radian units, are multiplied by 50 in order to allow interpretation in terms of millimeters of displacement for a circle of diameter 100 mm[77]. Motion parameters are relative to the reference volume which is the first volume in each scan session. Framewise displacement (FD), calculated as the sum of the absolute differences of the motion parameters for successive pairs of volumes, is also plotted. The mean FD across volumes is indicated in the plot titles, and corresponds to the summary metric shown in **Figure 2D, middle** (the inset shows results from subject 5). *B*, Distributions of FD. From the 1.8-mm data preparation, we plot histograms of FD observed across all volumes in all NSD runs (top row) and all resting-state (RS) runs (bottom row). Apparent head motion due to the interaction of respiration and the main magnetic field[127,128] is present in the NSD data and can be seen in the anterior-to-posterior translation parameter which corresponds to the EPI phase-encode direction (see panel A, subjects 3 and 5). Thus, in addition to the original FD values (blue histograms), we also plot a modified version of FD (orange histograms) in which we simply omit the anterior-to-posterior translation parameter. This is likely to provide more accurate estimates of actual head motion, though more accurate compensation might be achieved by frequency-based filtering based on the actual respiratory behavior of each subject[127,128]. Using a threshold of 0.15 (vertical red lines), we report the percentage of volumes whose modified FD exceeds this threshold (red inset numbers). Overall, the results indicate that most of the data are free of large head motions (with the possible exception of subject 8) and that the NSD and RS runs have comparable levels of head motion.


Subject 1
R
L
1
2
ON-OFF R^2 = 3%, tSNR = 65
600
550
500
ON-OFF R^2 = 59%, tSNR = 37
700
600
Subject 2
ON-OFF R^2 = 10%, tSNR = 53
520
500
480
ON-OFF R^2 = 49%, tSNR = 13
600
400
Subject 3
ON-OFF R^2 = 0%, tSNR = 58
320
300
ON-OFF R^2 = 48%, tSNR = 16
400
350
300
Subject 4
ON-OFF R^2 = 1%, tSNR = 51
600
550
500
ON-OFF R^2 = 46%, tSNR = 7
400
300
200
Subject 5
ON-OFF R^2 = 1%, tSNR = 57
800
750
700
ON-OFF R^2 = 60%, tSNR = 13
1000
800
600
Subject 6
ON-OFF R^2 = 8%, tSNR = 68
750
700
650
ON-OFF R^2 = 57%, tSNR = 48
900
850
800
750
Subject 7
ON-OFF R^2 = 1%, tSNR = 59
750
700
650
ON-OFF R^2 = 46%, tSNR = 34
1300
1200
1100
0
100
200
300
Time (s)
Subject 8
ON-OFF R^2 = 5%, tSNR = 21
1000
900
800
ON-OFF R^2 = 40%, tSNR = 11
800
600
400
0
100
200
300
Time (s)
0
Raw signal intensity
2000
0
Variance explained by ON-OFF model (%)
40
Raw data
ON-OFF model fit
Stimulus onsets

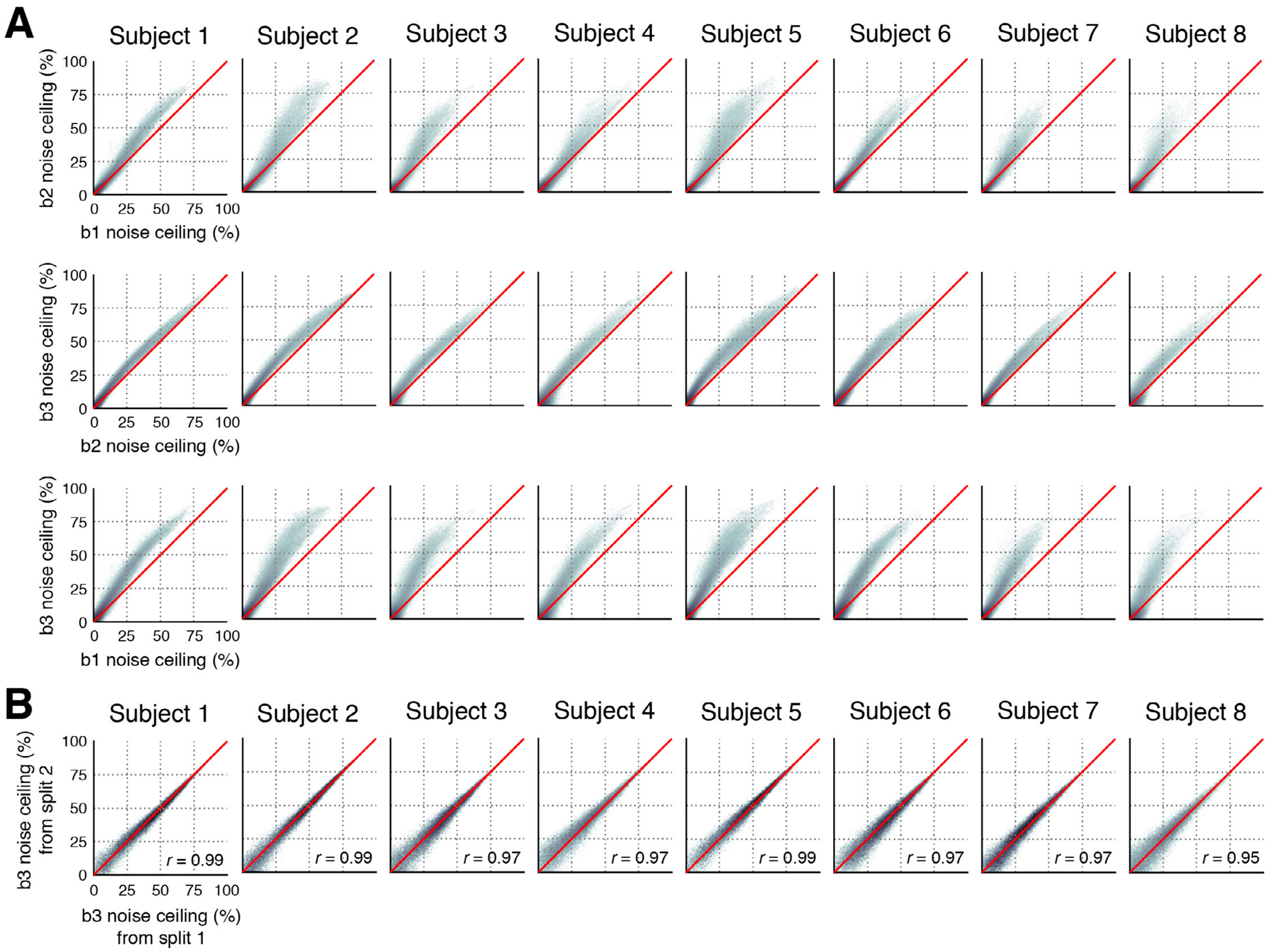
Supplementary Figure 3. Inspection of raw time-series data. To assist interpretation of the raw data shown in this figure, we fit a simple ON-OFF GLM model (see Methods) that assumes a fixed response to each presented image. For each subject, we show results for one axial slice in one NSD run (nsd20 scan session, run01, slice 42 of 84). The image on the left shows raw fMRI data (the first acquired volume in the run). The image in the middle shows the amount of variance explained (R^2) by the ON-OFF model. We select two voxels for detailed inspection: the voxel with the highest R^2 (labeled '2') and a control voxel located 18 mm (a distance of 10 voxels) towards the anterior direction (labeled '1'). The plots on the right show raw time-series data for these two voxels. Thin gray vertical lines mark stimulus onsets. Several observations can be made. First, in each subject, the raw time-series data for the high R^2 voxel (bottom plot) show clear stimulus-evoked signals: the blank trials that occur intermittently during the run lead to decreases in signal intensity that are captured by the ON-OFF model fit. Note that during periods of time involving only stimulus trials, there are in fact small modulations present in the ON-OFF model fit, and these correspond to the onsets of individual images. Second, in each subject, the raw time-series data for the control voxel (top plot) show little discernible stimulus-evoked signals, thereby providing an important comparison. Note that real differences in neural activity evoked by different images are expected to manifest as signal fluctuations in the data, and thus may account for some of the observed time-series fluctuations. Also, note that since motion correction has not been performed for these raw data inspections, it is likely that the observed slow signal drifts are due, in part, to small shifts in head position. Overall, these results confirm the quality of the NSD data by demonstrating that stimulus-evoked signals can be readily observed in the raw time-series data.



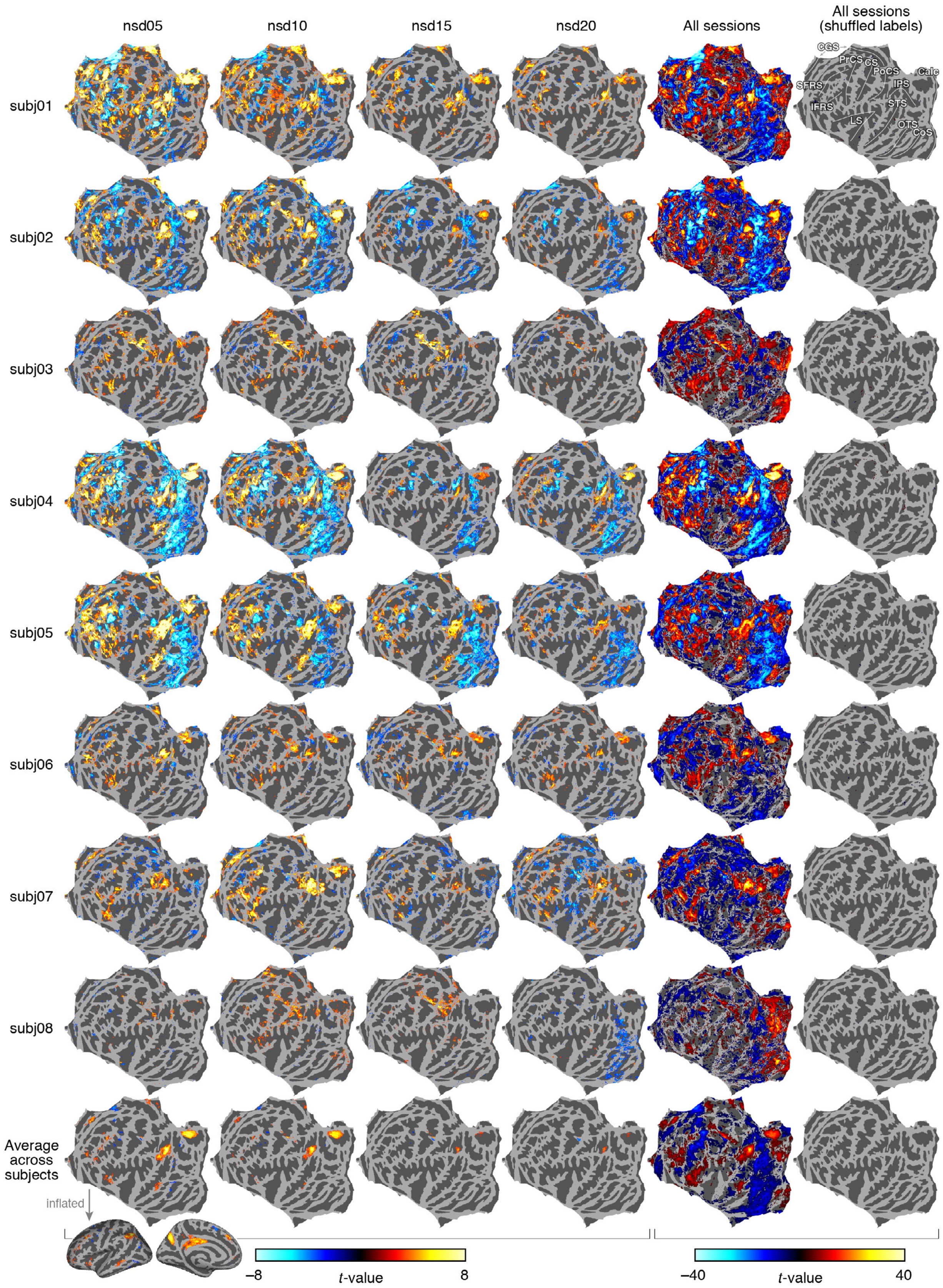
Supplementary Figure 4. Quantification of functional imaging stability. We took the mean fMRI volume (1-mm preparation) in each scan session, bias-corrected the volume by dividing by a fitted 3D polynomial (*autoqc_nsd_grand.m*), and then computed pairwise correlation across sessions. Dotted white lines mark increments of five NSD scan sessions. Inspection of similarity of the mean EPI volume across sessions reveals a few minor anomalies. We investigated these cases further and determined the following: the nsd36 scan session in subject 7 involved a poor scanner shim, which was largely but not fully corrected by the fieldmap-based processing; the nsd01 scan session in subject 8 involved an unusually large amount of head motion, which resulted in some residual spatial distortion; and the nsd12–nsd16 scan sessions in subject 8 involved a temporary sinus infection near frontal cortex that manifested as bright signal intensities in the EPI volumes outside the brain but otherwise did not cause any data problems. For visual inspection of these effects, see **Supplementary Videos 6–7**.


a
dMRI:
Full Sequence
AP
PA
Anatomical
(T1w)
Denoising,
degibbs ringing,
susc., motion, &
eddy correction,
DESiGNER
preprocessing
debiasing,
anatomical
registration
Preprocessed
dMRI:
Full Sequence
Extract run
volumes
Preprocessed
dMRI:
Run 1
Preprocessed
dMRI:
Run 2
b
× b=0
b=1500,3000
Run 1
Run 2
subj01
subj02
subj03
subj04
subj05
subj06
subj07
subj08
Participants
10
40
70
Signal-to-noise (SNR)
c
Run 1
Run 2
Run 1
Run 2
Run 1
Run 2
Associative
Projection
Commissural
d
MD
0.95
0.70
Run 2
Run 1
0.70
0.95
r = 0.982
rmse = 0.0082
ODI
0.45
0.10
Run 2
Run 1
0.10
0.45
r = 0.988
rmse = 0.0063
NDI
0.80
0.45
Run 2
Run 1
0.45
0.80
r = 0.981
rmse = 0.0134
ISOVF
0.14
0.0
Run 2
Run 1
0.0
0.14
r = 0.918
rmse = 0.0118
MK
1.10
0.70
Run 2
Run 1
0.70
1.10
r = 0.938
rmse = 0.0251
AK
0.85
0.55
Run 2
Run 1
0.55
0.85
r = 0.988
rmse = 0.0073
Fiber
Count
35000
0
Run 2
Run 1
0
35000
r = 0.996
rmse = 556.9
Fiber
Length
180
40
Run 2
Run 1
40
180
r = 0.997
rmse = 3.05
Fiber
Volume
100000
20000
Run 2
Run 1
20000
100000
r = 0.995
rmse = 1989.2
subj01
subj02
subj03
subj04
subj05
subj06
subj07
subj08
e
Run 1
Run 2
SEM
4.3
4.9
Volume (mm^3)
(log$_{10}$)
0
4.5
Streamline Count
(log$_{10}$)
60
180
Streamline Length
(mm)
0.40
0.65
FA
0.60
0.80
MD
0.125
0.30
ODI
0.50
0.75
NDI
0.80
1.10
MK
0.60
0.80
AK
f
Run 1
Run 2
Run 1
Run 2
Run 1
Run 2
FA
ODI
MK
FA
0.30
0.10
Run 2
Run 1
0.10
0.30
r = 0.878
rmse = 0.0132
ODI
0.60
0.20
Run 1
0.20
0.60
r = 0.951
rmse = 0.0137
MK
1.00
0.40
Run 2
Run 1
0.40
1.00
r = 0.930
rmse = 0.0245
subj01
subj02
subj03
subj04
subj05
subj06
subj07
subj08

Supplementary Figure 5. Diffusion processing for investigation of white-matter connectivity. *A*, Schematic of the diffusion pre-processing pipeline. Diffusion volumes were corrected for noise, Gibbs ringing, susceptibility, motion, eddy currents, and bias fields before being co-registered to the T_1 anatomy. Following pre-processing, the data were organized into two runs (corresponding to the 99-direction and 100-direction scans, respectively). *B*, Signal-to-noise ratio, computed in the corpus callosum (dots and error bars indicate mean and standard deviation across volumes, respectively; for Runs 1 and 2, error bars reflect 184 and 186 volumes, respectively). *C*, White-matter tract segmentation from an example subject (subject 7). White-matter tracts are organized based on typical anatomical and functional definitions into associative (left), projection (middle), and callosal (right) tracts and overlaid on the T_1 anatomy. *D*, Reliability of MD, ODI, NDI, ISOVF, Mean Kurtosis (MK), Axial Kurtosis (AK), and fiber count, length, and volume. Each dot indicates results averaged along a single tract. Pearson's correlation (*r*) and root-mean-squared error (RMSE) for each measure are indicated in the inset. *E*, Macrostructural and microstructural properties observed for different tracts. Error bars indicate ± 1 SEM across 8 subjects. *F*, Microstructural properties of cortical regions. Shown are tensor (FA; left), NODDI (ODI; middle), and kurtosis (MK; right) results mapped to the cortical surface of the example subject, with dorsal (top) and ventral (bottom) viewpoints of occipital cortex. Quantitative results are shown on the right, where each dot indicates results obtained for a single region in the HCP-MMP1 parcellation.

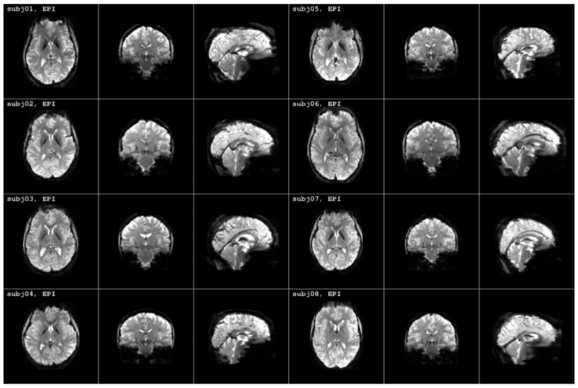


Supplementary Figure 6. Additional details regarding noise ceiling estimates. This figure provides additional detail on noise ceiling results shown in **Figure 3F–G**. All results reflect vertices within the nsdgeneral ROI. *A*, Detailed comparison of noise ceiling results for different beta versions. Each subplot is a 2D histogram comparing noise ceilings for two different beta versions. Improvements in noise ceilings are consistent across voxels and subjects. *B*, Reliability of noise ceiling estimates. Here we show split-half noise ceiling estimates for beta version 3. Each subplot is a 2D histogram comparing noise ceiling estimates calculated from two halves of the data from a given subject. The inset indicates the correlation between the two sets of estimates. Noise ceiling estimates are highly stable owing to the large number of images that inform the noise ceiling estimates.

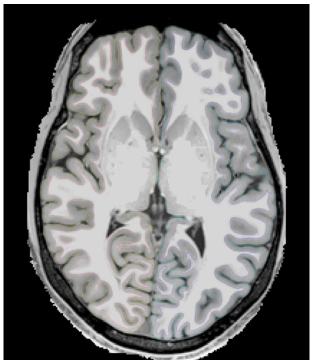


Supplementary Figure 7. Recognition memory effects in the NSD data. Same format as **Figure 4B**, but showing results for all individual subjects. Positive values indicate BOLD responses are greater for hits than for correct rejections, whereas negative values indicate BOLD responses are greater for correct rejections than for hits. The observed decrease in the magnitudes of the *t*-values (e.g. from nsd05 to nsd20) likely reflects a decrease in the subjects' recognition accuracy over the course of the experiment.

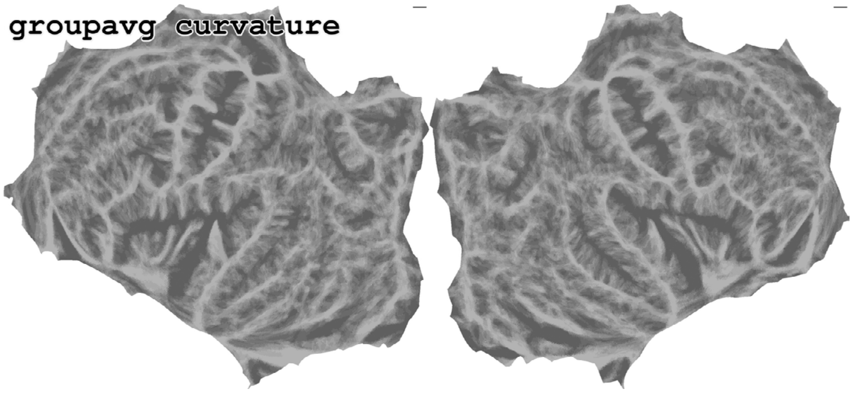
Supplementary Videos



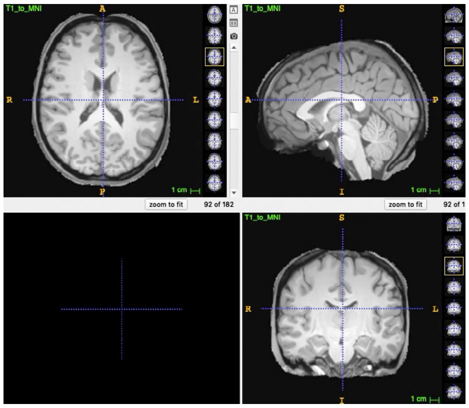
Supplementary Video 1. Inspection of image quality and co-registration quality. Videos available online (https://osf.io/tg5dw/ - T1-T2-EPI.mp4), https://osf.io/g86ep/ - T2-SWI.mp4, https://osf.io/s7b2a/ - T1-TOF.mp4). Three videos are provided. One video cycles between the T_1, T_2, and EPI volumes, another cycles between the T_2 and SWI volumes, and the third cycles between the T_1 and TOF volumes. All volumes have been transformed to a common anatomical space (set by the T_1 volume) in the course of data pre-processing.



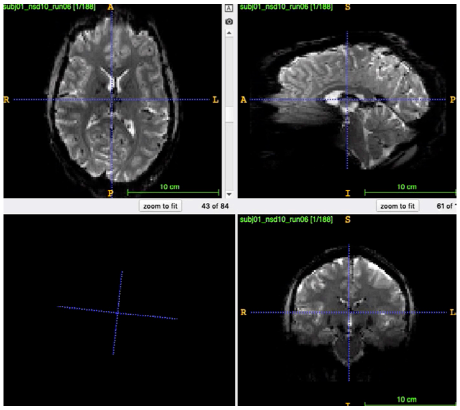
Supplementary Video 2. Inspection of cortical surfaces. Videos available online (https://osf.io/zyb3t/ - subj{01–08}_{axial,coronal,sagittal}.mp4). These videos show the FreeSurfer cortical surface reconstructions superimposed on the T_1 volume. Left hemisphere white and pial surfaces are colored blue and cyan, respectively; right hemisphere white and pial surfaces are colored red and yellow, respectively. Blue voxels indicate locations that have been judged to have surface imperfections.



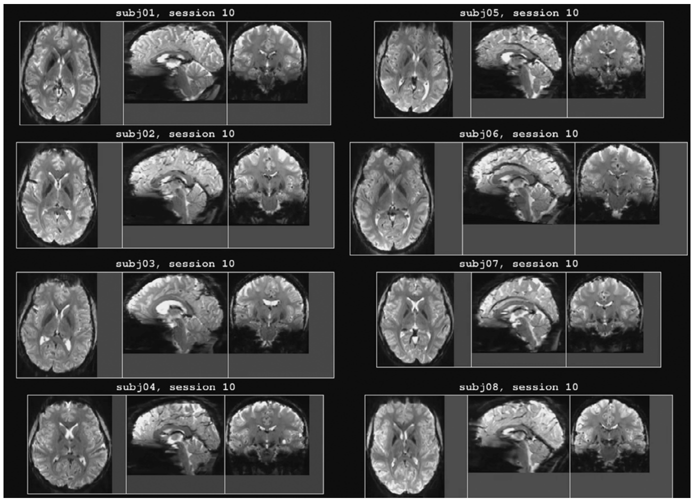
Supplementary Video 3. Inspection of fsaverage alignment. Video available online (https://osf.io/gh5bs/ - fsaveragecheck.mp4). This video cycles through (i) the binarized curvature of each of the NSD subjects mapped via nearest-neighbor interpolation to *fsaverage*, (ii) the average of this binarized curvature across subjects, and (iii) the *fsaverage* binarized curvature. The video is useful for assessing the quality of the folding-based alignment performed by FreeSurfer. Notice that the group-average curvature resembles the *fsaverage* curvature, as expected.



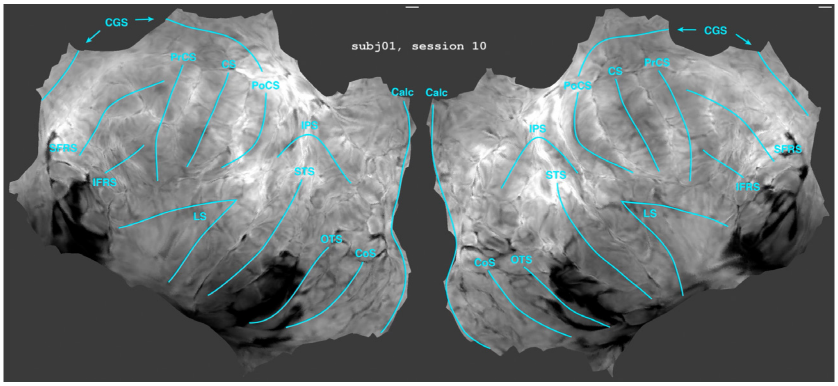
Supplementary Video 4. Inspection of MNI alignment. Video available online (https://osf.io/p3zqm/ - MNIcheck.mp4). This video cycles through the T_1 volumes of the NSD subjects after nonlinear warping to MNI space and the MNI template volume. The video is useful for assessing the quality of the nonlinear volume-based alignment.


subj01_nsd10_run06 [1/188]
A
R
L
P
10 cm
zoom to fit
43 of 84
subj01_nsd10_run06 [1/188]
S
A
P
I
10 cm
zoom to fit
61 of
subj01_nsd10_run06 [1/188]
S
R
L
I
10 cm

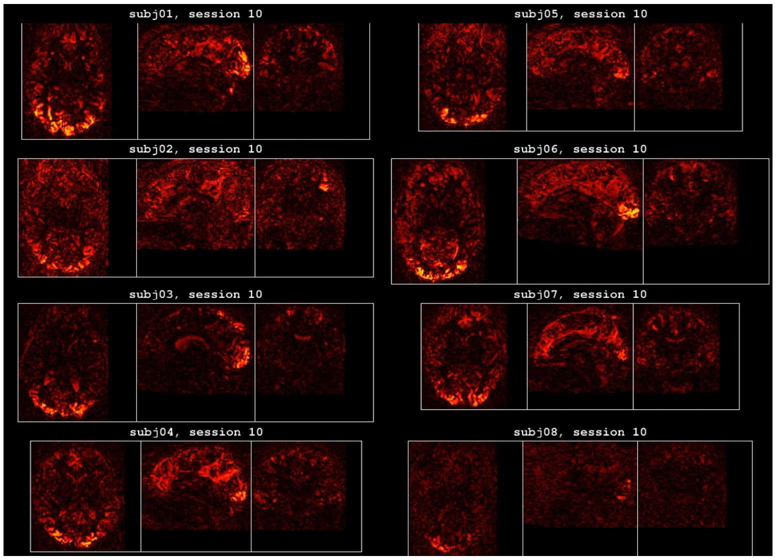
Supplementary Video 5. Inspection of raw and pre-processed EPI volumes. Videos available online (https://osf.io/zyb3t/ - subj{01–08}_nsd10_run06_{raw,pp}.mp4). These videos quickly scroll through all EPI volumes in a sample run. This is useful for assessing quality and stability of the functional imaging.



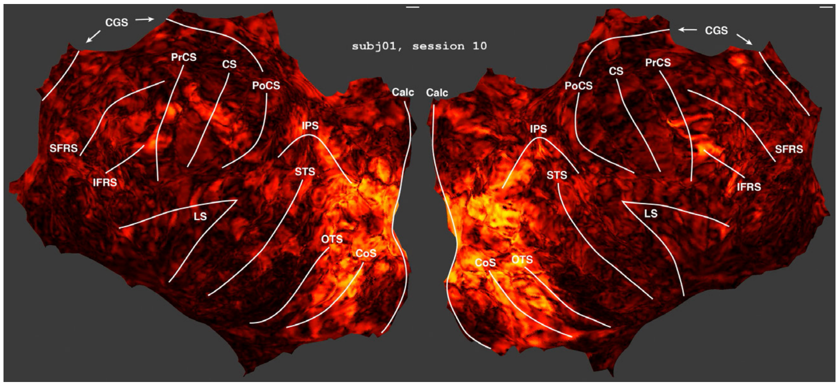
Supplementary Video 6. Inspection of mean EPI across scan sessions (volume visualization). Video available online (https://osf.io/ydf9j/ - grandmean.mp4). This video assesses the results of pre-processing the fMRI data. Each frame shows the mean EPI volume from a single scan session (1-mm data preparation). Note that session 0 corresponds to the prffloc scan session and the last two scan sessions from each subject correspond to the nsdsynthetic and nsdimagery scan sessions. This video is useful for assessing overall image quality and the stability of functional imaging across scan sessions. For quantitative analysis, see **Supplementary Figure 4**.



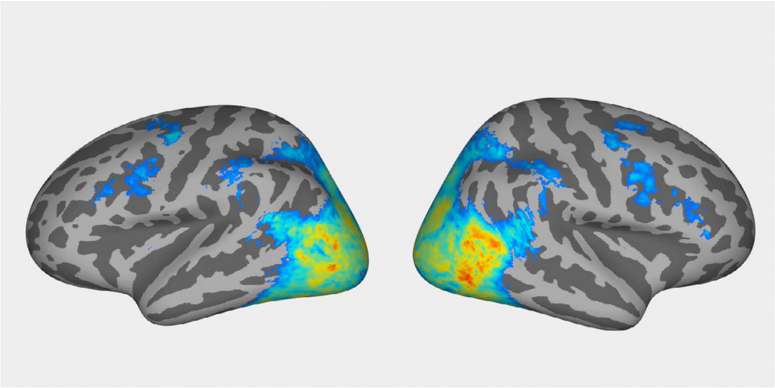
Supplementary Video 7. Inspection of mean EPI across scan sessions (surface visualization). Video available online (https://osf.io/ytjk4/ - grandmeansurface.mp4). This is similar in spirit to **Supplementary Video 6**, except that the mean EPI volumes have been projected onto each subject's cortical surface and then transferred to the *fsaverage* surface.



Supplementary Video 8. Inspection of BOLD signal strength across scan sessions (volume visualization). Video available online (https://osf.io/kwxta/ - grandR2.mp4). Each frame shows the amount of variance explained by the ON-OFF GLM model (1-mm data preparation; fixed color range). This video is useful for assessing the overall strength and stability of BOLD responses in the NSD dataset.



Supplementary Video 9. Inspection of BOLD signal strength across scan sessions (surface visualization). Video available online (https://osf.io/gu9wx/ - grandR2surface.mp4). This is similar in spirit to **Supplementary Video 8**, except that the variance explained volumes have been projected onto each subject's cortical surface and then transferred to the *fsaverage* surface.



Supplementary Video 10. Inflated surface visualization of noise ceilings. Video available online (https://osf.io/z3wxn/ - b3noiseceiling.mp4). This video shows the group-average b3 noise ceiling (see **Figure 3F**) on a rotating, inflated *fsaverage* surface. Values below 15% are thresholded away in order to show the underlying curvature. This video is useful for identifying brain regions whose activity is strongly related to the sensory content presented in the NSD experiment.

Supplementary References

88. Laumann, T. O. *et al.* Functional System and Areal Organization of a Highly Sampled Individual Human Brain. *Neuron* **87**, 657–670 (2015).
89. Chen, J. E. *et al.* Resting-state “physiological networks”. *NeuroImage* **213**, 116707 (2020).
90. Lynch, C. J. *et al.* Prevalent and sex-biased breathing patterns modify functional connectivity MRI in young adults. *Nature Communications* **11**, 5290 (2020).
91. Ades-Aron, B. *et al.* Evaluation of the accuracy and precision of the diffusion parameter EStImation with Gibbs and NoisE removal pipeline. *Neuroimage* **183**, 532–543 (2018).
92. McPherson, B. C. & Pestilli, F. A single mode of population covariation associates brain networks structure and behavior and predicts individual subjects’ age. *Commun Biol* **4**, 1–16 (2021).
93. Pestilli, F., Yeatman, J. D., Rokem, A., Kay, K. N. & Wandell, B. A. Evaluation and statistical inference for human connectomes. *Nature Methods* **11**, 1058–1063 (2014).
94. Daducci, A. *et al.* Accelerated Microstructure Imaging via Convex Optimization (AMICO) from diffusion MRI data. *NeuroImage* **105**, 32–44 (2015).
95. Jensen, J. H. & Helpern, J. A. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. *NMR in Biomedicine* **23**, 698–710 (2010).
96. Pierpaoli, C., Jezzard, P., Basser, P. J., Barnett, A. & Di Chiro, G. Diffusion tensor MR imaging of the human brain. *Radiology* **201**, 637–648 (1996).
97. Tournier, J.-D., Calamante, F. & Connelly, A. Robust determination of the fibre orientation distribution in diffusion MRI: Non-negativity constrained super-resolved spherical deconvolution. *NeuroImage* **35**, 1459–1472 (2007).
98. Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage* **61**, 1000–1016 (2012).
99. Smith, R. E., Tournier, J.-D., Calamante, F. & Connelly, A. Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information. *NeuroImage* **62**, 1924–1938 (2012).
100. Yeatman, J. D., Dougherty, R. F., Myall, N. J., Wandell, B. A. & Feldman, H. M. Tract Profiles of White Matter Properties: Automating Fiber-Tract Quantification. *PLOS ONE* **7**, e49790 (2012).
101. Hagmann, P. *et al.* Mapping the Structural Core of Human Cerebral Cortex. *PLOS Biology* **6**, e159 (2008).
102. Bassett, D. S. & Sporns, O. Network neuroscience. *Nature Neuroscience* **20**, 353–364 (2017).
103. Rokem, A. *et al.* The visual white matter: The application of diffusion MRI and fiber tractography to vision science. *Journal of Vision* **17**, 4–4 (2017).
104. Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J. & van Lier, R. Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience. *Sci Rep* **9**, 17456 (2019).
105. Frey, M., Nau, M. & Doeller, C. F. MR-based camera-less eye tracking using deep neural networks. *bioRxiv* 2020.11.30.401323 (2020) doi:10.1101/2020.11.30.401323.
106. Son, J. *et al.* Evaluating fMRI-Based Estimation of Eye Gaze During Naturalistic Viewing. *Cerebral Cortex* **30**, 1171–1184 (2020).
107. Avesani, P. *et al.* The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services. *Scientific Data* **6**, 69 (2019).
108. Tournier, J.-D. *et al.* MRtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *NeuroImage* **202**, 116137 (2019).
109. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
110. Garyfallidis, E. *et al.* Dipy, a library for the analysis of diffusion MRI data. *Front. Neuroinform.* **8**, (2014).
111. Fukutomi, H. *et al.* Neurite imaging reveals microstructural variations in human cerebral cortical gray matter. *NeuroImage* **182**, 488–499 (2018).
112. Tournier, J.-D., Calamante, F. & Connelly, A. Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. *NeuroImage* **35**, 1459–1472 (2007).
113. Smith, R. E., Tournier, J.-D., Calamante, F. & Connelly, A. SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *NeuroImage* **119**, 338–351 (2015).
114. Bullock, D. *et al.* Associative white matter connecting the dorsal and ventral posterior human cortex. *Brain Struct Funct* **224**, 2631–2660 (2019).
115. Jezzard, P. Correction of geometric distortion in fMRI data. *NeuroImage* **62**, 648–651 (2012).
116. Gratton, C. *et al.* Removal of high frequency contamination from motion estimates in single-band fMRI saves data without biasing functional connectivity. *Neuroimage* **217**, 116866 (2020).
117. Charest, I., Kriegeskorte, N. & Kay, K. N. GLMdenoise improves multivariate pattern analysis of fMRI data. *NeuroImage* **183**, 606–616 (2018).
118. Arcaro, M. J., Pinsk, M. A. & Kastner, S. The Anatomical and Functional Organization of the Human Visual Pulvinar. *J. Neurosci.* **35**, 9848–9871 (2015).
119. Berron, D. *et al.* A protocol for manual segmentation of medial temporal lobe subregions in 7 Tesla MRI. *Neuroimage Clin* **15**, 466–482 (2017).

120. Winawer, J. & Witthoft, N. Identification of the ventral occipital visual field maps in the human brain. *F1000Res* **6**, 1526 (2017).

121. Gomez, J., Natu, V., Jeska, B., Barnett, M. & Grill-Spector, K. Development differentially sculpts receptive fields across early and high-level human visual cortex. *Nature Communications* **9**, 788 (2018).

122. Kay, K. N. & Yeatman, J. D. Bottom-up and top-down computations in word- and face-selective cortex. *Elife* **6**, e22341 (2017).

123. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic Maps of Visual Topography in Human Cortex. *Cereb. Cortex* **25**, 3911–3931 (2015).

124. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

125. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi:10.1109/CVPR.2009.5206848.

126. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017).

127. Fair, D. A. *et al.* Correction of respiratory artifacts in MRI head motion estimates. *NeuroImage* **208**, 116400 (2020).

128. Power, J. D. *et al.* Distinctions among real and apparent respiratory motions in human fMRI data. *Neuroimage* **201**, 116041 (2019).