Food for thought: selectivity for food in human ventral visual cortex

Nidhi Jain[1], Aria Wang[5,6], Margaret M. Henderson[5,6], Ruogu Lin[6], Jacob S. Prince[2,4], Michael J. Tarr[2,5], and Leila Wehbe[*,5,6]

[1]Computer Science Department, Carnegie Mellon University
[2]Department of Psychology, Carnegie Mellon University
[3]Computational Biology Department, Carnegie Mellon University
[4]Department of Psychology, Harvard University
[5]Neuroscience Institute, Carnegie Mellon University
[6]Machine Learning Department, Carnegie Mellon University

* correspondence should be addressed to: lwehbe@cmu.edu

Abstract

Ventral visual cortex contains regions of selectivity for domains of ecological importance. Food is an ecologically and evolutionarily important category, whose high degree of visual variability may make the identification of selectivity more challenging. We investigated neural responsiveness to food using natural images combined with large-scale human neuroimaging. Leveraging the improved sensitivity of modern designs and statistical analysis methods, we identified two food-selective regions in the ventral visual cortex. Our results were robust across 8 subjects, multiple independent sets of images and multiple analysis methods. Additionally, these results were not due to stimulus properties or saliency. The identification of food-selective regions stands alongside prior findings of functional selectivity and provides an important addition to our understanding of the organization of knowledge within the human visual system.

Introduction

The representation of high-level visual information in the human brain has been marked by the phenomenon of selectivity for visual categories or properties of high ecological importance. Focusing on ventral visual cortex, there are multiple brain regions that show preferential responses to categories such as faces[1,2], bodies[3], places[4], and words[5], and to broad organizational principles such as animacy[6], real-world size[6], and "reach space"[7]. Independent of any particular theory on the origins and specificity of these functional brain regions[8,9], the prevailing view is that the likely role of these regions is to instantiate processes and representations for categories and properties that are highly relevant for day-to-day behavior. In a similar vein, food is a category that is relevant to evolution – the need to find nourishment is more ancient than social interaction and, arguably, more fundamental to survival. It is therefore surprising that food has not been consistently identified as a visual category for which localized, selective neural responses are observed.

The visual presentation of food images prompts a range of brain responses[10,11], including affective, sensory, and cognitive effects. However, agreement on neuroanatomical locations of food-related activation across studies using food images has been low to moderate[12]. Only 41% of 17 experiments in a meta analysis contributed to food-related clusters in the bilateral fusiform gyrus and left orbitofrontal cortex[12]. And in cases where statistically significant activations in response to food have been observed, they typically have been attributed to increased attention to food images arising from subjects' mental states and/or physiological factors[12,13]. One study of selectivity across a range of proposed categories found no robust selectivity for either fruits or vegetables in occipitotemporal cortex[14]. As such, to date there has been no robust demonstration of food selectivity in the human visual system.

One factor that may influence prior studies of food-elicited neural responses is context. We posit that the apparent inconsistency in detecting food-selective responses is, in part, a result of relying on unrealistic, isolated food and non-food images (e.g., Downing et al.[14]). A second contributing factor may be that prior studies have used a limited number of food images – insufficient to capture the large variety of visual properties of food or of the natural contexts in which it appears. Unlike faces, bodies, or word stimuli, food images vary widely in low- to mid-level visual characteristics, such as curvature, shape, texture, color and the organization of the parts into a whole. Consistent with these arguments, when using naturalistic food images, Tsourides et al.[15] observed a neural correlate, as measured by magnetoencephalography (MEG), at 85 msec for the visual food/non-food distinction.

Within-category variability in food appearance may contribute in several ways that render identifying food-selectivity more challenging than those selective for other ecologically

important categories. First, food images may not generate robust activation patterns in visual cortex without the presence of associated content (e.g. plates, tables, silverware). Second, because of the high visual variability for food as a category, detecting significant food-driven responses may require more sensitive designs than standard neuroimaging designs (which typically rely on small numbers of similar images and trials per condition).

Our study addresses these issues by using two large-scale datasets. Real-world images, drawn from the the Microsoft COCO dataset[16], were used for both the food and non-food conditions. Functional MRI (fMRI) data was collected at a massive scale as part of the Natural Scenes Dataset (NSD)[17], thereby improving our ability to detect effects across conditions. To preview our most important result, we reliably identify two distinct regions in ventral high-level visual cortex that are preferentially responsive to food images. These two strips surround the Fusiform Face Area (FFA) and are aligned on the anterior to posterior axis. We replicate these regions across subjects while controlling for other aspects of images that are thought to be coded in the ventral visual system, such as image perspective. In that food is incontrovertibly an ecologically critical category, this finding is consistent with earlier findings of selectivity in the perception of faces, bodies, and places. Multivariate pattern analyses suggest a rich organization of information within food-selective visual cortex, possibly reflecting gradients along which food is combined with other ecologically relevant categories.

Results

To investigate responsiveness to food in a large-scale natural setting, we used the Natural Scenes Dataset (NSD)[17], which consists of high-resolution fMRI responses to naturalistic scenes. NSD contains fMRI data from 8 screened subjects (S1-S8) who each viewed 9,000-10,000 scene images Of the 70,566 total unique images viewed across subjects, for purposes of consistency we focused on the 1,000 images that were viewed by all 8 subjects.

Though COCO images already include labels for many categories, including some types of food, there is important information not captured by these labels, such as whether an image contains human faces. We methodologically relabeled by hand the 1,000 shared images, based on 3 main attributes: location, content, and image perspective. We used the hierarchical structure shown in Fig. 1B (refer to *Methods* for labeling details, and Fig. 1A for examples). Image perspective was included because there is evidence that objects shown at human-reachable distances have a distinct representational signature in the brain[7,18] and food is often viewed at reachable distances.

Using these labeled images, we constructed a standard linear model that expresses brain

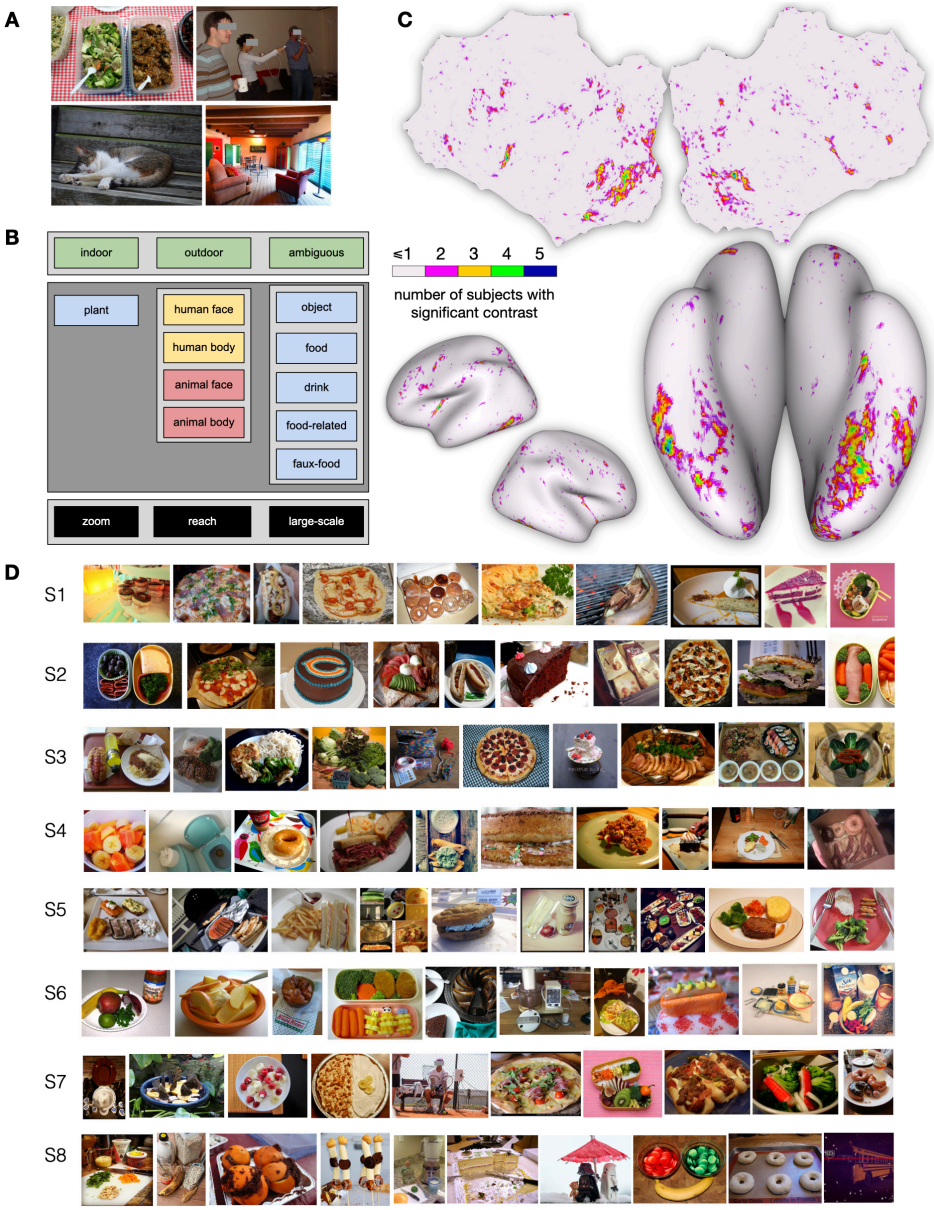

Figure 1: The 1,000 images viewed by all 8 subjects in NSD were manually relabeled to investigate responsiveness to naturalistic food images. (A) Example images labeled as (clockwise, from upper left): {outdoor, food, food-related, reach} {indoor, human face, human body, object, large-scale}, {indoor, object, large-scale}, {outdoor, animal face, animal body, object, zoom}. (B) The labeling taxonomy, including attributes of location (top), content (middle), and image perspective (bottom). (C) Flattened, inflated lateral, and inflated bottom views of the MNI surface indicating voxels with higher activity for food than all non-food labels for the 1,000 shared images. The subject count for a significant contrast was obtained at each MNI voxel. Voxels more responsive to food are found in the frontal, insular, and dorsal visual cortex, with the highest concentration across subjects occurring in the fusiform visual cortex. Both hemispheres show two strips within the fusiform that are separated by a gap that lies on the posterior-to-anterior axis. (D) Top 10 images per subject (S1-S8) leading to the largest responses in the food area. These images, which overwhelmingly depict food, were unique for each subject and were not in the set used to localize the food-selective region.

activity as a combination of the attributes assigned to each image. This model identified voxels that are more responsive to food than other categories, based on a t-test comparing the weights for food versus all other labels (Fig. 1C). Across the cortex, there are several regions showing significantly higher activation for food than non-food categories ($p<0.05$, FDR corrected), including some areas in parietal and frontal cortex, as well as on the ventral surface of the occipital lobe. We focus on ventral visual cortex due to the long history of mapping category-selective responses in this brain region. Across all subjects, we consistently find two food-selective strips in the ventral visual cortex that surround FFA on the lateral and medial sides. (Fig. 1C shows the count of subjects for whom these contrasts are significant at each MNI voxel, and the contrast strength is shown for individual subjects in Figs. 2A and S2A). Considering only unique images that were viewed by a given subject, Figure 1D shows the top 10 activating images for the food-selective voxels for that subject. These images overwhelmingly depict food. These images were not used to identify the food regions, and thus reinforce the generality of food selectivity across independent image sets. Note that these identified regions persist even when removing all images with the "reach" (Fig. S3) or "zoom" (Fig. S4) annotations – demonstrating that food-selective responses are not dependent on food being shown at a particular distance[7].

Given that food-selective regions appear adjacent to the FFA, we focused on the spatial relationship between food-selective and face-selective populations on the ventral surface. We compared the t-statistics for a contrast of food vs. non-food and t-statistics from a contrast of faces vs. non-faces for S1-S8 individually (Figs. 2A and S2A). The faces vs. non-faces contrast reveals a voxel cluster overlapping with the FFA[1,2] (Figs. 2A and S2A). FFA was localized for each subject through a separate visual category localizer experiment. (The faces vs. non-faces comparison also makes the methodological point that established category-selective regions can be reliably localized in a large-scale event-related design using stimuli embedded in complex, real-world scenes. This stands in contrast to typical localizer designs and decontextualized images[19]). The regions with higher activity for food are spatially distinct from the ones with higher activity for faces. This pattern persists when comparing food or faces to non-face and non-food images only (Fig. S5), indicating that the regions that have high activity for food and faces have highly independent or non-overlapping spatial extents.

We further investigated how food representations might be distributed across multiple voxels, using searchlight classification[20] (Figs. 2B and S2B). Training a decoder to classify food versus other categories revealed that food was decodable across a wide area of the ventral surface. The regions from which food information was decodable are a union of the regions that are high for food vs. all and the regions that are high for faces vs. all. This finding is

consistent with the idea that voxels primarily selective for other categories, such as faces, may contain information relevant to the detection of food[21].

We have focused on identifying food-selective regions through responses to the 1,000 shared images and our hand-labeled annotations. For the approximately 9,000 remaining images per subject that were not manually labeled, we can still take advantage of COCO annotations[16] (including specific types of food) to further investigate brain responses to food and validate our findings on an independent set of images. We built an encoding model using the 80 object labels provided by COCO and obtained the resulting voxel-wise weights for food labels. We find that the voxels having the highest weights for several individual food sub-categories (i.e., *cake, sandwich, pizza, and broccoli*) fall within previously identified food-selective regions (weights for S1 in Fig. 3A). Next, we investigated the specific contribution of food images to these voxel responses by comparing two encoding models: one including the 67 non-food COCO labels, and the other including both food and non-food labels. We compared the R^2 values of the two models on held-out data (Fig. 3B and Fig. S6). Many voxels on the ventral surface show improved prediction performance due to the inclusion of food labels, suggesting that modeling the presence of food beyond other categories was required to accurately predict the voxel responses. These voxels are distributed in roughly the same spatial pattern as the voxels with high-valued weights for individual food categories and our previously identified food regions, further supporting the generality of our results.

To better understand the organization of information in our food-selective visual areas, we isolated food-selective voxels using a mask of the ventral visual cortex based on corresponding ROIs from the HCP atlas[22] (see *Methods*). The resulting "food relevant" voxel masks, which were used for the following analyses, are shown in Figures 2C and S2C. To understand the representational structure of these regions, we ran a principal components analysis (PCA) on the responses from all subjects to the shared food images. The PCA produces for each voxel a set of principal component scores that capture the projection of its high-dimensional response profile across all images onto a lower dimensional subspace. The axes of this subspace – shared semantic axes – should correspond to the dimensions in food image space that are most strongly reflected in the voxel responses (Fig. 4A). In Figure 4B and C, we visualize the top and bottom images for each PC. The first three PCs are each associated with distinct groups of voxels. PC1 is characterized by small positive patches around the center of each food-preferring strip on the ventral surface, with more negative values close to the edges of each strip. Negative and positive scores for PC2 differentiate the lateral and medial strips of the food-selective region. PC3 scores are generally more spatially diffuse, but in the right hemisphere, PC3 scores are more negative near the FFA (i.e., medial side of the lateral strip, lateral side of the medial strip). Based on inspection of the top and bottom images associated

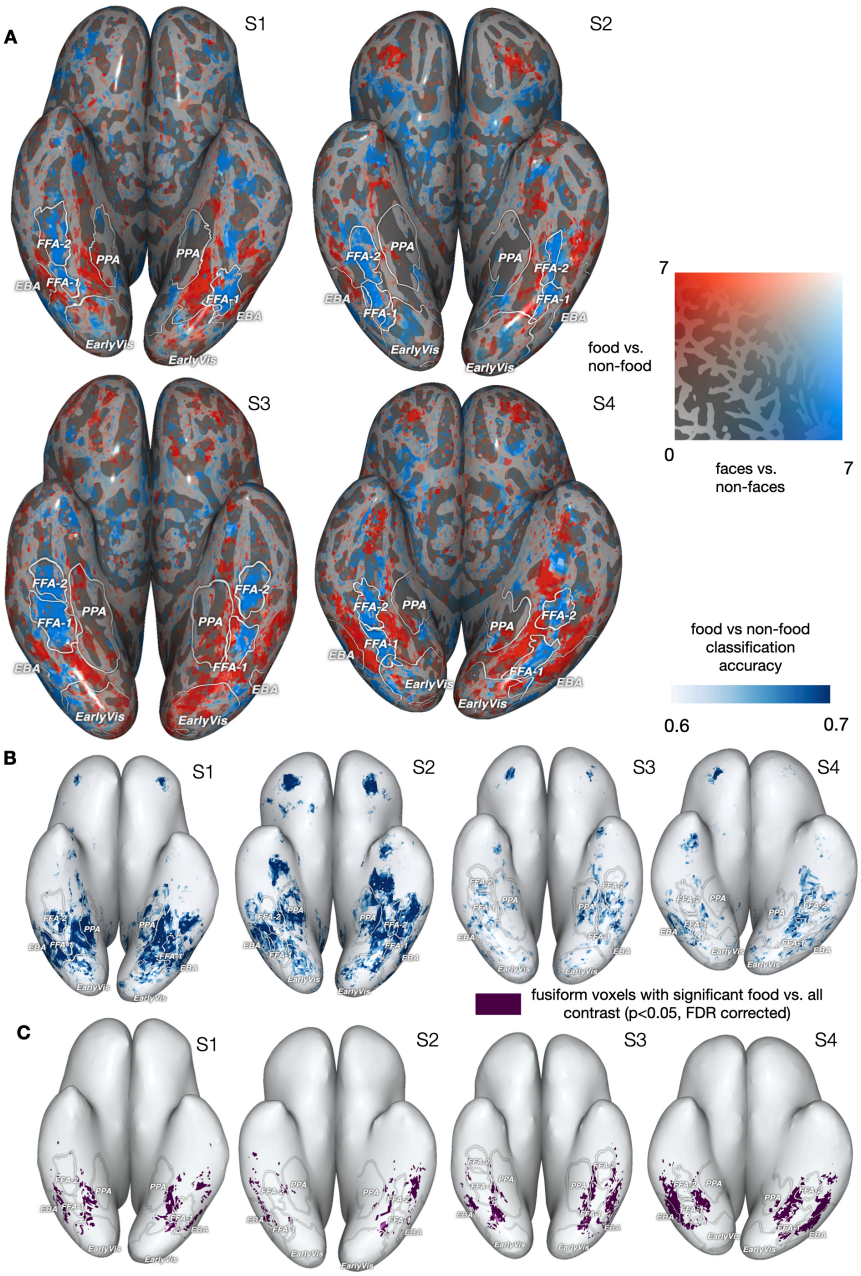

Figure 2: Food-selective regions at the individual subject level. (A) Comparing the spatial localization of food- and face-selective neural populations on the ventral surface, for S1-S4 (see Fig. S2 for S5-S8). Voxels' t-statistics from two 1-sided t-tests comparing food vs. non-food (red) and face vs. non-face (blue). The regions identified by each contrast are largely non-overlapping. This pattern is maintained for food vs. non-(food and face) and face vs. non-(face and food) (Fig. S5). (B) Classification accuracy for multivariate searchlight decoding of food vs. non-food images for S1-S4, with darker voxels signifying higher accuracy. These regions encompass the two sets of regions corresponding to high values for the food vs. non-food and the face vs. non-face contrasts (respectively red and blue in panel A). (C) Spatial mask for food-selective regions used in subsequent analyses for S1-S4 (highlighting ventral visual responses). The mask is the overlap between the region that is identified from the t-test for food vs. non-food (panel A, red) at $p<0.05$ (FDR corrected) and relevant neuroanatomically localized regions using the HCP atlas[22] (see *Methods*).

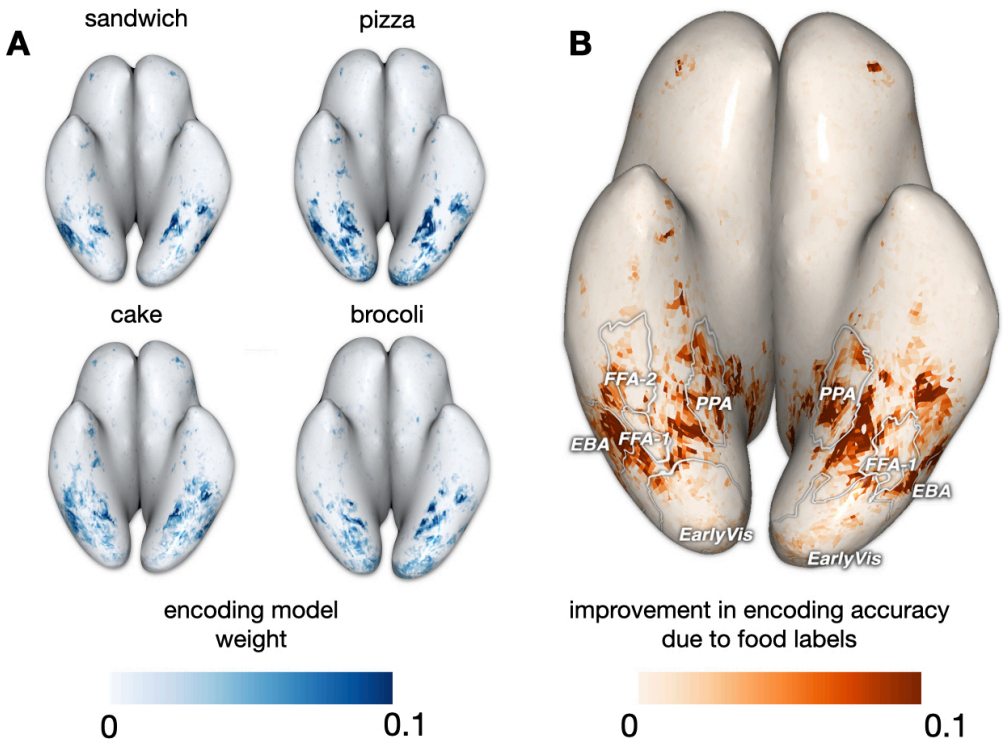

Figure 3: A consistent set of food-selective regions can be identified across independent image sets with different labeling schemes. We used the set of images for each subject that were not included in previous analyses, and an encoding model built from the 80 COCO object labels. (A) Voxel-wise encoding model weights for four food sub-categories from the original COCO dataset, shown for S1. We see variability in the weights, such as (perhaps, not surprisingly) pizza yielding higher weights in some areas than broccoli. (B) We compared predictive accuracy of an encoding model with all COCO labels (including 13 food and 67 non-food labels) to an encoding model with only the 67 non-food COCO labels. On S1's native surface, there is an improvement in validation set R^2 values when including the food labels (R^2 for the full model; R^2 for the model with food removed), with S1-S8 results in Fig. S6. Weights corresponding to individual food labels (A) and the pattern of improvement in R^2 (B) highlight similar food-selective regions. Such consistent results lend further support for these regions being robustly food selective.

with each PC, PC1 captures the prominence of food in an image, distinguishing images with food as a key focus in the foreground versus those with food as a background element. PC2 distinguishes food images based on overall scale, differentiating close-up images that focus on a few food objects from larger-scale images of food-related scenes (Fig. 4B). This is consistent with the pattern of positive scores for this PC on the medial side of the food-selective area, close to the PPA. PC3 distinguishes food images based on social attributes, separating food images that include few people from images of multiple people eating or preparing food, with social settings being at the end of the spectrum (Fig. 4C). Some amount of person or animacy-related information also appears to be reflected in the first two PCs (top right vs. bottom left images in Fig. 4B). Such results highlight the ecological importance of food as a category, as well as how high-level knowledge structures arise from the interaction between food and other ecologically important categories within the ventral visual cortex.

Finally, to explore the contributions of features based on different sources of information to the visual representation of food, we clustered food images according to their voxel responses in our food-selective regions. This analysis produces image clusters that are not easily characterized in terms of visual features, viewpoints or semantic attributes (Fig. S7A). We also constructed image clusters using two neural-network models – CLIP[23] and ResNet-18[24] – from which we derived semantic and visual embeddings that did not include the associated brain activity for the images. CLIP is trained on both images and text captions, enabling us to extract features that capture the high-level semantics of the images. ResNet-18, trained solely on images and their associated object labels, yields features with less emphasis on scene semantics. As shown in Figure S7, the clusters arising from CLIP capture semantic classes of food (e.g., fruits, deserts or meals; Fig. S7B) while the clusters arising from ResNet-18 appear more visually organized and more focused on individual elements (e.g., broccoli, pizza; Fig. S7C). Comparing the similarity of the cluster assignments of images for each of the three clustering procedures, neither CLIP or ResNet-18 clusters show any clear correspondence with our voxel-based clusters. The lack of correspondence in our clustering results suggests that the responses in food-selective areas do not organize easily into clusters based on scene semantics or object semantics.

Discussion

How are knowledge representations organized in the human brain? Within the visual system, one of the hallmarks of the past several decades has been *category selectivity* for faces, bodies, places, and words[1-5]. Consistent with the ecological importance of these categories, we predicted and found selectivity for another ecologically relevant category, food, within

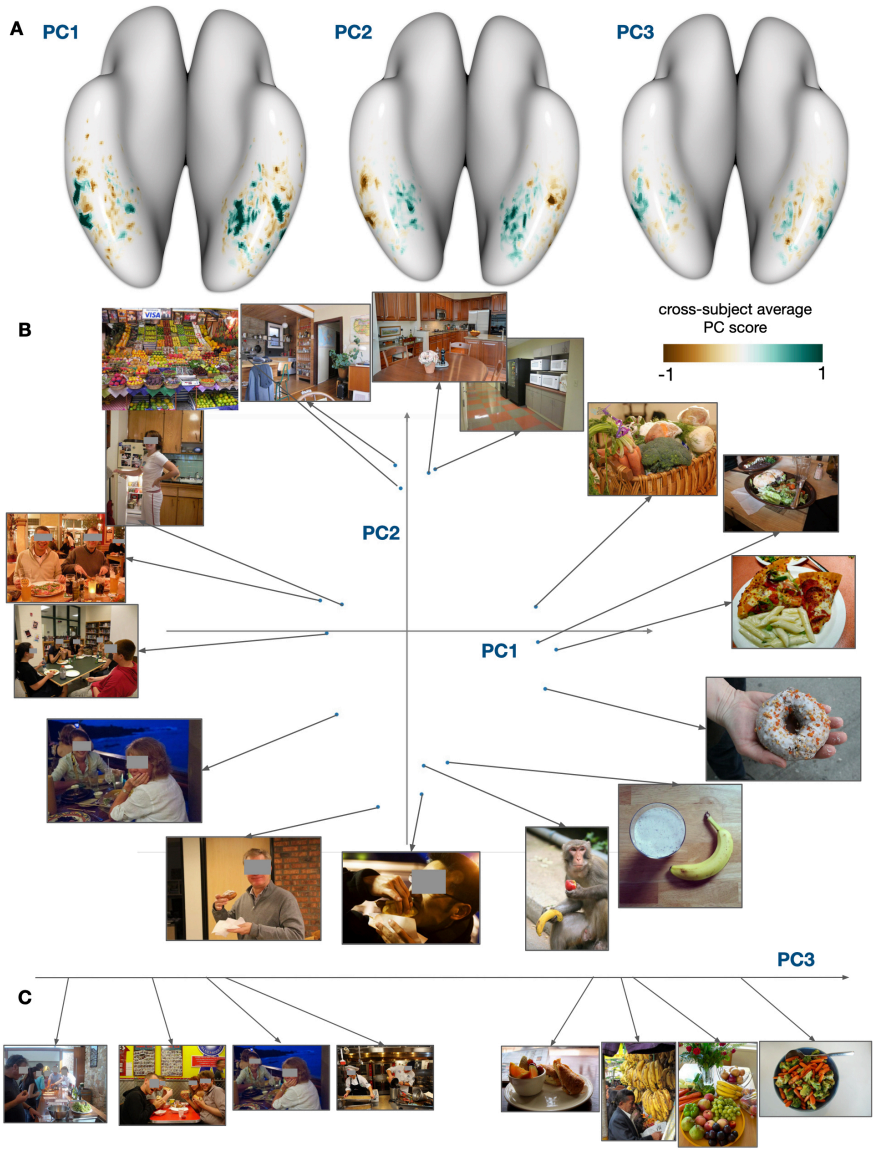

Figure 4: PCA of responses from food-selective regions provides insight into their functional structure. (A) Average principal component score across subjects for PC1, PC2, and PC3, shown on the MNI surface. Blue-green indicates high, brown indicates low PC scores. These top three PCs explain, respectively, 34.31%, 12.68%, and 11.16% of the variance. In (B) and (C), we show the images that lead to the highest and lowest activations in each PC. We include the 4 top and bottom images for ease of visualization. Top images for PC1 and PC2 are plotted in a 2D space (B), with the points connected to each image indicating its position in the space. In (C), we plot the top and bottom images for PC3 along a linear axis. Several patterns emerge here: PC1 scores yield small positive patches around the center of each food-preferring strip with more negative values close to the edges of each strip, and may capture the prominence of food in an image, separating images with focus on food in the foreground from those with food in the background. PC2 scores are higher medially (closer to PPA) and lower laterally, and seem to distinguish large-scale images of food-related places from close-by images of food and people eating food. PC3 scores in the right hemisphere food regions are lower at the center of the two strips, in the areas that border the FFA, while the left hemisphere does not show a clear pattern. PC3 appears to distinguish non-social food settings from social food settings. These results highlight that the combination of food with other ecologically important categories, including people (both faces and bodies) and places, creates a richer co-organization that reveals itself as gradients across the cortex.

the visual system, using a large-scale event-related fMRI design and images of real-world scenes[17]. We demonstrate their robustness by identifying the results using a traditional analysis and an encoding model analysis. We also verify the robustness of food-selectivity by showing consistent food-selective responses across independent sets of images, and we begin to characterize the fine-grained structure of representations within the food category itself.

Our approach and results allow us to rule out several alternative explanations for the finding of food selectivity. It is not likely that food selectivity reflects preferential responses to "reachspaces"[7], rather than food *per se*. This is ruled out on the basis that our labeling taxonomy allowed us to control for image perspective (i.e., including *reach* as a label). Specifically, we found that food-selectivity remained stable even after removing the *reach* labeled images. Another possible alternative is that food-selectivity reflects preferential responses to small vs. big real-world object size[6], again, rather than food *per se*. However, the representation of real-world object size manifests as *big* flanking the medial side of the FFA and *small* flanking the lateral side of the FFA. This explanation can be ruled out in that our observed food selective responses co-locate more with big, as opposed to small, regions, yet food categories, particularly prepared foods, have small real-world size. Finally, it is not likely that low- or mid-level visual features (i.e. color, spatial frequency) underlie our pattern of results. This is supported by the fact that food selectivity was primarily found in higher visual areas, rather than early visual areas (Fig. 2). Further, as discussed previously, the visual variability of food makes it unlikely that there is a set of low- or mid-level visual features or high-level shape structures that consistently correspond to food (in contrast, see[25–27]). These conclusions are consistent with a recent MEG study which excluded low-level visual features as an explanation for food selectivity[15]. Finally, it is unlikely that food selectivity can be solely attributed to greater attention or higher intrinsic visual salience for food relative to non-food[28]. Both human faces and bodies are subject to the same kinds of saliency effects[29], yet attentional/saliency differences are not the preferred explanation for face or body selectivity[30]. Moreover, within our study, faces and bodies comprised a reasonable proportion of the non-food contrast images, yet food selectivity was robust across these comparison categories.

Of note, we also observed food selectivity in the parietal and frontal cortices, though the localization of these regions was less consistent over subjects (Fig. 2). Although we focused on the ventral visual cortex regions based on the large body of past work investigating category-selective responses in this portion of the visual system, other brain regions may also play a role in processing food information during visually-guided behavior. The dorsal visual areas in particular may process the actions or affordances associated with food (i.e., cooking/eating), as suggested by past work showing that object representations in dorsal visual

cortex tend to be action-oriented[31,32]. Activation in frontal cortex appears to overlap roughly with orbitofrontal regions (rightmost map in Fig. S1), which may reflect the involvement of these areas in processing reward information associated with certain types of food[12,33,34]. Food selectivity was also observed in a number of subjects in the insular cortex, which has previously been implicatged in taste processing[12,34]. Future work should investigate the role of these areas, perhaps using manipulations that vary reward or action representations evoked by food stimuli.

While a finding of food selectivity naturally emerges from considering ecologically important visual categories, this leaves open the question as to how such selectivity arises in the human brain? We speculate that, similar to human language, domain-relevant perceptual inputs related to food can vary widely depending on the cultural and physical environment. Learned representations for food are only loosely constrained at the surface level, but still reflect common underlying mechanisms that have emerged over the course of evolution due to reward and the selection for learning abilities that flexibly responded to variations in inputs (the "Baldwin Effect"[35,36]). Thus, as a core property of knowledge organization, food selectivity is likely to have emerged as a neural preference shaped heavily by semantic associations and context.

Acknowledgements

Collection of the NSD dataset was supported by NSF IIS-1822683 and NSF IIS-1822929. The authors thank the NSD team for collecting and sharing the dataset. MMH was funded by a Distinguished Postdoctoral Fellowship from the Carnegie Mellon Neuroscience Institute. JP was supported by NSF BCS-1658278. The authors thank the following people for contributing ideas and commentary to this project: Laurie M. Heller, Andrew Luo.

Methods

fMRI data. We used the Natural Scenes Dataset (NSD)[17], consisting of high-resolution fMRI responses to natural scenes. The detailed experimental procedure are described by Allen et al.[17]. The naturalistic scene images were pulled from the annotated Microsoft Common Objects in Context (COCO) dataset[16]. 8 subjects each viewed between 9,000-10,000 natural scene images over the course of a year, each repeated 3 times. Of the 70,566 total images presented, 1,000 are viewed by all subjects. The data were collected during 30-40 scan sessions. Images were square cropped, presented at a size of 8.4° × 8.4° and for 3 s with 1-s

gaps in between images. The subjects were instructed to fixate on a central point and to press a button after each image if they had seen it previously.

The functional MRI data were acquired at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. The preprocessing steps included a temporal interpolation (correcting for slice time differences) and a spatial interpolation (correcting for head motion). Single-trial beta weights were estimated with a general linear model. FreeSurfer[37,38] was used to generate cortical surface reconstructions to which the beta weights were mapped. The beta weights corresponding to each image were averaged across repetitions of the image (3 repetitions of each image), resulting in one averaged fMRI response to each image per voxel, in each subject. The dataset also included several visual ROIs that were identified using separate functional localization experiments. We drew the boundaries of those ROIs for each subject on their native surface for better visualization and interpretation of the results. All brain visualization were produced using the Pycortex software[39].

Image labeling. The authors (n=7) performed manual image labeling of the 1,000 shared images based on each image's depicted location, image perspective and content. Location refers to whether the image is indoor or outdoor (or ambiguous), content refers to the categories of objects in the image (including the binary existence of food), and image perspective refers to the approximate scale of the image, discretized into *zoom*, *reach* or *large-scale*. *Zoom* refers to a very close shot, thereby likely concentrated on one object and excluding surrounding information. *Reach* images display objects at a human-reachable distance, and may activate representations related to object affordances[7,18]. *Large-scale* images encompass the remaining images, which include an image of a typical scene as opposed to one or more close-up objects. Images could only be assigned one label for location and perspective, but could be assigned multiple content labels. More details about this image labeling are described in the Figure 1A and B. Labeling was performed using the Computer Vision Annotation Tool[40]. In order to avoid variation in labels and ensure consistency, we performed several rounds of labeling and verification across multiple raters; each image was seen by a least two raters. Disagreements were discussed in the group of raters until unified labeling assignments were reached.

Encoding models. We constructed two different encoding models. The first was based on our hand-labeled annotations of the 1,000 shared images (Fig. 2). Encoding all 16 hand-labels into a single binary vector per image, we utilized voxel-wise ordinary least squares (OLS) encoding models to predict each individual voxel response to a given stimulus. Identifying voxels more responsive to Category A over Category B was done using a 1-sided t-test

between the respective learned model coefficients for each of the two categories. Note that this analysis collapses across the three "attributes" used in our labeling taxonomy (i.e., food is compared against object categories like faces, as well as against location labels like indoor). We used these methods to identify voxels that are more responsive to food than other labels, as well as for face versus other labels. We obtained a p-value from the t-value, then corrected for multiple comparisons across all voxels using the Benjamini-Hochberg False Discovery Rate procedure (FDR)[41], which is appropriate for fMRI results due to the assumption that they show positive dependence[42,43]. The significance of the contrast was computed at the subject level, the results were converted to MNI space, and plotted in Fig. 1C. See Figure S1 for an un-thresholded version of this map.

Our second encoding model was based on COCO object category labels, and made use of the set of images that were unique to each subject (3). The purpose of this model was to verify that our proposed food region derived from the 1,000 shared images is consistent across the larger set of images that also includes images not used in the first analysis. We used the 80 COCO object category annotations provided in the dataset, specifically each COCO label's corresponding bounding box proportion relative to the image (i.e., proportion of the image covered by the category of interest), as input to a ridge regression encoding model. We built and tested the model via 10-fold cross-validation, where R^2 was computed on a tenth of the data not used for training at each fold, and the 10 resulting R^2 values were averaged. The penalty parameter for each voxel was chosen independently by nested $10-$fold cross-validation. When determining which images were used to fit the encoding model, we create a set of images that contained half food and half non-food images. We considered images to include food if their maximum food label proportion exceeded a threshold of 0.15. We identified 940 such images, and randomly selected 940 non-food images, together creating a total input set of 1880 images. We built two models, one with all the labels, and one with all the labels that were not food (67 in total). We then computed the voxel-wise R^2 improvement from including food labels in the regression. In addition to helping identify voxels that responded most to inclusion of food, this encoding model also helped us visualize food sub-category activations. We observed the voxel-wise learned weights corresponding to specific COCO food labels (i.e. cake, sandwich) to uncover potential food sub-category patterns.

Decoding models. While an encoding model is able to provide some insight into single-voxel selectivity through response predictions, a decoding model can uncover distributed pattern-level representations of visual features. To observe representations at the population level, we used a searchlight decoding method[20]. Specifically, for each voxel in the cortical

sheet, we defined a searchlight sphere that consisted of 27 nearby voxels, and we trained a decoder to classify the existence of food based on the pattern of activation across these voxels. We used 5-fold cross validation via Support Vector Classification, with our input image set consisting of 108 food images and 108 randomly selected non-food images from our 1,000 shared images. High decoding accuracy from this method suggests that an area encodes food-related information at the pattern level, which our model is able to exploit in order to classify the existence of food.

Determining the ventral visual food selective regions. To generate a mask that only included the ventral visual food selective region, we first manually selected apparent relevant ROIs via the Glasser HCP Atlas[22]. We use the concatenation of sub-areas TE2p, PH, VVC, v8, PIT, and VMV3 to create our mask. After converting the mask for this anatomical area into each subject's native space, we identified the intersection of this mask with the identified food region from a food vs non-food significance test (Fig. 2 shows the final mask definition).

Principal Component Analysis (PCA). We ran PCA to better understand possible structure and/or correspondence in these food-selective regions. Using the food mask above that consists only of our proposed food region, we selected 'food-relevant' voxels for each subject. Then, we ran PCA on a matrix of concatenated 'food-relevant' voxels for all subjects (rows) by the activity related to shared food images (columns), reducing along the image dimension (the columns). We extracted the top principal axes of this matrix, and projected our initial data matrix onto the calculated lower-dimensional space to obtain the voxel-wise PC scores on the brain. To compare the voxel-wise PC scores across subjects, we converted the scores for each subject to the MNI template and average the scores across subjects for each MNI voxel. We identified the most positive and negative contributing images to each axis by computing the dot-product between the PC score and the activity related to an image, to assess whether the representations of each principle axis were cohesive or semantically interpretable.

Clustering analyses. We ran a K-means clustering analysis to better investigate visual and semantic patterns in the food selective regions. As a point of comparison with the voxel clustering results, we also clustered visual and semantic embeddings of these images derived from deep neural networks. To compute the clusters for one subject, we picked 940 food images. Voxel embeddings were calculated for each individual subject, using responses from voxels within the ventral food mask. To obtain visual and semantic embeddings for these same 940 images we used two trained deep neural networks: CLIP and ResNet-18[23,24]. CLIP,

trained on both images and text, allows us to extract features arising from a contrastive learning paradigm with dual semantic and visual constraints. We used the pretrained ViTB32 model, which was trained to align image and text embeddings within a shared space. Within this model, we extracted the features given an input image from the vision module of the model. Given an image, we call these corresponding CLIP features the CLIP embedding.

ResNet-18, trained on solely images, provides a visual feature-based embedding with no language component. Given an image, we ran a ResNet-18 model pretrained on ImageNet to extract the features from the average pool layer immediately preceding the final fully-connected layer[44]. We refer to these extracted features for a given image as the corresponding ResNet embedding of that image.

To cluster embeddings, we used K-means clustering algorithm with Euclidian distance. We consider a range of K values and for each, observe the average Euclidian distance from each data point to their corresponding cluster centroid. Next, we selected the first K value that led to the drop in the average distance for voxel embeddings beyond which the decrease plateaus (the elbow method). This value was 4. We use this same $K = 4$ for all three embedding clusterings.

To compare different clustering assignments, we constructed for each clustering procedure a 940×940 matrix where the rows and columns correspond to the 940 images. Each cell in this matrix is an indicator value where $\text{matrix}_{i,j}$ is 1 if the two images i and j are in the same cluster, and 0 otherwise. We then used Pearson correlation to compute the correlations between two clustering assignments. To visualize each cluster, we chose the closest images to the centroid of that cluster.

References

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

[2] N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–4311, 1997.

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

[4] P E Downing, Y Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

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

[6] Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

[7] Emilie L. Josephs and Talia Konkle. Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences*, 117(47):29354–29362, 2020.

[8] N Kanwisher. Domain specificity in face perception. *Nat. Neurosci.*, 3(8):759–763, 2000.

[9] M J Tarr and I Gauthier. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nat Neurosci*, 3(8):764–769, 2000.

[10] Jing Chen, Esther K. Papies, and Lawrence W. Barsalou. A core eating network and its modulations underlie diverse eating phenomena. *Brain and Cognition*, 110:20–42, 2016.

[11] Raffaella I. Rumiati and Francesco Foroni. We are what we eat: How food is represented in our mind/brain. *Psychon. Bull. Rev.*, 23(4):1043–1054, 2016.

[12] VanderLaan. The first taste is always with the eyes: A meta-analysis on the neural correlates of processing visual food cues. *NeuroImage*, 55(1):296–303, 2011.

[13] Ruud van den Bos and Denise de Ridder. Evolved to satisfy our immediate needs: Self-control and the rewarding properties of food. *Appetite*, 47(1):24–29, 2006.

[14] P E Downing, A W Chan, M V Peelen, C M Dodds, and N Kanwisher. Domain specificity in visual cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 16:1453–1461, 10 2006.

[15] Kleovoulos Tsourides, Shahriar Shariat, Hossein Nejati, Tapan K. Gandhi, Annie Cardinaux, Christopher T. Simons, Ngai-Man Cheung, Vladimir Pavlovic, and Pawan Sinha. Neural correlates of the food/non-food visual distinction. *Biological Psychology*, 115:35–42, 2016.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[17] Allen. A massive 7t fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat Neurosci*, 25(1):116–126, 2022.

[18] Emilie L Josephs, Haoyun Zhao, and Talia Konkle. The world within reach: an image database of reach-relevant environments. *Journal of Vision*, 21(7):14–14, 2021.

[19] Anthony Stigliani, Kevin S. Weiner, and Kalanit Grill-Spector. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35(36):12412–12424, 2015.

[20] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, 2006.

[21] James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[22] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, Stephen M Smith, and David C Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024, 2018.

[26] Shahin Nasr and Roger B H Tootell. A cardinal orientation bias in scene-selective visual cortex. *J Neurosci*, 32(43):14921–14926, 2012.

[27] Xiaomin Yue, Irene S Pourladian, Roger B H Tootell, and Leslie G Ungerleider. Curvature-processing network in macaque visual cortex. *Proc Natl Acad Sci U S A*, 111(33):E3467–75, 2014.

[28] Sanjay Kumar, Suzanne Higgs, Femke Rutters, and Glyn W. Humphreys. Biased towards food: Electrophysiological evidence for biased attention to food stimuli. *Brain Cogn.*, 110:85–93, dec 2016.

[29] Stephen R.H. Langton, Anna S. Law, A. Mike Burton, and Stefan R. Schweinberger. Attention capture by faces. *Cognition*, 107(1):330–342, 2008.

[30] Maura L. Furey, Topi Tanskanen, Michael S. Beauchamp, Sari Avikainen, Kimmo Uutela, Riitta Hari, and James V. Haxby. Dissociation of face-selective cortical responses by attention. *Proceedings of the National Academy of Sciences*, 103(4):1065–1070, 2006.

[31] Jorge Almeida, Bradford Z. Mahon, and Alfonso Caramazza. The role of the dorsal visual processing stream in tool identification. *Psychological Science*, 21(6):772–778, 2010.

[32] Maryam Vaziri-Pashkam and Yaoda Xu. Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *Journal of Neuroscience*, 37(36):8767–8782, 2017.

[33] Edmund T. Rolls. The Orbitofrontal Cortex and Reward. *Cerebral Cortex*, 10(3):284–294, 03 2000.

[34] Edmund T Rolls. Brain mechanisms underlying flavour and appetite. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1471):1123–1136, 2006.

[35] J. Mark Baldwin. A new factor in evolution. *The American Naturalist*, 30(354):441–451, 1896.

[36] Patrick Bateson. The Active Role of Behaviour in Evolution. *Biology and Philosophy*, 19(2):283–298, 2004.

[37] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

[38] Bruce Fischl, Martin I. Sereno, and Anders M. Dale. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999.

[39] James S. Gao, Alexander G. Huth, Mark D. Lescroart, and Jack L. Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in Neuroinformatics*, 9, 2015.

[40] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, August 2020.

[41] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[42] Christopher R Genovese. A bayesian time-course model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 95(451):691–703, 2000.

[43] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

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