Characterizing the Ventral Visual Stream with Response-Optimized Neural Encoding Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Decades of experimental research based on simple, abstract stimuli has revealed the coding principles of the ventral visual processing hierarchy, from the presence of edge detectors in the primary visual cortex to the selectivity for complex visual categories in the anterior ventral stream. However, these studies are, by construction, constrained by their *a priori* hypotheses. Furthermore, beyond the early stages, precise neuronal tuning properties and representational transformations along the ventral visual pathway remain poorly understood. In this work, we propose to employ response-optimized encoding models trained solely to predict the functional MRI activation, in order to gain insights into the tuning properties and representational transformations in the series of areas along the ventral visual pathway. We demonstrate the strong generalization abilities of these models on artificial stimuli and novel datasets. Intriguingly, we find that response-optimized models trained towards the ventral-occipital and lateral-occipital areas, but not early visual areas, can recapitulate complex visual behaviors like object categorization and perceived image-similarity in humans. We further probe the trained networks to reveal representational biases in different visual areas and generate experimentally testable hypotheses. Our analyses suggest a shape-based processing along the ventral visual stream and provide a unified picture of multiple neural phenomena characterized over the last decades with controlled fMRI studies.

1 Introduction

Understanding the nature of representations and computations in the visual system has been a longstanding goal in neuroscience. Substantial progress been made in this quest for early stages of the visual cortex by presenting model organisms with simple, abstract stimuli like noise or sine-wave gratings and studying the evoked responses, such as the peak of their tuning curves. These carefully crafted experiments, for instance, revealed the presence of edge detectors in the primary visual cortex [1] and sparked the search for the 'optimal input' beyond early visual areas across the visual cortical hierarchy. Over the last two decades, through numerous experiments, recording and analysis techniques, a conceptual understanding has emerged that early visual areas extract low-level features like edges or curves, mid-level regions extract complex local shapes, and high-level regions encode and represent different semantic categories, like faces or scenes [2, 3, 4, 5, 6]. While the underlying representational transformations that enable flexible and efficient visual perception remain unknown, there is accumulating evidence that visual categorization occurs in the ventral temporal cortex in humans [7]. This understanding of the visual system was enabled by synthesizing findings across several different hypothesis-driven experimental studies, most of them employing artificial stimuli atypical of real-word vision, thus lacking ethological validity.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Recently, deep convolutional neural networks (CNNs) optimized for computer vision (CV) tasks, have emerged as the most accurate models of the primate ventral visual stream [8]. Furthermore, these CV task-optimized models offer a hierarchical correspondence to visual areas along the ventral stream: early CNN layers best predict V1, while intermediate and late layers best predict V4 and IT [8, 9]. They provide intelligibility in the form of the goal-driven perspective, highlighting the functional role of representations [10]. However, we are far from a complete understanding of the top-down computational objectives (the battery of ethologically-relevant tasks) that may have shaped neural representations through evolution or experience-based learning; in the absence of right hypotheses, the task-optimized approach can fall short.

In this study, we leverage recent advances in large-scale fMRI data collection [11] to train hypothesis-agnostic computational models directly on neural data with novel predictive precision. Importantly, the stimulus set employed in this study contains crowded images of multiple common objects in their natural contexts, thus being more typical of everyday scenes and allowing us to characterize neural representations and computations in rich, naturalistic conditions. We demonstrate the remarkable ability of these "response-optimized" neural encoding models in reproducing the stimulus-response relationship in voxels along the ventral visual stream for arbitrary images, equaling or surpassing the unprecedented prediction performance of task-optimized models. We show that these models generalize well to contexts that are vastly different from their training conditions (artificial stimuli, novel subjects, etc.), providing a strong quantitative validation for the response-optimization approach of building encoding models for neurobiological visual systems beyond the retina [12]. Our approach in this paper, however, goes beyond prediction accuracy and aims to gain insights about neuronal tuning properties. Specifically, we implement neural encoding models that disentangle the processing of the "what" and the "where." This allows us to reveal the selectivity of individual brain voxels along both location and appearance dimensions. We first demonstrate that the estimated spatial topographic organization from our neural encoding models exhibits remarkable agreement with the fine-scaled retinotopic organization of the cortex, which is typically measured with dedicated, painstaking experimental paradigms. We analyze the features that emerge in the neural encoding models to understand the tuning properties of cortical voxels. This analysis suggests shape-biased tuning across the ventral visual stream, increasing in complexity from Gabor-like features in V1, curves and angles in V2 to concentric (circular) and complex shape processing in V4 and higher-order lateral-occipital and ventral-occipital ROIs respectively. We also go beyond single-voxel analysis to characterize how *representations* evolve along the ventral visual pathway. Probing the information encoded in model-predicted patterns of activity reveals the monotonically increasing separability of category information along the ventral stream hierarchy, strongly supporting the computational goal of the ventral visual stream as a visual categorization system. We further demonstrate the correspondence between model representations (of high-level visual areas) and mental representations (as assessed with human similarity judgements), suggesting that these response-optimized models inherently capture human psychological judgements, without any explicit supervision to do so. Furthermore, a principle components analysis reveals that models of high-level visual ROIs capture complex stimulus dimensions like faces, bodies and shapes.

Our paper illustrates how neural encoding models trained on hypothesis-agnostic, yet naturalistic datasets can be used to confirm neural phenomena characterized painstakingly using controlled experimental paradigms that might be hard to replicate in individual subjects. More importantly, however, we show how *in silico experiments* on neural encoding models can generate novel hypotheses and provide a powerful framework for building broad, generalizable theories about neural information processing.

2 Materials and Methods

Natural Scenes Dataset A detailed description of the Natural Scenes Dataset (NSD [1]) is provided elsewhere [11] (see also Supplemental Methods). The NSD dataset contains measurements of fMRI responses from 8 participants who each viewed 9,000–10,000 natural scenes (22,000–30,000 trials) over the course of 30–40 sessions. A special set of 1,000 images were shared across subjects; the remaining were mutually exclusive across subjects. We focused on modeling responses within 5 visual cortical ROIs in the study. Three ROIs belonging to the retinoptic early visual cortex, namely, V1v, V2v and hV4 were defined using a population receptive field (pRF) localizer scan session for each

[1] http://naturalscenesdataset.org

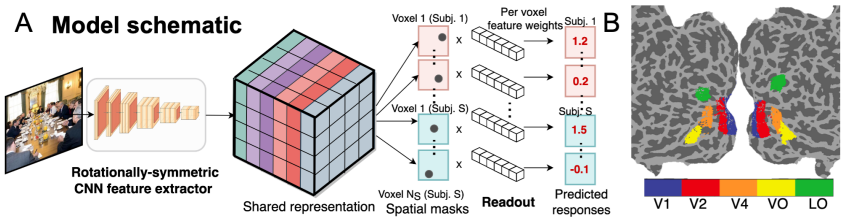

Figure 1: **A** Schematic of the neural encoding model architecture. **B** Zoomed-in view of the visual cortex on flatmap, showing the 5 visual ROIs.

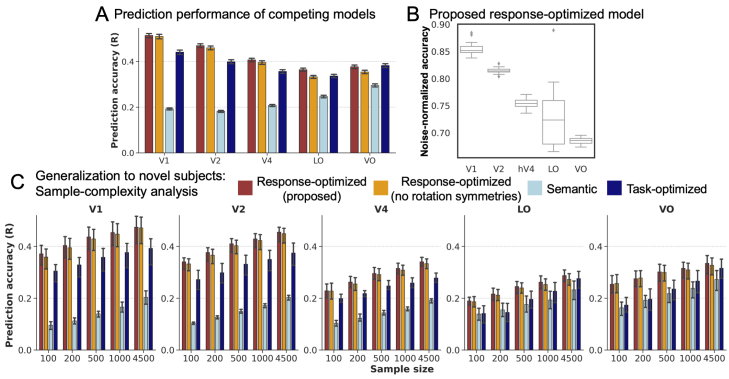

Figure 2: **Quantitative analysis. A** Prediction performance of competing models across the 5 visual areas. **B** Boxplot depicting the noise-normalizing predictivity of the proposed response-optimized model. **C** Generalization of competing models to novel subjects as a function of sample size used for fitting their respective response functions (or readouts). Error bars in barplots depict the 95% confidence-interval (CI) around estimated mean.

subject [11]. Two higher order ROIs, namely ventral-occipital areas (VO1-2) and lateral-occipital areas (LO1-2) were delineated using a visual probabilistic atlas [13](Figure 1B).

Response-optimized neural encoding model We trained separate voxel-level neural encoding models for each of the above regions with the same backbone architecture. Keeping the architecture constant across visual areas ensured that differences in emergent representations are driven mainly by the nature of response patterns. The predictive model comprises a convolutional neural network (CNN) *feature extractor* shared across all subjects, but unique to 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. The linear readout is *factorized* into *spatial* and *feature* dimensions following [14] (Figure 1A). This allows us to separate spatial tuning (the "where") from feature tuning (the "what"). The shared CNN feature extractor consists of four sequential convolutional blocks, with each block comprising the following feedforward computations: two 3×3 convolutional layers, each followed by an inner batch norm, and a ReLU; and an anti-aliased AvgPool operation (stride = 2) at the end. Instead of regular convolutions, we employ E(2)-steerable convolutions in all our models to extract features invariant to orientation [15]. This modeling choice is inspired by neural computations in early visual areas where groups of neurons are known to perform similar computations, e.g., edge or curve detection, at different orientations. Each convolutional layer contains 48 feature sets extracted at 8 orientations. Weights of the readout are a computed as outer products between a spatial filter and a feature vector. The spatial filter further has a positivity constraint (enforced using rectification) and is length-normalized for each voxel.

Training and testing response-optimized encoding models We used 4 NSD subjects for training and reserved the other 4 subjects for studying generalization. The 4 training subjects collectively saw 37,000 natural scene images, including the 1,000 shared images. 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 neural encoding model were optimized jointly to minimize the masked *mean squared error* between the predicted and measured fMRI response. This loss allows us to propagate errors through the shared CNN even if the subjects are not exposed to common stimuli since we can exclude the subjects/voxels for which the target response is not present from the loss calculation. We quantify 'predictive accuracy' on the test images as the Pearson's correlation coefficient between the predicted and measured fMRI response at each voxel. We also compare this against the noise-ceiling (see Suppl. Methods for further details).

Baseline models **(a) Semantic Model:** We fit a non-image-computable semantic model using category annotations for NSD images (obtained from the MS-COCO database). The input to the categorical model is 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. We fitted l_2 regularized linear regression models on this semantic space to predict voxel responses for held-out stimuli (see Suppl. Methods). **(b) Task-optimized Model:** We followed the standard procedure for fitting task-optimized models to voxel responses. We employed an AlexNet architecture [16] optimized for object recognition on the large-scale ImageNet dataset [17]. We extracted features from pre-trained layers of this network (conv1-fc7) and fitted l_2 regularized linear regression models separately for each layer. The regularization parameter was optimized independently for each subject, each layer and for voxels in each visual area by testing among 8 log-spaced values in [1e-4, 1e4]. For each visual area, we identified the best layer that predicts responses (averaged across all voxels and subjects) on the validation set: conv3 for V1, conv4 for V2 and V4, and fc7 for LO and VO. These models were used to quantify test performance. **(c) Response-optimized (no rotation symmetries):** Here, we train response-optimized models with the same architecture as the proposed models but without sharing weights across filter orientations.

Generalization to new subjects We tested the generalization of response-optimized models to 4 test subjects in NSD by freezing the weights of the shared CNN and only training linear readouts. We varied the size of the training set for optimizing linear readout weights from a mere 100 stimulus-response pairs to a large set of 4,500 pairs from novel subjects (see Suppl. Methods).

Characterizing spatial tuning of early visual areas The factorized readout allow us to separate spatial and feature selectivity. We assessed whether the spatial selectivity of our model trained on natural stimuli agrees with measurements from carefully controlled experiments that use artificial stimuli. The pRF is defined as the region of the visual field within which a stimulus results in increased aggregated activity across a population of neurons, as reflected in fMRI. It is commonly estimated in a retinotopy experiment using pRF localizer scans. Based on the analysis of these data, each brain voxel's retinotopic organization can then be described with three parameters: the pRF location in polar coordinates (polar angle and eccentricity), and size. Using the voxel-wise spatial mask of the readout in our encoding model, we estimated the pRF location as the average of the coordinates weighted by the learned spatial mask value at each grid location. To compare the estimated eccentricity against the corresponding measurements from pRF localizers, henceforth called *'eccentricity agreement'*, we calculated the Pearson's correlation coefficient between the measured and predicted eccentricity arrays in each ROI. To compare the estimated polar angle values against corresponding measurements from pRF localizer scans, henceforth called *'polar angle agreement'*, we adopted a measure from circular statistics called circular correlation coefficient [18], due to the circular nature of angular data (see Supp. Methods).

Characterizing neural representations by behavioral comparisons: **(a) Assessing separability of category and semantic information along the visual hierarchy** In addition to probing the learned readout layer that gives insights into spatial tuning properties of voxels, we were interested in the analysis of the response-optimized feature extractor for each visual area to understand *featural* tuning. To achieve this, we used representational similarity analysis (RSA) [19]. Here, we correlated the patterns of predicted activity between multiple exemplars from different categories of stimuli to obtain a representational similarity matrix (RSM) for each visual area. Further, we quantified the separability of categorical information within these similarity matrices by computing a correlation coefficient (Kendall's τ) between model RSMs against a ground truth adjacency matrix defined by object category labels. Specifically, the elements of this matrix are 1 if the corresponding two images belong to different categories and 0 if the images belong to the same category. This analysis was first performed using a subset of 104 images from the THINGS database [20] that belong to a pre-defined set of categories for ease of visualization. These categories were chosen in accordance with previous fMRI studies employing RSA [21, 22] and include the following: {face, hand, elephant, cat, plant, fruit, car, tool}. Subsequently, we also validated the obtained results using another large-scale vision dataset, namely ImageNet-16 [23]. We sampled 500 stimuli from each ImageNet-16 category, resulting in a total of 8,000 images and computed the similarity between model RSMs and the categorical adjacency matrix following the same approach described for the THINGS dataset. To assess visual information flow in relation to high-level semantic understanding, we computed pairwise similarity between predicted responses for held-out NSD images to construct

model-predicted RSMs for each visual area. We constructed a semantic RSM by correlating the 80-D binary category vectors (which indicate whether each category was present in the image or not) for every pair of these NSD images and quantified agreement with model-predicted RSMs using Kendall's τ (henceforth, termed the 'NSD: Semantic' analysis). We computed this agreement using *measured* RSMs for all visual areas as well. In RSA described here and anywhere else, error bars (95% CI) are computed by performing bootstrap sampling over voxels.

(b) Assessing alignment with human perception We compared model representations against behavioral data from a Multiple Arrangement task where participants interpreted 100 NSD images and arranged them according to the similarity of their content (this dataset called 'nsdmeadows' was released with NSD [11]). Critically, these images were not seen by any of the response-optimized models. We constructed a model Representational Dissimilarity Matrix (RDM) for each visual ROI using the correlation distance metric for quantifying pairwise dissimilarities between model-predicted voxel responses. We compared model RDMs against human pairwise dissimilarity judgements using the Kendall's τ coefficient and call this the 'NSD: perception' analysis.

Characterizing shape bias of model representations We first train a linear layer on top of the predicted responses of response-optimized models to map these responses to visual categories in the ImageNet-16 database using a small subset of this database. Next, we evaluate the classification performance of these linear classifiers on held-out images from the ImageNet-16 dataset and Silhouettes dataset [24]. Finally, we quantify the shape bias of all models using 1,200 images from the texture-shape cue conflict dataset [24] (see Suppl. Methods).

Characterizing voxel tuning through feature visualization We performed a *feature visualization* analysis to find visual patterns that would maximally activate individual model neurons emulating brain voxels in different response-optimized encoding models. Following [25], we start with a random noise input x_0 for each model neuron i and iteratively update the input along the gradient $\frac{\partial a_i}{\partial x}$ to synthesize inputs that would result in higher and higher predicted activation a_i for that neuron (see Suppl. Methods).

3 Results

Response-optimized models approach the noise ceiling for prediction performance and generalize to novel subjects with low sample complexity We observed a remarkably high noise-normalized prediction accuracy ($\sim$70-85%) using the proposed rotationally-symmetric response-optimized model in all 5 visual ROIs, including the higher-order ROIs (LO1-2 and VO1-2) (Figure 2B). Next, we investigated the possible quantitative advantages of learning a rotationally-symmetric feature space against the representations from a standard convolutional architecture with no weight sharing across orientations. Importantly, we observe that the proposed rotationally-symmetric convolutional architecture performs on par with a standard CNN (Figure 2A), even outperforming the latter in certain visual ROIs, despite containing fewer parameters. This suggests that sharing weights across filter orientations provides a strong inductive bias, allowing us to fit expressive models efficiently. We further compared these models against semantic and task-optimized models, and found that response-optimized models consistently outperform these baselines in early visual areas. In high-level visual areas, the response-optimized model outperforms the semantic model, while performing on par with task-optimized model.

Next, we assessed how each of these predictive models generalizes to the remaining set of 4 test subjects. This analysis intended to compare the statistical efficiency afforded by different representations [26]. As such, we examined the trend of how much stimulus-response data the models need to achieve a certain level of prediction performance [26]. Here, we freeze the shared feature extractor network and vary the amount of stimulus-response pairs from the 4 new subjects to train the linear readouts. We do the same for the task-optimized baseline, where the linear prediction model is fit on new subjects. As shown in Figure 2, the response-optimized models can achieve a significantly high accuracy even when only just 5% of the training examples from novel subjects ($\sim$ 500 samples) are used to train the readout. This remarkable generalization of response-optimized networks, but not task-optimized networks, to novel subjects suggests our models contain appropriate "inductive biases" that improve sample efficiency.

Computational models generalize to OOD datasets We next investigated out-of-distribution (OOD) test settings to understand the generalization behavior of response-optimized models. The first OOD setting we consider is the NSD-synthetic experiment, comprising fMRI recordings to synthetic stimuli (samples shown in Figure 3) from the same NSD subjects. As shown in Figure 3, response-optimized models significantly outperform task-optimized models in this OOD domain ('artificial' stimuli), suggesting that the mapping from task-optimized DNN features to voxel responses, was domain/dataset-specific. Response-optimized models, on the other hand, allow for flexible OOD generalization. Generalization to other novel fMRI datasets, including [27, 28, 29] is shown and discussed in the Suppl. Results.

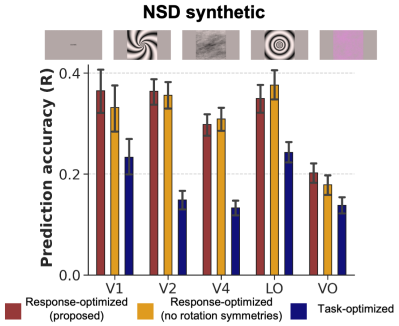

Figure 3: **Generalization to synthetic stimuli.** OOD generalization of competing models to synthetic (non-naturalistic) stimuli from *nsdsynthetic* experiment. Sample stimuli are shown on the top. Error bars depict the 95% CI around estimated mean.

Learned spatial masks reproduce the fine-grained retinotopic organization of early visual cortex Next, we estimate the population receptive field from the learned spatial masks and compare the estimated parameters against the measurements from the independent pRF localizer scan session. It is important to note that the encoding models do not account for the spatial locations of target voxels. Despite this, the estimated spatial topographic organization from the encoding models exhibits remarkable levels of agreement with the fine-scaled retinotopic organization of the cortex without any explicit supervision to do so (Figure 4A,B): *eccentricity agreement* $\in$ [0.48-0.79] for all 4 subjects whereas the *polar angle agreement* $\in$ [0.70-0.82]. We further compared this agreement on the estimated pRF parameters for the held-out set of 4 NSD subjects by varying the quantity of data used to estimate their linear readouts from the fixed convolutional core of the response-optimized model (pre-trained on the initial set of 4 subjects). We find that while the agreement increases with the sample size used to fit readouts, even with much fewer samples (e.g. 100), we can still characterize their retinotopic organization (Figure 4C). Previously, retinotopic organization of the human visual cortex has been dominantly studied with spatially modulated stimuli. While encoding models have been shown to be broadly useful for characterizing retinotopy [30, 31], a precise quantification of the agreement between model-estimated and localizer-estimated retinotopic parameters, as well as the effect of sample sizes on this agreement, has been lacking. Here, we present an alternate approach based on naturalistic stimulation and response-optimized encoding models to delineate subject-specific retinotopic maps and precisely quantify the agreement. Remarkably, the retintopic organization estimated from the proposed encoding models is even more spatially smooth than the localizer-based organization, despite not being informed by anatomy.

Response-optimized models characterize transformations of image representations along the visual cortical hierarchy

(a) Monotonically increasing alignment with human perception along the ventral visual stream We evaluated models on behavioral data from a Multiple Arrangement task where participants arranged NSD images according to the similarity of their content. This behavioral analysis of neural network models (Figure 5A) suggests that higher visual areas (LO and VO) more strongly capture human perceptual patterns and image semantics than early visual areas and this quantitative agreement increases monotonically along the ventral visual hierarchy ('NSD perception' and 'NSD semantics' analysis discussed in Methods). Importantly, responses patterns in their respective models group semantically similar categories together, as shown in the t-SNE plot for model-predicted RDM (region VO). For instance, faces are all clustered in the bottom right of the t-SNE plot and scenes group together in the left (Figure 5A), despite wide variations in the visual appearance of these categories. This is remarkable, given that the models received no supervision from perceptual/semantic measures. Human behavioral data ('nsdmeadows') appears very similar and is shown in Suppl. Results. This suggests that these models, solely optimized for high-level visual cortical response prediction – without any semantic supervision – can even serve as powerful models for uncovering the structure

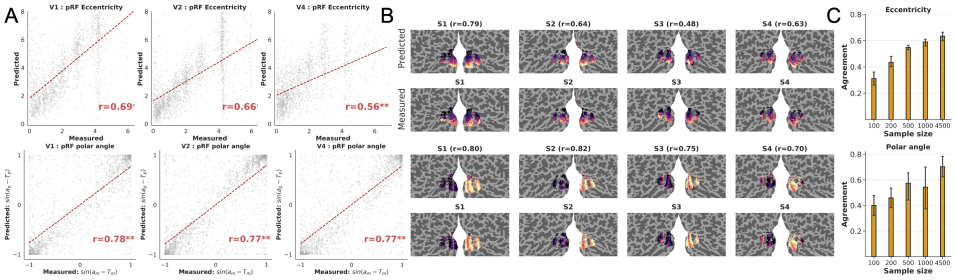

Figure 4: **pRF estimation with response-optimized encoding models**. **A** Scatter plots showing predicted and localizer-estimated retinotopic parameters for all voxels in all 4 subjects (see Suppl. Methods for polar angle metrics shown on scatter plots). Inset correlations are computed using voxels from all 4 subjects. **B** Predicted and localizer-estimated retinotopic parameters visualized on cortical flatmaps. Bolded values on top show subject-specific agreement. **C** Agreement between estimated and measured retinotopic maps as a function of training examples (stimulus-response pairs) from novel subjects used to train their linear readout. Error bars depict the 95% CI around estimated mean.

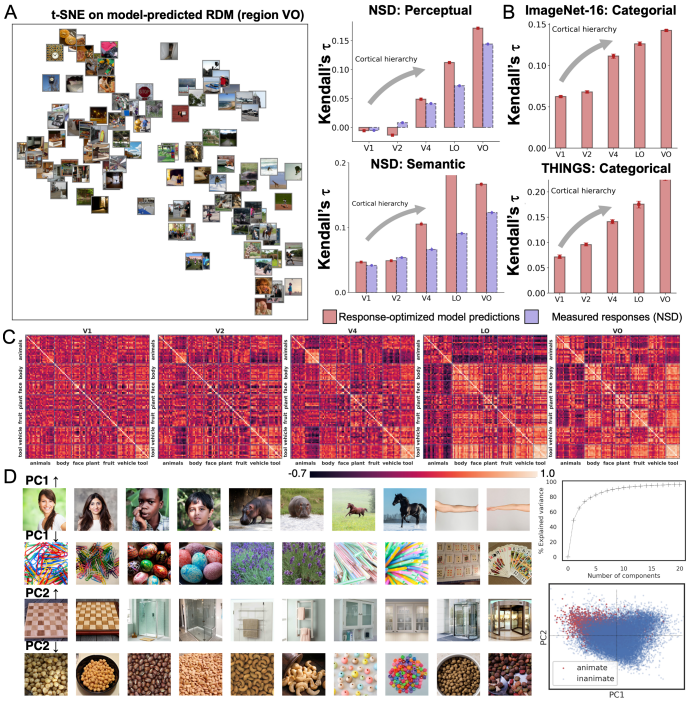

Figure 5: **A** highlights that models of high-level visual areas (LO and VO) are more aligned with human perception and image semantics than early visual ROI (V1-V4) models. **B** and **C** highlight the **separability of category information** along the ventral visual stream. **B** Results of the Representational Similarity Analysis (RSA) framework applied to the THINGS and ImageNet-16 datasets containing different categories of objects. **C** Matrices of all pairwise similarities between model-predicted responses for sampled images in the THINGS dataset. **D** [Left] Top 2 images for the 5 most and least activating concepts of the first two PCs of model-predicted **VO** responses. [Upper right] Total explained variance as a function of the number of PCs. [Lower right] All images from the THINGS dataset projected onto the first two principal dimensions of the response and colored by their labels.

of mental representations. Intriguingly, the representational structure of predicted responses exhibits stronger agreement with perceived similarity than the representational structure of measured responses. This suggests that both the perceptual and semantic information are likely more explicit in predicted responses (even for held-out data) than measured responses. One possible explanation is that fMRI measures of brain activity have undesired noise components and predictive models, in particular, response-optimized models as shown here, can *denoise* this fMRI activity.

(b) Monotonically increasing separability of category information along the ventral visual stream Task-optimized computational models have suggested that the ventral stream hierarchy 'untangles' representations through a sequence of processing steps, so that representations of objects and categories that are inseparable at the initial processing stage (e.g., V1) become untangled at the last stage of the hierarchy [32, 10]. Here, we present a strong evidence for this increased separability through hypothesis-agnostic response-optimized models. Previous studies have shown that the representational geometry for the same set of stimuli varies systematically across different cortical areas, providing a useful signature for how information is transformed across different cortical processing streams [32]. Here, we employ the representational similarity analysis (RSA) framework

to relate emergent representations in response-optimized models of difference brain regions against a ground truth adjacency matrix defined by object category labels. Keeping model architecture the same, simply by changing the response targets for an encoding model from voxels in visual areas V1 through LO-VO, we observe a drastic change in the geometry of the extracted representation. The increasingly prominent block-diagonal structure in the RSMs along the ventral visual stream (Figure 5C), consistent with categorical distinctions in the THINGS dataset, highlights that the distributed activity patterns to exemplars of the same category become increasingly more similar; and distributed patterns to exemplars of different categories become progressively dissimilar along the processing stream, strongly supporting the computational goal of the ventral visual stream as a visual categorization system. We also quantified the separability of categorical information in different visual areas (Figure 5B). Here again, we see that this finding regarding increasing separability of categorical information holds across both datasets containing different visual categories.

Response-optimized models reveal high-level visual information within ventral-occipital cortex Next, we wanted to discover what kinds of features emerge in the ventral-occipital region model and whether they are capable of discriminating between complex categories. Instead of looking at all model neurons emulating voxels, we built a compact neural representational space by passing a large stimulus set ($\sim$27,000 images from the THINGS database [20]) through the predictive models and then performing Principal Components Analysis (PCA) on the predicted responses of this network (see Suppl. Methods). Visualizing the images from top concept categories that elicit highest and lowest activation of individual principal components (PCs) reveals complex stimulus dimensions: the first principal component, which explained $\sim 49\%$ of the variance in responses strongly separates faces, animals and bodies from inanimate objects, despite vast variations in the visual appearance of images within sub-categories. The second principal component, accounting for $\sim 16\%$ variance, distinguished objects based on their shape (circular versus rectilinear objects), suggesting that such shape features are important for VO. Importantly, these features do not emerge when the same analysis is applied to early visual ROIs (V1-V4).

Response-optimized models exhibit a shape bias A striking known discrepancy between human and machine vision is the shape vs. texture bias: human classification decisions are shape-biased whereas CNNs trained for CV tasks tend to classify images by texture [24, 33]. We study whether the shape bias of humans is reflected in response-optimized models of cortical regions that subserve visual categorization in humans. First, we assess the transfer performance of response-optimized models for classification on held-out stimuli from the ImageNet-16 dataset as well as the Silhouette dataset described in [24] (see Suppl. Methods). We find that the classification performance is significantly better than what would be expected from the same network architecture with a random initialization (Figure 6B), suggesting that a brain response-optimization objective alone can yield effective representations. Further, this classification performance is better using models of intermediate and high-level visual areas (V4, LO and VO) than early visual areas, providing further support for the separability of categorical information in high-level, but not early visual cortex [7]. The LO model has the highest classification performance on the Silhouette dataset (which solely contain shape information), consistent with and providing further evidence for the previous hypotheses that LO tracks object shape information [34]. We then evaluate the performance of this mapped response-optimized model on the diagnostic cue conflict dataset to assess whether the model makes its categorization decisions based on shape or texture. We follow [24] and define *shape bias* as the percentage of times the model classifies an image based on its shape over texture, provided it returned either the correct shape or correct texture label for the image. A model is shape-biased if its *shape bias*$> 50\%$. We observed that *all* response-optimized models exhibit a shape bias (Figure 6B [bottom]), with this metric ranging from 56.7-69.4$\%$, suggesting that representations within the ventral visual stream are likely biased towards shapes than textures. This is in stark contrast to standard CV models which show a texture bias. Further, as shown in Figure 6B (bottom), it is not the mapping to neural response but the way features are optimized that explains the differing shape/texture bias of task-optimized and response-optimized models.

We also visualized the features in different response-optimized models using an image synthesis procedure (described in Methods). Visual inspection shows that the features increase in complexity as we go from early to high-level visual ROI models: simple edge-like features or high-low frequency detectors are seen in V1, curves and edges are seen in V2, more complex curvatures in V4 and finally, fully-formed complex shapes like concentric hexagons are seen in LO and VO (Figure 6A).

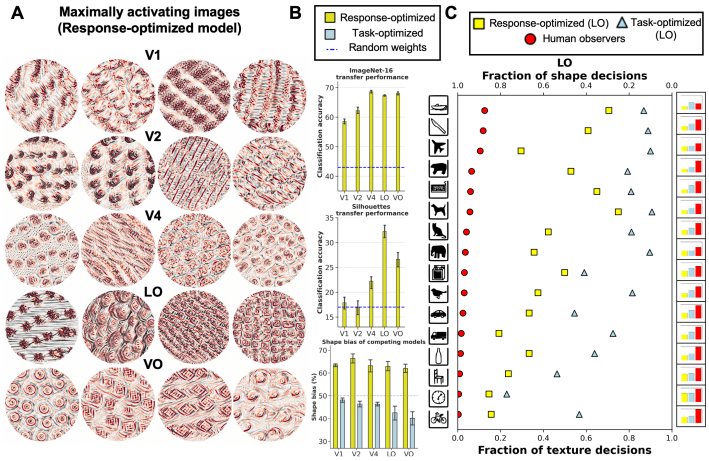

Figure 6: **A Maximally activating images** for 4 model neurons (emulating brain voxels) in each visual ROI model, generated using the image synthesis procedure. **B Shape bias analysis**. Top and middle barplots show the transfer performance of response-optimized models on the ImageNet-16 and Silhouette datasets. Bottom plot shows the shape bias of all response-optimized and task-optimized models. **C** shows the fraction of shape vs. texture decisions for stimuli with cue conflict. Bar plots on the right display the proportion of correct decisions (either shape or texture recognized correctly) as a fraction of all trials for human observers as well as response-optimized and task-optimized models of LO.

4 Discussion

Here, we set out to answer the question: How far can deep learning on neural responses alone get us in predicting neural responses to ID and OOD test sets and understanding the ventral visual pathway? Thus, far, visual neuroscience had been largely measurement-limited, constraining modeling efforts that bring neuroscience data to bear on the model development process. The large sample size of Natural Scenes Dataset presented an opportunity to change this status quo. We capitalized on the natural variation in NSD to train deep neural networks solely with supervision from brain response prediction. We studied model generalization in different settings: from synthetic stimuli to novel fMRI datasets with widely different preprocessing pipelines. We demonstrated success in recapitulating key features of early visual areas (e.g. their precise retinotopic organization), representational transformations along the ventral visual stream along with demonstrating that high-level visual features like faces, places and complex shapes are represented in high-level visual areas. Remarkably, we also demonstrated that despite no supervision from human behavioral measures, these models can capture human behavioral abilities, including not only object categorization, but also other fine-grained measures such as perceived image similarity. These hypothesis-agnostic models, we believe, can complement task-optimized models in furthering our understanding of human vision; the response-optimization approach, in principle, allows exploring infinite paths in the model space, and when optimized properly, converging towards one that explains maximum possible variance in the data. Since any features that emerge in these optimized networks are a result of optimization to match neural targets, unconfounded by any other supervision signal, these models can provide stronger tests about the visual representations within different visual areas.

In terms of human object categorization behavior (as tested on ImageNet-16), our models lack significantly compared to ImageNet-supervised DNN models. While part of this may be due to the domain shift (NSD/COCO $\rightarrow$ ImageNet), the restricted transfer learning setting (a single linear layer to map predicted voxel responses to category labels), small size of the transfer set or simply the coarse spatial resolution of fMRI, this gap is still worth exploring. In this study, we focused on the use of DNN models in understanding the ventral visual stream but using neural data to build better models for other communities (e.g. computer vision) is also scientifically fascinating and has the potential to develop reinforcing cycles. It will be interesting to see whether regularizing object categorization models with model-predicted neural responses can grant them brain-like biases (e.g. shape bias) and capabilities (e.g. OOD generalization). Studies have shown that explicitly encouraging neural networks to build representations that are like neural representations or principles in V1 can improve their robustness under stimulus noise and adversarial attacks [35, 36]. In a similar vein, regularization with high-level visual cortical activity might help close the gap between human and machine vision, since this activity is the neural basis for human visual intelligence.

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] Anitha Pasupathy and Charles E Connor. Population coding of shape in area v4. *Nature neuroscience*, 5(12):1332–1338, 2002.

[3] Xiaomin Yue, Sophia Robert, and Leslie G Ungerleider. Curvature processing in human visual cortical areas. *NeuroImage*, 222:117295, 2020.

[4] Scott L Brincat and Charles E Connor. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature neuroscience*, 7(8):880–886, 2004.

[5] Nicole C Rust and James J DiCarlo. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–12995, 2010.

[6] Paul E Downing, AW-Y Chan, MV Peelen, CM Dodds, and N Kanwisher. Domain specificity in visual cortex. *Cerebral cortex*, 16(10):1453–1461, 2006.

[7] 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.

[8] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.*, 111(23):8619–8624, Jun 2014.

[9] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[10] D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, Mar 2016.

[11] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[12] Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen A Baccus, and Surya Ganguli. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *arXiv preprint arXiv:1912.06207*, 2019.

[13] Liang Wang, Ryan EB Mruczek, Michael J Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral cortex*, 25(10):3911–3931, 2015.

[14] 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.

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

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

[17] Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pages 248–255, 06 2009.

[18] S Rao Jammalamadaka and Ambar Sengupta. *Topics in circular statistics*, volume 5. world scientific, 2001.

[19] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[20] Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

[21] Nikolaus Kriegeskorte. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, 3:35, 2009.

[22] Yaoda Xu and Maryam Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1):1–16, 2021.

[23] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[24] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

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

[26] Anna A Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. Is it that simple? linear mapping models in cognitive neuroscience. *bioRxiv*, 2021.

[27] Radoslaw Martin Cichy, Gemma Roig, and Aude Oliva. The algonauts project. *Nature Machine Intelligence*, 1(12):613–613, 2019.

[28] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. Bold5000, a public fmri dataset while viewing 5000 visual images. *Scientific data*, 6(1):1–18, 2019.

[29] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):1–12, 2022.

[30] Ghislain St-Yves and Thomas Naselaris. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, 180:188–202, 2018.

[31] 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.

[32] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10, 2014.

[33] Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018.

[34] Kalanit Grill-Spector, Zoe Kourtzi, and Nancy Kanwisher. The lateral occipital complex and its role in object recognition. *Vision research*, 41(10-11):1409–1422, 2001.

[35] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

[36] Zhe Li, Wieland Brendel, Edgar Walker, Erick Cobos, Taliah Muhammad, Jacob Reimer, Matthias Bethge, Fabian Sinz, Zachary Pitkow, and Andreas Tolias. Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32, 2019.

[37] Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] They are described within the Discussion section
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We don't foresee any negative societal impacts of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] I've read all guidelines
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No theoretical results are discussed in this paper.
   (b) Did you include complete proofs of all theoretical results? [N/A] No theoretical results are discussed in this paper.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Yes, the code is included in the Supplementary Material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes, all training details are included in the Methdos and Supplementary Methods.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, they are displayed in all figures.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Yes. All models were trained, validated and analysed on an NVIDIA Titan Xp GPU. Training each response-optimized model on 1 GPU took $\sim$ 8 hours. Additional details are listed in the Supplementary.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] e.g. [23, 37]
   (b) Did you mention the license of the assets? [No] The assets employed in this study were made publicly available by the creators
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The assets employed in this study were made publicly available by the creators
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The assets don't contain any offensive content to our knowledge.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix