High-level visual areas act like domain-general filters with strong selectivity and functional specialization

Meenakshi Khosla[1] and Leila Wehbe[2,3]

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139
[2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213
[3]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213

March 16, 2022

Abstract

Investigation of the visual system has mainly relied on a-priori hypotheses to restrict experimental stimuli or models used to analyze experimental data. Hypotheses are an essential part of scientific inquiry, but an exclusively hypothesis-driven approach might lead to confirmation bias towards existing theories and away from novel discoveries not predicted by them. This paper uses a hypothesis-neutral computational approach to study four high-level visual regions of interest (ROIs) selective to faces, places, letters, or body parts. We leverage the unprecedented scale and quality of the Natural Scenes Dataset to constrain neural network models of these ROIs with functional Magnetic Resonance Imaging (fMRI) measurements. We show that using only the stimulus images and the associated activity in an ROI, we are able to train from scratch a neural network that can predict the activity in each voxel of that ROI with an accuracy that beats state-of-the-art models. Moreover, once trained, the ROI-specific networks can reveal what kinds of functional properties emerge spontaneously in their training. Strikingly, despite no category-level supervision, the units in the trained networks act strongly as detectors for semantic concepts like 'faces' or 'words', thereby providing substantial pieces of evidence for categorical selectivity in these visual areas. Importantly, this selectivity is maintained when training the networks with selective deprivations in the training diet, by excluding images that contain their preferred category. The resulting selectivity in the trained networks strongly suggests that the visual areas do not function as exclusive category detectors but are also sensitive to visual patterns that are typical to their preferred categories, even in the absence of these categories. Finally, we show that our response-optimized networks have distinct functional properties. Together, our findings suggest that response-optimized models combined with model interpretability techniques can serve as a powerful and unifying computational framework for probing the nature of representations and computations in the brain.

1 Introduction

Large strides have been made in understanding early visual cortex by presenting model organisms with abstract stimuli like oriented edges or sine-wave gratings and studying the properties of evoked neuronal responses [1, 2]. However, detailed characterization of neuronal responses in high-level areas has been more difficult, partly because it requires determining what stimuli to present to probe the response properties of these areas. This is a circular problem: understanding what an area encodes requires presenting the optimal stimulus but the selectivity of the visual area remains hidden until the optimal pattern is presented [3].

One solution is to hypothesize the stimulus attributes represented in a region and then design experiments with carefully selected stimuli to test the hypothesis. Indeed, with this deductive approach, several category-selective regions have been reliably identified within the human ventral temporal cortex (VTC) and macaque inferotemporal cortex (IT), including regions responding selectively to faces [4, 5, 6], places [7, 8, 9], bodies [10, 11, 12, 6], tools [13], words [14, 15], and other categories [16]. However, this approach is not exhaustive or scalable; response properties within large swathes of the sensory cortex remain elusive.

An alternative approach is the use of naturalistic images and videos to drive activity in the visual cortex, followed by the use of encoding models to test different hypotheses about voxel-level tuning [17, 18, 19, 20]. This inductive system identification approach decouples data collection from hypothesis testing, allowing multiple candidate models to be tested, post-hoc, on the same dataset. Models are constructed to test a specific hypothesis about brain function, and are typically adjudicated based on their prediction accuracy on held-out datasets [21]. Therefore, this approach still relies on the prior specification of competing hypotheses and their formulation in terms of explicit quantitative functional forms (e.g. low-level gabor wavelet pyramid, motion-energy pyramid, semantic encoding models etc.).

Recently, representations extracted from deep neural networks have set new standards for predicting neural responses along the ventral visual pathway in humans and non-human primates [22, 23]. Optimizing for tasks like object recognition can lead to the emergence of representations that accurately predict the ventral visual pathway, with different tasks leading to different alignment with individual brain regions [24]. This suggests that artificial and biological networks could share computational goals, offering a new way to test computational hypotheses about the brain.

All the approaches above rely on strong underlying hypotheses. In the deductive approach, the hypothesis governs the selection of the narrow range of stimuli. In the inductive system identification approach, hypotheses are specified via an encoding model. When an encoding model uses representations from task-optimized deep neural networks, the hypothesis is specified by the network task. In contrast, hypothesis-neutral approaches with minimal a priori assumptions are likely to be flexible and effective in revealing tuning properties throughout the visual system. One such approach is response optimization, i.e., fitting model parameters to reproduce the brain response related to stimulus directly. When successful in predicting new responses and generalizing to unseen contexts, the response-optimization approach can facilitate the discovery of unknown neuronal tuning properties and provide strong tests for existing theories.

To date, the amount of data available to fit response-optimized models was often deemed insufficient. However, recent advances in large-scale data collection present an opportunity

to change this status-quo. Indeed, several recent studies have successfully built response-optimized models of early visual cortex [25, 26, 27, 28]. However, it is unclear how such an approach generalizes to high-level visual areas and what it might reveal about them.

We adopt a hypothesis-neutral approach and systematically characterize selectivity in four human higher-level visual regions of interest (ROIs) via response-optimized models. We focus on the fusiform face area (FFA) [4], the extrastriate body area (EBA) [10], the visual word form area (VWFA) [14], and the retrosplenial cortex (RSC, a visual place selective area) [29]. First, we leverage the unprecedented scale and quality of the massive 7T fMRI Natural Scenes Dataset (NSD) [27] to train a deep neural network model to predict activity of voxels in each ROI. Each network is randomly initialized (not pre-trained on any task) and is optimized to predict the activity of all voxels in the ROI from the stimulus image. Our models achieve high prediction performance, on par or outperforming state-of-the-art task-optimized models. Then, we ask what kinds of functional properties emerge spontaneously in our response-optimized models. We examine the trained networks through structural analysis (feature *visualizations*) as well as functional analysis (feature *verbalization*) by running high-throughput experiments with these models on large-scale probe datasets and dissecting the evoked network activations [30, 31]. Strikingly, despite no category-level supervision (since the networks are solely optimized for brain response prediction), the units (neurons) in the optimized networks act as detectors for high-level visual concepts like 'faces' (in the FFA model) or 'words' (in the VWFA model), thereby providing one of the strongest evidences for categorical selectivity in these ROIs till date.

The observed strong semantic selectivity in model neurons raises another important question: are the ROIs simply functioning as processing units for their preferred category (e.g., simply deciding whether the stimulus contains a face or a word) or are they a by-product of a non-category-specific visual processing mechanism? To probe this, we create selective deprivations in the visual diet of these response-optimized networks and study the selectivity of model neurons in the resulting 'deprived' networks. We find that the resulting models still demonstrate high selectivity for the preferred category, even in the absence of any experience with the preferred category. This suggests that the same "filters" used to encode the preferred category are used to encode other non-specific natural images as well. The results presented here further indicate that category-selective voxels do not respond to their preferred category by the unique structure of that category but perhaps by some other general structural characteristics that other stimuli may share with the preferred category.

Beyond characterizing tuning properties of individual voxels, we further demonstrate that the proposed models generalize remarkably and selectively to different perceptual tasks: representations from the model of the fusiform face area (FFA) can predict facial identity while those from the retrosplenial cortex (RSC, a visual place selective area) can discriminate between spatial layouts of different indoor scenes, revealing important functional distinctions between different ROIs. Together with this new class of data-driven models for higher order visual areas and novel model interpretability techniques, our study illustrates that response-driven deep neural network models of visual cortex can serve as powerful and unifying tools for probing the nature of representations and computations in the brain.

Results

We train deep neural network models directly to predict the brain activity related to viewing natural images. We break from the current computational neuroscience approach of extracting image representations from networks previously optimized on large image databases, with tasks such as object classification. Instead, we optimize a network starting from scratch to directly predict the recorded activity in the voxels of a given high-level ROI. The network therefore learns to represent images in a way that is optimal for predicting voxel activity in that ROI, capturing the important dimensions of variance and the tuning of the voxels. We train a separate model for each of four ROIs: FFA, EBA, VWFA and RSC. We capitalize on the natural variation in the rich Natural Scenes Dataset (NSD) [27] to train these models directly on stimulus-response pairs from a wide range of naturalistic scenarios. Notably, the dataset contains complex and sometimes crowded images of various everyday objects in their natural contexts at varied viewpoints. The stimulus set is thus more typical of real-world vision and allows us to characterize neural representations and computations in ethological conditions.

Accurate predictions on complex, cluttered scenes; rapid generalization to new subjects

Our models learn deep convolutional feature spaces that are shared across thousands of voxels over multiple subjects. We utilize a rotation-equivariant Convolutional Neural Network (CNN) architecture to learn these feature spaces directly from fMRI data. This architectural design choice enables the model to learn identical features at multiple orientations and spatial locations, mimicking the response properties of neurons in early and intermediate visual areas, which are known to capture similar features like edges and curves but at different orientations and locations in the visual field [32, 33, 34, 35]. We employ a linear *readout* model on top of this feature space to predict the responses of individual voxels in the ROIs. The linear readout is *factorized* into *spatial* and *feature* dimensions following popular methods for neural system identification in mouse visual cortex [36]. This factorization separates receptive field location (i.e., what portion of the visual space is the voxel most sensitive to?) from feature tuning (i.e. what features of the visual input is the voxel sensitive to?). Sharing the entire representational network across subjects, sharing convolutional filters weights across visual field locations (translation equivariance) and orientations (rotation equivariance) and a factorized readout jointly enable the sample-efficient training of the response-optimized models. Our baseline is a task-optimized model trained to perform object classification on the large-scale ImageNet dataset (see Fig.1 and the Methods section for details). To assess the prediction performance of this model, we used the standard methodology of modeling the voxel response as a linear weighting of the task-optimized network's output units (from the best-performing layer that is determined using a separate validation set). Further, just as in our response-optimized models, this linear mapping is factorized into spatial and feature dimensions as this was found to significantly improve performance over the traditional non-sparse readout method.

We compared the performance of response-optimized and task-optimized models (see fig.1). The models are trained jointly on four subjects and their performance is estimated on

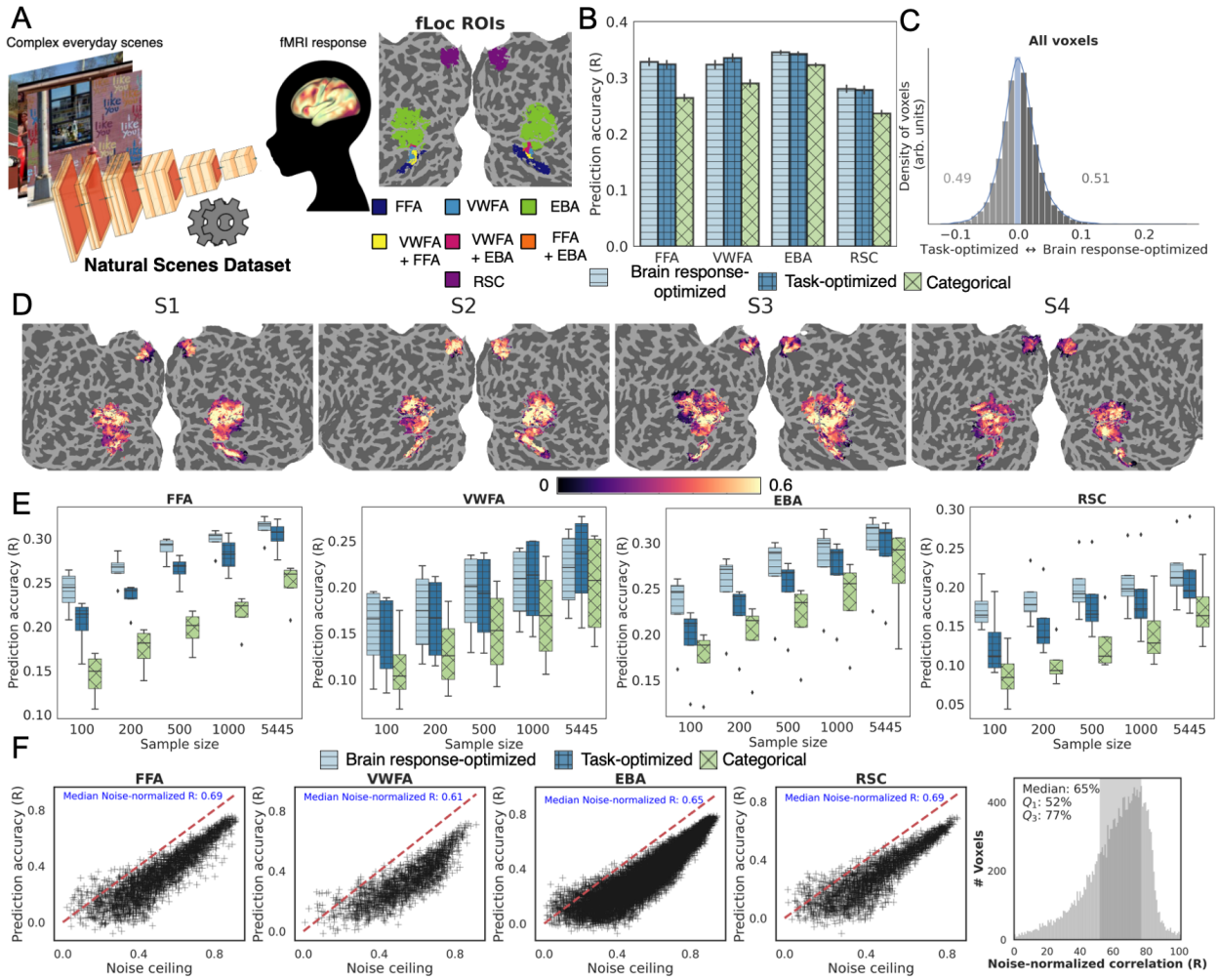

Figure 1: **Quantitative Results. A** depicts a schematic of the experimental paradigm and a cortical flatmap of the 4 visual ROIs studied here, which have some noticeable overlaps. **B** shows the prediction accuracy of the proposed and baseline models (a task-optimized model trained on object classification on ImageNet and a simple categorical model) as estimated by the Pearson correlation coefficient (R) between predicted and held-out responses for the same subjects on which the models were trained. **C** Voxel-wise distribution of the difference between prediction accuracies of the response-optimized and task-optimized models. The inset shows the proportion of voxels that are better predicted by each. The response-optimized models achieve parity with the task-optimized model trained on a million ImageNet images (no difference was found through a permutation test, $p>0.01$ for all 4 ROIs). **D** Cortical flatmap illustrating the prediction accuracy achieved by the response-optimized model in all voxels of the four ROIs. High predictive accuracy (>0.6 unnormalized correlation) is achieved within large swathes of these ROIs. **E** Generalization performance of all models to new subjects as assessed by varying the amount of stimulus-response pairs used to train the linear readout. Response-optimized models generalize much more efficiently to novel subjects than task-optimized or categorical models. **F** Un-normalized prediction accuracy (R) of every voxel against the corresponding noise ceiling. Noise-normalized prediction accuracy is reported in the inset. Much of the variance in predictive accuracy across voxels is driven by their noise ceiling. Response-optimized models attained approximately 61-70% of the noise ceiling, functioning as one of the most quantitatively precise voxel-level models of these higher-order regions.

a held-out set of 1,000 images seen by these same subjects. For these subjects, the response-optimized models attained approximately 65% of the noise ceiling (median correlation), with noise-normalized correlations lying between 52-77% for 50% of the voxels, yielding one of the most computationally precise voxel-level models of these higher-order regions. Further, response-optimized networks achieve parity with task-optimized networks on the same subjects (FFA: p = 0.274, VWFA: p=0.014, RSC: p=0.709, EBA: p = 0.510. p-values calculated by two-side permutation test with N=1,000). Similar to results previously reported using recordings from non-human primates [22], our models are far more predictive than the category ideal observer model which employed the category membership of labeled objects in the image (another, simpler baseline than task-optimized models).

Next, we assessed how these predictive models generalize to the remaining set of four subjects that were not used to train the model. For this analysis, we train each network to predict activity for the remaining subjects by only optimizing the weights of the final linear readout while keeping the rest of the network fixed. We vary the amount of stimulus/responses pairs from the new subjects to train the readouts, from only 100 samples to a large set of 5,445 stimulus-response pairs. The influence of neural dataset size on predictive accuracy can complement prediction performance when evaluating the quality of two competing models. We consider that the best representation is the one that enables the most sample-efficient learning of the readout model for new subjects. Importantly, the difference in performance between response-optimized and task-optimized networks becomes even more striking as we limit the stimulus-response pairs (see fig. 1.E). In FFA, EBA and RSC, the average performance for the response-optimized networks is already at more than 78% of its final value after just 100 training samples, compared with $60-67\%$ and $50-65\%$ for the task-optimized and categorical models respectively. In FFA and EBA, we need 500 samples for the task-optimized network to achieve a comparable performance to the response-optimized network with 200 samples. This remarkable generalization of response-optimized networks suggests that they are able to sufficiently constrain the space of possible solutions in the right manner so that the readouts for new subjects can be learned with few samples.

Selectivity, Tolerance and Clutter-invariance revealed by network dissection

Here we demonstrate that the high prediction accuracy and generalization of our hypothesis-neutral response-optimized networks do not come at the cost of model intelligibility. Instead, we show that the response-optimized models possess both empirical and aesthetic virtues, being computationally precise and elegant at the same time. Notably, features that emerge in the trained networks result from optimization to match ROI responses. After training the networks, we can probe their learned features, understand the computations they perform and, consequently, understand the characteristics of the ROI responses they model.

We adapt the recently proposed technique of network dissection [30, 31] to generate "verbal" explanations for the responses of different voxels. The technique measures the degree of alignment between a voxel's response properties and an extensive visual concept dictionary, spanning objects and fine-grained visual concepts like parts of objects, colors, materials, and textures. We quantify the agreement between each concept and individual voxel using

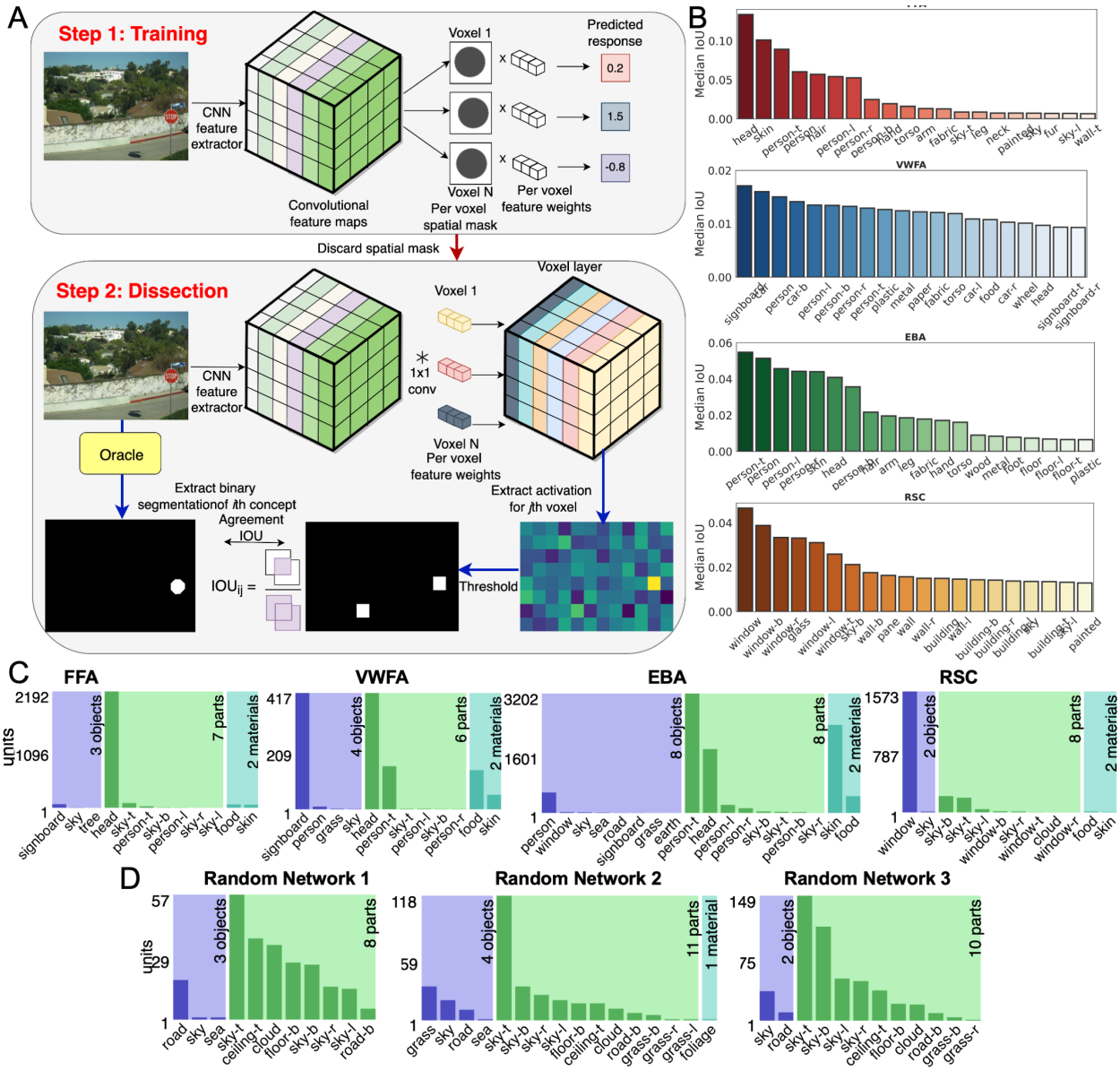

Figure 2: **Network Dissection Results. A** Schematic of the model conversion procedure used to obtain the dissection results. The spatial mask is discarded and the learned feature tuning of every voxel is employed to create an additional 1x1 convolutional layer, so that every voxel is represented by an independent unit in this convolutional layer. **B** demonstrates the median IoU metric across all voxels belonging to an ROI for the top 20 visual concepts rank ordered by median IoU. The top concepts for FFA, VWFA and EBA discovered using this hypothesis-neutral approach ('heads', 'signboards' and 'person' respectively) align remarkably well with the known domain-specificty of voxels in these regions. **C** shows the matched concepts for every response-optimized model, i.e., the number of units in the 'voxel' layer that showed high alignment (IoU>0.04) with a human-interpretable visual concept. Multiple units (i.e., multiple voxels) are associated with the same high-level concept. Contrary to object recognition networks which are trained with explicit label supervision and which show a broad diversity of detectors, the matched visual concepts in these response-optimized networks are highly specific and aligned with the previously hypothesized functional role of these ROIs. **D** shows the matched concepts for 3 networks with the same architecture as response-optimized models but random weights. The detection of these concepts (e.g. sky, grass, road) is likely driven by low-level cues, like color instead of complex features.

the Intersection over Union (IoU) metric following [30]. The IoU metric is computed on an independent, large-scale natural image database comprising a diverse set of real-world environments (indoor, urban, and natural). An alignment is computed between two maps for each image in the database. The first map indicates the high-level concept corresponding to each pixel in the image (pixel-level labeling is performed by humans or a high-performing segmentation network). The second map indicates the spatial regions within the image that are highly activated by the convolutional filter corresponding to a voxel (see Methods section for further details). After computing the alignment across images, the result is an alignment value for every voxel-concept pair. It is essential to note this methodological framework's subtle yet profound implications. Previously, response profiles have been primarily defined using image-level category labels. Our proposed approach enables us to model voxels as convolutional filters and systematically identify the image properties that they respond to without an a-priori hypothesis specification. Our approach enables us to characterize not just 'which' images activate a particular brain voxel but also 'what' in those images drives the response, providing a rich characterization of neural responses to crowded natural scenes.

Fig. 2(B) shows the results of this dissection procedure: for the top 20 concepts in each ROI, the median IoU is shown across all voxels in that ROI. Following [30], we use an IoU threshold of 0.04 to detect 'matching', i.e., if a voxel's corresponding filter exhibits a high agreement with a concept map (exceeding 0.04 IoU threshold), we say that the particular voxel *detects* or *encodes* that concept. We show in fig. 2[C] the concepts for which a high agreement is found and the count of the voxels that achieve a high agreement with that concept (each voxel is only counted once, against the concept with the top IoU). Fig. 2[D] shows the same measures for untrained networks having the same architecture as response-optimized models, but with random weights. Finally, fig. 3 shows a visualization of the part of highly activating images that leads to the high activation. For each ROI and its preferred concept (according to the median IoU metric), the five top voxels are chosen, and for those, the top 100 images activating images are picked from the large scale dataset. For randomly selected images from this set, the area that leads to maximum activation of the voxel is shown.

Fig. 2 shows that the FFA network's favorite concept is 'head' (the large scale dataset didn't have a face label, and faces are labeled as parts of heads). The head concept has the highest median IoU of 0.125, more than other top concepts for FFA and the other ROIs. Some FFA voxels predictors have IoU as high as 0.20 with the 'head' concept. We can contrast this with the 'head' detectors (model units with IoU>0.04 against the 'head' concept) that emerge spontaneously in the last convolutional layer of a standard AlexNet trained on image categorization (ImageNet)[1]. In this task-optimized model, the best detector yields an IoU of 0.15 against the 'head' concept maps with the median IoU among all head detectors being 0.08. Fig. 3[A] illustrates how the chosen voxels in FFA act as head (face) detector, with the parts of images driving the predicted response being the faces.

The VWFA network has high median IoU with signboards (containing letters) even though the number of images with writing is small in the NSD dataset. Fig. 3[B] shows that signboards with very different lettering and signs (different backgrounds, fonts, styles, scales, colors, orientations etc) drive the predicted response in the top voxels, even though

[1]Pre-trained model downloaded from here: `https://pytorch.org/hub/pytorch_vision_alexnet/`

the scenes are often cluttered with myriad objects at once as shown. This result replicates the highly invariant orthographic processing in the VWFA [37]. The VWFA also has high alignment with the concepts of head and person, which is due to the anatomical overlap of localized VWFA with FFA that we discuss in a following section. EBA has high IoU with people, head and skin. The IoU with the people concept is highest, highlighting that EBA cares about body parts. Interestingly, for RSC, we observe a large number of 'window' detectors after applying the dissection procedure. Navigational affordances are important for scene perception, and windows might be particularly indicative of such affordances in indoor scenes and critical to functional scene understanding (for e.g., windows indicate an obstructed path where movements are blocked) [38]. Other concepts evoking high response within the RSC (e.g., 'wall', 'glass', 'building') all relate to scene perception and navigation, contrary to concepts discovered within the models optimized for the other three ROIs, highlighting the functional differences between these visual areas. Despite no access to category labels during training, our networks gain a strong semantic selectivity for high-level visual concepts.

Maximally exciting images reveal structured, high-level features consistent with the previously hypothesized functional role of each ROI

In classical neuroscience, identifying the optimal visual stimuli (the peak of the tuning curve) for different neurons has been instrumental in understanding the selectivity of these neurons and how they contribute to perception. Our models allow us to follow this approach for complex stimuli. We perform an unconstrained optimization over input noise to discover input images that would result in maximal (predicted) excitation of individual voxels. We refer to these images as the maximally exciting inputs. While optimization without naturalistic constraints may impose its own set of challenges including generation of hard-to-recognize visual features, visualization with regularization or naturalistic constraints may not be truly faithful to the model. We favor the former approach and do not restrict the space of maximally exciting inputs to the naturalistic domain. Instead, we let the optimization process evolve complex visual inputs without constraints.

In fig. 4, we show that maximally exciting images for different voxels in the same ROI capture very similar visual properties. Out of all the features that could emerge from unconstrained optimization in a network trained on cluttered natural scenes, from simple features such as rounded shapes or eyes to possibly more complex high-level features, face-like images with ovelapping small and large circles almost exclusively pop up for all FFA voxels, providing a strong support for the hypothesis that full 'face' features lead to increased activation in FFA. Similary, for EBA, we observe elongated curved shapes, loosely similar in form to body parts such as arms or legs. Maximally exciting inputs for the VWFA resemble orthographic units comprising curves and lines of different stroke widths like the visual form of letters. Finally, the maximally exciting inputs for voxels within RSC are reminiscent of windows (in accordance with the Network dissection results) in different reference frames, which may be linked to RSC's role in spatial cognition [39]. We also observe rectilinear features in the maximally activating images for RSC, consistent with the previously found rectilinear pref-

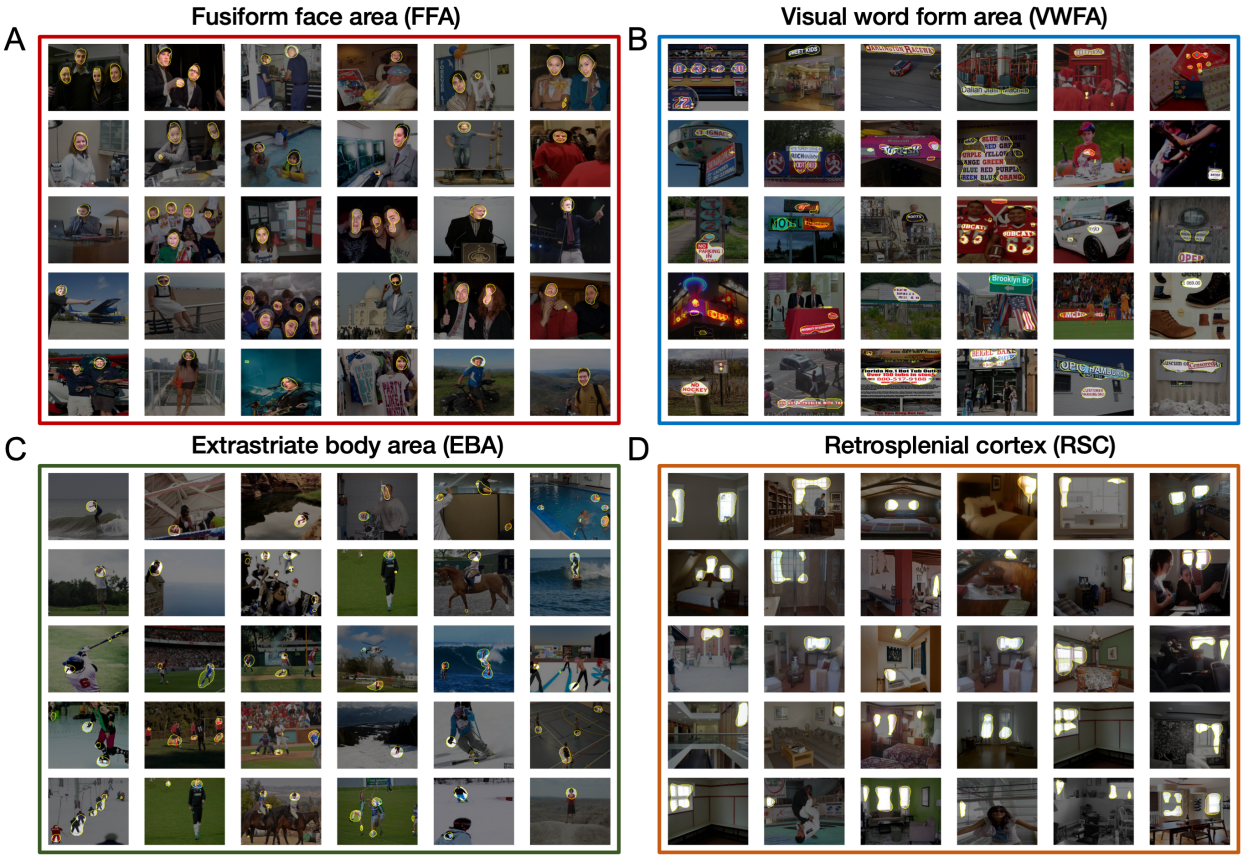

Figure 3: **Segmentation maps.** Activation of a single voxel (converted to a 1x1 convolutional filter) in response to an input image is visualized as the region in the input space that elicits the highest (top 1% quantile level) activation in the corresponding filter output. Top five voxels as ranked by the IoU with the preferred concept for every ROI ('heads', 'signboards', 'person' and 'windows' for FFA, VWFA, EBA and RSC respectively) are identified and input images are randomly selected for each voxel among the top 100 most activating images for that particular voxel to maximize diversity across voxels. Each row corresponds to a distinct voxel within the respective ROIs.

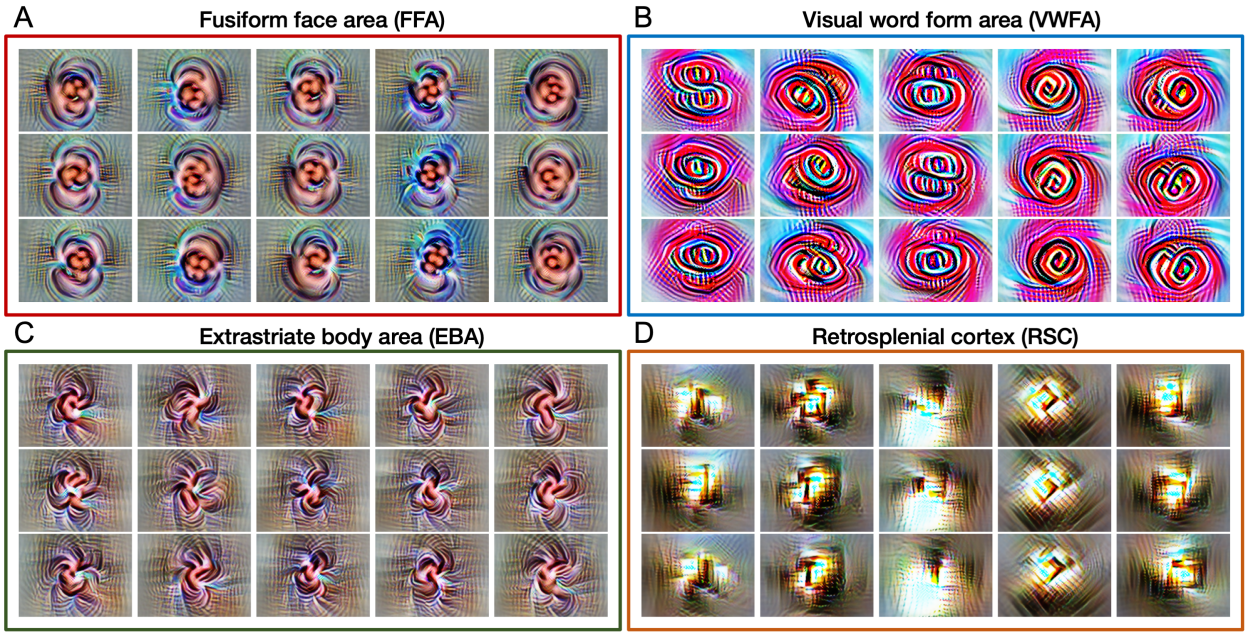

Figure 4: **Maximally exciting inputs.** Most activating images as discovered by the optimal stimulus identification procedure, wherein an unconstrained optimization is performed over random input noise to discover the input that would result in maximal excitation of individual voxels. Images are shown for 5 randomly selected voxels per ROI starting from 3 random initial points per voxel. Each column is a randomly chosen voxel and each row is a different initialization. This structural analysis reveals that the featural drivers of individual voxels (the *visualizations*) are consistent with the *verbalizations* for those voxels as expressed by the network dissection procedure. More specifically, face-like visual forms, curved features akin to orthographic symbols, skin-colored shapes reminiscent of some body parts and window-like rectilinear features emerge spontaneously for FFA, VWFA, EBA and RSC respectively.

erence of scene-selective areas [40], while the EBA and the FFA only captured curvilinear features. Our results therefore add proof to this hypothesis by learning the model that best predicts the data instead of starting *apriori* with the hypothesis.

End-to-end models capture tuning differences between voxels in the same ROI

To investigate if the proposed models are indeed capturing meaningful differences between voxels, we computed *spatial generalizability matrices* by correlating the predicted response of each voxel against the measured response of every other voxel to obtain an N×N correlation matrix for shared models, where N is the total number of voxels across all participants [41, 42, 43]. These matrices reveal the similarity of the tuning of each pair of voxels. To account for higher variability in measured versus predicted response, we normalize the rows and columns of this correlation matrix following [44]. The diagonal dominance in these identifiability matrices, as shown in Figure 5[A], suggests that predicted responses are most similar to

the same voxel's measured responses, indicating that all models are successfully able to capture meaningful voxel-level idiosyncracies, albeit to different extents. The generalization matrices reveal the presence of several distinct clusters (at least two) for all ROIs, such that the models of voxels in one cluster are highly predictive of responses of voxels in the same cluster (both within, and across participants), but do not generalize to other clusters. Importantly, this clustering structure is prevalent across participants (specifically for FFA, EBA and VWFA, with more variability for RSC), indicating a shared organization. We leave a thorough characterization of the differences between these clusters for future work.

Next, we turn our focus to voxels that were strongly selective for a different semantic concept than the conjectured preferred category for the visual area they belonged to. We specifically examine the 'head' selective model neurons identified by the network dissection procedure within VWFA and EBA. We assessed the correlation between the quantitative agreement of these voxels with the 'head' concept over the 'signboard' or 'person' concept (IoU, as evaluated by our proposed dissection procedure) and the degree of their face-selectivity over selectivity for words or body parts (as quantified with the independent functional localizer experiment); our results suggested a very strong correspondence between the two (Pearson's R$\sim$0.34-0.66, $p<0.001$ for all 4 subjects and both comparisons, Figure 5B), despite the very different experimental paradigms involved in the two quantifications. Further, in addition to comparing this relative selectivity, we also compared the absolute 'head' selectivity of voxels in EBA and VWFA against the face-selectivity (t-value) measured with the localizer agreement. Again, we see a striking pattern of similarity in these estimated and measured (Pearson's R$\sim$0.6-0.7 ($p<0.001$); see Figure 5[C] for qualitative match and Appendix for quantitative agreement). Our computational models are thus able to successfully capture the spatially overlapping representations of semantic categories and graded functional organization within human extrastriate cortex.

High selectivity persists in 'face-deprived' and 'body-deprived' networks

The strong semantic selectivity in our response-optimized models raises an important question: are the ROIs they model simply functioning as detectors for their preferred category? For example, training a neural network to predict FFA would then be equivalent to giving it images associated with a label that indicates the presence of a face. An alternative hypothesis is that category-selective ROIs are sensitive to visual properties that are typical of their preferred category, even in the absence of that category. In that case, training a neural network to predict FFA would provide it with a more complex signal that allows it to pick up on the sensitivity of the FFA to those visual properties.

To differentiate between these alternatives, we focus on FFA and EBA and train response-optimized models with the same architecture above but with a visual training diet that is entirely deprived of images containing the 'person' category. Surprisingly, we find that, despite not seeing *any* image with human faces or bodies during training, the networks optimized to predict FFA and EBA retain their categorical selectivity, as shown in Figure 6. The training diet deprivation did result in a drop in the agreement of model voxels with their respective preferred category (e.g., in FFA, the median IoU with 'head' dropped from

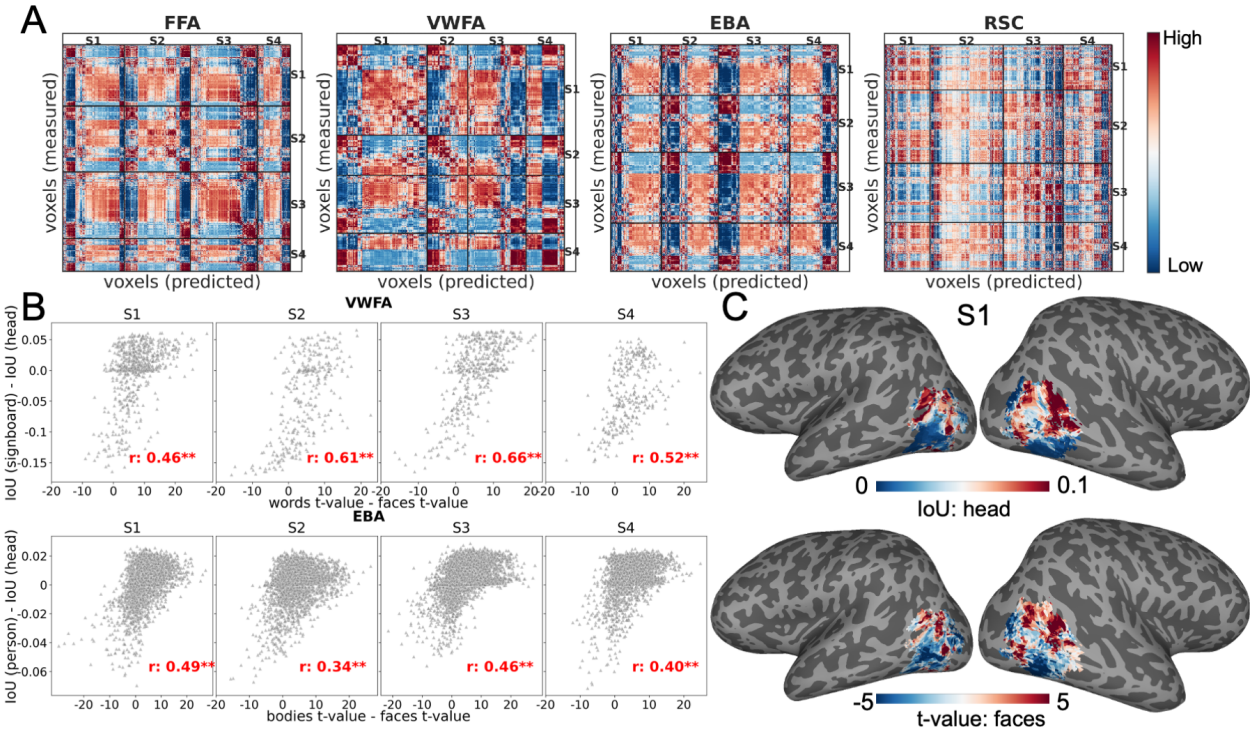

Figure 5: **Voxel-level identifiability A** shows *spatial generalization matrices* for every ROI computed by correlating the predicted response of every voxel against the measured response of every other voxel belonging to the same ROI (within and across subjects). Black lines mark subject boundaries. **B** Scatter plot depicting the difference between IoU of the preferred category (loosely, 'signboard' and 'person' for VWFA and EBA, respectively) and faces against the difference between the corresponding t-stat values estimated by the functional localizer experiment. Each point is an individual voxel belonging to the same ROI. **C** shows the cortical surface plot of face selectivity as measured by the localizer experiment against the IoU with the 'head' concept as quantified by the dissection procedure for one subject (Similar plots for remaining subjects are shown in the Appendix)

~ 0.13 in the non-deprived network to ~ 0.08 in the deprived network), indicating a slightly reduced selectivity. However, the preferred concept was still overwhelmingly 'head', and the resulting segmentation pictures look qualitatively similar to those in fig. 3. In other words, we found that networks trained with a visual diet deprived of their preferred category can still extrapolate the responses to the preferred category.

Indeed, we observe in fig. 6[C] and [D] that models do not incur a severe loss in prediction performance when data from a new domain (i.e. 'faces' and 'bodies' category) is presented at test time. Actually, the FFA model achieves even slightly better prediction performance on held-out images with faces. This example of systematic generalization in the developed response-optimized models is also interesting from the perspective of modern deep learning, which is often criticized for its failure to generalize in this systematic, out-of-distribution way. The ability of these models to generalize to new domains further also validates their proposed usage as virtual stand-ins for fMRI experiments, supporting their ability to act as *model organisms* for large-scale fMRI experiments.

Models reveal important functional distinctions between different regions

The previous experiment highlights that response-optimized models do not merely act as detectors for the preferred category of their respective brain voxels but also capture more complex tuning properties relevant to their preferred category but not uniquely exhibited by it. Here, we ask if the models *meaningfully* discriminate between stimuli belonging to their preferred set? More specifically, we want to test existing accounts of functional specialization which implicate FFA in face perception, particularly face identity discrimination [45, 46] and RSC in spatial cognition [39].

We first perform a face discrimination task using the response-optimized models of all four ROIs. We extract for each model the predicted voxel-wise response for a small subset of facial images from the CelebA dataset, comprising 20 identities, each with 100 train, 30 validation and 30 test images [47, 48]. We compare the face recognition accuracy of these model predictions against the recognition accuracy of a general-purpose representation from a CNN trained to perform image classification on the large-scale ImageNet dataset. We also compare the face recognition accuracy of these models with that of a representation from a nework trained explicitly to do facial recognition on a large set of VGG face identities [49]. For each of the representations above (the four ROIs' predictions and the general-purpose and face specialized representations), we train a linear function to predict the CelebA identities. We find that FFA predictions significantly outperform the predictions from all other ROI models (discrimination score of 85% compared to 79-80%), even outperforming the highly transferable representation of ImageNet trained networks at 78% (however, still falling short against features from networks trained to discriminate a large number of identities at 96%). This result supports and provides additional evidence for the role of FFA in face identification and the ability of our model to pick up these functional capacities.

Next, we perform a room layout prediction task, again using the voxel-wise predictions generated by the four ROI models and the representations generated by the general-purpose CNN trained on ImageNet. The objective of this task is to predict the correct layout type

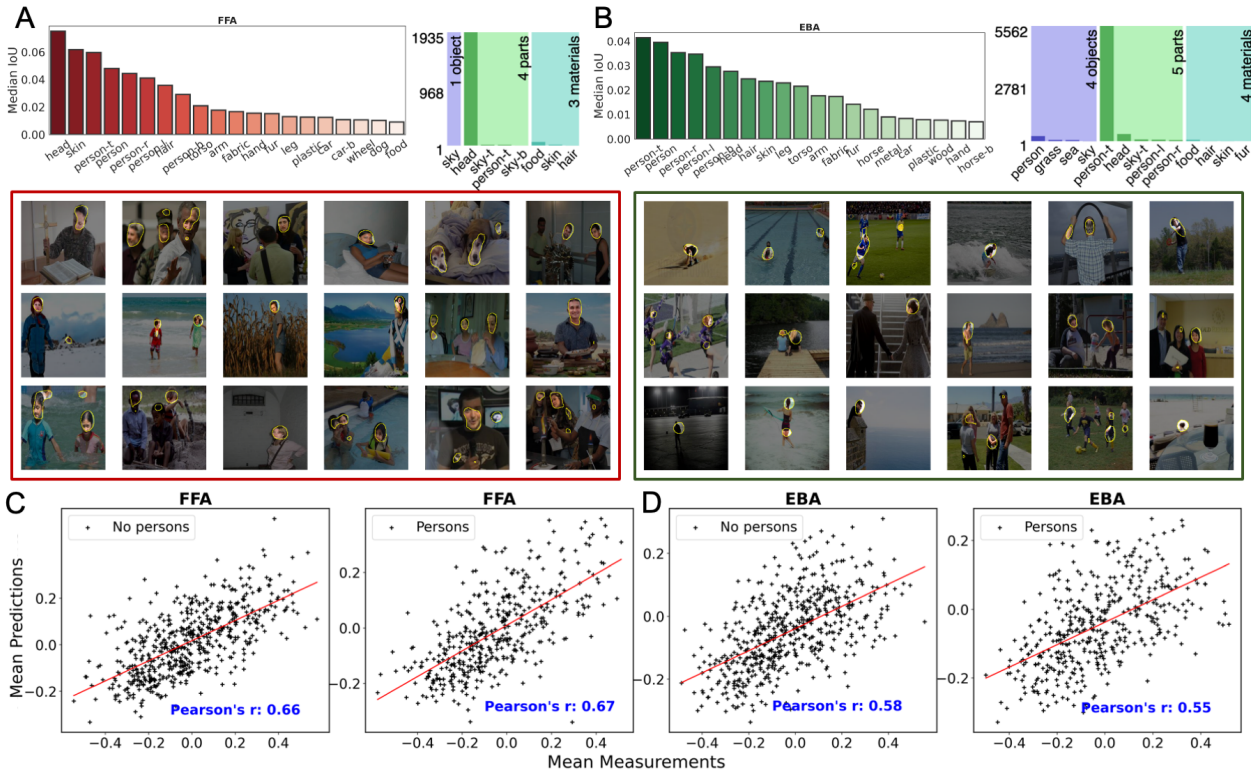

Figure 6: **Selective deprivation in visual diet.** **A** and **B** show results of the dissection procedure applied to response-optimized models of FFA and EBA respectively, when trained using a visual diet entirely deprived of the 'person' category; Left panels show the median IoU computed across all voxels in that ROI against the top 20 concepts identified using the median IoU metric. Right panels show the matched visual concepts for the respective ROIs. An IoU threshold of 0.04 is applied to detect matching. Activation of 3 randomly selected concept detectors in each ROI model to top images are shown below. **C** and **D** depict the mean measured response across all voxels for every test image against the corresponding mean predicted response, separated into 'persons' and 'no persons' categories for FFA and EBA respectively. Pearson's correlation coefficient between the mean predicted and measured values is reported inside each scatter plot.

from the 11 categories described in [50] defined using a keypoint-based parametrization (Figure 7[B]). The dataset comprises 4,000 natural scenes across diverse indoor scene categories from the SUN database [51], that are split into sets of 3200 training, 400 validation and 400 test images. We follow the same procedure for quantifying the layout estimation accuracy as before, i.e., training a linear classifier on top of each representation. Here, we observe a different trend. The RSC predictions outperform the other ROI models. They perform just as well as the ImageNet trained representations in solving this task (recall that the RSC network was trained with $\sim$ 35,000 images and their associated brain responses, while the ImageNet network is trained on a million images). This result is highly suggestive of the role of RSC in scene understanding, particularly aspects relevant to spatial navigation, and of our model's ability to pick up these functional capacities.

Discussion

In this paper, we exploit the ability of data-driven, hypothesis-neutral deep neural networks to model the responses of high-level visual ROIs. Through a large, rich stimulus set afforded by the Natural Scenes Dataset (NSD), we offer new evidence that generalizes decades worth of hypothesis-driven results to ethologically valid settings.

Brain response predictivity We found that response-optimized deep neural network models—trained solely with supervision from fMRI activity—accurately predict activity related to new images in multiple visual category-selective ROIs. The performance of these models rivals the predictive performance of state-of-the-art task-optimized models used to predict brain activity [22]. Many model networks can be consistent with brain activity data as demonstrated by the number of recent papers in this area [52]. It can be argued that a good representation would allow for efficient learning (in terms of the number of data points required) of a linear predictor of brain responses. Thus we ran an analysis in which we evaluated the trained models on the ability of their representations to predict data for new subjects while restricting the training set size. We found that the response-optimized models were better able to generalize to novel subjects at small sample sizes than task-optimized models. This result suggests that end-to-end optimization solely driven by response measurements can yield better correspondence to brain data than networks optimized on behaviorally relevant tasks with millions of images.

Verbalization of voxel selectivity with network dissection Neuroscience is gradually adopting naturalistic stimuli not amenable to parametrization and neural networks to model brain responses. This trend requires concomitant methodological advances to derive trustable conceptual understanding from brain response models. The perspective is that accurate and generalizable computational models of brain function can serve as a reasonable proxy to biological visual systems and can be probed or interrogated to understand better the properties of the system they model. Here, we developed a systematic methodological framework to understand the tuning properties learned by our response-optimized networks. Our framework aligns the recent progress in understanding neural networks using large-scale annotated datasets and network dissection procedures [30] with the longstanding goal of

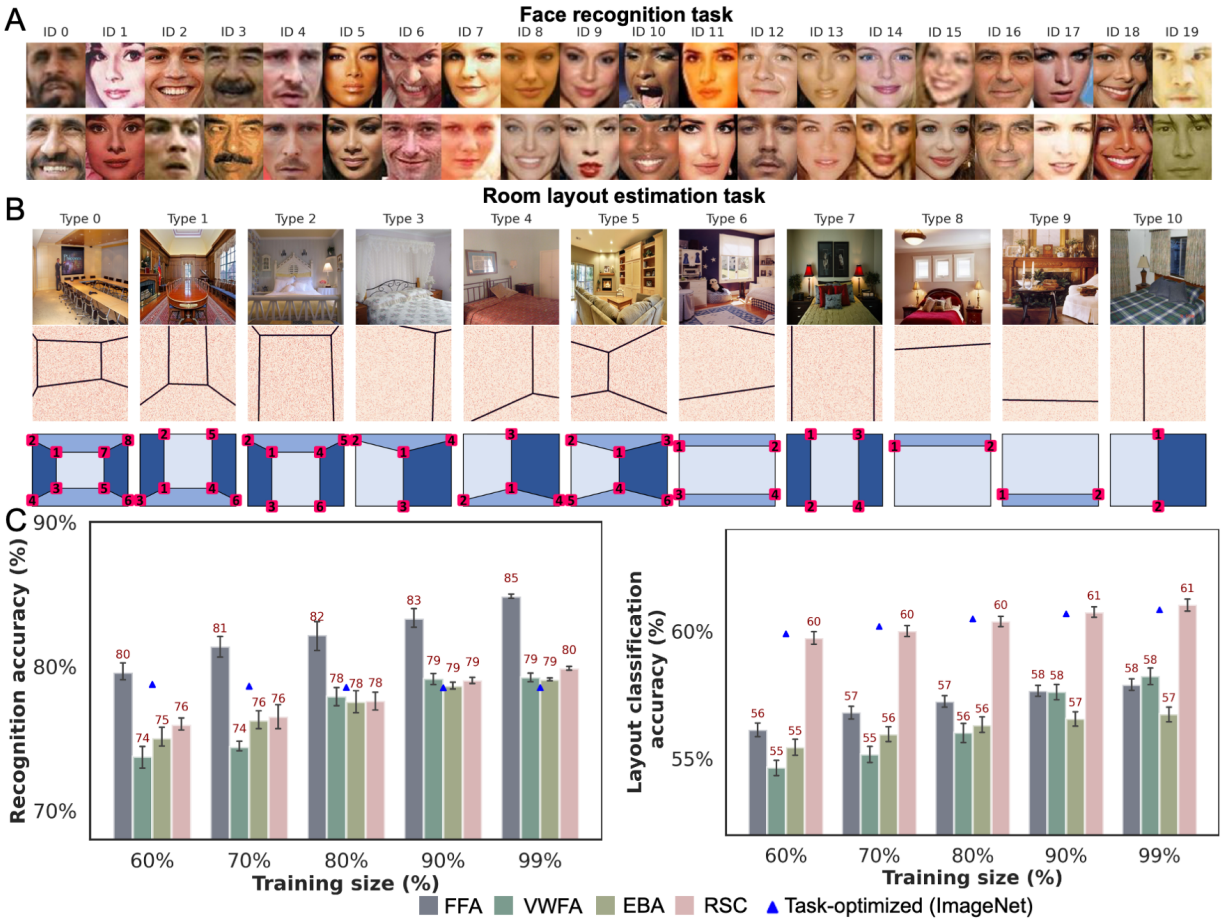

Figure 7: **Task performance of neural representations.** **A** and **B** show sample stimuli along with their classification labels for the two visual discrimination tasks considered in this study, namely face identity recognition and room layout classification. In **B**, the ground truth layout planar segmentation of indoor scenes in the room layout estimation task are depicted below the respective scenes and the definition of their corresponding room layout types are illustrated in the bottom panel. **C** depicts the transfer performance of neural representations from response-optimized networks of each individual ROI on the two tasks. Intriguingly, representations from FFA achieve the best face individuation accuracy and those from RSC perform best on the layout estimation task, remarkably consistent with the previously hypothesized functional roles of these visual areas.

understanding the brain by characterizing neuronal tuning properties. This alignment procedure capitalizes on the usage of a factorized readout that disentangles the spatial ("where") and feature ("what") dimensions of the voxel's response properties and fully characterize the "what" dimension using dissection procedures applied on the 'model neuron' corresponding to the voxel.

Using complex stimuli like cluttered natural scenes with multiple objects in context to probe neuronal response properties poses multiple challenges. For instance, if a set of images that highly activates a voxel is identified, several competing interpretations can be ascribed to that voxel's tuning function. There can be multiple low-level (e.g., visual features) or high-level (semantic properties) consistent with the identified images. Indeed, as stimuli become more complex, it is increasingly more challenging to recognize the consistent pattern implicit among any given set of images due to the tangling of visual features (textures, shapes, edges, etc.). The dissection procedure in this study helps simplify this problem by not just recognizing 'which' natural images elicit higher responses but also systematically and quantitatively characterizing 'what' part of each image leads to increased responses. Effectively, we can verbalize the functional role of voxels in each candidate region. We demonstrated that emergent concepts in response-optimized networks are highly specific and aligned with each regions' previously hypothesized functional role. E.g., we reported an exclusive presence of face detectors in the model for FFA. If the learned category-selectivity were in reality due to low- or mid-level features correlated with the category of interest, those features would have likely become apparent because of the dataset used in the network dissection analysis. That dataset is one of the largest densely-annotated natural image datasets in computer vision research. For instance, if the selectivity of FFA to faces could be explained by the unique structure of eyes or skin-colored textures, response-optimized models would have caught on to these simpler features. The segmentation-based dissection procedure would have revealed this pattern of selectivity as it scores units against category-level labels and thousands of other concepts, including different textures and object parts (including eyes, nose, etc.). Nonetheless, with our dissection procedure, we see an almost exclusive selectivity for full faces in the FFA voxels. The results of our dissection procedure also highlight that, while solely optimized for brain response prediction, the proposed computational models also solve the challenge of perceptual invariance. The models respond selectively to their preferred category despite tremendous variation in the precise physical characteristics of the preferred objects.

Visualization of voxel selectivity with image synthesis Next, we used an optimization-based image synthesis technique to construct the stimulus that causes synthetic neurons modeling individual voxels to activate maximally. Several recent studies have attempted to describe the tuning properties captured by DNN models of visual cortex using such image syntheses algorithms [53, 23, 54], and have even demonstrated the abilities of these methods for controlling neural firing activity in mouse primary visual cortex and macaque V4 [55, 56]. However, several distinctions between groundlaying work in this direction and our results are worth consideration. Unlike some prior studies that use the hypothesis space of generative adversarial networks, we do not impose a 'naturalness' prior on the optimized images. We neither employ a task-optimized model trained on large-scale object databases to derive the

features that map onto brain activity. Both these choices may bias the features to contain high-level semantic content (especially since task-optimized models are trained with explicit category-level pressures), even if the neurons/voxels primarily encode low-level or mid-level features that are not naturalistic or neatly verbalizable. Moreover, existing fMRI studies apply such techniques at the ROI level. Despite the fact that our optimal inputs were synthesized de novo from random pixel noise to activate single voxels in our study, we find that the optimized images still contain human recognizable complex patterns consistent with the hypothesized functional role of these voxels. The images appear to capture critical functional properties of voxels in these high-level visual regions. We argue that encoding models fitted directly to neural data with no prior training, offer a useful, complementary perspective for understanding the visual system to hypotheses-driven, task-optimized models. Any set of features that emerge in the empirically-estimated response-optimized networks are optimized to explain representations in the brain, and are not tangled with confounds from top-down constraints unrelated to neural activity.

Specific versus non-specific mechanisms in shaping categorical selectivity Next, we analyzed the role of visual experience in shaping response selectivity by training response-optimized models with a visual diet completely deprived of faces or bodies. Despite this selective deprivation, units in models of FFA and EBA retained strong selectivity for their preferred semantic content. How do we interpret these findings? The models effectively became face and body selectors, which means that they could generalize their predictions to stimuli that were not included in training and learn that those stimuli will cause maximal firing. The models could infer the preferred category of their corresponding ROI through training with a dataset devoid of this preferred category. The models were able to go this generalization because of the properties of the representations of FFA and EBA voxels. These voxels appear to respond to visual configurations characteristic to faces or body parts, respectively. These configurations could be present in parts of images that do not correspond to a person but happen to have face-like (e.g., circles) or body-like (e.g., elongated, curved shapes). One potential mechanistic explanation is that these visual category-selective ROIs act like filters that constantly look for matches to specific visual properties. While faces or bodies can activate these filters maximally, other visual patterns that resemble them also activate the filters, perhaps to a lesser extent. For the FFA, these other visual patterns could act like pareidolia, or perhaps a subtle version of pareidolia that is hard to detect by humans but that the FFA could still pick up on.

Our computational models' systematic generalization and extrapolation ability suggests that they may enable us to discover functional specialization in underexplored parts of the visual system. Most claims regarding functional specialization and domain-specificity in the brain are grounded in hypothesis-driven contrast-based experiments, which might not include stimuli relevant for these regions. Even a large dataset such as NSD might exclude certain categories that are important for a given brain region. After training our models on the under-explored regions, the generalization ability portrayed here can help us characterize selectivity post-hoc.

Our results also relate to important questions related to statistical learning in artificial and natural systems. In fact, out-of-distribution generalization is an important unsolved

problem in many parts of machine learning. But here we find that our models are robustly able to generalize to unseen categories, which only pays tribute to the strong semantic selectivity of the ROIs themselves, and the fact that they response to features characteristic of their preferred category even in the absence of that category. Another question is whether developing a selectivity for faces requires experience with the unique structure of faces? Existing behavioral evidence from some face deprivation studies [57] suggests that face-processing abilities can persist without any face-specific experience. While subsequent studies have provided counter-evidence in favor of the necessity of face experience for face-domain formation, the dispute remains far from settled given current evidence. Our experiments highlight that face and body selectivity can emerge spontaneously in computational models with no face and body experience, and this high selectivity is maintained across diverse naturalistic variations (see segmentation maps in Figure 6). Though one important caveat is that our models were trained using supervision from brain regions that have experience faces and bodies.

Link between emergent representations and their functional properties We have shown that our models learn to accurately predict their respective ROIs, and that the representations they learn allow them to select for a small number of preferred categories. We further put these representations to a stricter test and evaluated their functional capabilities. We tested existing functional specialization accounts which implicate FFA in face identification [5] and RSC in spatial cognition [39] by simulating these fine-grained discrimination tasks using the representations from response-optimized models of all ROIs. We found that the representations from the FFA network beat the other representations and a representation from a task-optimized network at the face identification task. We also found that the RSA network beat the other representations and a representation from a task-optimized network at the spatial task. The observation that these complex visual capacities are realized spontaneously in neural networks optimized solely to predict brain activity suggests that these networks are learning a representation that is faithful to the information represented by their respective regions.

Conclusion Probing computational models of the brain imposes several challenges that may lead to confounded conclusions about neural representations and computations. The most critical confound is that any conceptual insights gained from a computational model are helpful insofar as the model is a good approximation of the biological system. The prediction accuracy of response-optimized DNN models is not yet perfect. This may be due to significant limitations in the model architecture, insufficient stimulus-response data for fitting complex neural network models, supervision from noisy fMRI signals, or a combination of these and other factors. The work presented in this study, nevertheless, was able to (1) replicate decades of hypothesis-driven work, (2) show robustness of the learned representations even in the absence of the category of interest, and (3) show that these representations could achieve important functional roles characteristic of their respective ROIs. This work also reveals a new empirical space to improve the ability to predict brain activity, by considering different model architectures, other high-level cortical regions (particularly along the dorsal visual pathway where the current task-optimized models have not yielded the same level of

predictive success), or other imaging techniques (fMRI, MEG, EEG) etc.

An important caveat of network dissection, as employed in this study, is that it studies units in isolation; in several cases, semantic concepts may be encoded by a combination of multiple units (voxels). We could extend network dissection techniques to understand the properties of simulated population responses in the underexplored regions of the visual cortex, whose precise functional characterization remains elusive. Even though single neurons or single voxels don't appear to exhibit high selectivity for object categories in those areas of the brain, we can ask whether populations of neurons or voxels in these regions encode and represent human-understandable concepts using novel network interpretability techniques [58].

Our work demonstrates that a less hypothesis-committed approach can complement hypothesis-driven study of the visual cortex in meaningful ways. This empirical approach, enabled by the new data revolution in neuroscience and large-scale compilation and dissemination of neural data, can offer a complementary basis for building broader theories about neural computations, that generalize to a range of ethologically relevant scenarios.

Materials and Methods

Natural Scenes Dataset

A detailed description of the Natural Scenes Dataset (NSD; http://naturalscenesdataset.org) is provided elsewhere [27]. Here, we just briefly summarize the data acquisition and preprocessing steps. The NSD dataset contains measurements of fMRI responses from 8 participants who each viewed 9,000–10,000 distinct color natural scenes (22,000–30,000 trials) over the course of 30–40 scan sessions. Scanning was conducted at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. Images were taken from the Microsoft Common Objects in Context (COCO) database cite Lin 2014, square cropped, and presented at a size of 8.4° x 8.4°. A special set of 1,000 images were shared across subjects; the remaining images were mutually exclusive across subjects. Images were presented for 3 s with 1-s gaps in between images. Subjects fixated centrally and performed a long-term continuous recognition task on the images. The fMRI data were pre-processed by performing one temporal interpolation (to correct for slice time differences) and one spatial interpolation (to correct for head motion). A general linear model was then used to estimate single-trial beta weights. Cortical surface reconstructions were generated using FreeSurfer, and both volume- and surface-based versions of the beta weights were created. The 4 ROIs considered in this study, namely, the Fusiform face area (FFA, includes FFA1 and FFA2), Extrastriate body area (EBA), Visual word form area (VWFA) and Retrosplenial cortex (RSC), were manually drawn based on the results of the functional localizer (fLoc) experiment after a liberal thresholding procedure.

Response-optimized encoding model architecture

We trained separate voxel-level predictive models for each of the above category-selective regions with the same backbone architecture. The predictive model comprises a shared con-

volutional neural network *core* common across all subjects that represents the feature space unique for specific visual areas. We employ a linear *readout* model on top of the feature space to predict the responses of individual voxels in a specific region of interest under the assumption that the feature space likely represents the input received by these areas and these regions perform close-to-linear transformations on this input. A linear readout on a shared feature space is further based upon the often made assumption that the activity across a set of neurons or voxels in one individual can be related to the activity of the second individual in the homologous functional region by a linear transform [59]. Further, the linear readout is also *factorized* into *spatial* and *feature* dimensions following popular methods for neural system identification. This allows us to separate spatial tuning or receptive field locations (i.e., what portion of the sensory space is the voxel most sensitive to?) from feature tuning (i.e. what features of the visual input is the voxel sensitive to?). The base feature extraction network or the core thus performs all nonlinear transformations to convert the raw sensory stimuli (i.e., pixels) into a representation characteristic of a particular visual area, whereas the readout linearly maps this extracted representation into voxel responses. The core consists of four sequential convolutional blocks, with each block comprising the following feedforward computations: two convolutional layers each followed by an inner batch norm and nonlinear activation (ReLU) operations and an anti-aliased AvgPool operation at the end. Instead of regular convolutions, we employ E(2)-steerable convolutions in the core of all our models to compute orientation dependent activations for many different orientations, thereby achieving joint equivariance under translations and rotations by design [60, 61, 35]. This enables us to apply filters not just in every spatial location, as in a standard convolutional layer, but also in every orientation, increasing parameter sharing and improving the statistical efficiency of deep learning. This modeling choice is also inspired by neural computations in early visual areas where it is hypothesized that groups of neurons perform similar computations at different orientations, e.g., edge or curve detection at different orientations. From an implementation perspective, the filters in these equaivariant convolutional operations are constructed as a linear combination of a fixed system of atomic filters, which helps avoid artifacts and enables arbitrary angular resolution with respect to sampled filter rotation [60]. The readout contains all voxel-specific parameters and maps the extracted representation to individual voxel responses. Weights of the readout are a sum of outer products between a spatial filter and a feature vector. The spatial filter further had a positivity constraint (enforced using rectification) and was normalized independently for each voxel by dividing each spatial weight by the square-root of the sum of squared spatial weights across all locations.

Training and testing models

Combined across all 4 subjects, the dataset comprises 37,000 natural scene images, among which 1,000 images are shared across all subjects and the rest are mutually exclusive. We used the 1,000 shared images for testing our models and split the remaining stimulus set into 35,000 training and 2,000 validation images. All parameters of the response-optimized model were optimized jointly to minimize the *mean squared error* between the predicted and measured response. Since for every image in the training set, the response is measured from only a single subject and not all subjects, we use a masked mean squared loss to train

the model across multiple subjects. Let $r_{s,v}^{pred.}$ and $r_{s,v}^{meas.}$ denote the predicted and measured response of voxel v in subject s to image i, respectively. Then, the loss function during training is given, as

$$\mathcal{L} = \sum_{i \in \text{Batch}} \sum_{s=1}^{S} \sum_{v=1}^{n_s} 1_{i \in I_s} (r_{s,v}^{pred} - r_{s,v}^{meas.})^2,$$

where $1_{i \in I_s}$ is the indicator variable specifying if image i was shown to subject s. The proposed method allows us to propagate errors through the shared network even if the subjects are not exposed to common stimuli since we can always exclude the subjects/voxels for which the response is not present from mean error calculation within each batch. The shared network thus benefits from diverse, varying stimuli across subjects with less extensive constraints on data collection from single subjects. Models were trained for a maximum of 100 epochs using Adam with a learning rate of 1e-4, a batch size of 16 and early stopping (patience = 20) based on the Pearson's correlation coefficient between the predicted and measures responses on the validation set; validation curves were monitored to ensure convergence.

We measure performance ('predictive accuracy') on the 1,000 test images by computing the Pearson's correlation coefficient between the predicted and measured fMRI response at each voxel.

Baseline models

Task-optimized models : We compared response-optimized models against standard task-optimized models which have shown state-of-the-art performance in predicting neural responses in the primate visual cortex. In all comparisons, we employed an AlexNet architecture [62] optimized for object recognition on the large-scale ImageNet dataset [63]. We extracted features from intermediate layers of this network and employed the same spatial x feature factorized readout as used in the response-optimized networks to linearly map layer activations to brain voxel responses in each region. We selected the model layer that maximally predicted the brain responses in each region on a validation set(Conv-5 for all considered high-level visual areas). The readout parameters for task-optimized models were optimized independently for each visual region using the same training protocol as the response-optimized models. Thus, the readout models were trained for a maximum of 100 epochs using Adam with a learning rate of 1e-4 and a batch size of 16. We further applied an early stopping criterion (patience = 20) based on the Pearson's correlation coefficient between the predicted and measures responses on the validation set.

Categorical models : Category ideal observer models, employ the category membership of labeled objects in the image to predict the responses evoked by the image. Unlike task-optimized and the proposed response-optimized models, categorical models are not image-computable and rely on annotations generate by human observers. These oracle models have absolute access to the categories present in an image and have previously been shown to

explain substantial variance in image representations in both macaque and human IT [64, 22]. One might expect their performance to be even higher for explaining image representations in category-selective visual clusters in high-level cortex. We obtained object category labels for every NSD image from the MS COCO database [65]. The input to the categorical model is thus an 80-D binary vector corresponding to the 80 object categories annotated in the database, where each element indicates whether the corresponding category was present in the image or absent (note that NSD images contain multiple objects per image). We fitted l_2 regularized linear regression models (known as ridge regression) on this representational space to find weights corresponding to different categories for every voxel. The regularization parameter was optimized independently for each subject and for voxels in each visual area by testing among 8 log-spaced values in [1e-4, 1e4].

Quantifying the semantic selectivity of voxels

High-level visual concepts are generally verbalizable, and throughout this paper, we refer to voxels that encode and represent these concepts as 'semantically selective'. To quantify the selectivity of voxels for different human-interpretable concept categories, we adapt the previously proposed framework of 'Network dissection' to our brain response-optimized models [30, 31]. We see this as a fine-grained approach to characterize voxels that looks at not just the image-level category labels but rather dense pixel-level segmentations across thousands of cluttered natural scenes to characterize a voxel. The probe dataset used for quantifying the semantic selectivity of voxels comprises 36,500 held-out images from the validation set of the large-scale Places365 dataset. The reference segmentation for these probe images comes from the Unified Perceptual Parsing image segmentation network [66] previously trained on 20,000 scene-centric images from the ADE20k dataset [67]. The latter is exhaustively and densely annotated with objects, parts of objects and in some cases, even parts of parts. This reference segmentation assigns every pixel a semantic label from a large vocabulary of human-interpretable concepts, comprising 335 object classes, 1,452 object parts and 25 materials. A unique advantage of the factorized readout employed in this study is that it allows us to disentangle spatial selectivity from feature selectivity. To enable the network dissection procedure to be applicable to our models, we first discard the learned spatial selectivity of every voxel and use the learned feature tuning of every voxel to create an additional 1x1 convolutional layer, so that every voxel is represented by an independent unit in this convolutional layer. These units are used to characterize the semantic selectivities of voxels irrespective of the position of the respective semantic categories in the visual field. This yields a model that is entirely convolutional and dissecting the last layer of this model (which has as many units as the number of voxels in the ROI) reveals the semantic selectivities of all voxels. As proposed in [30], the selectivity of a particular unit (or voxel in our case) is quantified by computing the Intersection over Union (IoU) of the corresponding thresholded activations of that unit for a large number of images from the probe dataset against the reference segmentation. A voxel is termed as semantically selectivity for a *concept* if its IoU with the reference segmentation of that concept is greater than 0.04. Further details about the dissection procedure employed in this study are described elsewhere [31].

Synthesizing maximally exciting inputs

We performed a qualitative *feature visualization* analysis to find the visual pattern that would maximally activate individual model neurons emulating brain voxels. Neural networks are differentiable with respect to their inputs. Starting from a random noise input, we use these gradients to iteratively move the input towards the goal of maximizing activation in individual model neurons. This visualization technique is commonly employed in neural network interpretability research to find the featural drivers of model neurons [68]. Most visualization techniques further employ an *image prior* in the form of a regulariser to restrict the maximally exciting input to a suitable subset of the image space [69, 70]. Formally, the goal of finding the maximally exciting input (MEI) x^* is then expressed as the following optimization problem.

$$x^* = \underset{x \in R^{H \times W \times C}}{\arg\max}\ A_{ij}(\theta, x) + \mathcal{R}(x)$$

where $A_(i, j)(\theta, x)$ denotes the activation of unit i from layer j in the neural network to input x (H: Height, W: Width, C: Channels), and θ denotes the parameters of the network. The latter are fixed during the above optimization procedure. $\mathcal{R}(x)$ denotes regulariser . In order to generate MEI for the jth voxel, we set i to the network output layer and j to be the index of the model neuron in the output layer that emulates voxel j. The above optimization problem is, in general, a non-convex optimization problem but we can find (at the very least) a local minimum by performing gradient ascent in the input space and updating x iteratively in the direction of the gradient of $A_{ij}(\theta, x) + \mathcal{R}(x)$

Transfer learning on fine-grained visual disrimination tasks

To formally test existing functional specialization accounts which implicate FFA in face perception and RSC in spatial cognition, we simulated face discrimination and spatial layout prediction tasks with independent stimuli in response-optimized models of all brain regions. For the face identity discimination task, we included stimuli from the MiniCelebA dataset which comprises facial images of 20 identities, each having 100/30/30 train/validation/test images [47]. For the spatial layout estimation task, the stimuli include 4,000 diverse indoor scenes from the SUN database [51]. These stimuli were split into sets of 3200 training, 400 validation and 400 test images. Each stimulus image has a corresponding label for the room layout type, where the layout categories were defined using a keypoint-based parametrization (as illustrated in Figure 7). This helps us frame the room layout estimation task as a classification problem. We use response-optimized models to extract the *predicted* responses of voxels in every region for stimuli from these fine-grained visual categorization tasks. To ensure that the differences in performance of response-optimized models on fine-grained visual categorization are not driven by the differences in the number of voxels in every region, we selected the top 512 voxels in every region based on test correlations (i.e., correlation between the predicted and measured responses on 1,000 test images). This number was chosen to match the dimensionality of representations from the pre-trained VGG16 architecture [49] optimized for face-recognition on the large-scale VGGFace2 dataset [71]. We consider the performance of this representation on the MiniCelebA dataset as an estimate of the upper

bound on performance expected by these models on face recognition as we did not have access to human face recognition performance on this dataset. We also consider the 512-D dimensional representation from a VGG16 architecture trained on image categorization using the large-scale ImageNet database as an additional baseline for both visual discrimination tasks. This baseline was chosen because ImageNet-trained networks yield highly transferable representations that perform well in a range of vision-based tasks [72]. We fitted l_2 regularized linear classification models (known as ridge classifiers) on these different representational spaces to predict the class label (e.g. facial identity or room layout type) of held-out stimuli. We vary the size of each transfer dataset from 60-100% of the maximum training set size and report the corresponding classification accuracy on the held-out set as a function of the transfer dataset size. For each of the above representational models (response-optimized, ImageNet-optimized or VGGFace2-optimized), the regularization parameter was optimized independently for each task (face recognition or spatial layout estimation) and each training set size (60-100%) by testing among 10 log-spaced values in [1e-5, 1e5]. We selected the regularization parameter value that yielded best classification accuracy on the validation dataset.

Noise ceiling estimation

Imperfect predictions of models are not solely due to model imperfections, but may arise due to the inherent noise in the fMRI signal, which biases the prediction accuracy downward. Noise ceiling for every voxel represents the performance of the "true" model underlying the generation of the responses (the best achievable accuracy) given the noise in the measurements. They were computed using the standard procedure followed in [27] by considering the variability in voxel responses across repeat scans. The dataset contains 3 different responses to each stimulus image for every voxel. In the estimation framework, the variance of the responses, $\sigma^2_{\text{response}}$, are split into two components, the measurement noise σ^2_{noise} and the variability between images of the noise free responses σ^2_{signal}.

$$\hat{\sigma}^2_{\text{response}} = \hat{\sigma}^2_{\text{signal}} + \hat{\sigma}^2_{\text{noise}}$$

An estimate of the variability of the noise is given as $\hat{\sigma}^2_{\text{noise}} = \frac{1}{n}\sum_{i=1}^{n} \text{Var}(\beta_i)$, where i denotes the image (among n images) and $\text{Var}(\beta_i)$ denotes the variance of the response across repetitions of the same image. An estimate of the variability of the noise free signal is then given as,

$$\hat{\sigma}^2_{\text{signal}} = \hat{\sigma}^2_{\text{response}} - \hat{\sigma}^2_{\text{noise}}$$

Since the measured responses were z-scored, $\hat{\sigma}^2_{\text{response}} = 1$ and $\hat{\sigma}^2_{\text{signal}} = 1 - \hat{\sigma}^2_{\text{noise}}$. The noise ceiling (n.c.) expressed in correlation units is thus given as $n.c. = \sqrt{\frac{\hat{\sigma}^2_{\text{signal}}}{\hat{\sigma}^2_{\text{signal}} + \hat{\sigma}^2_{\text{noise}}}}$. The models were evaluated in terms of their ability to explain the average response across 3 trials (i.e., repetitions) of the stimulus. To account for this trial averaging, the noise ceiling is expressed as $n.c. = \sqrt{\frac{\hat{\sigma}^2_{\text{signal}}}{\hat{\sigma}^2_{\text{signal}} + \hat{\sigma}^2_{\text{noise}}/n}}$. We computed noise ceiling using this formulation for every voxel in each subject and expressed the noise-normalized prediction accuracy (R) as a percentage of this noise ceiling.

References

[1] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[2] Jack L Gallant, Jochen Braun, and David C Van Essen. Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science*, 259(5091):100–103, 1993.

[3] Corey M Ziemba and Jeremy Freeman. Representing "stuff" in visual cortex. *Proceedings of the National Academy of Sciences*, 112(4):942–943, 2015.

[4] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

[5] Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

[6] Mark A Pinsk, Kevin DeSimone, Tirin Moore, Charles G Gross, and Sabine Kastner. Representations of faces and body parts in macaque temporal cortex: a functional mri study. *Proceedings of the National Academy of Sciences*, 102(19):6996–7001, 2005.

[7] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[8] Shahin Nasr, Ning Liu, Kathryn J Devaney, Xiaomin Yue, Reza Rajimehr, Leslie G Ungerleider, and Roger BH Tootell. Scene-selective cortical regions in human and nonhuman primates. *Journal of Neuroscience*, 31(39):13771–13785, 2011.

[9] Andrew H Bell, Nicholas J Malecek, Elyse L Morin, Fadila Hadj-Bouziane, Roger BH Tootell, and Leslie G Ungerleider. Relationship between functional magnetic resonance imaging-identified regions and neuronal category selectivity. *Journal of Neuroscience*, 31(34):12229–12240, 2011.

[10] Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

[11] Marius V Peelen and Paul E Downing. Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, 93(1):603–608, 2005.

[12] Kevin S Weiner and Kalanit Grill-Spector. Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *Neuroimage*, 52(4):1559–1573, 2010.

[13] Linda L Chao, James V Haxby, and Alex Martin. Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature neuroscience*, 2(10):913–919, 1999.

[14] Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

[15] Krishna Srihasam, Joseph B Mandeville, Istvan A Morocz, Kevin J Sullivan, and Margaret S Livingstone. Behavioral and anatomical consequences of early versus late symbol training in macaques. *Neuron*, 73(3):608–619, 2012.

[16] Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

[17] Michael C-K Wu, Stephen V David, and Jack L Gallant. Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, 29:477–505, 2006.

[18] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[19] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

[20] Mark D Lescroart and Jack L Gallant. Human scene-selective areas represent 3d configurations of surfaces. *Neuron*, 101(1):178–192, 2019.

[21] Stephen V David and Jack L Gallant. Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, 16(2-3):239–260, 2005.

[22] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[23] N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):1–14, 2021.

[24] Aria Wang, Michael Tarr, and Leila Wehbe. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Michael Oliver and Jack Gallant. A deep convolutional energy model of v4 responses to natural movies. *Journal of Vision*, 16(12):876–876, 2016.

[26] Katja Seeliger, Luca Ambrogioni, Yağmur Güçlütürk, Leonieke M van den Bulk, Umut Güçlü, and MAJ van Gerven. End-to-end neural system identification with neural information flow. *PLOS Computational Biology*, 17(2):e1008558, 2021.

[27] Emily Jean Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Logan T Dowdle, Bradley Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, et al. A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *bioRxiv*, 2021.

[28] Ghislain St-Yves, Emily J. Allen, Yihan Wu, Kendrick Kay, and Thomas Naselaris. Brain-optimized neural networks learn non-hierarchical models of representation in human visual cortex. *bioRxiv*, 2022.

[29] Korbinian Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth, 1909.

[30] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[31] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

[32] David H Hubel, Torsten N Wiesel, and Michael P Stryker. Anatomical demonstration of orientation columns in macaque monkey. *Journal of Comparative Neurology*, 177(3):361–379, 1978.

[33] RB Tootell and Susan L Hamilton. Functional anatomy of the second visual area (v2) in the macaque. *Journal of Neuroscience*, 9(8):2620–2644, 1989.

[34] Xiaomin Yue, Irene S Pourladian, Roger BH Tootell, and Leslie G Ungerleider. Curvature-processing network in macaque visual cortex. *Proceedings of the National Academy of Sciences*, 111(33):E3467–E3475, 2014.

[35] Alexander S Ecker, Fabian H Sinz, Emmanouil Froudarakis, Paul G Fahey, Santiago A Cadena, Edgar Y Walker, Erick Cobos, Jacob Reimer, Andreas S Tolias, and Matthias Bethge. A rotation-equivariant convolutional neural network model of primary visual cortex. *arXiv preprint arXiv:1809.10504*, 2018.

[36] David A Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating what and where. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3509–3519, 2017.

[37] Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6):254–262, 2011.

[38] Michael F Bonner and Russell A Epstein. Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, 114(18):4793–4798, 2017.

[39] Anna S Mitchell, Rafal Czajkowski, Ningyu Zhang, Kate Jeffery, and Andrew JD Nelson. Retrosplenial cortex and its role in spatial cognition. *Brain and neuroscience advances*, 2:2398212818757098, 2018.

[40] Shahin Nasr, Cesar E Echavarria, and Roger BH Tootell. Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *Journal of Neuroscience*, 34(20):6721–6735, 2014.

[41] Jean-Rémi King and Stanislas Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210, 2014.

[42] Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals new aspects of meaning composition. *bioRxiv*, 2020.

[43] Mariya Toneva, Jennifer Williams, Anand Bollu, Christoph Dann, and Leila Wehbe. Same cause; different effects in the brain. *Proceeedings of the conference on Causal Learning and Reasoning*, 2022.

[44] Ido Tavor, O Parker Jones, Rogier B Mars, SM Smith, TE Behrens, and Saad Jbabdi. Task-free mri predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220, 2016.

[45] Kalanit Grill-Spector, Nicholas Knouf, and Nancy Kanwisher. The fusiform face area subserves face perception, not generic within-category identification. *Nature neuroscience*, 7(5):555–562, 2004.

[46] Josef Parvizi, Corentin Jacques, Brett L Foster, Nathan Withoft, Vinitha Rangarajan, Kevin S Weiner, and Kalanit Grill-Spector. Electrical stimulation of human fusiform face-selective regions distorts face perception. *Journal of Neuroscience*, 32(43):14915–14920, 2012.

[47] Dat Thanh Tran, Serkan Kiranyaz, Moncef Gabbouj, and Alexandros Iosifidis. Pygop: A python library for generalized operational perceptron algorithms. *Knowledge-Based Systems*, 182:104801, 2019.

[48] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[49] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[50] Yinda Zhang, Fisher Yu, Shuran Song, Pingmei Xu, Ari Seff, and Jianxiong Xiao. Large-scale scene understanding challenge: Room layout estimation. In *CVPR Workshop*, 2015.

[51] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[52] Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

[53] Zijin Gu, Keith Wakefield Jamison, Meenakshi Khosla, Emily J Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, Mert R Sabuncu, and Amy Kuceyeski. Neurogen: activation optimized image synthesis for discovery neuroscience. *NeuroImage*, 247:118812, 2022.

[54] Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

[55] Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

[56] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

[57] Yoichi Sugita. Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences*, 105(1):394–398, 2008.

[58] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8730–8738, 2018.

[59] J Swaroop Guntupalli, Michael Hanke, Yaroslav O Halchenko, Andrew C Connolly, Peter J Ramadge, and James V Haxby. A model of representational spaces in human cortex. *Cerebral cortex*, 26(6):2919–2934, 2016.

[60] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018.

[61] Maurice Weiler and Gabriele Cesa. General $e(2)$-equivariant steerable cnns. *arXiv preprint arXiv:1911.08251*, 2019.

[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[63] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[64] Kamila Jozwik, Nikolaus Kriegeskorte, Radoslaw Martin Cichy, and Marieke Mur. Deep convolutional neural networks, features, and categories perform similarly at explaining primate high-level visual representations. 2019.

[65] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[66] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[67] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[68] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[69] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[70] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[71] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[72] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.