Brain Dissection: fMRI-trained Networks Reveal Spatial Selectivity in the Processing of Natural Images

Gabriel H. Sarch[†]
Machine Learning Department
Carnegie Mellon University

Michael J. Tarr
Neuroscience Institute
Carnegie Mellon University

Katerina Fragkiadaki[*]
Machine Learning Department
Carnegie Mellon University

Leila Wehbe[*]
Machine Learning Department
Carnegie Mellon University

[*]Equal advising
[†]Correspondence to gsarch@andrew.cmu.edu

Abstract

The alignment between deep neural network (DNN) features and cortical responses currently provides the most accurate quantitative explanation for higher visual areas [1, 2, 3, 4]. At the same time, these model features have been critiqued as uninterpretable explanations, trading one black box (the human brain) for another (a neural network). In this paper, we train networks to directly predict, from scratch, brain responses to images from a large-scale dataset of natural scenes [5]. We then employ "network dissection" [6], a method used for enhancing neural network interpretability by identifying and localizing the most significant features in images for individual units of a trained network, and which has been used to study category selectivity in the human brain [7]. We adapt this approach to create a hypothesis-neutral model that is then used to explore the tuning properties of specific visual regions beyond category selectivity, which we call "brain dissection". We use brain dissection to examine a range of ecologically important, intermediate properties, including depth, surface normals, curvature, and object relations and find consistent feature selectivity differences across sub-regions of the parietal, lateral, and ventral visual streams. For example, in the three scene-selective network we find that RSC prefers far depths and in-plane horizontal surface normals, while OPA and PPA prefer near and mid depths and vertical surface normals, indicating a change in the spatial coordinate system used for scene representations across RSC and OPA/PPA. Such findings contribute to a deeper, more fine-grained understanding of the functional characteristics of human visual cortex when viewing natural scenes. Project website: `https://brain-dissection.github.io/`.

1 Introduction

Humans possess an extraordinary ability to process and interpret visual information; however, our understanding of how such information is represented and processed within the brain remains incomplete. While advances in modeling visual cortex have been made through the use of abstract stimuli and the study of evoked neuronal responses, these efforts have primarily focused on low-level visual features [8] or high-level categorical selectivity [9]. In reality, the brain utilizes numerous ecologically-relevant intermediate features to perceive the world, such as 3D information, physical relationships, and object properties [10, 11].

Preprint. Under review.

Characterizing high-level areas in the visual cortex presents challenges due to the circular problem of finding optimal stimuli. Although deductive [12, 13, 14, 15, 16] and inductive [17, 18, 19, 20] approaches have identified a wide range of functionally-defined brain regions, these methods are not scalable for exploring a wider range of brain areas and diverse features. To address these limitations, we leverage "network dissection" [6], a method that enables exploring the feature selectivity of neural network units at scale. Combined with brain response-optimized networks trained on fMRI data for subjects viewing natural images, this method offers a hypothesis-neutral approach to uncovering feature tuning properties across the visual system [21, 22]. While Khosla and Wehbe [21] demonstrate the effectiveness of network dissection for validating known category selectivities, here we report that this combined approach, *brain dissection*, can be extended to examine more granular features, such as depth, curvature, and object relations.

We apply brain dissection across the visual cortex to uncover functional differences between regions. In particular, the human visual system is believed to be organized into distinctive processing streams: the parietal, lateral, and ventral visual streams. While previous experimental and modeling work has made progress in characterizing each of these streams [23, 24, 25, 1, 26, 27, 28], at a fine-grained level, their distinct roles in visual processing remain under-specified. Functional networks within these streams likewise remain functionally under-specified. For example, precise differences among the three components of the "scene network", the occipital place area (OPA), parahippocampal place area (PPA), and the retrosplenial complex (RSC) [29] remain unclear for naturalistic, large-scale viewing of scenes with prior work providing primarily broad functional characterizations in terms of the representation of scene layout [30, 31, 32], 3D features [20, 33, 34, 35], scene objects [15, 21], and navigational affordances [36, 37].

Utilizing a hypothesis-neutral approach with brain response optimization, we demonstrate feature selectivity differences across multiple visual streams and scene regions, encompassing a wide variety of highly specific feature properties. Importantly, our findings are based on subjects viewing natural images and do not require a separate experiment for each category or visual feature.

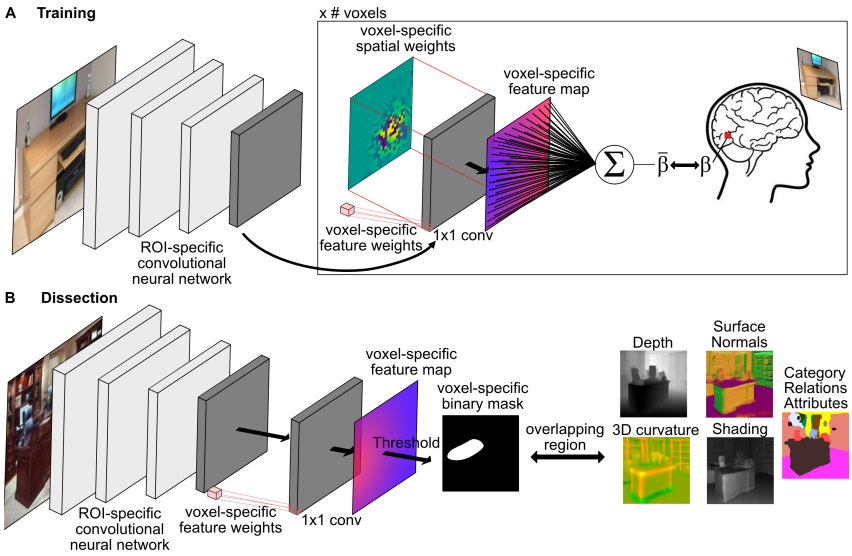

Figure 1: **A. fMRI Response-Optimized Training Procedure.** All voxels in an ROI share a convolutional backbone. Each voxel is assigned a set of learnable spatial and feature weights. The voxel weights are applied to the output features of the backbone network to predict the voxel response for an input image. Training minimizes the mean squared error between the predicted and actual voxel response of NSD images. **B. Dissection Procedure.** For each voxel in an ROI and test image, we obtain the voxel's selectivity binary mask by applying the voxel's feature weights to the output of the ROI-specific backbone and thresholding the resultant feature map. We examine the overlapping region between the mask and a measurement of interest (e.g., depth map).

2 Brain Dissection

Brain dissection provides a hypothesis-neutral approach for identifying the most significant image features for predicting the response of a specific voxel. The method trains a convolutional neural network tailored to predict voxel responses within a defined sub-region. By training a backbone

network for this sub-region and incorporating a linear readout for each voxel, the network learns to extract image features crucial for predicting each voxel (Fig. 1A). After training, we "dissect" the network by inputting held-out images and extracting regions that the network considers most relevant for each voxel (Fig. 1B). Subsequently, we analyze properties of the voxel-selective regions within the images. Our training and dissection procedure closely follows Khosla and Wehbe [21].

2.1 fMRI Response-Optimized Network Training

We trained a set of neural network encoding models to predict fMRI voxel responses to NSD images (Fig. 1A). Visual cortex was divided into sub-regions based on the Human Connectome Project (HCP) atlas [38], with a separate convolutional neural network (CNN) backbone shared across all voxels within each sub-region. For each voxel in a sub-region, a linear readout was applied to the backbone output features to predict individual voxel responses to the images. The linear readout was factorized into voxel-specific spatial and feature dimensions, as described by Klindt et al. [39]. Spatial weights correspond to the most responsive locations of a voxel in an image, while feature weights represent the visual information to which a voxel is most responsive.

Given the activations of the last layer of the ROI-specific backbone network ($A \in \mathbb{R}^{C \times H \times W}$), the spatial weights for a voxel in the ROI ($w_s \in \mathbb{R}^{H \times W}$) were multiplied channel-wise to A. The feature weights ($w_f \in \mathbb{R}^{C}$) were applied to each spatial position of A. The output is a voxel-specific feature map, which was summed and added to a learned bias to obtain the predicted voxel response. The network was trained using mean-squared error with the true voxel response. We trained each network using NSD data from 8 subjects, including 68,400 training images, and 3,600 validation images.

2.2 Dissection Procedure

During dissection (see Fig. 1B), the spatial weights were discarded and only the feature weights were used to focus the analysis on visual selectivity for each voxel. For each voxel k, we obtained its corresponding ROI-specific backbone network f_{roi} (with $k \in roi$) and voxel-specific feature weights $w_f^k \in \mathbb{R}^C$. For every input image x in the evaluation dataset, we passed the image through f_{roi} to obtain activations A. The feature weights ($w_f^k \in \mathbb{R}^C$) were then applied to each spatial position of A using a 1x1 convolution. The output is a voxel-specific feature map $A_k \in \mathbb{R}^{H \times W}$. The top quantile level T_k is determined such that $P(A_k > T_k)$ = 0.01 over every spatial location of the activation map.

To compare a low-resolution voxel activation map to the input-resolution annotation L_c for some concept or measure c (e.g., L_c could be a category mask, depth map, surface normal map, etc.), the activation map is scaled up to the mask resolution S_k from A_k using bilinear interpolation. S_k is then thresholded into a binary segmentation: M_k = $S_k \geq T_k$, selecting all regions for which the activation exceeds the threshold T_k.

As in Bau et al. [6], to score each voxel k for for a discrete category concept c, we compute the data-set-wide intersection over union (IOU) score:

$$IOU_{k,c} = \frac{\sum_X |M_k(x) \cap L_c(x)|}{\sum_X |M_k(x) \cup L_c(x)|} \tag{1}$$

To score the selectivity of each voxel k against a continuous concept c (e.g., depth), we compute the median value within the region defined by M_k, and take the average median across all images as the selective value for that voxel:

$$V_k = \frac{\sum_X median(L_c(x)[M_k(x)])}{\sum |X|} \tag{2}$$

where $L_c(x)[M_k(x)]$ represents the values in $L_c(x)$ defined by the mask $M_k(x)$.

3 Experiments

3.1 Brain Data

fMRI dataset. We used the Natural Scenes Dataset (NSD) [5], which consists of high-resolution fMRI responses to naturalistic images from Microsoft COCO [40]. NSD contains 7T fMRI responses (1.8 mm, 1.6 s) from 8 subjects who each viewed 9,000–10,000 distinct color natural scenes (22,000–30,000 trials). Subjects fixated centrally and performed a long-term continuous image recognition task. The noise ceiling (NC) was estimated in each voxel as described in Allen et al. [5]. We only include voxels with NC $\geq$ 10% variance. 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 under more ecological conditions.

Regions of Interest (ROIs). We used the "streams" anatomical atlas from NSD to define three ROIs that cover the parietal, lateral, and ventral visual streams (mid- and high-level ROIs were grouped). We then split each of the ROIs into sub-regions using the HCP atlas, including any ROI in the HCP atlas with at least 50% voxel overlap with any of the streams. We added a fourth ROI, which we call "Medial", and took all HCP regions overlapping with RSC. We further split the "Ventral" ROI into "Ventrolateral" and "Ventromedial", where "Ventromedial" included HCP regions overlapping with PPA. A separate network was trained for each of the HCP sub-regions. We also examined three scene ROIs – RSC, OPA, and PPA – as defined by the category functional localizer.

3.2 Measures

Unless otherwise specified, we evaluate on the Places365 dataset [41], and use the top 1000 images per sub-region per subject as determined by the predicted responses to the images by the sub-region network. Thus, the evaluation takes into account: 1) the top-most-activating images for a region; 2) the areas of the image that most contribute to the response as determined by network dissection. We perform brain dissection on the following measures:

Depth. We analyze the selectivity of depth, which refers to the perceived distance between the viewer and different regions of the scene. To obtain depth maps for each evaluation image in Places365, we use the ZoeDepth metric depth estimation network [42], which has been shown to exhibit exceptional zero-shot generalization performance on indoor and outdoor scenes. We evaluate two measures of depth: 1) *metric depth*, which represents the absolute depth in meters for each pixel in the image, and 2) *relative depth*, which denotes depth maps normalized to a range of 0-1 for each image. Consequently, absolute depth captures the overall nearness or farness across the images, while relative depth focuses on the nearness or farness within an individual image.

Surface Normals. Surface normals, vectors perpendicular to surfaces, are essential for spatial cognition, as they indicate surface orientation and reflect the spatial coordinate system. Using the XTC network [43], we estimate normals for each Places365 image. The x, y, and z components are segmented into 512 bins (8 per component; $8^3 = 512$ combinations), and voxel selectivity is reported as the bin number.

3D Curvature. Curvature features provide information about 3D structure and are invariant under rigid transformations. We use the XTC network [43] for estimating principal curvatures.

Shading. One way to infer scene geometry is "shape from shading" [44, 45] using the intrinsic image decomposition $I = AS$, where S is a shading function parameterized by lighting and depth. We use the XTC network [43] for estimating shading.

Objects, Attributes, Relations. Scenes are characterized by their object relationships, attributes, and categories. We run an evaluation on the GQA dataset [46]. Based on the Visual Genome dataset [47], GQA annotated each image with a scene graph of bounding boxes and labels for 1703 object categories, 310 object relations, and 617 object attributes. Segment Anything [48] was used to obtain segmentation masks from the bounding boxes for each of the annotations.

4 Results

We employed brain dissection to investigate the scene features represented throughout visual cortex. Our investigation focused on the Ventromedial, Ventrolateral, Lateral, Parietal, and Medial regions, as well as the scene network – PPA, OPA, and RSC. Our initial examination centered on identifying and quantifying the differences between these regions based on four spatial measures, namely, depth, surface normals, 3D curvature, and shading. These measures were chosen to cover a range of ecologically-relevant spatial features. Following our spatial analysis, we explored how these identified differences are reflected in the processing of object categories, attributes, and spatial relations. This aspect of our study sheds light on the relationship between the observed variations in spatial features and the resultant impact on higher-order visual processing – specifically how we perceive and interpret objects and their relationships within a given scene. Finally, we quantified the spatial clustering of voxels for each measure to better understand the spatial organization for different visual features.

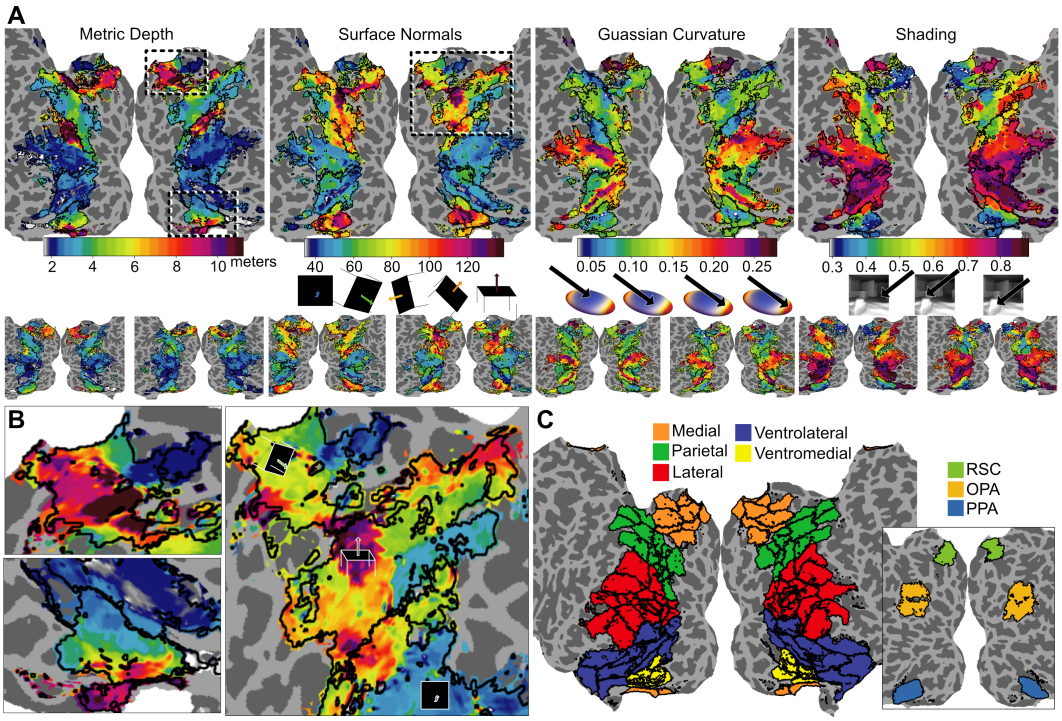

Figure 2: **A.** Depth, surface normals, Gaussian curvature, and shading on flatmaps colored by selectivity per voxel for S1 (large), S2, and S5 (maps for the remaining five subjects are in supplemental). **B.** Enlarged portions of flatmaps from A. Medial areas show depth preference for far-reaching depths (top left). There is a gradient from near to far depths from lateral to medial ventral (bottom left). Parietal and Medial areas show clusters of "upright" and "right/left" surface normals, while Lateral areas prefer "camera facing" surface normals (right).

4.1 3D Spatial Dissection

4.1.1 Brain Flatmaps

To gain a better understanding of how 3D structure is represented in the brain, we used brain dissection to quantify the selectivity for each voxel across the high-level visual cortex for the four spatial measures. This process allowed us to create a detailed selectivity map for each voxel that reveals how 2D inputs are transformed into 3D representations that enable reasoning about the physical world. Figure 2 displays these selectivity maps on a cortical flatmap of the brain. We observe that Lateral areas demonstrate selectivity for close distances, aligning with the overlap of Lateral ROIs with regions selective for bodies, frequently featured in the foreground of images. Notably, we observe that the depth selectivity for the Lateral and Ventrolateral ROIs are similar, and stand in contrast with the rest of the ROIs. This pattern is also true for the other spatial metrics (Surface Normals, Gaussian Curvature and Shading) we study, where the Lateral and Ventrolateral appear as a distinct cluster from the Ventromedial, Parietal and Medial ROIs. One possibility is that these two ROIs represent spatial information in a manner that supports behaviors that require interacting with nearby objects and with other people. In contrast, we observed high selectivity for far distances in the Medial and Ventromedial areas, suggesting a specialized role for behaviors that require processing and understanding far-reaching aspects of a visual scene (e.g., navigation). Additionally, we observed a strong gradient in the selectivity for depth in Ventral areas. The gradient transitioned from near-preferring depths in the Ventrolateral ROI to far-preferring depths in the Ventromedial ROI. Such a gradient is likely to be important for building up scene representations that capture a wide variety of behaviorally relevant affordances. Finally, we observed distinct patterns related to the processing of surface normals. Large areas of the Lateral and Ventrolateral regions showed a preference for surface normals in the negative X direction, pointing back at the camera (again, consistent with the preference of these areas for bodies and objects). In contrast, the Parietal and Medial areas predominantly preferred surface normals oriented in the positive Z direction (pointing upwards) and along the Y direction (pointing right and left). This pattern of preferences may reflect differences in the coordinate systems being used to represent spatial information in these regions. Surface normals

oriented at the camera indicate an egocentric frame of reference, while surface normals oriented from the ground plane indicate an allocentric frame of reference.

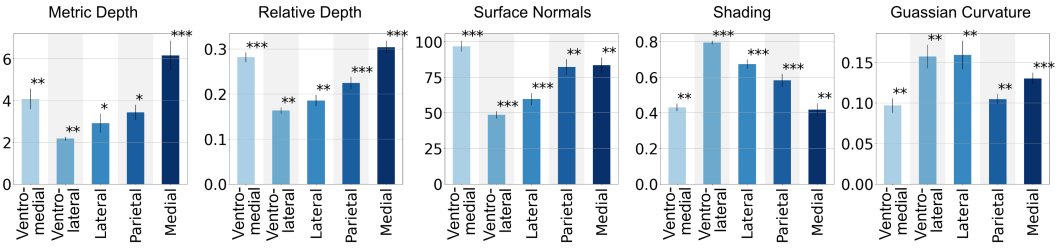

Figure 3: Mean $\pm$ standard deviation ($N = 8$ subjects) for each of the high-level visual ROIs. *** = significantly different from all other ROIs, ** = significantly different from all but one other ROI, * = significantly different from all but two other ROIs ($p < 0.01$; *post hoc* Tukey test; $N = 8$ subjects).

4.1.2 Spatial Dissection of High-level Visual Areas

We quantify the differences between the high-level visual ROIs in terms of mean $\pm$ standard deviation for four 3D spatial measures: depth (metric & relative), surface normals, shading, and Gaussian curvature (Fig. 3). We observe striking differences in depth preferences between the Medial and Ventrolateral ROIs. Medial and Ventromedial areas exhibit significantly larger absolute and relative depth selectivity compared to the Ventrolateral and Lateral regions ($p < 0.01$; *post hoc* Tukey test; $N = 8$). Conversely, Ventrolateral and Lateral areas have significantly lower average absolute depths than the Medial and Ventromedial areas, as depicted in Figure 3 ($p < 0.01$; *post hoc* Tukey test; $N = 8$). This disparity underscores the specialized functional roles these regions have in processing natural images with differing depth distributions, preparing the perceiver to act on both nearby and distant objects. Concerning the processing of surface normals, we observe that the Ventromedial region significantly prefers normals indicative of surfaces pointing in the "up" direction ($p < 0.01$; *post hoc* Tukey test; $N = 8$). Parietal and Medial areas show a stronger preference for surface normals oriented to the left and right compared to the other ROIs ($p < 0.01$; *post hoc* Tukey test; $N = 8$). In contrast, the Ventrolateral and Lateral regions have a preference for surface normals pointing back at the camera ($p < 0.01$; *post hoc* Tukey test; $N = 8$). Again, camera oriented surface normals suggest a representation using an egocentric frame of reference, while up surface normals suggest an allocentric frame of reference. While left/right oriented surface normals are less readily interpretable, one possibility is that they denote a representation grounded in a bodycentric frame of reference, thereby supporting manual interactions with surrounding objects.

Examining the shading preferences across the ROIs, we again notice a clear distinction. Ventrolateral areas prefer the lightest shading, while Medial and Ventromedial areas prefer the darkest shading ($p < 0.01$; *post hoc* Tukey test; $N = 8$). This observation reaffirms the notion of functional segregation in the visual cortex, with different regions specializing in processing specific visual attributes, in the case of shading, perhaps reflecting additional cues to scene depth.

Lastly, we consider Gaussian curvature, a measure that captures the degree of 3D curvature of a surface. Higher Gaussian curvature implies a more spherical geometry. Ventrolateral and Lateral areas significantly prefer higher Gaussian curvature (Fig. 3), indicating a preference for more spherical surfaces ($p < 0.01$; *post hoc* Tukey test; $N = 8$). In contrast, Parietal and Ventromedial areas prefer lower Gaussian curvature, suggesting a preference for flatter surfaces ($p < 0.01$; *post hoc* Tukey test; $N = 8$). Ventrolateral and Lateral areas may be encoding spatial features relevant to object appearance for both recognition and grasping, while Parietal and Ventromedial areas may be encoding scene layout in terms of large-scale surfaces and scene elements. Such findings add another layer to our understanding of how different areas of the visual cortex process and interpret 3D scene structure.

4.1.3 Spatial Dissection of Scene Areas

In this section, we delve into the differences between the scene ROIs (PPA, OPA, RSC) with respect to the four 3D spatial measures, providing a quantified analysis of their specialized roles in scene perception. We illustrate the voxel-selectivity on brain flatmaps (Fig. 4A) and plot the mean $\pm$ standard deviation ($N = 8$) for each of the each ROIs and measures (Fig. 4B). These distributions and mappings elucidate the distinct patterns of selectivity across the scene ROIs.

There is a considerable increase in both metric and relative depth for RSC compared to both PPA and OPA, as evidenced in Figure 4B ($p < 0.001$; *post hoc* Tukey test; $N = 8$). This pattern indicates

that RSC in processing far-reaching information in scenes, consistent with the claim that RSC is preferentially involved in navigation [49]. Additionally, we observe a distinct shift in surface normal preference, with RSC showing a preference for more right/left surface normal bins compared to OPA and PPA, which favor upright surface normals ($p < 0.001$; *post hoc* Tukey test; $N = 8$). This distinction underscores the functional segregation between these scene ROIs in interpreting surface orientation and suggest, again, that RSC may represent information supporting navigation, while OPA and PPA may represent information more relevant to interacting within a local region of space.

To illustrate these observations, we identified voxels that have the closest selectivity to the ROI mean for depth or surface normals for PPA, OPA, and RSC, and plotted the voxel's top five images by predicted response (Fig. 5). We include corresponding dissection masks below these on images and either depth or surface normal images. RSC exhibits a preference for far depths and more "right/left" surface normals, indicative of vertical structures relevant to navigation. In contrast, OPA and PPA exhibit a preference for "up" surface normals which are relevant to representing the local scene layout. Once more, these results highlight the distinct functional roles these areas play in scene perception.

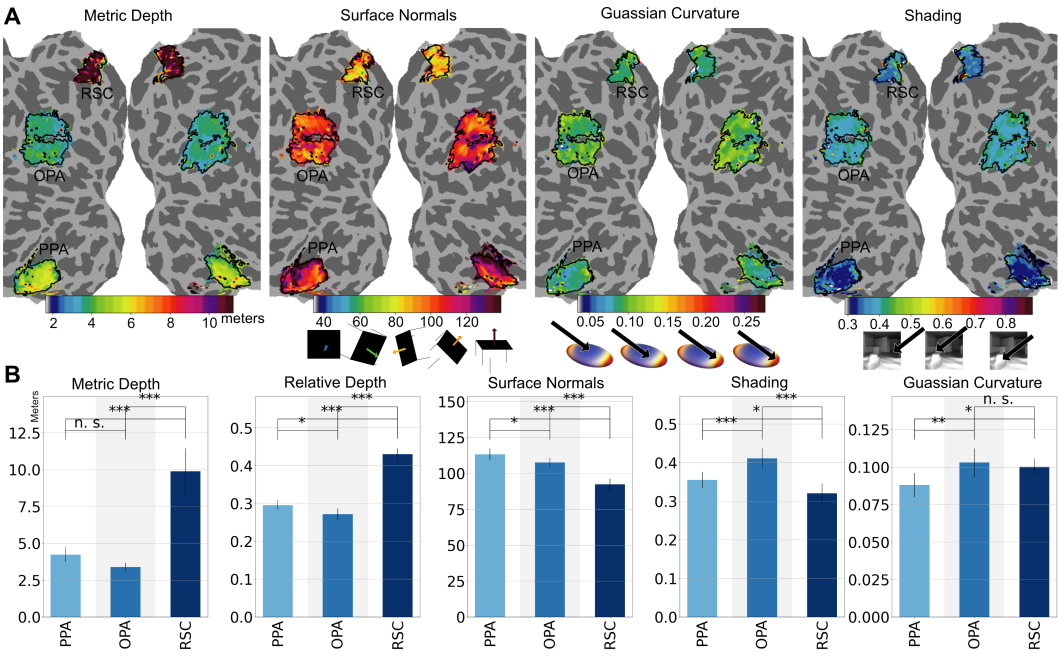

Figure 4: **A.** Flatmaps showing voxel-preferences in OPA, PPA, and RSC for the four spatial measures. **B.** Mean $\pm$ standard deviation ($N = 8$ subjects) for each of the high-level visual ROIs. *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, n.s. = not significant $p > 0.05$ (*post hoc* Tukey test; $N = 8$ subjects).

4.2 Object Relations, Attributes and Categories

We utilize the GQA dataset [46] in concert with brain dissection to gain insights into the selectivity differences for object relationship, attributes, and categories within an image. The median IOU (mean $\pm$ standard deviation across 8 subjects) for the top 20 concepts for each ROI is reported in Figure 6. With respect to high-level visual ROIs, we observe distinct preferences: Parietal areas prefer inanimate and scene objects, such as walls and tables, that are more relevant to both navigation and physically interacting with objects. In contrast, Lateral areas exhibit a preference for person and animate objects, while Ventrolateral areas favor persons and objects – representations that support social interactions and scene understanding. When examining object relations, Parietal areas prefer spatial relations such as 'in front of,' 'near,' 'in,' and 'above,' while Lateral areas favor relations involving people, such as 'wearing,' 'holding,' and 'with'. In terms of attributes, Parietal areas show a preference for colors and materials, whereas Lateral areas favor animate attributes involving people, such as 'playing,' 'standing,' and 'happy'. These preferences are largely aligned with prior results showing Parietal preferentially processes spatial visual information [23], Lateral preferentially processes social visual information [25], and Ventral preferentially processes objects and faces [24].

As illustrated in Figure 6, within scene ROIs, OPA shows greater selectivity for indoor scene elements, such as chairs, tables, and floors. PPA exhibits a mix of indoor and outdoor preferences, whereas

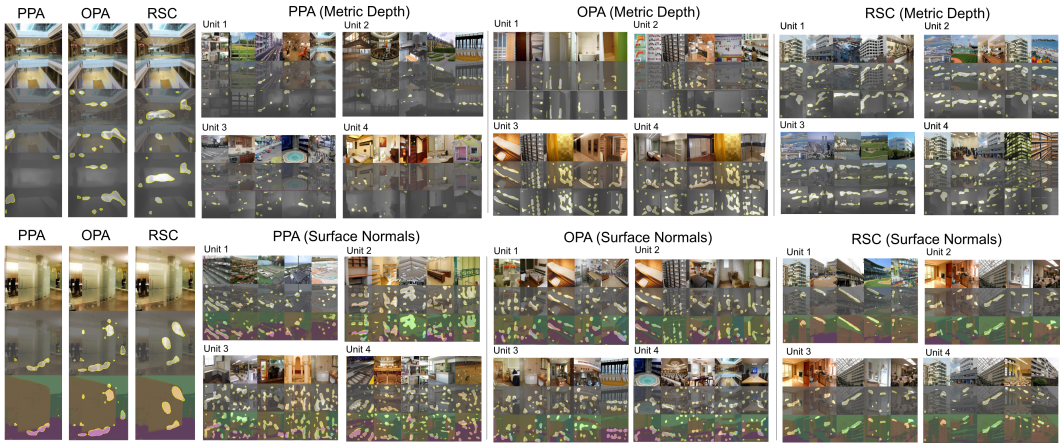

Figure 5: Voxels that have the closest selectivity (by euclidean distance) to the ROI mean of depth (top) or surface normals (bottom) for PPA, OPA, and RSC. An example image is shown with a representative voxel (left) and the top five images by predicted response for the top four voxels (right) for each ROI. Dissection masks are visualized on the images and either the depth or surface normal images. RSC prefers far-reaching depth and vertical structures, while OPA and PPA prefers more local depths and a mix of vertical and horizontal structures.

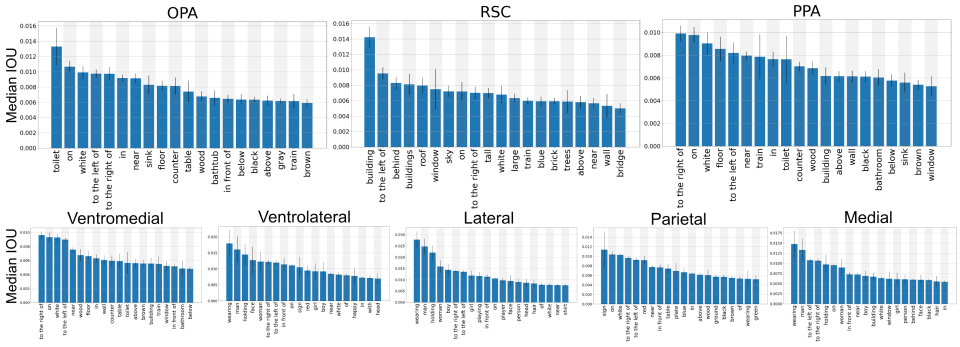

Figure 6: Median IOU (mean $\pm$ standard deviation; $N = 8$ subjects) for the GQA dataset [46] for each of the scene ROIs (top) and high-level visual ROIs (bottom).

RSC shows a clear preference for outdoor scene elements like trees, buildings, and the sky, once again reinforcing a dichotomy between local space interactions and navigation. Furthermore, when considering attributes, RSC again demonstrates more selectivity for "outdoor" attributes compared to PPA and OPA, which show higher selectivity for indoor elements. These findings underscore the distinct functional implications of these ROI selectivity profiles for processing visual scenes.

4.3 Topographic Organization of the Spatial Preferences

We investigated the topographical organization of the spatial measures within each ROI. High-level visual areas are topographically organized, comprising multiple hierarchically organized areas that are selective for specific domains such as faces and scenes [50, 51, 12, 13, 14, 15, 16, 52]. However, whether and how this topographical organization applies to the spatial measures used here remains unclear. Given a specific type of spatial information, we explored: (a) if the voxels in an ROI are differently tuned for this type of information (e.g., sub-groups of voxels are tuned for different surface normal orientations); and if so, (b) does the tuning follow a smooth topographical organization with relatively large clusters processing similar information. To quantify the degree of clustering by selectivity similarity, for each measure we computed the average cluster proportion (cluster size normalized by total voxels in the ROI) (considering clusters > 5 voxels) that each voxel belongs to, and report the results compared to a random permutation of the voxels (Fig. 7). We employed connected components [53] on the 3D voxel data to determine whether two neighboring clusters belong to the same cluster. For continuous measures, such as depth, we consider two voxels the same if they fall within δ of each other. We used $1/100^{th}$ of the data range across all ROIs as the δ for each of the continuous measures. For discrete concepts, such as object category, we consider two neighbors to be in the same cluster if they belong to the same discrete category.

A cluster analysis was run for each ROI. We report the average cluster size for a given measure compared to the average cluster size for the measure when the voxels are randomly permuted ((Fig. 7; $p < 0.05$ paired t-test; $N = 8$)). We visualize the spatial clusters on flatmaps: All of the high level visual areas showed clustering significantly different from random for all spatial measure, indicating that their voxels are tuned for different values of each measure in a way that is topographically organized. At the level of the scene areas however, RSC was not distinguishable from chance for metric depth, relative depth, shading, gaussian curvature and category. This suggests that except for surface normals, the voxels in RSC are tuned similarly. Similarly, the results suggest that the voxels in PPA are tuned differently only for metric depth, and the voxels in OPA are tuned differently for metric depth, relative depth and surface normals. These findings highlight the robust topographical organization within the ROIs, extending beyond traditional domain-specific areas. Critically, this extended organization highlights the fact that, as a whole, visual cortex is representing a broad range of ecologically relevant properties that support complex behavior [54].

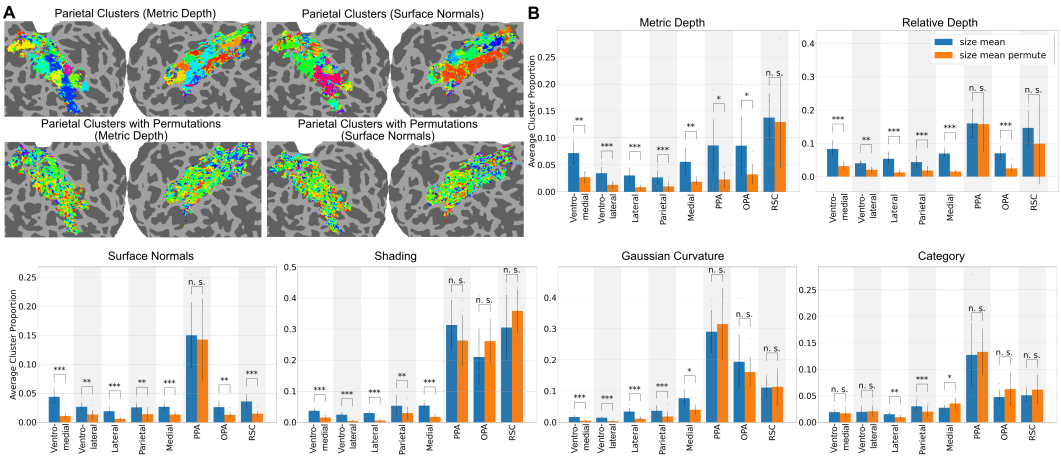

Figure 7: **A.** Metric depth (left) and surface normals (right) spatial clusters on flatmaps for example subject S1, where each color is a different cluster with (bottom) and without (top) random spatial permutations. **B.** Bar plots showing mean $\pm$ standard deviation of average cluster size per voxel for each ROI. *** = $p < 0.001$, ** = $p < 0.01$, * = $p < 0.05$, n.s. = not significant $p > 0.05$ (paired t-test; $N = 8$).

5 Conclusion - Discussion

A clearer understanding of the representational progression in natural vision requires developing hypothesis-neutral models and interpretable methods. We leverage advances in quantifying the feature selectivity of artificial neural networks [6] merged with response-optimized models trained on fMRI data [21] to better characterize the spatial properties of a broad range of visual brain regions. This method, "brain dissection", enables a more granular examination of feature preferences within multiple visual processing streams, as well as within functional sub-regions such as the scene network.

Limitations. Our approach relies on network dissection and therefore inherits some of its limitations. Network dissection considers each output unit in the network (in our case corresponding to an fMRI voxel) in isolation during interpretation, thus it is not capable of identifying the selectivity of a group of output units. We thus identify the selectivity of individual voxels, which is in line with the voxel-wise modeling school of thought, but not with multivariate pattern analysis. Furthermore, in natural images, there are inherent correlations between specific categories and their spatial attributes (e.g., scenes tend to be rectilinear, pointing up, with far away features, while bodies tend to be close up, facing the camera, etc. [10, 11]). We mitigate this limitation by focusing on ROIs that are widely held to encode spatial properties.

Dissection of spatial properties. In contrast to earlier studies [21], our work focuses on characterizing visual sub-region preferences in terms of spatial properties, as well as high-level categories (e.g., faces, places, etc.). The dissection of spatial properties provides complementary perspectives on visual processing: the relationships, orientations, and positions of objects and surfaces within a scene in terms of spatial features (see also [10]). Importantly, different regions of the visual cortex show striking differences in their processing preferences for these features. Notably, Medial and Ventromedial areas show larger absolute and relative depth preferences, flatter surfaces, and darker

shading, while Ventrolateral and Lateral regions prefer lower average absolute depths, curvier objects, and lighter shading. This extends to distinct preferences for object relationships, attributes, and categories, with Parietal areas favoring inanimate objects and spatial relations, Lateral areas preferring person and animate objects, and Ventral areas favoring both persons and objects, aligning with previous literature. At a more fine-grained level, we also found that the components of the scene network exhibit differences in their feature preferences. RSC appears to have a unique role in processing far spatial information and outdoor scenes, showing a preference for greater metric and relative depths, and right/left surface normals indicative of vertical structures and a possibly bodycentric representation of space. In contrast, PPA and OPA show a stronger preference for "up" surface normals, underscoring their contributions to a possibly allocentric representation of space.

A comprehensive understanding of natural vision should cover early visual processing, ecologically-based intermediate features, and high-level, semantically-grounded representations. Current knowledge largely focuses on the first and third areas. Brain dissection offers potential for uncovering data-driven insights into the full visual processing hierarchy.

References

[1] 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.

[2] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[3] Aria Y. Wang, Kendrick Kay, Thomas Naselaris, Michael J. Tarr, and Leila Wehbe. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *bioRxiv*, 2022.

[4] Colin Conwell, Jacob Prince, George Alvarez, Talia Konkle, and Kendrick Kay. Opportunistic experiments on a large-scale survey of diverse artificial vision models in prediction of 7t human fMRI data. In *Conference on Cognitive Computational Neuroscience*, 2022.

[5] 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, J Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, 2022. doi: 10.1038/s41593-021-00962-x.

[6] 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.

[7] Meenakshi Khosla, N. Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32:1–13, 2022.

[8] Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do We Know What the Early Visual System Does? *Journal of Neuroscience*, 25(46):10577–10597, 2005.

[9] Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annual Review of Neuroscience*, 27:649–677, 2004.

[10] Margaret M. Henderson, Michael J. Tarr, and Leila Wehbe. A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *Journal of Neuroscience*, 2023.

[11] Margaret M. Henderson, Michael J. Tarr, and Leila Wehbe. Low-level tuning biases in higher visual cortex reflect the semantic informativeness of visual features. *Journal of Vision*, 23(4): 8–8, 2023.

[12] Kalanit Grill-Spector, Tammar Kushnir, Talma Hendler, and Rafael Malach. The dynamics of object-selective activation correlate with recognition performance in humans. *Nature neuroscience*, 3(8):837–843, 2000.

[13] J Sergent, S Ohta, and B MacDonald. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115:15–36, 1992.

[14] 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.

[15] Geoffrey K Aguirre, Eric Zarahn, and Mark D'Esposito. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron*, 21(2):373–383, 1998.

[16] Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7):293–299, 2003.

[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] Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, pages 2022–03, 2022.

[22] Meenakshi Khosla, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Characterizing the ventral visual stream with response-optimized neural encoding models. *Advances in Neural Information Processing Systems*, 35:9389–9402, 2022.

[23] Leslie G Ungerleider. Two cortical visual systems. *Analysis of visual behavior*, pages 549–586, 1982.

[24] Bhavin R Sheth and Ryan Young. Two visual pathways in primates based on sampling of space: exploitation and exploration of visual information. *Frontiers in integrative neuroscience*, 10:37, 2016.

[25] Emalie McMahon, Michael F Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. 2023.

[26] Vladislav Ayzenberg and Marlene Behrmann. The where, what, and how of object recognition. *Trends in Cognitive Sciences*, 2023.

[27] Dawn Finzi, Daniel LK Yamins, Kendrick Kay, and Kalanit Grill-Spector. Do deep convolutional neural networks accurately model representations beyond the ventral stream? In *2022 Conference on Cognitive Computational Neuroscience*, 2022.

[28] Gabriel H Sarch, Hsiao-Yu Fish Tung, Aria W Wang, Jacob S Prince, and Michael J Tarr. 3d view prediction models of the dorsal visual stream. In *2023 Conference on Cognitive Computational Neuroscience*, 2023.

[29] Daniel D Dilks, Joshua B Julian, Alexander M Paunov, and Nancy Kanwisher. The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, 33 (4):1331–1336, 2013.

[30] Merim Bilalić, Tobias Lindig, and Luca Turella. Parsing rooms: the role of the ppa and rsc in perceiving object relations and spatial layout. *Brain Structure and Function*, 224:2505–2524, 2019.

[31] Dwight J Kravitz, Cynthia S Peng, and Chris I Baker. Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *Journal of Neuroscience*, 31(20): 7322–7333, 2011.

[32] Linda Henriksson, Marieke Mur, and Nikolaus Kriegeskorte. Rapid invariant encoding of scene layout in human opa. *Neuron*, 103(1):161–171, 2019.

[33] Moira R Dillon, Andrew S Persichetti, Elizabeth S Spelke, and Daniel D Dilks. Places in the brain: Bridging layout and object geometry in scene-selective cortex. *Cerebral Cortex*, 28(7): 2365–2374, 2018.

[34] Samoni Nag, Daniel Berman, and Julie D Golomb. Category-selective areas in human visual cortex exhibit preferences for stimulus depth. *Neuroimage*, 196:289–301, 2019.

[35] Soojin Park and Marvin M Chun. Different roles of the parahippocampal place area (ppa) and retrosplenial cortex (rsc) in panoramic scene perception. *Neuroimage*, 47(4):1747–1756, 2009.

[36] 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.

[37] Jeongho Park and Soojin Park. Coding of navigational distance and functional constraint of boundaries in the human scene-selective cortex. *Journal of Neuroscience*, 40(18):3621–3630, 2020.

[38] Chu-Chung Huang, Edmund T Rolls, Jianfeng Feng, and Ching-Po Lin. An extended human connectome project multimodal parcellation atlas of the human cortex and subcortical areas. *Brain Structure and Function*, 227(3):763–778, 2022.

[39] David Klindt, Alexander S Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what" and "where". *Advances in Neural Information Processing Systems*, 30, 2017.

[40] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL `http://arxiv.org/abs/1405.0312`.

[41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[42] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[43] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020.

[44] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014.

[45] Edward H Adelson and Alex P Pentland. The perception of shading and reflectance. *Perception as Bayesian inference*, 409:423, 1996.

[46] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. *arXiv preprint arXiv:1907.12133*, 2019.

[47] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[48] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[49] Merim Bilalić, Tobias Lindig, and Luca Turella. Parsing rooms: the role of the PPA and RSC in perceiving object relations and spatial layout. *Brain Struct. Funct.*, 224(7):2505–2524, 2019.

[50] Nicholas M Blauch, Marlene Behrmann, and David C Plaut. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, 119(3):e2112566119, 2022.

[51] Dawn Finzi, Eshed Margalit, Kendrick Kay, Daniel LK Yamins, and Kalanit Grill-Spector. Topographic DCNNs trained on a single self-supervised task capture the functional organization of cortex into visual processing streams. In *SVRHM 2022 Workshop @ NeurIPS*, 2022. URL `https://openreview.net/forum?id=E1iY-d13smd`.

[52] Nidhi Jain, Aria Wang, Margaret M Henderson, Ruogu Lin, Jacob S Prince, Michael J Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Commun. Biol.*, 6(1):175, 2023.

[53] Azriel Rosenfeld and John L Pfaltz. Sequential operations in digital picture processing. *Journal of the ACM (JACM)*, 13(4):471–494, 1966.

[54] Stefania Bracci and Hans P. Op de Beeck. Understanding human object vision: A picture is worth a thousand representations. *Annual Review of Psychology*, 74(1):113–135, 2023.