Title:

Brain Feature Maps Reveal Progressive Animal-Feature Representations in the Ventral Stream

Authors: Zhang Z.[1], Hartmann T.S.[2], Born R.T.[2], Livingstone M.S.[2], Ponce, C.R.[2]

1. Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA
2. Department of Neurobiology, Harvard Medical School, Boston MA

Abstract

What are the fundamental principles that inform representation in the primate visual brain? While *objects* have become an intuitive framework for studying neurons in many parts of cortex, it is possible that neurons follow a more expressive organizational principle, such as encoding generic features present across textures, places, and objects. In this study, we used multi-electrode arrays to record from neurons in early (V1/V2), middle (V4), and late (posterior inferotemporal cortex (PIT)) areas across the visual hierarchy, estimating the local operation of each neuron over the entire visual scene. These estimates, called "heatmaps," highlight regions of strong activity over the natural scene. We found that while populations of neurons with foveal receptive fields across V1/V2, V4, and PIT responded over the full scene, they focused on salient sub-regions within object outlines. The best-captured object feature class belonged to *animals*, not general objects, as a trend that increased over the visual hierarchy. These results show that the monkey ventral stream is partially organized to encode local animal features over objects, even as early as primary visual cortex.

Introduction

When viewing a natural image, we effortlessly perceive many features — the layout of the scene, the identity and position of objects, and the surfaces against which they stand. How do neurons in the brain respond to such diverse and complex features? In early studies, inferotemporal cortex (IT) neurons were described as acting in a distributed fashion, responding to general "aspects of shape, texture, or colour"[1] that lacked a simple linguistic description. However, it was also common to find neurons responding to "faces or hands"[1], and these neurons inspired many studies using semantic categories[2–6]. The use of categories raises a contrast in understanding visual perception, specifically, on the strengths and limitations of supervised and unsupervised approaches. In decoding neuronal responses, supervised methods rely on predefined labels or categories, focusing on named distinctions such as *places* or *objects*. While this paradigm allows for important discoveries, it risks introducing bias, narrowing the field's focus to a few hypotheses or semantic categories. Further, many such decoding studies often rely on classification based on whole images, implicitly assuming that every part of the image contributed equally to the classification. These limitations can be complemented by unsupervised approaches, which allow neuronal responses to reveal pattern representations that are not constrained by pre-defined labels (patterns such as those present within textures, objects, and scenes). Both approaches are required to fully understand how the visual system encodes naturalistic scenes. To provide a complementary view of population-level coding in visual cortex, in this study, we adopted an unsupervised, local-to-global approach common in machine learning: *attribution* maps or *heatmaps*. Inspired by operations in convolutional neural networks (CNNs) like weight sharing, where a filter is convolved across an input scene, or more direct approaches such as Grad-CAM (gradient-weighted class activation mapping[7]), heatmaps visualize overall activity and identify the specific regions of an image that elicit the strongest response. In this way, these heatmaps serve as brain feature maps, revealing how a neuron, neuronal micro-clusters

(such as cortical columns or *multiunit* recordings), or entire population respond to visual stimuli. Heatmaps do not start with predefined labels but instead highlight scene areas that elicit the highest activity, allowing for an unbiased examination that could corroborate or expand our understanding of neuronal tuning[8].

Here, we investigated how neuronal populations in V1/V2, V4, and PIT responded to entire natural scenes by measuring activity in response to different regions of the same scene. By examining single- and multi-unit neuronal heatmaps within each area, we identified the visual attributes evoking the strongest responses and asked: how do neuronal sites and populations capture entire objects versus backgrounds? Our findings reveal a continuum along the primate visual hierarchy. Neurons responded to the whole scene, yet populations were also highly activated by local regions within objects, with increasing selectivity for animal features, such as faces and hands, along higher-order areas. This trend aligned with the hierarchical tuning of the primate visual system but not all artificial networks. For comparison, we generated heatmaps in state-of-the-art neural networks (convolutional-, recurrent convolutional neural networks, and vision transformers) and related them to algorithms predicting viewing behavior. While some artificial networks showed a similar progression from local to global features, their representations diverged significantly from biological heatmaps, particularly at higher layers, suggesting that future computational models of primate vision will be better aligned if they match the ethological biases of the primate brain. Finally, by letting neurons respond freely to objects, object parts, backgrounds, and textures, neuronal heatmaps reveal that object identification was only a partial representational goal of these neuronal sub-populations. Instead, the heatmaps more frequently captured local regions present in objects, particularly in animals, rather than representing the entire objects. Overall, this study underscores the value of unsupervised approaches in capturing the breadth of neuronal representations.

Results

We recorded neuronal population responses in V1/V2, V4, and PIT. In each session, we mapped each site's receptive field (RF). To create a heatmap, we simulated the "weight sharing" operation using RFs from either V1/V2, V4, or PIT (one population heatmap per region), as was done in a previous study[8]. Specifically, instead of moving the filter around the image as done in convolutional networks, we moved the image relative to the fixed RF: the animals performed a fixation task while a 16° x 16° natural scene image (embedded on 30° x 30° Brownian noise) was flashed at different positions within a 9° x 9° grid (spacing of 2°) relative to the center of the population RF (**Fig. 1, a-c**). Each scene included a variety of natural and artificial settings as well as computer-generated textures. The scenes were shown for 100 ms ON and 100 to 200 ms OFF, with 3-14 flashes between juice rewards; different scenes were interleaved within a trial (**Fig. 1c**). We measured the neuronal firing rate for 200 ms, from 50 ms before image onset to 50 ms after image offset. The raw measured output at each time step was a set of firing rates arranged into a 2-D grid (a heatmap), where each row and column index corresponded to the location of the neuron's RF over the scene, and the value was the firing rate evoked by that part of the scene (**Fig. 1d, e**). These heatmaps were resized to the shape of the original stimulus scenes (16° x 16°) and superimposed for visual inspection (**Fig. 1f, g**). The resulting heatmap revealed the neurons' preferred shapes in the scenes, highlighting the specific parts they were most activated by, eliminating investigator bias and stimulus pre-processing. We used six monkeys with chronic microelectrode arrays in V1/V2, V4, and PIT (two monkeys in all three areas, one monkey in PIT only, and three monkeys in V1 only), with up to 64 channels total per monkey (**Fig. 1h**). Additionally, two extra monkeys implanted in V1/V2, V4, and PIT, were used to test generalizability (**Supplementary Fig. 1**).

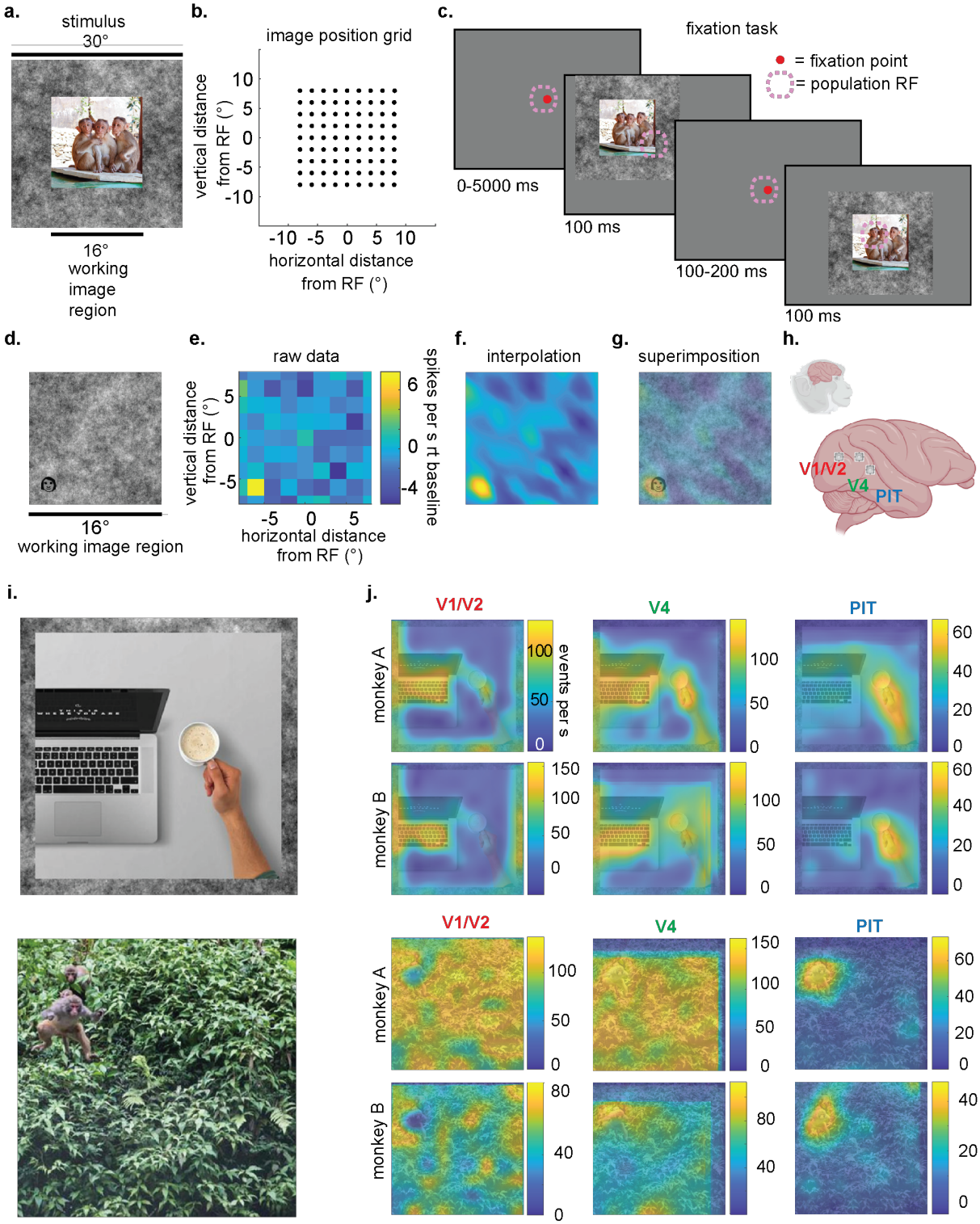

Fig. 1. Heatmaps. a. Example of a stimulus: a 16° x 16° test natural scene (the working image region) embedded within a larger 30° x 30° background of brown noise to prevent edge artifacts and motion cues. **b.** Grid of 9° x 9° = 81 stimulus positions relative to the population RF, with the center at (0, 0)°. **c.** Fixation task: monkeys fixated on a central red dot while stimuli were presented for 100 ms at different grid positions relative to the population RF (pink dashed circle). **d-g.** The simplest case of a heatmap evoked by a working image region. **d.** The working image region contains a 2°-wide cartoon face placed within a brown noise background (only the working image region is shown, which remains 16° x 16° as in panel a.). This particular stimulus serves to locate offsets between RF and the grid center. **e.** Spike rate response elicited by every one of the 81 positions within the working image region: this is the raw heatmap. **f.** Heatmap resized and interpolated (via `imresize.m`, MATLAB) to the size of the working image region (16°x16°) and **g.** superimposed on the stimulus. **h.** Schematic showing the approximate location of implanted chronic arrays for V1/V2, V4, and PIT regions. **i.** Examples of natural scenes used as stimuli. **j.** Averaged heatmaps of neuronal populations in V1/V2 (column 1), V4 (column 2), and PIT (column 3) in monkey A and monkey B generated from the natural scenes shown in **i.** The heatmaps may show a gap between the edge of the image: this reflects correction for the center position of the population RF, as these differed as a function of visual area**.** Two monkeys with implants in V1/V2, V4, and PIT are featured in this figure, although major results were replicated across all monkeys.

Neurons across the hierarchy showed increasing focus on animals.

Every single- and multiunit (a site), r_u, in an array region $r \in \{\mathrm{V1/V2, V4, PIT}\}$ generated one heatmap $H_{r_u}(x_k)$ for each natural scene $x_k \in I$, where I is the set of all K scenes (**Fig. 1i**). The heatmap $H_{r_u}(x_k)$ (a 2D matrix of shape $[h, w]$) represented the time-averaged activity of the unit r_u, over the 200 ms window. To facilitate comparisons across regions, each heatmap $H_{r_u}(x_k)$, initially in the range of $[H_{r_u}^{\min}(x_k), H_{r_u}^{\max}(x_k)]$, was normalized to the range of [0, 1]. Finally, to assess regional differences in tuning, the normalized heatmaps across s sites within the same array region r were averaged, resulting in a neuronal population heatmap $H_r(x_k)$ for each scene x_k (**Eq. 1**).

$$H_r(x_k) = \frac{1}{s}\sum_{u=1}^{s} \frac{H_{r_u}(x_k) - H_{r_u}^{\min}(x_k)}{H_{r_u}^{\max}(x_k) - H_{r_u}^{\min}(x_k)} \tag{1}$$

We found that within each region, the neuronal population responded to every part of the scene, but certain regions of the scene evoked stronger activity than others. In particular, we found that V1/V2 neurons mainly responded strongly to background features and spatial-frequency regions. In contrast, PIT neurons primarily responded strongly to animal features. Meanwhile, V4 neurons demonstrated a mixed response, responding to both features. For example, when presented with a scene containing a laptop computer next to a hand holding a cup, V1/V2 neurons responded more to the black and white keyboard — a pattern with repetitive arrangement of keys, and less to the hand holding the cup — a human body part. In contrast, PIT neurons responded more to the hand holding the cup, and less to the keyboard. V4 neurons showed responses to all of these objects (**Fig. 1j, rows 1-2**). As a second example, when presented with a scene of a distant macaque jumping in front of a rich green canopy, V1/V2 neurons responded more to the canopy, PIT neurons responded more to the monkey, and V4 neurons responded to both (with a subtle emphasis on the macaque, **Fig. 1j, rows 3-4**). Generally, we found that as we moved deeper along the ventral pathway, population responses signaled animal features, such as monkey and human faces, hands, and bodies in the natural images. However, it is important to reiterate that neuronal populations *responded across the whole scene*, with populations differing in the specific locations that they emphasized overall.

While this may suggest a general tuning towards objects within the scene, the fact that a *keyboard* evoked less population activity than a *hand* (both being "objects") suggested that more elemental features, rather than object identity alone, drove these responses. To quantify how the responses of different areas changed based on the presence of object type and shape, we obtained an independent assessment of the scene's visual content. Specifically, we created binary masks $M(x_k)$ to represent different information within the scene x_k. These masks allowed us to measure how the neuronal heatmaps overlapped with each mask. These masks were generated independently of neuronal data, using both algorithmic and crowd-sourcing approaches, and they linked specific scene regions with object labels. After obtaining a set of masks per scene, we merged masks of subordinate labels into single masks of superordinate categories (such as *moss* into *plants*, or *snouts*, *monkeys,* etc. into *animals*), resulting in a total of 27 possible semantic masks per scene (**Fig. 2a, Methods**). Next, we superimposed each mask $M(x_k, i, j)$ at each position (i, j) over the population heatmaps $H_r(x_k, t, i, j)$, which were computed at each 10 ms time step t over the 200 ms response window. This resulted in a total of $T = 20$ time steps ($t = 1, \ldots, T$). For each scene x_k and time t, we measured the average population neuronal response inside the mask $R_{in,r}(x_k, t)$ and outside the mask $R_{out,r}(x_k, t)$, in each region r (V1/V2, V4, and PIT) (**Fig. 2b, Eq. 2, 3**).

$$R_{in,r}(x_k, t) = \frac{1}{hw}\sum_{M(x_k,i,j)=1} H_r(x_k, t, i, j) \tag{2}$$

$$R_{out,r}(x_k,t) = \frac{1}{hw} \sum_{M(x_k,i,j)=0} H_r(x_k,t,i,j) \quad (3)$$

We first examined how the *animal* mask captured the responses of these areas. We quantified this effect by computing the mean activity difference $D_r(x_k,t) = R_{in,r}(x_k,t) - R_{out,r}(x_k,t)$ and measuring the area under the curve (AUC) of $D_r(x_k,t)$. As a key control, we shuffled the response regions for each scene to obtain a shuffled heatmap $H_{r(shuffled)}(x_k,t)$ and also measured the AUC of $D_{r_{shuffled}}(x_k,t)$.

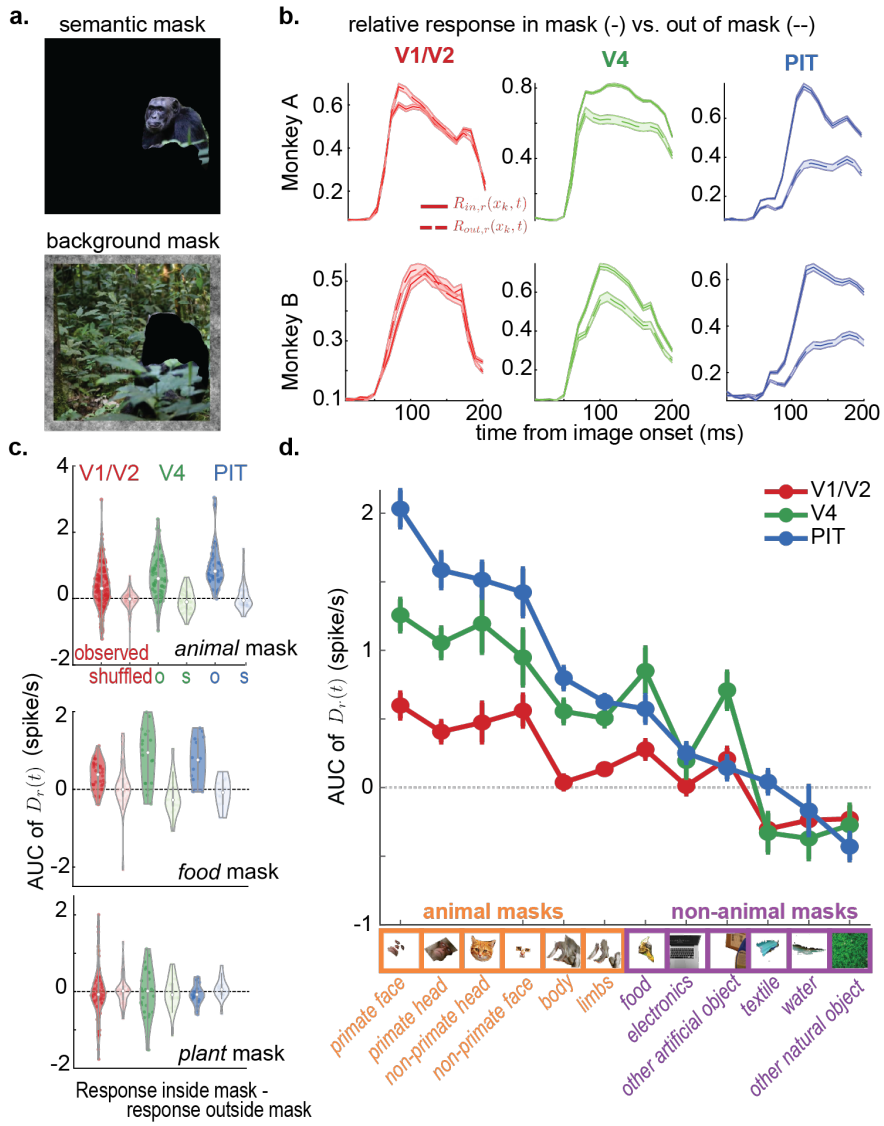

Fig. 2. Neurons across the Hierarchy Showed Increasing Focus on Animal Features. a. Example of the *Animal* masks derived from the scene of a chimpanzee in a jungle (top) and the background mask of the same image (N = 2 animals for illustration). **b.** Population heatmap activity inside (solid lines) and outside of the mask (dashed lines, ±SEM), for V1/V2 (red), V4 (green) and PIT (blue), plotted as a function of image onset (N = 2 animals for illustration). **c.** Aggregate difference between the curves, measured as the area under the curve (AUC) for the difference curve defined by the in- vs. outside response curves $AUC(D_r(x_k,t))$, for V1/V2, V4 and PIT (red, green, blue, dark colors, *N = 283*, comprising responses of neuronal populations for six animals, up to 36 pictures); the observed values in dark colors are compared to the shuffled-map AUC values (lighter colors). Top plot: *Animal* mask, middle plot: *food* mask, bottom plot: *plant* mask. **d.** Neuronal response $AUC(D_r(x_k,t))$ per visual area, in six monkeys, for animal (orange) and non-animal (purple) representative categories. Insets show examples of the content of the animal and non-animal masks.

Animal masks identified regions evoking stronger neuronal activity for PIT than for V4, and stronger for V4 than for V1/V2. Still, these masks isolated V1/V2 responses better than they did for

shuffled controls (**Fig. 2c, top**, observed AUC median ± SE; V1/V2: 0.19 ± 0.01, V4: 0.63 ± 0.06, PIT: 0.73 ± 0.08; shuffled maps: V1/V2: -0.0083 ± -0.0002, V4: 0.005 ± 0.004, PIT: -0.030 ± -0.001). This suggests that the *animal* masks represented a mix of low- and high-level saliency features, such as those detected by shallow algorithms like graph-based visual saliency metrics[9], but also by deeper models such as CNNs. Generally, AUC values for *animal* masks were higher along the ventral stream than AUC values for non-*animal* masks (**Fig. 2c, middle and bottom**). Specifically, the highest differences in values were for primate *face* and *head*, followed by non-*primate face* and *head,* and then by the *body*, *limbs* (**Fig. 2d**). Inter-areal differences in AUC decreased in non-*animal* masks, *food*, *textiles*, *electronics*, and *water* (as examples, separately considering *food, books, environment, plant,* and other labels; AUC values, median ± SE; V1/V2: 0.022 ± 0.003, V4: 0.16 ± 0.01, PIT: 0.006 ± 0.003; AUC for shuffled maps: V1/V2, 0.0057 ± 0.0003, V4: -0.031 ± 0.002, PIT: 0.002 ± 0.002). These results suggest that the monkey visual ventral pathway, as early as V1, is organized, at least in part, to increasingly encode background-isolated regions containing animal features. The fact that results differed as a function of mask content suggests this trend was not a trivial consequence of the analysis or from architectural interareal differences such as receptive field size. One important caveat is that the parts of the lateral ventral stream from which we recorded emphasize the central visual field; a different set of representational patterns might be found when recording from neurons representing eccentric regions of the visual field, as these neurons show tuning for less curved features[10,11]**.**

Populations responded to objects, but prioritized local features.

Above, we showed that areas across the ventral stream progressively prioritized the processing of animal information, with the strongest focus observed in PIT. Conversely, areas like V1 and V4 showed strong processing activity to more background and inanimate visual features. Upon visual inspection, we observed that the neuronal heatmaps highlighted certain subregions within the broader semantic masks, instead of fully encompassing the entire object outline as defined by human participants and object-trained deep networks. To delve deeper into this observation, we further investigated whether neuronal populations processed *category* information by selectively responding to localized visual features, to the overall segmented object, or a combination of both.

For this analysis, we used two monkeys with arrays implanted in V1/V2, V4, and PIT (monkeys A, and B). To quantify the spatial extent of neuronal activity, we introduced a metric called a *feature-to-object ratio* (FTO), an estimate of the neuronal activity's coverage within an object. Here, *objects* are defined as semantic classes that could be outlined by masks $M_o(x_k)$ of the object o of the scene x_k, such as *monkey* or *car* (**Fig. 3a**). If the peak neuronal activity were uniformly distributed within the object, the FTO would attain a value of 1 (**Fig. 3b**); conversely, if none of the peak responses fell inside the object outline, the FTO would be 0. To compute the FTO, we first normalized the neuronal activity heatmaps $H_{r_u}(x_k)$ of each unit r_u to the range of [0, 1] and obtained the population heatmaps for each region r using **Eq. 1**. The FTO for each object o across all K scenes, was computed by multiplying each binary mask $M_o(x_k)$ with the heatmap $H_r(x_k)$ in region r at every position of (i, j). The sum of these values was normalized by the size of the mask, across scenes (**Eq. 4**). To control the relative frequency of response hotspots within a scene, we shuffled the activity within the map, as in the above analyses.

$$FTO_r(o) = \frac{1}{\sum_{k=1}^{K}\sum_{i,j} M_o(x_k, i, j)} \sum_{M_o(x_k,i,j)=1} H_r(x_k, i, j) \tag{4}$$

We found that the FTO ratio was less than 1 for all visual areas. We also found that the median FTO ratio to *animal* masks depended on visual area, increasing from V1/V2 to PIT (V1/V2, V4, PIT values median ± SE, Monkey A: 0.5144 ± 0.0001, 0.5295 ± 0.0002, 0.5514 ± 0.0001, ANOVA: F = 5.713, p = 3.4 x10^{-3}, DF = 2,2169); Monkey B: 0.4493 ± 0.0002, 0.4845 ± 0.0003, 0.5286 ± 0.0002; ANOVA: F = 25.305 p =1.5 x10^{-11}, DF = 2,1747) (**Fig. 3c,d**). We fit a linear model to

the median FTO ratio across V1, V4, and PIT and compared the slope k. The trend across visual areas depended on specific categories and segmentation mask size: for example, the median $k(FTO(head))$ was larger than for $k(FTO(animal)$, which was larger than $k(FTO(other\ natural\ objects))$ (k(FTO) median ± SE in head, eyes, animal, other natural objects in Monkey A and B: 0.057 ± 0.006, 0.048 ± 0.008, 0.032 ± 0.005, -0.011 ± 0.006; shuffled control: -0.026 ± 0.004, -0.012 ± 0.005, -0.025 ± 0.006, -0.030 ± 0.006; **Fig. 3e**). Finally, we replicated these analyses using two new monkeys (monkeys C and D), also implanted in V1/V2, V4, and PIT. These monkeys showed the same effect (**Supplementary Fig. 1**). This result suggests that neuronal populations ranging from V1/V2, V4, and PIT were more turned to specific localized features within objects than to the overall identity of the animal objects.

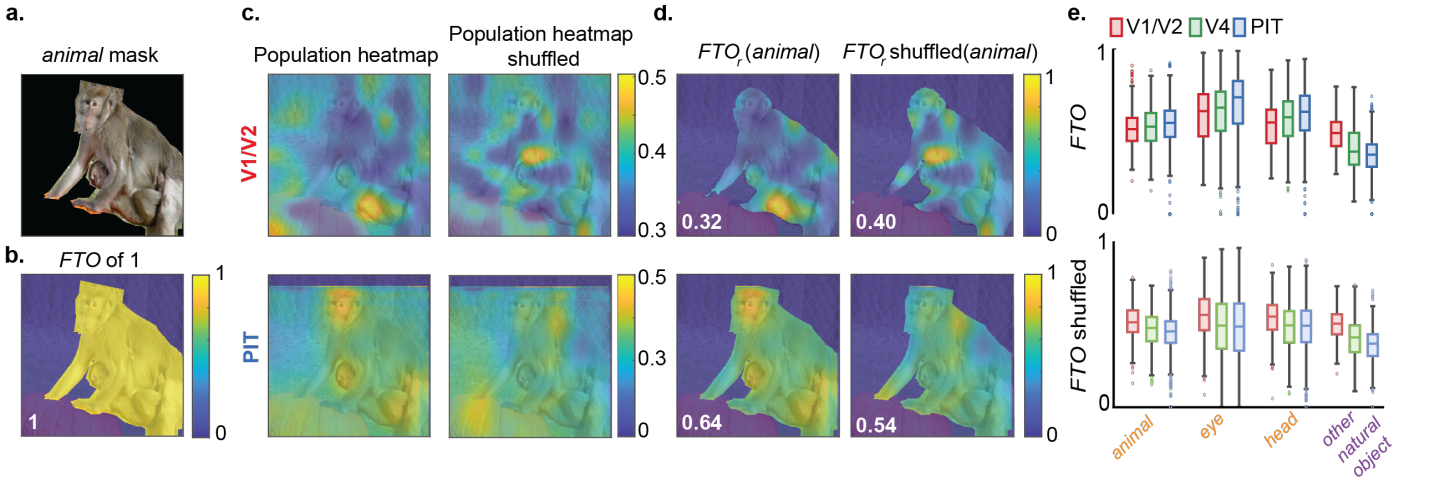

Fig 3. Feature to Object. a. animal mask of a natural scene of a mother monkey and a baby monkey. **b.** Illustration of neuronal activity was distributed evenly within the object outline resulting in FTO of 1 **c.** PIT and V1 population heatmap and shuffled heatmap in Monkey A. **d.** FTO(*animal*) of PIT and V1/V2 heatmaps in b. **e.** Top: FTO ratio of Monkey A in V1/V2 (red), V4 (green), PIT (blue) for *animal*, *eye*, *head*, and *other natural object*; the observed FTO values in dark colors. Bottom: In comparison, the FTO values in shuffled-map (lighter colors) for the example. Monkeys A and B used for this analysis. Results replicated in monkeys C and D.

Single site basis for the animal-feature focus

As noted above, a key result of this study is that animal-feature focus increased along the hierarchy of ventral visual areas, from V1/V2 to V4, and PIT. These results describe the tuning of neurons (with perifoveal receptive fields) as a function of brain region, specifically, by averaging the activity of all neuronal sites within each given image region (i.e., population heatmaps). Next, we examined whether *individual* neuronal sites also showed stronger responses to animal-related features within natural scenes. Specifically, we explored if this trend arose from one or more of the following mechanisms: a) an increase in the number of animal-tuned sites, b) an increase in sharper tuning among these sites, or c) both.

We began by asking whether the population heatmap reflected individual site heatmaps, or whether it emerged only from the combination of many diverse individual heatmaps. In other words, how similar was each site's heatmap to that of its local population? To address this, we calculated the individual-to-population Pearson correlation between each site's heatmap and the population heatmap (excluding the site in question to avoid circularity). A strong correlation (close to 1) would indicate that on average, individual site maps were highly similar to the population map. In contrast, a weaker correlation (close to zero) would suggest a greater diversity among individual maps, consistent with the hypothesis that the population map arises from only the combined activity of individual maps signaling objects or object parts. Across regions (V1/V2, V4, and PIT) and monkeys, individual sites showed a relatively diverse level of feature selectivity that aligned moderately with the population heatmaps. Individual V1/V2 maps showed a mean correlation of 0.371 ± 0.002 (N = 15,996 values), V4 maps, 0.352 ± 0.003 (N = 8,250), and PIT maps, 0.221 ± 0.002 (N = 14,882).

There was a reliable effect of region (V1/V2, V4, or PIT) on these correlations (one-way ANOVA F = 1251.2, $p < 0.0001$, **Fig. 4a-b**). Per visual examination, we noted that some individual heatmaps had a lower correlation to the population heatmap but also showed relatively weak firing rate responses. This suggested a potential issue regarding channel quality. To investigate this, we compared the individual-to-population correlations to a "channel quality" metric. This metric was a split-half correlation, calculated by holding out one experiment, dividing each channel's responses to a given scene into two halves, generating separate heatmaps for each half, and measuring the correlation between the two heatmaps. Across all monkeys and regions (V1/V2, V4, and PIT), the mean split-half correlation was 0.414 ± 0.002. Channels with higher signal-to-noise ratios (SNRs) showed greater individual-to-population correlations (Pearson correlation coefficient: 0.29, $p < 0.0001$, N = 38,010 comparisons). This supports the interpretation that the population heatmaps reflected a distributed code. While many individual sites showed strong similarity to the population heatmap, most sites

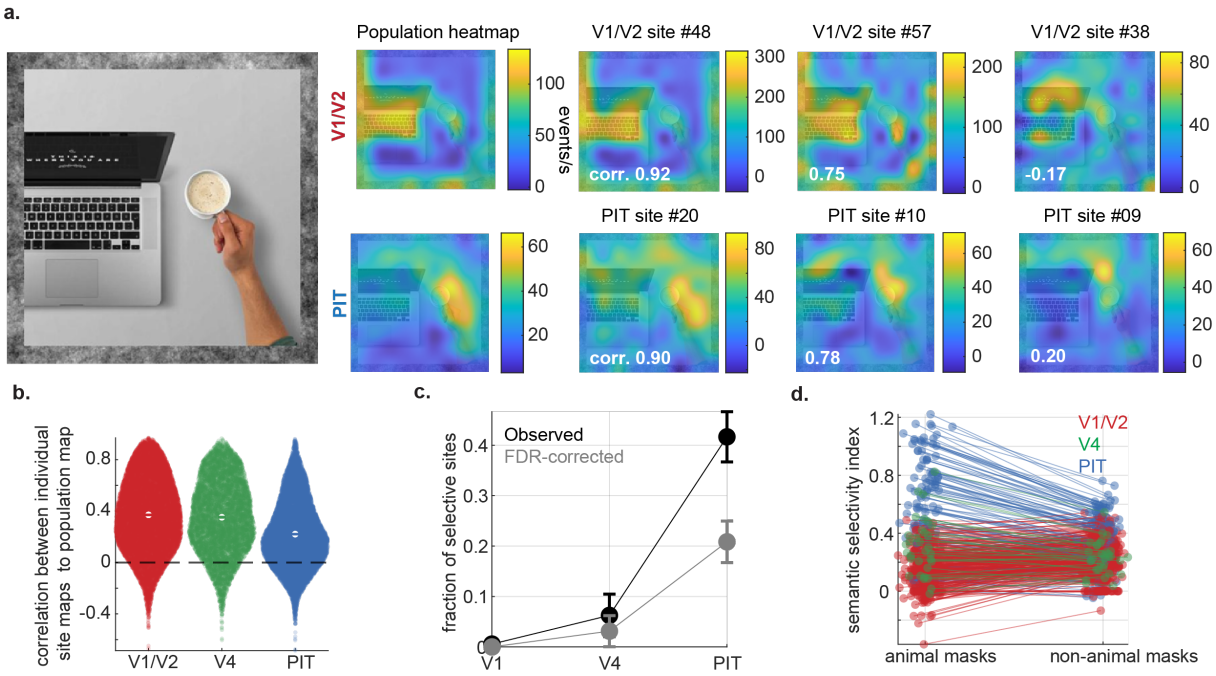

Fig. 4. Population and Individual-site Heatmaps. **a.** Example stimulus image (left). Population-level heatmaps for V1/V2 (top left) and PIT (bottom left) shown alongside representative individual site-level heatmaps for V1/V2 (top row) and PIT (bottom row). Each site's heatmap shows its correlation to the population map (Monkey A). **b.** *Swarm plots* showing the distribution of correlations between individual site-level maps and their respective population maps for V1/V2, V4, and PIT*, across all animals. Central value shows the mean correlation.* **c.** Fraction of *animal-feature preferring* sites across V1/V2, V4, and PIT*.* Selectivity *was* defined by a significant semantic selectivity index (SSI) favoring animal-related masks ($p < 0.01$*, observed fractions in* black*, gray shows fractions after* FDR-correc*tion of p values)*. **d.** Semantic selectivity indices (SSI) for animal-related and non-animal-related masks across individual neuronal sites in V1/V2(red), V4(green), and PIT(blue). Lines connect paired animal and non-animal SSIs for *a given* site. All six monkeys were used for these analyses.

appeared to capture different parts of the same scene. Collectively, these sites responded to all aspects of the scene, with an increased focus on various specific animal- and object-part features, particularly in PIT.

Next, we measured the fraction of animal-feature-preferring sites within each region. For each neuronal site, its heatmap was normalized between [0, 1] and multiplied by semantic masks (e.g., *animal*, *plant*, *electronics*) to isolate responses within the defined regions. The activity within these masks was normalized by mask area, and responses to animal-related masks were compared to those for other semantic masks using a Wilcoxon rank-sum test, testing whether the distribution of responses within the animal-related masks was significantly different from the responses within other masks. We chose a strict threshold of p value = 0.01 to designate sites as "preferring" animal features

and counted the number of sites per region across all monkeys. False-discovery rate (FDR) correction was also applied to account for multiple comparisons. We found that the fraction of sites preferring animal-related features increased throughout the ventral hierarchy: the fraction of animal-feature sites was 0.006 ± 0.004 in V1/V2, 0.062 ± 0.044 in V4, and 0.417 ± 0.051 in PIT (after FDR correction: 0 in V1/V2, 0.031 ± 0.030 in V4, and 0.208 ± 0.041 in PIT, **Fig. 4c**). These findings support the hypothesis that the increasing representation of animal-related features along the ventral stream was partially based on an increasing number of selective sites.

Finally, we asked if neuronal sites were getting more selective for animal features (perhaps by acquiring sparser "tuning widths," as it were. This is tricky to do in experiments without ordered numerical axes such as orientation or direction. As an approximation, we measured how strongly a given neuronal site responded to features within a specific semantic mask, compared to outside the mask, for each scene. We defined a semantic selectivity index (SSI) for this analysis. The index is computed by normalizing each individual site's heatmap to a range of [0, 1], then applying a semantic mask (*e.g., animal*) to isolate the region of the heatmap corresponding to the feature of interest. The neuronal activity inside the mask was computed by multiplying the binary mask with the heatmap. Similarly, the activity outside the mask was determined by applying the complement of the mask to the heatmap. The SSI was then defined as the difference between these two values. Positive SSI values indicate that the site responded more strongly to the semantic feature, while negative values suggest reduced or suppressed activity to that feature. Using previously defined notations, the SSI for each site r_u is the mean activity difference, $R_{in,r_u}(x_k,t) - R_{out,r_u}(x_k,t)$, averaged over time (t) across all scenes (k) as in **Eq. 5**.

$$SSI_{r_u}(o) = \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T}[R_{in,r_u}(x_k,t) - R_{out,r_u}(x_k,t)] \tag{5}$$

Since not every scene contained features for every semantic mask, we grouped all SSIs into two broad categories: animal-related and non-animal-related. This ensured that all images could contribute to the individual neuronal site's analysis. The animal-related group included masks corresponding to biologically relevant categories such as *bird*, *monkey*, *quadruped*, *human*, and *face*. The non-animal-related group, in contrast, included masks for semantic features unrelated to animal entities, such as *plant*, *vehicle*, *furniture*, and other non-*animal* objects. We found that, for individual sites in V1/V2, animal-, and non-animal masks showed comparable responses favoring non-animal masks, with median SSIs of 0.14± 0.01 and 0.23 ± 0.01 (SE per bootstrap, N = 358, $p < 0.0001$, Wilcoxon rank sum test). In contrast, V4 and PIT showed higher SSIs for animal-related features (V4, median SSI for animal and non-animal masks: 0.30 ± 0.047, 0.23 ± 0.02, N = 32, $p = 0.01$; PIT, median SSI for animal and non-animal masks: 0.68 ± 0.04, 0.39 ± 0.03, N = 96, $p < 0.0001$, **Fig. 4d**). This suggested that individual neuronal sites showed increasing selectivity to animal features over non-animal features depending on their locations within the ventral stream.

In summary, we conclude that the animal-feature focus along the ventral stream resulted from two mechanisms: an increase in the number of sites that are more strongly activated by animal features, and for a given neuronal site, a disproportionate increase in the activation towards animal features, although the site also responds robustly to features in other types of masks.

Some deep networks showed an animal-feature focus.

Next, we asked if any state-of-the-art visual hierarchical system developed a focus on animal features as the primate visual system did. We searched for this phenomenon in artificial neural networks (ANNs), selecting 18 models with architectural diversity, including convolutional neural networks (CNNs), convolutional recurrent neural networks (CRNNs), and vision transformers (ViTs), all pretrained on ImageNet[12]. We also wanted to assess whether a model's capability to match neuronal response patterns could anticipate this phenomenon. To achieve this, we ensured that the

selected models displayed a diverse range of scores in the Brain-Score benchmark[16]. To simulate the neurophysiological experiments above, we tested each model with the same set of scenes I and sampled three layers at different depths in each artificial model ("early," "middle," and "late") to emulate the three cortical stages (V1/V2, V4, PIT). By systematically sampling layers from distinct depths of each model, we ensured that the selected layers represented progressively deeper levels of feature abstraction, analogous to the functional roles of cortical stages. Once each layer was selected, we randomly sampled s = 32 units within that layer, simulating the sampling of the microelectrode arrays (**Fig. 5a,** see **Methods**). Similar to **Eq.1**, we defined $F_l(x_k)$ as the average normalized activated feature map for a natural scene x_k. This is calculated as the mean activation over all units l_u in each ANN layer l, as shown in **Eq. 6**.

$$F_l(x_k) = \frac{1}{s}\sum_{u=1}^{s}\frac{H_{l_u}(x_k) - H_{l_u}^{\min}(x_k)}{H_{l_u}^{\max}(x_k) - H_{l_u}^{\min}(x_k)} \tag{6}$$

As in the *in-vivo* experiments, we quantified mean activation inside the mask $R_{in,l}$ and the mean activation outside the mask $R_{out,l}$ for each layer l in every model m. We then measured the area under the curve (AUC) for the difference ($D_l = R_{in,l} - R_{out,l}$) (**Fig. 5b, Supplementary Table A**).

We found that as a group, ANNs showed a wide diversity of behaviors. Heatmaps in some ANNs, such as robust ResNet-50 and CORnet-S, displayed the same increasing focus toward animal features, whereas other ANNs such as Xception and AlexNet did not. To quantify whether an ANN trend resembled those observed in the monkeys, we applied three criteria: (1) the ANN should show an increasing focus on animal features across layers, (2) this increasing trend would be of comparable magnitude to the neurophysiology data, and (3) not only with animal features but also subordinate categories (**Fig. 5c, Supplementary Table B**). We used a series of tests to narrow the list of ANNs that act more like the monkey brain.

To test criteria (1), we measured the ordinary least square (OLS) fitting slope k_a to animal-features across $D_{l_{early}}$, $D_{l_{middle}}$, $D_{l_{late}}$ for each monkey or model m, and also the slope k_{na} to non-animal-features (**Fig. 5b, e**). The neurophysiology data showed a positive slope across areas for animal features but a flat slope for non-animal features (OLS slope median ± SE: k_a: 0.38 ± 0.03, k_{na}: 0.07±0.05), indicating an increasing focus on animal features from V1/V2 to PIT. CORnet-S showed a positive trend that was similar to the monkey data (k_a: 0.48 ± 0.06 , k_{na}: 0.05 ± 0.04), whereas ViT-base, ViT-small, and SqueezeNet also showed positive trends but were less similar to the monkey brain (ViT-base, k_a: 0.98 ± 0.14 , k_{na}: -1.33 ± 0.25; ViT-small, k_a: 0.67 ± 0.07, k_{na} :0.15 ± 0.04; SqueezeNet, k_a: 0.26 ± 0.06, k_{na}: -0.13 ± 0.04). ANNs such as VOne-CORnet-S, Xception, and AlexNet showed flat trends to both animal- and non-animal features (VOne-CORnet-S, k_a: 0.08 ± 0.08, k_{na}: -0.11±0.05; Xception: k_a: -0.15 ± -0.04, k_{na}: -0.10 ± 0.02, AlexNet k_a: -0.09 ± 0.06, k_{na}: -0.11 ± 0.03).

Next, we compared the trend magnitude between the ANN- and neurophysiology results by asking if there were similar effect differences $R_{in,l}$ between $D_{l_{early}}$, $D_{l_{middle}}$, and between $D_{l_{middle}}$ and $D_{l_{late}}$. To do this, we combined two-sample tests of significance, a correlation of stage difference, and a multivariate ANOVA (MANOVA). Monkeys showed a difference $R_{in,l}$ between $D_{l_{early}}$, $D_{l_{middle}}$ and between $D_{l_{middle}}, D_{l_{late}}$ (two sample t-test, p: 5.98×10^{-9}, 5.72×10^{-2}, DF: 195, 143). Among the ANNs that met criteria (1) with positive k_a values, only VOneNet-Resnet50-at showed a significant change between both $D_{l_{early}}$, $D_{l_{middle}}$ and $D_{l_{middle}}, D_{l_{late}}$ (two-sample Student's t-test, p = 1.07×10^{-2}, 1.88×10^{-4}, DF, 58), while respectively, Resnet-50 robust, ViT-small, CORnet-S, SqueezeNet, Resnet-50 and ViT-base showed more similar trends as the monkeys (MANOVA predictor p, Resnet-50 robust: 2.2×10^{-8}, ViT-small: 5.8×10^{-12}, CORnet-S: 0.01, Squeeze Net: 2.2×10^{-5}, Resnet-50: 3.2×10^{-10}, ViT-base: 5.2×10^{-6}. We also ranked each network by its similarity to the monkey trends, using a Pearson correlation coefficient, and found that the best correlations were shown by ViT small (0.96), SqueezeNet and ViT base (0.88), and CORnet-S (0.82). Among the non- and anti-correlated networks were VOneNet-ResNet-50 (-0.09), AlexNet (-0.59), and ViT tiny (-0.83, **Fig 5d**). We found a

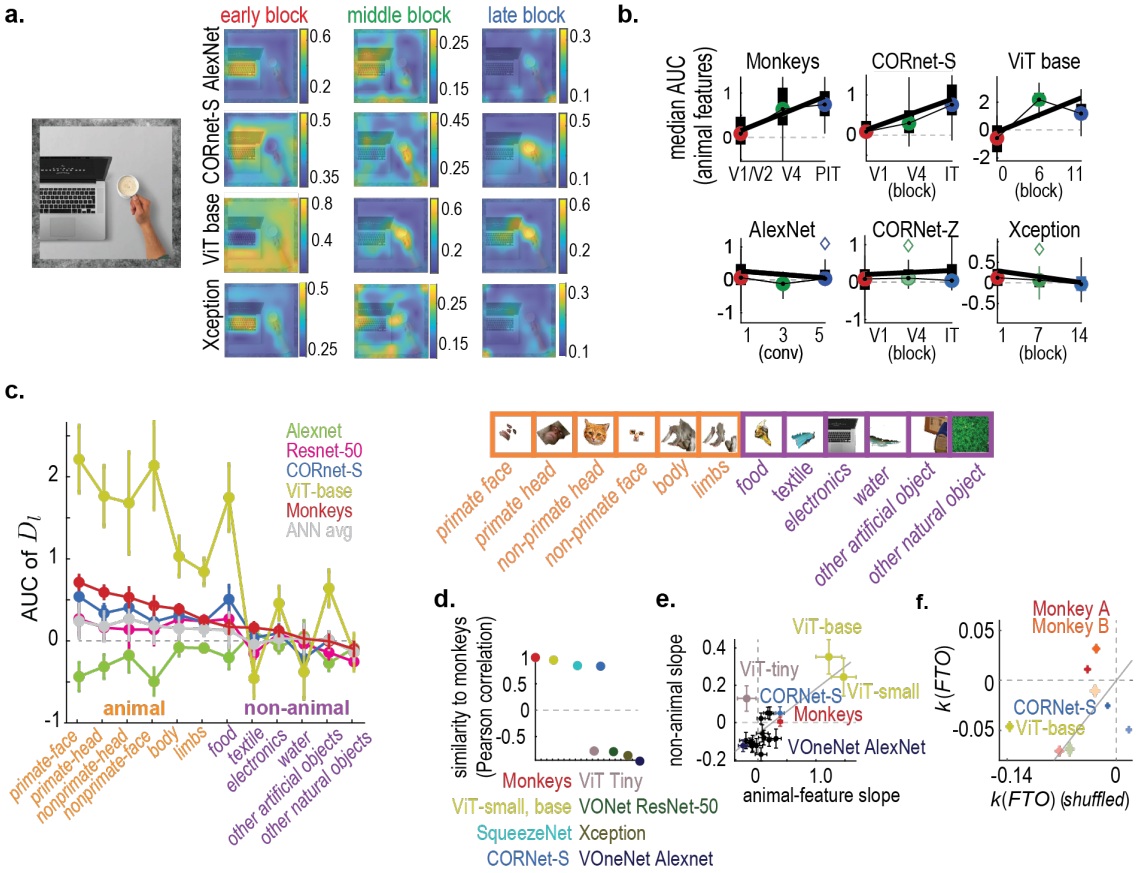

Fig. 5. Evaluation of Deep Neural Networks on Brain Similarity. a. Example of the heatmaps from monkeys from V1 to V4 to PIT and ANNs in early, middle, late layers to the scene of a person with partial face visible **b.** Mean (population heatmap) activity inside (solid lines) and outside of the mask (dashed lines, ±SEM), for V1/V2/early block (red), V4/middle block (green) and PIT/late block (blue) **c.** non-animal features vs animal features AUC difference for monkeys and ANNs, for representative categories. **d.** Pearson correlation coefficients measuring the similarity of each artificial neural network (ANN) to neural data. Each dot represents an individual model, ranked along the y-axis by its correlation with monkey trends. Higher values indicate stronger alignment with the monkey neural hierarchy. Dots are color-coded to match the top and bottom-ranked model names, for visual clarity. **e.** Scatterplot comparing the slopes of feature selectivity trends for animal-related features (x-axis) versus non-animal-related features (y-axis) across monkeys and artificial neural networks (ANNs). Each point represents a single model or the monkey data, with the x-axis capturing how feature selectivity for animals increases across layers and the y-axis capturing the trend for non-animal features. Models closer to the monkey point exhibit trends that better align with monkey neural data. Error bars indicate the variability of slopes for each feature type. The gray line shows the least-squares regression line to the scatter plot. **f.** Slope of FTO vs slope of FTO in control for animal masks (dark colors), and *other natural objects* (lighter colors) in Monkeys, CORnet S and ViT base. All of the monkeys were used in these analyses.

correlation between a network's trend to capture animal features over its hierarchy, and the same network's trend to capture non-animal features (linear fit slope = 0.24 (0.14-0.33, 95% CI), suggesting that part of the trend can be explained by a focus towards objects in general (**Fig. 5e, Supplementary Table C**). CORnet S and ViT base had a negative $k(FTO)$ for both animal- and natural texture features, in contrast to positive $k(FTO)$ for animal features in monkeys (**Fig. 5f, Supplementary Table D**). Overall, we find a high diversity in feature maps that emerges across ANNs, given the same objective task (*i.e.*, ImageNet classification); this diversity of feature maps provides a substrate that can be used for extensive projects aligning deep networks of vision with neuronal electrophysiology and perceptual reports[13].

Free viewing and automated saliency maps correlated with V4 and PIT heatmaps

These activity heatmaps showed the activity of a subpopulation of neurons over a large visual scene. One hypothesis is that monkeys might be more likely to make saccades to regions of the scene that also drive the mean population activity strongly. To determine if these heatmaps could serve as a basis for bottom-up saliency, we measured how monkeys looked at the visual scenes under free-viewing conditions, and composed *fixation maps* which we could correlate with the neuronal heatmaps. We measured the similarity of neuronal heatmaps to fixation maps and to predictions made by each artificial model.

In the experiments, we measured the monkeys' eye movements in a free-viewing task, using the same natural scenes I, presented at the same size as in the neuronal experiments. In this task, each monkey initiated a trial by fixating a small (0.2°-diameter) round target in the center of the monitor. After acquisition, the scene was flashed on for one second and the animal was then free to view any features in the scene. If they kept their gaze within the natural scene, they received a liquid reward. By tracking their gaze over the stimuli, we were able to create an eye-position heatmap per scene (fixation maps, **Fig. 6a-c**) and compared it with the corresponding neuronal activity heatmaps from different regions (**Fig. 6c, *ii***) for the same scenes. Finally, we processed the same visual scenes I using saliency models such as Graph-based visual saliency (GBVS[9]), Itti-Koch saliency[14], and *Fast, Accurate, and Size-Aware* salient object detection (FASA implemented in DeepGaze[15], **Fig. 6c, *iii***). We correlated all fixation- and saliency maps to each other and to the V1/V2, V4, and PIT population heatmaps, both observed and shuffled (**Fig. 6d**).

We found that, across monkeys, the V1/V2 heatmaps were most closely correlated to each other (Pearson coefficient, 0.751 ± 0.022) than were the V4 heatmaps to each other (V4: 0.670 ± 0.026) or the PIT heatmaps (PIT: 0.099 ± 0.037, **Fig. 6e**); shuffled maps showed much lower correlations (shuffled V1/V2: -0.054 ± 0.025, shuffled V4: 0.028 ± 0.028, shuffled PIT: -0.017 ± 0.020). The V1/V2 maps resembled the V4 maps more than the PIT maps (Pearson correlation between V1/V2–V4: 0.67 ± 0.03; between V1/V2–PIT, 0.10 ± 0.04 (Wilcoxon signed rank test, p: 2.5 × 10^{-7}, Z-value = 5.159, N = 36). No such area-to-area proximity was evident in the shuffled maps (shuffled V1/V2–V4, 0.03 ± 0.03, shuffled V1/V2–PIT, -0.02 ± 0.02, p = 0.16).

Next, we examined the fixation maps. The monkeys showed some correlation in their fixation maps, as compared across scenes (Pearson coefficient of 0.20 ± 0.04, p = 5.8 × 10^{-4}, Wilcoxon paired sign rank test, Z-value = 3.44, N = 36). Compared to the neuronal heatmaps, the fixation maps showed the weakest correlation to V1/V2 heatmaps (0.03 ± 0.02), an intermediate correlation to V4 maps (0.09 ± 0.020), and a higher correlation to PIT maps (0.13 ± 0.03); there was a reliable statistical difference in this trend (p = 6.2 × 10^{-3}, one-way ANOVA, area as factor, F = 5.3, DF = 2, N = 104, **Fig. 6f**). There was no statistical difference between the median correlation between observed V1/V2–fixation maps vs. the median correlation between shuffled V1/V2–fixation maps (Pearson correlation for observed was 0.03 ± 0.02 vs. shuffled -0.01 ± 0.02, p = 0.18, Wilcoxon sign rank test, Z-value = 1.34, N = 36). However, there were statistical differences between observed V4–fixation maps vs. shuffled V4-fixation maps (0.09 ± 0.02 vs. -0.03 ± 0.02, p = 6.6 × 10^{-4}, Wilcoxon sign rank test, Z-value = 3.41, N = 36) and between observed PIT-fixation maps vs. shuffled PIT-fixation maps (0.13 ± 0.03 vs. -0.00 ± 0.02, p = 3.8 × 10^{-4}, Wilcoxon sign rank test, Z-value = 3.55, N = 36). The fixation maps also showed a reliable similarity to the algorithmic saliency maps, with a stronger correlation to GBVS (0.17

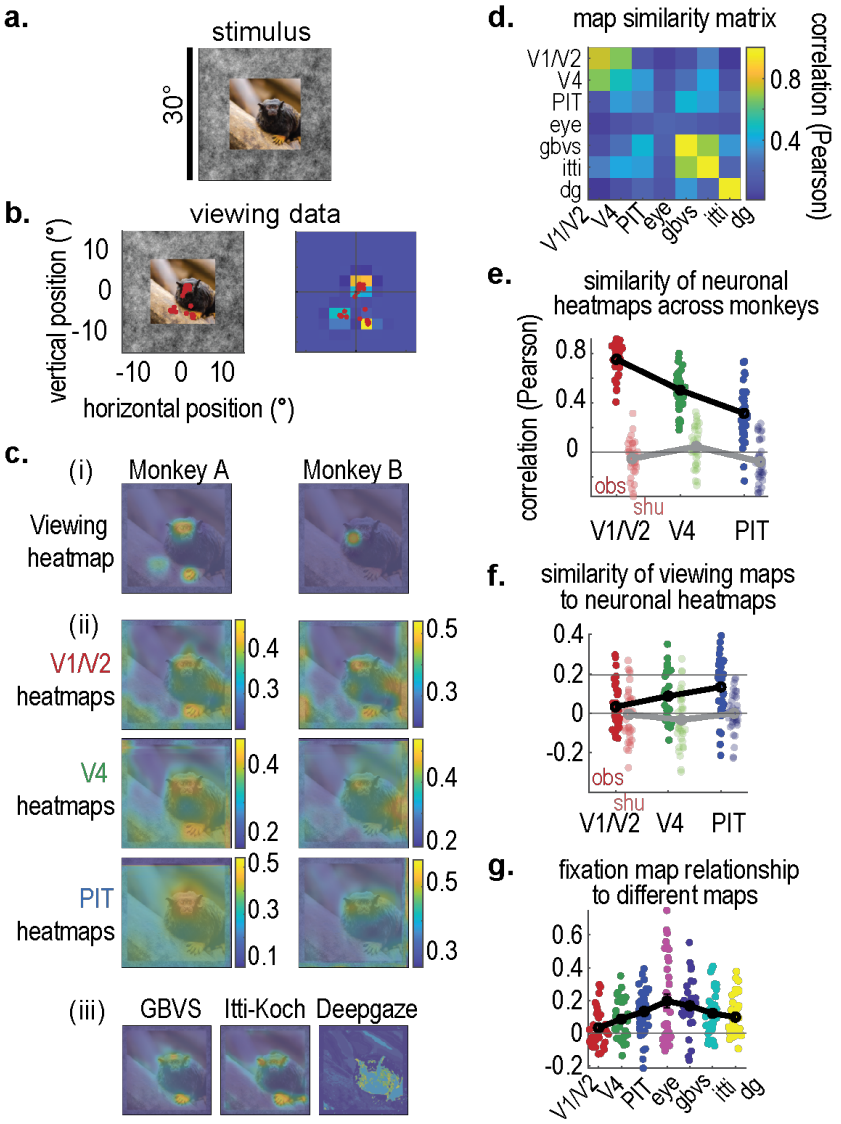

Fig. 6. Free viewing and Automated Saliency Maps. a. Example scene. **b.** Example of eye positions tracked over the scene to create fixation maps. **c.** Examples of viewing maps on the same scene (*i*), along with neuronal heatmaps from V1/V2, V4, and IT (*ii*), and algorithmic saliency maps (graph-based visual saliency, Itti-Koch saliency, and fast, accurate, and DeepGaze-FASA, *iii,* monkeys A and B used). **d.** Pearson correlation values relating all neuronal heatmaps (V1/V2, V4, IT), all fixation maps (*eye*), and all saliency maps (*gbvs, itti, dg*). The diagonal cells show the monkey-monkey correlations in neuronal heatmaps and the monkey-monkey correlations in fixation maps. **e.** Similarity of neuronal heatmaps across monkeys for V1/V2 maps, for V4 maps, and for IT maps (observed values; light red, light green, blue; shuffled neuronal heatmaps, dark red, green, and blue); each point shows one scene. **f.** Similarity of neuronal heatmaps to fixation maps, per area. **g.** Similarity between fixation maps to neuronal heatmaps and to saliency algorithms. Monkeys A and B were used for these analyses.

± 0.03, $p = 6.7 \times 10^{-6}$, Wilcoxon sign rank test, Z-value = 4.50, N = 36), then to the Itti-Koch maps (0.12 ± 0.02, $p = 5.2 \times 10^{-5}$, Wilcoxon sign rank test, Z-value = 4.05, N = 36), and lastly to the DeepGaze-II maps (0.10 ± 0.02, $p = 1.4 \times 10^{-4}$, Wilcoxon sign rank test, Z-value = 3.82, N = 36, **Fig. 6g**). PIT maps were generally better correlated to all saliency maps than were V4 and V1/V2 maps (PIT-saliency maps similarity: 0.34 ± 0.03, V4–saliency maps: 0.24 ± 0.03, V1/V2 -saliency maps: 0.13 ± 0.02, all values statistically different per one-way ANOVA with area as single factor, $p = 3.8 \times 10^{-8}$, F = 18.1, DF = 2, N = 314). This suggests that visual saliency algorithms overall simulate operations of more anterior visual cortex. In conclusion, we found a relationship between the monkeys' viewing behavior and their neuronal activity heatmaps. The monkeys' viewing patterns were more closely correlated with anterior visual cortex area heatmaps, and the same anterior visual cortex area heatmaps were also more correlated with all artificial saliency maps compared to posterior visual areas.

Discussion

Primates perceive and interact with objects, but how is this ability reflected in neuronal populations subserving vision? While the notion that the function of individual neurons maps to object classes may seem intuitive, the sheer diversity of objects suggests that a more efficient representational scheme is necessary, one that is distributed across shared local features[1]. Recent studies suggest that most visual cortex neurons represent lower-level features: when exposed to stimuli encompassing very large image spaces, such as those represented by deep generative networks, neurons show wider response ranges[16] and expansive tuning functions[17], all seemingly driven by specialized local features[18]. This aligns with early investigations by Tanaka and others (1996)[19], which showed that inferotemporal (IT) neurons encode moderately complex features rather than entire objects. It also potentially aligns with the theory that neurons represent features that are more pragmatic for guiding behavior[20] (for example, a monkey should notice and then orient its gaze to anything that might look like an eye, even if it is not, as it might signal a friend, foe, predator, or food). These features were simpler than those found in natural scenes, supporting the idea that objects and places are represented in a distributed manner across neuronal populations. This distributed coding is essential because under natural conditions, objects such as faces are typically larger than the receptive fields of individual neurons, even in anterior regions. There may be higher-level representations by neurons in the hippocampus, thalamus, or prefrontal cortex, but this remains to be shown using contemporary technical approaches that systematically deconstruct visual inputs[16,18].

However, the nature of this distributed function remains unclear. Given that neurons show evoked activity in response to nearly any kind of image (while reserving their stronger activity to particular image features), one might suppose that our perception of a full scene depends on neuronal populations being systematically excited by all features in the image. Alternatively, it is plausible that perception is constructed from sparser cues, akin to how visual information in the cortical representation of the blind spot is interpolated by neighboring visual features[21]. Under this view, the role of the visual system is not to explicitly represent objects or other high-level concepts, but to simply provide a *good-enough*, "pragmatic" ensemble of features that allow interactions with the world, from navigation to grasping and recognition[22].

In this study, we used a technique inspired by weight sharing in CNNs, where a given RF can be convolved across a full visual scene. We found that local neuronal populations (consisting of single units, multiunits, and hash) did respond to the entire visual scene, but often burst more frequently to more specific regions, depending on the visual area. For example, while V1 was driven widely by textured backgrounds, PIT neurons responded to features that signal animals or objects belonging to *animate*[4] sub-categories. While it is a well-known fact that monkey IT neurons respond strongly to face and hands[1], many studies using these *animate* stimuli have been interpreted as evidence that these neurons respond to these objects, without localizing the features that defined

these objects. Here, we found a hierarchy-wide organizing principle that informs us about the content of earlier cortical stages, such as V4 and V1/V2. Our results show a gradually increasing tendency for neural responses to emphasize the representation of features *contained* within monkey and human faces, hands, and bodies, but not covering the entire area defined by the semantic definition (spatial mask) of these objects.

This focus on animal features is consistent with the fact that macaque monkeys are foraging and social animals that spend a significant fraction of their time identifying conspecifics, classifying their actions, and avoiding threats from other animals[23,24] (including humans in laboratory settings). In the visual world, conspecifics and predators can be occluded by trees and rocks, so monkeys often need to be able to identify other animals relying on the patches of textures perceived, instead of the whole silhouette. High curvature is a likely low-level feature prevalent in these patches, present as the outline of eyes, faces, and fingers[25,26], but not as common in parches of artificial and other natural objects and textures. Face features are highly decodable and separable[25], thus the focus on animal features may be directly related to curvature tuning, particularly given the lateral position of our recordings.

Detecting animal features is a problem that is also faced by deep artificial networks of vision trained for classification. ANN capability is largely dependent on their training dataset. While datasets such as ImageNet[12] offer a diversity of scenes to simulate the natural environment, they are still far short in matching the rich streams of complex information that we experience in our daily lives. We found that only a few ImageNet[12]-trained ANNs come to share this similarity with the visual system, in learning animal-like textures, such as fur and color patterns, suggesting this to be a reliable mechanism for detection, perhaps as much as shape.

One major limitation of this study is that while weight sharing is a useful way to estimate how the visual scene is represented, a gold-standard experiment would see the collection of responses from a larger population of neurons, with receptive fields tiling the entirety of the visual scene. The fact that most of our neurons had foveal- and peri-foveal receptive fields illustrates this limit. In principle, a broader study could reveal subpopulations of neurons that respond equally strongly over the full area of a semantic mask, across variations in the semantically defined object's size, position, and rotation. While our opinion is that this is not a necessary feature of a distributed code for object recognition, this remains to be seen. High-density recording probes, such as Neuropixels[27], might allow for the realization of this experiment in the near future.

Acknowledgements. We thank Mary Carter and Elizabeth Cleaveland for their technical help with data acquisition and surgical procedures. Research reported in this publication was supported by the Common Fund of the National Institutes of Health under award 1DP2EY035176-01 (C.R.P.), by the National Science Foundation under CAREER Award No. 2143077 (C.R.P.; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation), by the David and Lucile Packard Foundation (2020-71377), by Core Grant 2P30EY012196-26 (M.S.L.), NIH EY16187, and EY025670 (M.S.L.), and the Ilfield Gift Fund (R.T.B.).

Author contributions. *Conceptualization:* Z.Z., T.S.H., C.R.P.*, Methodology:* Z.Z., C.R.P. *Experimentation, Visualization, Investigation:* T.S.H, Z.Z., C.R.P. *Resources and Investigation:* M.S.L, R.T.B., C.R.P. *Writing - Original draft preparation*: Z.Z., C.R.P. Supervision: C.R.P., M.S.L., R.T.B. *Writing- Reviewing and Editing*, Z.Z., T.S.H., M.S.L., R.T.B, C.R.P.

Methods

Animals and behavior. Six monkeys were implanted with microelectrode arrays (chronic). Two monkeys were implanted with 64 channels in three chronically implanted electrode arrays along the ventral stream (V1/V2, V4, posterior IT). One monkey was implanted with floating arrays

(Microprobes for Life Sciences, Gaithsburg FL) in PIT cortex. Three monkeys were implanted with Utah arrays (Blackrock Microsystems) in the right hemisphere operculum. The task for all monkeys was to maintain their gaze on a red fixation target (< 0.2° radius) while large pictures were flashed at different locations on the screen. If the animals maintained their gaze within a 2.2° radius of the fixation target (1° for V1/V2 -array monkeys), they received a juice reward. All procedures received approval by Institutional Animal Care and Use Committees at Harvard Medical School and Washington University School of Medicine, procedures conformed to NIH guidelines provided in the Guide for the Care and Use of Laboratory Animals.

Receptive field mapping. Each recording site's receptive field was mapped by flashing a picture of a grayscale cartoon face in a 16° x 16° grid at different locations relative to the fixation point with 2° spacing. The picture was flashed on for 100 ms and left off for 200 ms. The picture evoked firing rate responses from each array channel; responses were defined as the mean firing rate over 0-200 ms after picture onset minus the mean firing rate over the first 30 ms after picture onset. The subsequent 9 x 9 response matrices from every array channel were interpolated over 250 x 250 points. Each interpolated map was used to compute a receptive field, defined by regions where the mean firing rate exceeded the 99th percentile of all responses in the map. The largest region exceeding this threshold was identified using the MATLAB function *regionprops.m*, and its centroid location was used as the site's RF location. We defined the population RF as the mean array response. To characterize the reliability of individual RF estimates, we tested whether the site responded differentially to each position using a one-way ANOVA (picture position as level, $p < 0.05$ after false discovery rate correction). If the site was reliable, we used the site-specific RF estimate. Otherwise, we assigned it the population estimate. Since the RF center of multiunit sites in each array site did not always overlap with the center of the heatmap, we averaged the RF location per site, transformed it into a 2D vector in Cartesian space, and used it as a translation vector to shift the heatmap using the MATLAB function *imtranslate.m*.

Stimuli. The stimuli were $K = 36$ number of 16° x 16° natural-scene photographs embedded in a 30° x 30° brown-noise image. Each photograph x_k included natural (e.g., animals in the wild) and artificial settings (e.g., humans in laboratory settings) as well as computer-generated textures (e.g., bar fields). The scenes were presented in a 9° x 9° grid (spacing of 2°) with the center position located at the center of the population receptive field. The scenes were shown for 100-ms ON and 100 to 200-ms OFF, 3-14 flashes between juice rewards. Different scenes were interleaved within a trial. Every scene was presented at every position 30-37 times (V1/V2 experiments) or 11-13 times (PIT experiments). Each scene was later segmented independently using an automated semantic image segmentation model (DeepLab with ResNet-101 backbone[28], trained on COCO-Stuff 164k dataset[29]) and a human-based segmentation task (Amazon Mechanical Turk, MTurk). Additional 16 scenes added during revision were segmented using Segment Anything[30], and DeepLab[28]. All natural-scene photos were obtained from *www.pexels.com* and were free of copyright restrictions (one picture was provided by the Biomedical Primate Research Centre, Rijswijk, The Netherlands).

Neuronal heatmaps. Heatmaps were evoked by randomly flashing each image of the natural scene at different positions relative to the center of the population RF, while the monkeys performed a fixation task. In their raw form, feature maps were matrices of average spike rates (200-ms-long, from 50 ms before image onset to 50 ms after image offset, 20 time-steps) computed per location (9 x 9), which could then be interpolated to the size of the input scene. The 9 x 9 maps were created by passing the spike rates and stimulus positions to *griddata.m*, then flipping the image left-right and up-down, as image coordinates have their lowest index value at the upper-left pixel, unlike regular numeric matrices that rely on Cartesian coordinates. For sites with RFs that were not centered at the fovea, an offset was added to the heatmap based on the eccentricity of the RF (measured independently of the heatmap experiments). All responses were based on a mix of single- and multiunits (u_r) in regions r (e.g., V1/V2, V4, PIT), though most channels contained multiunits and visual hash.

For population analyses, we computed the heatmap $H_{u_r}(x_k)$ for each unit u_r in region r (e.g., V1/V2, V4, PIT) for a given scene x_k. Each unit's heatmap was normalized to the range [0,1] using the formula:

$$\frac{H_{r_u}(x_k) - H_{r_u}^{\min}(x_k)}{H_{r_u}^{\max}(x_k) - H_{r_u}^{\min}(x_k)}$$

where: $H_{r_u}^{\max}(x_k)$ and $H_{r_u}^{\min}(x_k)$ are the maximum and minimum activation values for unit u_r across the spatial dimensions (h x w) of the scene x_k, respectively.

To compute the population-level heatmap $H_r(x_k)$ for the region r, we averaged the normalized heatmaps across all s units in region. The resulting population heatmap is given by **Eq. 1,** reproduced here for reference.

$$H_r(x_k) = \frac{1}{s}\sum_{u=1}^{s}\frac{H_{r_u}(x_k) - H_{r_u}^{\min}(x_k)}{H_{r_u}^{\max}(x_k) - H_{r_u}^{\min}(x_k)}$$

To define a null hypothesis, for every time-averaged feature map, we scrambled the position information for every picture 999 times while keeping the array of responses the same across sites and across time.

Semantic masks. Each scene was segmented independently using both automated and human-based segmentation approaches. To get segmented scenes from DeepLab[28], a ResNet101-based model, we adapted the unofficial PyTorch implementation and extracted labels from the COCO-Stuff 164k dataset[29] within each scene. The pretrained weight *deeplabv2_resnet101_msc-cocostuff164k-100000* was downloaded from https://github.com/kazuto1011/deeplab-pytorch. Because there is not a *monkey* label in COCO-Stuff, and because DeepLab often classifies one object class (*tree*) with subordinate labels (*e.g., leaves*, *branch*), we merged the extracted labels into superordinate classes according to the COCO-Stuff label hierarchy. Ten *animal* labels and one *person* label were combined into a single *animal* class; six *water* and fluid-related labels into a single *water* class; nine *plant*-related labels such as *tree* and *moss* into a *plant* class; eleven textile labels such as *blanket* and *rug* into a *textiles* class; four raw-material labels such as *metal* and *plastic* into a *material* class; five structural and textual labels including *net* and *fence* into a *structural* class; and six electronics labels such as *keyboard* into an *electronics* class. These classes were used as masks to segment the natural scenes. For revision scenes, we used Segment Anything[30] to automatically generate a few masks that DeepLab did not detect.

In the human-guided approach, we first used Google Cloud Vision[11] to provide ten labels that characterized each photograph, with the goal of asking human annotators to provide segmentation masks that best illustrated each label. We removed labels such as *organism* if they seemed too technical. After integration, there were 3–6 labels per scene. Participants on MTurk[31] were asked to trace polygons around objects for each given label, each scene was randomly selected from the set of 36. Combinations of these maps were used to create higher-order categories such as “face,” “animacy,” and “environment.” There was a total of 27 possible semantic masks per scene. Each mask was defined by a binary matrix of the same size as the original natural scene.

To obtain the masks, we created a webpage for interactive semantic segmentation tasks on Amazon Mechanical Turk. The users were provided with the following information:

Title: Trace polygons around the border of given labels (< 5 minutes)

Description: Apply a polygon over each object described per label in a precise and smooth way. The Task requires usually around 3 - 5 minutes.

Time Allotted: 8 Min

Qualifications Required: HIT approval rate (%) is greater than 70.

Instructions:

- *This is a semantic segmentation task.*
- *To get started, first inspect the image. To the right, there is a list of labels indicating objects present in the image.*
    - *Task is to apply a polygon over each object described per label.*
    - *Click on any label in the box to activate that polygon. Left click to start annotating the image.*
    - *Add points to make a polygon (mask) around the relevant region.*
    - *When finished, press C to close the polygon.*
    - *To undo, click Undo. To redo one label, click Delete to remove a selected polygon. Click Annotate to go back to annotating.*
- *It is okay to overlay the same image region with multiple labels if you find it appropriate.*
- *Please complete all labels in order to proceed to the next task.*

All procedures received ethical approval by the Washington University Medical School Institutional Review Board.

Eq. 2-3 (reproduced below for reference) estimate how semantic masks contained neuronal or hidden unit activity over time.

$$R_{in,r}(x_k, t) = \frac{1}{hw} \sum_{M(x_k,i,j)=1} H_r(x_k, t, i, j)$$

$$R_{out,r}(x_k, t) = \frac{1}{hw} \sum_{M(x_k,i,j)=0} H_r(x_k, t, i, j)$$

These plots led to a *difference* curve $D_r(x_k, t) = R_{in,r}(x_k, t) - R_{out,r}(x_k, t)$, and the area under the curve of $D_r(x_k, t)$ was used to summarize the overall activity. To control for the overall sparsity of the map, we shuffled the response regions for each scene to obtain a shuffled heatmap $H_{r(shuffled)}(x_k, t)$ and also measured the AUC for $D_{r_{shuffled}}(x_k, t)$.

Computing feature-to-object ratio (FTO). Objects were defined by semantic masks discussed in the previous section. The intuition behind the *feature-to-object ratio* (FTO) is to inspect the ratio of the localized feature over the semantic object in the heatmaps. There are two extreme cases: if the activation has the highest response only within the semantic object, then it results in a FTO of 1. On the contrary, if the responses all fell outside of the mask or were all zero, then FTO will be 0. As mentioned above, each unit (u_r) in regions r (V1/V2, V4, PIT) generated a heatmap $H_{r_u}(x_k)$ for a scene x_k. The heatmap was normalized into [0, 1] using the formula

$$\frac{H_{r_u}(x_k) - H_{r_u}^{\min}(x_k)}{H_{r_u}^{\max}(x_k) - H_{r_u}^{\min}(x_k)}$$

The population heatmap $H_r(x_k)$ of a region r was computed by averaging across all units in r. To calculate the heatmap contained by a semantic mask $M_o(x_k)$, we multiplied the binary mark $M_o(x_k)$, with the heatmap $H_r(x_k)$ at each (i, j) position and summed the values inside the mask:

$$\sum_{M_o(x_k,i,j)=1} H_r(x_k, i, j)$$

Finally, we normalized this value by the semantic object size. FTO for each object o in each region r in brain was computed in **Eq. 4** (reproduced below)

$$FTO_r(o) = \frac{1}{\sum_{k=1}^{K}\sum_{i,j} M_o(x_k, i, j)} \sum_{M_o(x_k,i,j)=1} H_r(x_k, i, j)$$

where the denominator represents the total size of the object mask across all scenes ($k = 1, \ldots, K$), and the numerator is the heatmap value within the mask. For neural networks, heatmap $H_r(i)$ was replaced by a feature map $F_l(i)$ of a specific layer l.

As a control, FTO was computed for each shuffled heatmap. We then fitted a linear regression line to the $FTO(o)$ for each monkey in V1/V2, V4, and PIT and for each neural network in three selected layers discussed in detail below. The linear regression slope k for each $FTO(o)$ across regions (or ANN layers) was estimated. For the statistical test, a one-way ANOVA test was performed between the $k(FTO(o))$ objects, and the p values were reported.

Individual-site heatmap correlation to population heatmap. To measure the similarity of individual neuronal sites' heatmaps to the population-level heatmap, we iterated over every neuronal site. A mean region-level (e.g., V1/V2, V4, PIT) heatmap was generated for each recording array by averaging the heatmaps of all channels within the array, excluding the channel under evaluation. The site's heatmap was then correlated with the mean region-level map using Pearson's correlation coefficient. This was repeated for all images within an experiment, all sites, and all recording arrays. Before correlation, each heatmap was resized to a standard resolution of 400 × 400 pixels. Additionally, channel quality was evaluated using a split-half correlation method, where scene presentations from one held-out experiment were divided into two halves, and correlations were computed between heatmaps derived from each half. The relationship between channel quality and correlation to the mean map was quantified using Pearson's correlation. Statistical comparisons were performed using one-way ANOVA to test for differences in correlation values.

Fraction of animal-feature-preferring sites per area. Two types of semantic masks (*animal*, *plant*) were used to determine if a given site responded more reliably to features inside animal masks. Similar to above analysis, for each site, its heatmap was normalized to a range of [0, 1] to standardize responses across images and experiments. Each mask was applied to the heatmap, then the product was normalized by number of mask pixels that equals to 1. This metric was correlated to the AUC metric used above but was a simpler alternative for this series of analyses. To assess site selectivity for animate features, the response distributions for animal vs. non-animal masks were compared using Wilcoxon rank-sum tests. A strict significance threshold ($p < 0.01$) was applied to designate a site as "selective." False-discovery rate (FDR) correction was performed to account for multiple comparisons.

Individual site preferences. To determine whether a given neuronal site responded more to features within animal- vs. non-animal segmentation masks, we developed a semantic selectivity index (SSI). Each individual-site heatmap was normalized to a range of [0, 1], and SSIs were calculated by comparing the average response of neuronal sites within a given semantic mask to the average response outside the mask. The difference between these averages constituted the SSI for each mask. For each neuronal site, SSIs were organized into a matrix, with rows corresponding to images and columns to semantic masks. Because not every image had features for every given mask, we grouped all SSIs into two broad categories: animal-related and non-animal-related. This ensured that all images could contribute to the individual neuronal site's analysis. The animal-related group included masks corresponding to biologically relevant categories such as *animal*, *bird*, *monkey*, *quadruped*, *human*, and *face*. The non-animal-related group, by contrast, included masks for semantic features unrelated to animate entities, such as *plant*, *vehicle*, *furniture*, and other non-animate objects. We then compared these SSI groups within each visual area.

Convolutional neural networks. We selected 14 different neural network models partially based on their ranking on Brain-Score[32], going from 1–108. This set comprised *AlexNet*[33], VGG-16[34], CORnet-S and -Z[35], ResNet-50[36], ResNet-50-robust[37], EfficientNet-B0[38], VOneNet (ResNet-50 backbone, ResNet-50-robust backbone, ResNet 50 non-stochastic backbone, AlexNet backbone and CORnet-S backbone)[39], DenseNet-201[40], Xception[41], and SqueezeNet[42]. In each network, we evenly selected three building blocks, identified the last convolutional layer per building block, and applied corresponding batch normalization and rectification to such convolutional layer. CORnet comprises four building blocks, each simulating the visual information processing of V1, V2, V4, or IT. We thus used the *V1nonlin*, *V4nonlin*, and *ITnonlin* layers in CORnet-Z and *V1nonlin3*, *V4nonlin3*, and *ITnonlin3* layers in CORnet-S. AlexNet has five convolutional layers *conv1* to *conv5*, and we used all hidden units and applied rectification to *conv1*, *conv3*, and *conv5* layers. VGG-16 has 16 weight layers and uses sequential 2-3 convolution-rectification layer motifs, which allows for a deeper trainable network. We used all units in the last convolutional layer of each motif: *conv1_2*, *conv3_3*, and *conv5_3.* In ResNet-50, there are five building blocks compressing repeating layer motifs, therefore we used layers *activation_1_relu*, *res3d_branch2c*, and *res5c_branch2c*, each coming from one building block, and with rectification applied accordingly. The VOneNet models comprise a VOne Block, simulating primate V1, followed by classic neural network backbones. Therefore, we used the *voneblock* with the convolutional layers in the last two blocks of the backbone network. For example, VOneNet ResNet-50 series all have a VOne Block followed by two ResNet building blocks, so we selected layers the same way as ResNet-50, except for replacing the first one by *voneblock*. Likewise, we chose the layers in VOneNet (ResNet-50 robust backbone), VOneNet (CORnet-S backbone), VOneNet (ResNet 50 non-stochastic backbone) following a similar procedure. DenseNet-201 is a 201-layer-deep, densely connected convolutional neural network and its convolution operations are distributed in one convolutional block followed by two dense blocks. We thus chose conv1|*relu*, *conv3_block12_2_conv*, *conv5_block32_2_conv*. EfficientNet B0 is a convolutional neural network baseline that is designed by neural architecture search. It has one convolution and seven mobile inverted bottleneck *MBConv* blocks followed by the final classification blocks. We selected the last convolutional layers from blocks *efficientnet-b0*|*blocks_0*, *efficientnet-b0*|*blocks_7*, and '*efficientnet-b0*|*blocks_15* and applied batch normalization and sigmoid rectification to each one of them. *Xception* uses deep wise separable convolutions in the network, and it has 36 convolutions spread out in 14 modules. Here we selected layer *block1_conv2_act*, *block7_sepconv3_act*, and *block14_sepconv3_act*. SqueezeNet comprises a standalone convolution block, eight Fire modules, and ended with a last convolutional block. We evenly chose three blocks in the architecture and used the last convolutional layer (after rectification) in each block: *fire2_relu_squeeze1x1, fire6_relu_squeeze1x1, and relu_conv10.* Every ANN hidden unit feature map was subsampled in an evenly spaced 9 x 9 grid and then normalized to the range of [0, 1]. Since ANNs do not have temporal responses in the activation, we simulated the temporal dynamics in ANNs (for comparison purpose) using normalized peristimulus time histogram (PSTH) from *in-vivo* experiments described above. Specifically, we multiplied the activation value from each hidden unit in the ANN with a randomly sampled PSTH. We randomly sampled V1 PSTH for hidden units in early layer, V4 PSTH for units in mid layer, and PIT PSTH for those in the late layer.

Comparison between models and brain. A number of statistical tests were conducted to compare the area under the curve (AUC) of D and the linear regression slope of k for each semantic object between regions (array regions r in brain or layers l in neural networks). Two sample Student's t-test was performed between D in early and middle regions D_{early}, D_{middle} and in the middle and late regions D_{middle} and D_{late} and p values were reported. Spearman Correlation and Multivariate ANOVA (MANOVA) was conducted between D_{early}, D_{middle} and D_{late} between monkeys and neural networks with reported p values. Spearman correlation tests whether the strength and direction of monotonic association between D_{early}, D_{middle} and D_{late} between the monkey brain and neural networks are significant. MANOVA aims to test whether the mean vectors D_{block} for each block in $\{early, middle, late\}$ in brain or model are all equal to one another.

Eye movement and saliency maps. In the gaze position experiments, the monkeys sat in front of the monitor. To start each trial, they acquired fixation on a center spot, after which a natural scene was presented. These scenes were the same stimuli as used in the neurophysiology experiments (30° x 30° total size, with the central 18° x 18° containing the photograph and the surround containing brown noise). The scenes were fixed to the center of the monitor. The animals were permitted to move their gaze anywhere within the scene within 1000 ms. To compute the eye fixation maps, the eye position in each trial was propagated into the ClusterFix toolbox[43] to identify the first saccade after image onset. The eye position values (comprising a 1 x 2 tuple per ms) were binned into a 15 x 15 grid (15 bins were chosen for two reasons: the central canvas ratio was 18 ÷ 30 = 0.6, so it was computationally convenient to have a grid size resulting in an integer when multiplied by 0.6, and also because the raw neuronal heatmaps measured 9 x 9 bins, and 15 × 0.6 = 9 x 9 fixation maps. These maps could then be interpolated to the original scene size, as the heatmaps were as well.

Saliency algorithms highlighted the regions in visual scenes that were most likely to draw the attention of viewers, as measured perceptually or via saccades. We used three algorithms. The first was the Itti, Koch and Niebur algorithm[14], which worked by decomposing the visual image into feature channels using filter convolution (for orientation, color and intensity), implementing an antagonistic center-surround operation, normalizing each output, and then combining into a final saliency map. The second was the Graph-Based Visual Saliency model[9], which worked by extracting feature maps and using Markov chains over these maps, defining saliency as the equilibrium distribution over map regions. Both of these were applied using the code released under the *vision.caltech.edu* domain. The third was DeepGaze-II, which was based on Convolutional Neural Networks and adaptive gradient methods to estimate the gaze direction[15].

Low-level visual feature analysis. To define luminance maps, every scene (in RGB values) was converted into grayscale representations that approximate perceived brightness, by applying weighted contributions of the RGB channels as $L = 0.2989 \cdot (red) + 0.5870 \cdot (green) + 0.1140 \cdot (blue)$. The resulting map was then compared against the neuronal response maps. Contrast maps were also computed to quantify local variability in luminance, emphasizing regions with abrupt changes in brightness. Every scene (RBG) was first converted to grayscale, then local contrast was estimated as the standard deviation of pixel intensities within a (3 x 3)-pixel sliding window using the *stdfilt.m* function in MATLAB.

References

1. Gross, C. G. Representation of visual stimuli in inferior temporal cortex. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **335**, 3–10 (1992).
2. Perrett, D. I., Rolls, E. T. & Caan, W. Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* **47**, 329–342 (1982).
3. Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B. & Tootell, R. B. H. Faces and objects in macaque cerebral cortex. *Nat. Neurosci.* **6**, 989–95 (2003).
4. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–41 (2008).

5. Rust, N. C. & DiCarlo, J. J. Selectivity and tolerance (‘invariance’) both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci. Off. J. Soc. Neurosci.* **30**, 12978–95 (2010).
6. Hung, C.-C., Carlson, E. T. & Connor, C. E. Medial Axis Shape Coding in Macaque Inferotemporal Cortex. *Neuron* **74**, 1099–1113 (2012).
7. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).
8. Arcaro, M. J., Ponce, C. R. & Livingstone, M. The neurons that mistook a hat for a face. *eLife* **9**, (2020).
9. Harel, J., Koch, C. & Perona, P. Graph-Based Visual Saliency. *Proceedings of the 19th International Conference on Neural Information Processing Systems* 545–552 (2007).
10. Hasson, U., Levy, I., Behrmann, M., Hendler, T. & Malach, R. Eccentricity Bias as an Organizing Principle for Human High-Order Object Areas. *Neuron* **34**, 479–490 (2002).
11. Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R. & Livingstone, M. S. Seeing faces is necessary for face-domain formation. *Nat. Neurosci.* **20**, (2017).
12. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi:10.1109/CVPR.2009.5206848.
13. Linsley, D. *et al.* Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. Preprint at https://doi.org/10.48550/arXiv.2306.03779 (2023).
14. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).
15. Kümmerer, M., Wallis, T. S. A. & Bethge, M. DeepGaze II: Reading fixations from deep features trained on object recognition. (2016).
16. Ponce, C. R. *et al.* Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell* (2019) doi:10.1016/j.cell.2019.04.005.
17. Wang, B. & Ponce, C. R. Tuning landscapes of the ventral stream. *Cell Rep.* **41**, 111595 (2022).

18. Rose, O., Johnson, J., Wang, B. & Ponce, C. R. Visual prototypes in the ventral stream are attuned to complexity and gaze behavior. *Nat. Commun. 2021 121* **12**, 1–16 (2021).
19. Tanaka, K. Inferotemporal Cortex and Object Vision. *Annu. Rev. Neurosci.* **19**, 109–139 (1996).
20. Neural Mechanisms for Interacting with a World Full of Action Choices | Annual Reviews. https://www.annualreviews.org/content/journals/10.1146/annurev.neuro.051508.135409.
21. Komatsu, H., Kinoshita, M. & Murakami, I. Neural Responses in the Retinotopic Representation of the Blind Spot in the Macaque V1 to Stimuli for Perceptual Filling-In. *J. Neurosci.* **20**, 9310–9319 (2000).
22. Cisek, P. Resynthesizing behavior through phylogenetic refinement. *Atten. Percept. Psychophys.* **81**, 2265–2287 (2019).
23. Post, W. & Baulu, J. Time budgets of Macaca mulatta. *Primates* **19**, 125–140 (1978).
24. Son, V. D. Time Budgets of Macaca fascicularis in a Mangrove Forest , Vietnam. *Lab. Primate Newsl.* **43**, (2004).
25. Ponce, C. R., Lomber, S. G. S. G. & Livingstone, M. S. M. S. Posterior Inferotemporal Cortex Cells Use Multiple Input Pathways for Shape Encoding. *J. Neurosci.* **37**, 5019–5034 (2017).
26. Arcaro, M. J. & Livingstone, M. S. A hierarchical, retinotopic proto-organization of the primate visual system at birth. *eLife* **6**, (2017).
27. Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, eabf4588 (2021).
28. Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018).
29. Caesar, H., Uijlings, J. & Ferrari, V. COCO-Stuff: Thing and Stuff Classes in Context. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1209–1218 (IEEE Computer Society, 2018). doi:10.1109/CVPR.2018.00132.
30. Kirillov, A. *et al.* Segment Anything. Preprint at https://doi.org/10.48550/arXiv.2304.02643 (2023).

31. Crowston, K. Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars. in *Shaping the Future of ICT Research. Methods and Approaches* (eds. Bhattacherjee, A. & Fitzgerald, B.) 210–221 (Springer, Berlin, Heidelberg, 2012). doi:10.1007/978-3-642-35142-6_14.
32. Schrimpf, M. *et al.* Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv* 407007 (2018) doi:10.1101/407007.
33. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1097–1105 (2012).
34. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. (2014).
35. Kubilius, J. *et al.* CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv* 408385 (2018) doi:10.1101/408385.
36. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1512.03385 (2015).
37. Engstrom, L., Ilyas, A., Salman, H., Santurkar, S. & Tsipras, D. Robustness (Python Library). (2019).
38. Tan, M. & Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. in *Proceedings of the 36th International Conference on Machine Learning* 6105–6114 (PMLR, 2019).
39. Dapello, J. *et al.* Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. in *Advances in Neural Information Processing Systems* (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) vol. 33 13073–13087 (Curran Associates, Inc., 2020).
40. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017* **2017-Janua**, 2261–2269 (2016).

41. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. in 1800–1807 (IEEE Computer Society, 2017). doi:10.1109/CVPR.2017.195.
42. Iandola, F. N. *et al.* SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. Preprint at https://doi.org/10.48550/arXiv.1602.07360 (2016).
43. König, S. D. & Buffalo, E. A. A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *J. Neurosci. Methods* **227**, 121–131 (2014).