Macaque Ventral Visual Responses to Diverse Stimuli and during Natural Vision

A dissertation presented

by

Wu Xiao

to

The Department of Molecular and Cellular Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biology

Harvard University

Cambridge, Massachusetts

December 15, 2022

© 2023 Wu Xiao

All rights reserved.

Dissertation Advisors: Gabriel Kreiman, Margaret S. Livingstone Wu Xiao

Seeing Context:
Macaque Ventral Visual Responses to Diverse Stimuli and during Natural Vision

Abstract

Studies of vision in the primate brain must contend with a gap between reductionist experiments and the rich context of natural behaviors. I explored the selectivity of macaque ventral visual neurons within a broad stimulus space, leveraging contemporary developments in deep generative image models. Further, I analyzed detailed neuron response properties during free-viewing behavior, developing novel statistical analysis methods and computational models. I found that high-level visual neurons in the inferior temporal cortex, previously thought to code for semantic categories, responded just as strongly to ineffable visual features. During task-free active viewing, neurons throughout the ventral visual pathway responded in ways that comported with the classical view of these neurons as reactive feature detectors. The results suggest that the ventral visual cortex specializes in encoding the present visual input in a category-general way, even during active behavior.

Table of Contents

List of Figures

Acknowledgements

This thesis would not have been possible without the generous help of many people, to whom I am eternally grateful. I want to thank my collaborators and colleagues, who taught me so much and whose expertise took our projects much further than I alone could ever have: Carlos Ponce and Peter Schade introduced me to the world of electrophysiology. Alex Bardon and Yuan Li taught me how to run crowd-sourced psychophysics experiments. Till Hartmann broadened the scope of XDream to V1 and advised me on measuring delays in eye tracking signals. Pouya Bashivan taught me to build computational encoding models of neuronal responses. Qianli Liao and Honglin Chen helped me learn to train modern deep neural networks. Mengmi Zhang kept me organized in formatting my data for sharing. Michael Arcaro tutored me in AFNI and generously shared his scripts and stimulus sets. Saloni Sharma educated me about the parietal lobe and still graciously overlooks my constant forgetting. Kasper Vinken taught me by example to do statistics carefully and interpret the results cautiously. Many others provided essential help for realizing my projects. I am grateful to Pavel Gorelik for helping design and build a mechanical eye to calibrate eye trackers. John LeBlanc helped customize my rig and engineer the primate chairs. Ariana Sherdil advised me on how to train and care for monkeys. David Castro helped with electronics problems in the lab with infinite patience. John Assad and Eden Sayed advised me on using Neuropixels probes and equipment. I want to thank Pranav Misra, Shane Shang, and Jerry Wang for brooking long-winded, not-always-relevant philosophical discussions with me. I appreciate the patience and support of Susan Dymecki, Grace Gill, and Sheila Thomas, who helped me practice and polish my presentations. I thank all the people who asked incisive questions and gave insightful advice about my work: Marcelo Armendariz, Frederico Azevedo, Xavier Boix, Trenton Bricken, Stephen Casper, Arturo Deza, Giorgia Dellaferrera, Manana Hakobyan, Diana Hidalgo, Leyla Isik, Kohitij Kar, Chenguang Li, David Mazumder, Venki Murthy, Elisa Pavarino, Olivia Rose, Paula Sanchez, Martin Schrimpf, Morgan Talbot, Binxu Wang, Daan Wesselink, Yuchen Xiao, Jie Zheng, and many others whom I cannot name.

I am grateful to those who enriched my thinking by putting problems in a new light and brainstorming ideas with me, even if those ideas did not materialize into results included in this thesis. I want to thank Anh Nguyen and Jeff Clune, who enriched my thinking on the premises and purposes of feature visualization. Jan Drugowitsch guided me in exploring neural population codes to capture with preferred stimuli. SueYeon Chung helped me think about generalization in computational models. Kenneth Harris taught me how to use the tools of linear algebra to treat high-dimensional neuronal responses and adversarial examples. Khanh Dao Duc, Ethan Greenblatt, and Douglas Hofstadter graciously acted as sounding boards who helped me develop future research directions.

I am grateful to the organizers, speakers, teachers, and co-attendees at the CBMM summer course, Herchel Smith symposia, and the MCN workshop at Janelia Farm. Even more than other conferences, these occasions helped me see my research in a fresh light and re-invigorated me to take on new questions and old questions anew.

My gratitude goes to the MCO Ph.D. training program. Fanuel Muindi, Lindsay Guest, and Allie Pagano were ever friendly and helpful and made the program feel like home. I am also grateful to my close-knit cohort of twelve. Anna, Ceejay, Charlie, Hasreet, Heather, Jack, Limdi, Oscar, Rachel, Rockwell, and Roya, you are like family to me. I want to thank Mansi Srivastava for hosting my rotation and teaching me about regenerative biology. Even though I did not continue in this field, it continues to capture my imagination. I would also like to thank Armin Bahl and Marcela Bolaños for their generous mentorship during my first-year rotations.

I am grateful to the Herchel Smith Fellowship and the Stuart H.Q. & Victoria Quan Fellowship, who have generously supported my studies.

No words can fully express my gratitude to my advisors and dissertation advisory committee for their selfless attention, guidance, and help. Florian Engert graciously hosted me for an unforgettable eight weeks of rotation, overlooked my joining Gabriel and Marge's labs, and agreed to chair my committee with glee even when the rules left no one else eligible. Richard Born excused my ignorance about the voluminous literature on vision during eye movements, patiently guided me through it, and alerted me to the tracking latency artifact that led me astray for so long. L. Mahadevan kept my eyes open to interesting questions in my data beyond my fixations and magnanimously helped me think through my blue-sky research interests. Marge and Gabriel are my constant inspirations beyond the two principal components defining my research.

Finally, I wish to thank my friends and family for their unwavering support and endless patience throughout this journey.

To those who taught me how to learn, my mother foremost

To Marge and Gabriel, who let me explore

To Flora, who keeps me in touch with life besides science

It is possible to reduce the laws of nature to simple principles[.] [T]heir simplicity, and the technical difficulties, form a criterion of the beauty of our theories.

God does not care about our mathematical difficulties; He integrates empirically.

Albert Einstein, as quoted by Leopold Infeld

1 Introduction

We conceive of the infinitely complex world with our finite heads by simplifying phenomena into concepts, categories, and principles. In the same way, neuroscience aims to explain the brain by categorizing its activity into sensory, motor, executive, and memory functions, parceling its anatomy into regions, and classifying neurons into cellular and response types. The understanding of visual processing in the primate neocortex exemplifies this reductionist program. The primate visual cortex comprises two principal processing streams, the ventral 'what' and the dorsal 'where' stream. The two streams, respectively, analyze the form and location of visual objects (Ungerleider & Mishkin, 1982). In the ventral stream, research has categorized neurons into response types that fit into an overall picture of a sequential transformation of the visual input from low-level luminance variations to high-level representations—objects, categories, and identities. A hierarchy of cortical areas embodies this orderly sequence of processing. The primary visual cortex (V1), which serves both the dorsal and ventral pathways, contains form-analyzing neurons tuned to oriented edges (Hubel & Wiesel, 1962), spatial frequency (Campbell et al., 1969; De Valois et al., 1982), and color (Conway et al., 2002; Hubel & Wiesel, 1968). V4, an intermediate area in the ventral stream, contains neurons sensitive to textures and the local curvature of contours (Pasupathy & Connor, 1999). The inferior temporal cortex (IT), the highest stage of ventral visual processing, contains neurons selective for object categories such as faces, animals, and inanimate things (Desimone et al., 1984; Hung et al., 2005).

Just as we perceive the world through our lens of understanding, earlier literature influences subsequent studies by primarily imparting matters of fact and secondarily impacting ways of thinking. Early lesion studies suggested the necessity of IT not for seeing in general but for 'detecting the meaning of objects on visual criteria' (Gross, 1994; Klüver, 1951; Mishkin &

Pribram, 1954). This hard-won knowledge about the function of IT anchors the typical questions asked about the ventral stream. Studies have inquired into what categories are distinguished by IT: for example, faces, body parts, animacy, and tools (Desimone et al., 1984; Gross et al., 1979); what kind of encodings would allow for easy categorization: invariant or equivariant representations; how these representations can be derived: by hierarchical, multi-stage transformations of visual features (Riesenhuber & Poggio, 1999); and how feature selectivity organizes itself on the cortex: into modules in systematic maps that recapitulate throughout the processing hierarchy (Bao et al., 2020; Freiwald & Tsao, 2010). The emerging picture is that the ventral stream serves a transparent visual representation to higher cognitive processes for the latter to sample as required. This representation elucidates what information is latent in the pattern of light intensities on the retina to emphasize ethologically relevant data such as object category, shape, size, and identity. Incidentally, this ethological emphasis provides a boon to investigators hoping to explain IT functions using natural language. This picture paints the ventral visual stream as a specialized processing module—a reverse graphics card. Outstanding questions remain only about how the system develops, how it benefits (or not) from learning, and what exact computations interpose between pixels and invariant representations—to achieve the latter is a still open engineering challenge despite the impressive advances in computer vision made by deep learning.

Might the reverse graphics card hypothesis obscure essential aspects of ventral vision? Even in the primary visual cortex, arguably the best-understood cortical visual region, a tally reckoned that quantitative models could only capture a minority of the responses of a minority of neurons (Olshausen & Field, 2005). The models Olshausen and Field considered did not incorporate many then-known neuronal mechanisms, which, because they were difficult to measure jointly, could not be integrated fully into a computational model (Rust & Movshon, 2005).

Today, better quantitative models of ventral neurons incorporate deep artificial neural networks (Schrimpf et al., 2020; Yamins et al., 2014). Nevertheless, modern models only marginally improve on the fraction explained of neuronal responses in V1 (Cadena et al., 2019). Higher visual areas see more improvement for lack of good previous models. Still, current models only capture a slight majority of the stimulus-driven response variance. Artificial networks are problematic models of ventral vision in other respects. Despite their stellar task performance and ability to predict neuronal responses, artificial networks use starkly different visual features from biological neurons. Artificial networks are biased toward representing textures, as opposed to the shape bias of human perception (Geirhos et al., 2018). Moreover, this texture bias may correspond to an overemphasis on informative but uninterpretable speckles (Ilyas et al., 2019; Szegedy et al., 2013). In other words, models use unnamable features, contrary to what many propose for IT.

A second blind spot highlighted by Olshausen and Field is the lack of knowledge about neural responses to natural stimuli during natural behaviors. Studies of form vision usually use carefully controlled tasks and behaviors. An example is the rapid serial visual presentation (RSVP) task. In it, tens of randomly ordered images are presented per second while the animal maintains fixation on a spot. This task is optimal for studying the feature selectivity of neurons while minimizing context-, history-, and task-dependent effects. The RSVP task, like other reductionist experiments, is justified by the unspoken assumption that response properties can be studied in isolation, then composed without modification to explain more complex natural phenomena. However, the brain is a nonlinear and dynamical system. There are clues that natural stimuli, situations, and behaviors may reveal contextual effects that interact in nonobvious ways with classical properties gleaned from random images and restrictive behaviors. In the mouse, even the primary visual cortex hosts multimodal information such as running speed, facial

movements, and global brain states. These representations multiplex with stimulus selectivity on single trials and are only separable at the level of neuronal populations. Generations of investigators have tested primate vision during more natural viewing conditions (DiCarlo & Maunsell, 2000; Gallant et al., 1998; Leopold & Park, 2020; Livingstone et al., 1996; Sheinberg & Logothetis, 2001). Most studies support, in principle, that ventral visual neurons preserve classical response properties during more natural viewing. In practice, however, studies have traded off between the degree of naturalism and the strength of evidence for classical properties. For instance, there is no direct evidence to show that single neurons maintain fixed receptive fields and selectivity during the daily activities of a monkey. It seems probable that responses must be non-classical to some degree. Perception itself does not always track physical reality. Perception during binocular rivalry can change without any change in the physical stimulus. Perception can also fail conditionally, as in phenomena like intrasaccadic change blindness. Indeed, there is a surfeit of documented, disparate non-classical response properties in the visual system (Gilbert & Li, 2013). The question is one of quantification and synthesis. To what extent is visual neuronal activity affected by non-visual variables, i.e., information beyond the pattern of light arriving at the retina? How do the various non-classical response properties interact and unify to create perception and guide behavior? Natural contexts have a particular explanatory interest because, after all, the brain evolved to control the adaptive behaviors of an organism. The practical corollary is that natural contexts should be more likely to recruit behaviorally relevant neural processes.

In this thesis, I studied primate visual neuronal responses to more diverse stimuli and during more natural viewing behaviors. Published reports have described some of the work (Bardon et al., 2022; Ponce et al., 2019; Xiao & Kreiman, 2020); here, I will emphasize the main ideas and focus on the parts of collaborative work to which I contributed most substantively. For

reasons of space and theme, I will not cover some of my other studies during my Ph.D. (Xiao et al., 2018; Yuan et al., 2020; Zhang et al., 2022).

In Chapter 2, I helped develop an algorithm to define the stimulus preference of neurons independent of notions of categories, objects, or shapes. The method directly queried the target neuron in a closed loop and adapted the stimulus by searching in a broad and diverse generative model. In IT, this method enabled us to identify images that strongly activated neurons without clearly belonging to object categories, showing that IT neurons were selective for nonsemantic visual features. In Chapter 3, I examined ventral stream neuronal activity when monkeys freely viewed natural images. Neurons throughout the ventral stream evinced classical response properties during active vision, including feature selectivity, response latency, and spatially local receptive fields, all yoked to each fixation. Unlike conscious perception, ventral visual neurons changed their responses with eye movements several times per second. The free-viewing behavior provides a testing ground to compare and combine disparate theories of response properties into a unified model.

2 An unbiased method to uncover visual neuron stimulus preference

Introduction

In studying the responses of visual neurons, experimenters must choose some stimulus set with which to probe them. It is not feasible to catalog a neuron's responses to all stimuli, even for static images. At a rapid clip of seeing twenty images per second, an indefatigable monkey fixating eight hours a day for a year—provided that the same neuron can be held for so long, nontrivial though possible with current techniques (McMahon et al., 2014)— can only give neuronal responses to just over 200 million images. That is fewer than the number of different six-by-six punch cards (2^{36} = about 69 trillion) or four filled games of tic-tac-toe.

Thus, experimenters select their stimuli based on the current understanding of neuron function. For example, theories positing gnostic units invite investigators to use pictures of commonplace objects (Gross, 2002; Konorski, 1967). Students of 'face neurons,' a proposed class of neurons specialized for processing faces in supporting social interactions, examine the contrast between face and nonface object responses and scrutinize face features such as gender and inter-eye distance. Because shape helps categorize objects, researchers study contour and curvature variations in V4, an intermediate processing area. Likewise, studies also probe V4 and V2 with textures defined by the high-order statistics of light patterns.

Since theories are built on experimental results, using the same theories to inform stimulus choice leaves room for missed insights to emerge occasionally through serendipity. For example, anecdote has it that Hubel and Wiesel first discovered V1 orientation tuning via the unexpectedly rigorous responses to the edges of changing projector slides after a long search of LGN-like center-surround tuning yielded few spikes. Similar insights might await discovery in

higher visual areas containing yet more sparsely coding and sharply tuned neurons. Is there a way to test stimulus selectivity efficiently while relying as little as possible on the experimenter's preconceptions?

One solution is adaptive stimulus search: an algorithm that iteratively samples 'interesting' stimuli to adapt to neuronal activity recorded in a closed loop (Carlson et al., 2011; Yamane et al., 2008). For example, Carlson et al. (2011) studied feature tuning in V4 by optimizing the location and curvature of a contour feature on an object silhouette. The optimized stimuli reliably converged to the same features starting from different random initializations and activated the target neurons more strongly than the initial stimuli, showing that the search had discovered something tailored to the neurons. An adaptive search method depends on a stimulus parameterization to search through. The silhouette parameterization would not have been able to propose, for example, a face for a hypothetical face-selective neuron. Likewise, searching through the parameters of a face generator will never lead to an object silhouette with a sharp corner on the left if that is the preferred stimulus for a neuron. In other words, while stimulus search avoids committing to a fixed image list, using a domain-specific image parameterization nevertheless leads to the same problem of constraining stimuli to a theory. Image parameterization is the same problem as building generative models of images. The challenge lies in realizing a generative image model that can produce diverse stimuli not restricted to a particular theory.

A breakthrough in models that can produce diverse yet naturalistic images followed the advent of deep artificial neural networks and generative adversarial training (Denton et al., 2015; Goodfellow et al., 2020). Deep image generative models can synthesize complex, naturalistic, and diverse images from vectors of numbers. In many models, the same continuous vector space can represent pictures of many kinds of objects, the background context in which they occur, as

well as textures that do not quite depict entire objects. This ability to capture diverse visual patterns with natural statistics made deep generative models prime candidates to use in adaptive search methods for defining visual neuron stimulus preference.

In this chapter, I describe my collaborative work to design an algorithm for automatically uncovering preferred stimuli for visual neurons. The algorithm combined deep generative neural networks and algorithms for high-dimensional non-gradient optimization. For brevity, I refer to the combined algorithm as XDream, to stand for Extending DeepDream (Mordvintsev et al., 2015) for Real-time Activity Maximization. The work in most of this chapter already appears in a series of reports (Bardon et al., 2022; Ponce et al., 2019; Xiao & Kreiman, 2020). I will summarize and synthesize the published results focusing on my contributions. I present results that XDream could find effective stimuli for visually selective neurons in both artificial neural networks and the primate brain. XDream operated under the realistic conditions of limited experimental time and no *a priori* knowledge of the target neuron's tuning. Not constrained by a particular theory, XDream could synthesize preferred stimuli for neurons from V1 to the inferior temporal cortex (IT). The synthetic images for IT neurons showed that IT stimulus preference was broader than object categories. For example, for face neurons, XDream-generated stimuli drove comparable activity to faces but were quantifiably unlike faces.

Meanwhile, several other groups developed a related but distinct method for identifying optimal stimuli. Visual neuron tuning is analogous to the feature selectivity of units in artificial neural networks. Finding the (locally) optimal input for artificial units is a tractable problem. Because the activation of an artificial unit is a differentiable function, it can be efficiently optimized to find a highly activating input by following the function gradient. This procedure cannot find preferred stimuli for biological neurons because determining the derivatives of neuronal output

functions requires complete connectomics knowledge—i.e., knowing all the connections and strengths between the retina and the neuron under study. Currently, whole-brain connectomes are available in *C. elegans* (White et al., 1986), emerging in the fruit fly (Dorkenwald et al., 2022; Scheffer et al., 2020), but remain out of reach for larger brains such as those of mice, monkeys, and humans. Even a complete electron microscopy connectome may still not contain enough information to define the functional strength of synaptic connections. Instead of seeking the veridical weight matrix generating neuronal responses, studies have shown that artificial units can approximate neuron response functions (Schrimpf et al., 2020; Yamins et al., 2014). These artificial network-based models are end-to-end differentiable and can thus *substitute* for biological neurons in determining stimulus preference. The optimized image only needs to be verified with the neurons.

Using substitute models to propose effective stimuli is complementary to using closed-loop adaptive algorithms to search for the same. Comparing and potentially combining both methods could lead to a more thorough picture of the stimulus preference of a neuron. The substitute model and adaptive search methods, which arose contemporaneously, are superficially similar: Both use deep learning tools and have analogs in computer vision research (Mordvintsev et al., 2015; Nguyen et al., 2016). Nevertheless, key differences distinguish the two methods. The substitute model method depends on a good model of neuron tuning. Therefore, this method is only as effective as the substitute model can accurately predict neuron responses, especially to novel synthetic stimuli. Rather than relying on an encoding model of neuron responses, adaptive search directly queries the neuron in a closed loop, instead using a generative image model to balance stimulus diversity with search tractability. The substitute model method can also use a (differentiable) generative model to parametrize the image but has only used low-level image

regularizations in practice. In the last part of this chapter, I compare the adaptive search and substitute model methods concurrently with the same neurons.

Results

Development and validation of XDream in silico

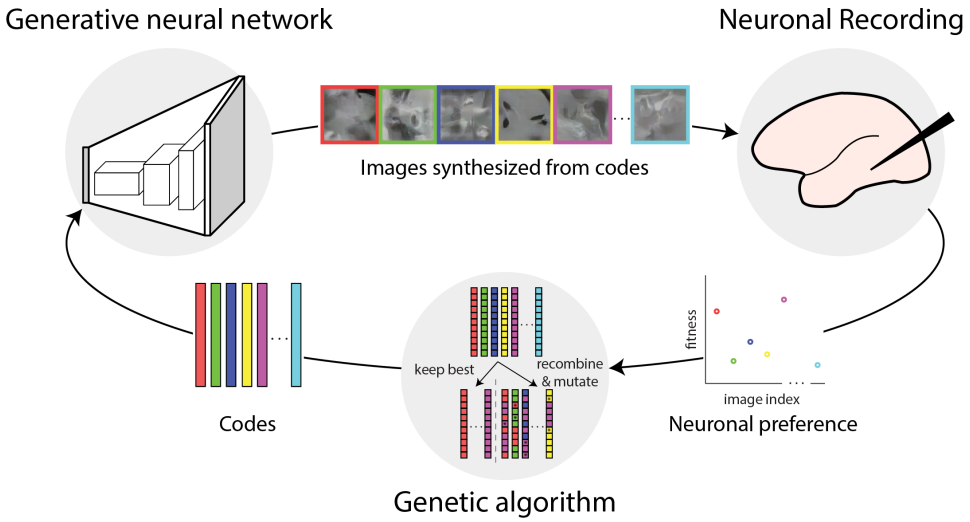

Figure 2.1. Overview of the XDream algorithm for uncovering preferred stimuli.
The XDream algorithm optimized stimuli to drive selective neuronal responses. It comprised 1) a generative neural network, which synthesized images from a compressed representation ('image codes') and 2) a genetic algorithm, which optimized image codes given neuronal responses. Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

We implemented an algorithm to find preferred stimuli for visual neurons from a large and diverse image space not confined by a prior theory of what features neurons should respond to. The algorithm used black-box optimization in a closed loop with neuron recording to search for image parameters to maximize neuronal responses to the corresponding image (**Figure 2.1**). The image space was a generative adversarial network pretrained on ImageNet. We used a custom genetic algorithm as the optimization algorithm. We refer to the combined algorithm as XDream, for Extending DeepDream (Mordvintsev et al., 2015) for Real-time Activity Maximization.

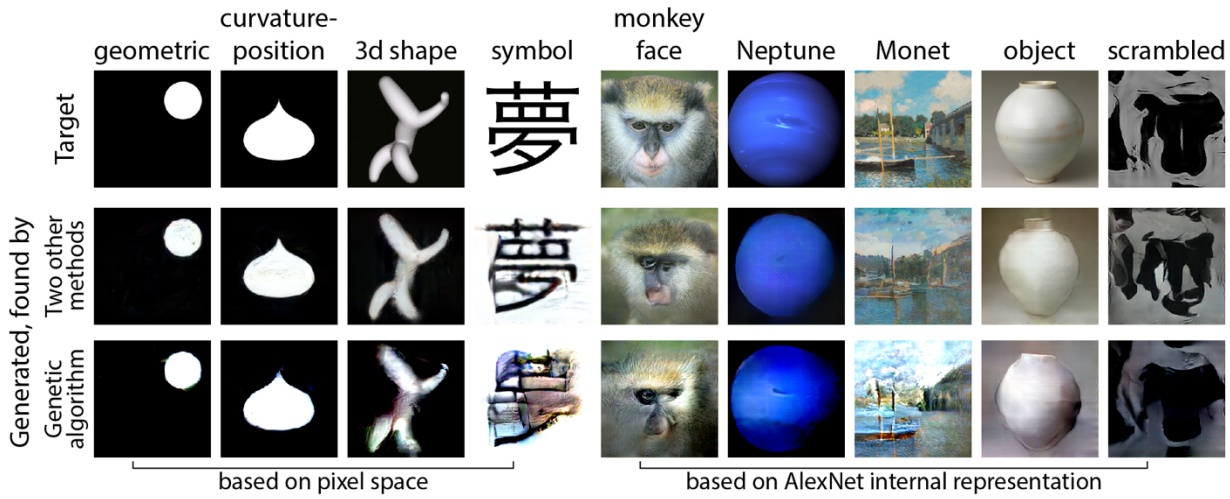

Figure 2.2. Testing the expressive range of the generative model underlying XDream. We evaluated whether the generative network underlying XDream could approximately reconstruct diverse target images and whether XDream, guided by various fitness functions, could find the image codes that approximated the target images. The first row shows the target images. The second row shows network-synthesized images where we used 'strong' algorithms (see text) to search for image codes to reconstruct the target image. The third row shows network-synthesized stimuli where we used XDream to search for image codes. The 'strong' algorithms required explicit knowledge about the generative network, whereas XDream did not use this knowledge. Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

We first validated whether the specific generative model we used, DeePSiM-fc6 (Dosovitskiy & Brox, 2016), could synthesize assorted images from vector inputs termed 'image codes.' We first asked whether there existed image codes that could approximately reconstruct a variety of target images (**Figure 2.2**, first four columns). Because we knew the target images and the full definition of the generative model, we could use two strong methods to find optimal image codes: first, by converting the image to a code using a forward model (Dosovitskiy & Brox, 2016); second, by optimizing the image code using backpropagation to minimize a cost function (Nguyen et al., 2016), here defined as the Euclidean distance from each target image in pixel values. Between them, these two strong methods found image codes that, when converted by the generative model into synthetic images, closely matched target images that depicted simple filled 2D shapes, 3D objects, or line patterns (**Figure 2.2**, first two rows, first four columns).

Having established that the generative model was expressive enough to synthesize various images from suitable codes, we next asked whether we could find these codes by optimizing a loss function under neuronal recording-like conditions. Neuronal responses do not come with a forward model or permit gradient-based optimization. Thus, we used a non-gradient-based genetic algorithm to search for image codes. We used the XDream algorithm to optimize the Euclidean distance from each target image, iterating for 200 generations of 40 images each. Eight thousand image presentations were realistic for an RSVP experiment: With each image shown for 300 ms, the session would take 40 minutes of total viewing. With this number of presentations, XDream evolved synthetic images closely resembling the targets (**Figure 2.2**, third row, first four columns) and discovered solutions that were almost as good as those found by the two strong methods described above.

These toy problems illustrated both the abilities and limitations of XDream. As with prior work on adaptive stimulus search, XDream was contingent on the image parameterization used. In our setting, this image parameterization was a deep generative adversarial network trained on diverse natural images. Although the generative network could mimic all kinds of images ranging from a picture of Neptune to a painting by Monet, the network likely could not generate all arbitrary pixel combinations, even though this statement would be hard to prove formally.

With the limits of the generative model in mind, we simulated a condition one step closer to recording neurons. We re-defined the Euclidean distance loss function in the internal representation space of AlexNet (Krizhevsky et al., 2017) instead of in pixel space. AlexNet, a convolutional neural network trained to classify images, contained internal representations similar to ventral visual representations (Schrimpf et al., 2020). The original target image still minimizes this new distance function. However, unlike pixel space, a distance function in AlexNet

representation space no longer necessarily had a unique image as the global optimum because AlexNet representations were nonlinear. The loss function was also no longer convex (in pixel space), making the optimization harder. Both these properties—nonunique solutions and nonconvex landscape—were likely to characterize the tuning of biological neurons. In this case, the combined XDream algorithm could still approximately recover the target images using an experimentally feasible number of image presentations (**Figure 2.2**, third row, last five columns) and performed nearly as well as backpropagation-based optimization (**Figure 2.2**, second row, last five columns).

Lastly, we validated XDream by testing its ability to synthesize preferred stimuli for single units in AlexNet. To start, we tested units in the first and last layers of AlexNet. First-layer units were linear filters whose global optima were well-defined. Within the valid range of input values, the global optimum was the image whose pixels saturated each element of the $11 \times 11 \times 3$ filter weights across positions and color channels. For the example units (**Figure 2.3a**), XDream-optimized stimuli resembled the expected global optimum filter and emphasized the more heavily weighted filter positions.

Figure 2.3. *In silico* test of XDream on AlexNet units.
We used XDream to identify strongly activating input for units in CaffeNet, an instance of AlexNet and a model of visual neurons. **a**, We tested five units in the first convolutional layer (conv1). The first row visualizes each unit's globally optimal input, an 11 × 11-pixel image patch containing saturated pixels (red, blue, green, cyan, magenta, yellow, white, or black). A gray mask was blended in to indicate the contribution of each pixel to the output. The second row shows one example evolved image for each unit. The figure shows the center 11 × 11 pixels within the unit receptive fields of images evolved at a native resolution of 256 × 256 pixels. **b**, Summary of an example unit evolution in the classification layer. The histograms show response distribution to training images (ImageNet) and evolved images. The white-to-black gradient for evolved images indicates progress during the evolution. The images correspond to the top three highly activating ImageNet examples and the best evolved image; the numbers indicate activation of the target unit. **c**, We calculated the activation ratio between the best evolved and ImageNet images. The violin plots show the distribution of this activation ratio across the 100 units and four layers in AlexNet we tested. Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

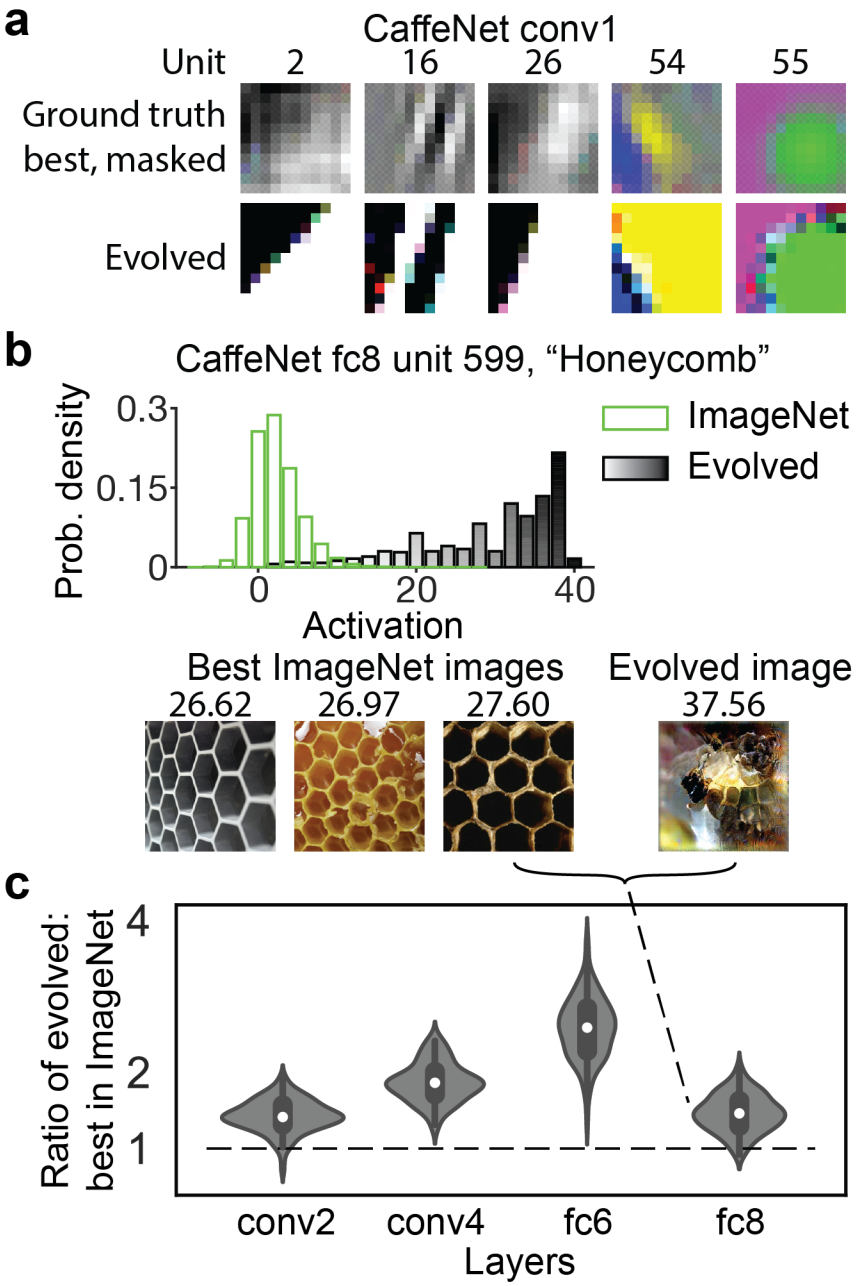

Figure 2.3 (Continued).

Last-layer object-detector units were highly nonlinear functions. Their globally optimal images were ill-defined even though the weights of AlexNet were fully known. For these and any other units in an ANN, one could find strongly activating inputs using backpropagation. However, doing so naively would create unrecognizable noise, a surprising discovery termed adversarial examples (Szegedy et al., 2013). There is active debate about whether adversarial examples reveal the insights or design failures of ANNs (Ilyas et al., 2019). We sidestepped the issue of adversarial examples by comparing XDream-optimized images to the best ImageNet images (Deng et al., 2009), the training set of AlexNet. For an example unit that represented the category, 'honeycomb,' XDream led to a synthetic image that activated the unit about 1.37 times as strongly as did the best of 1.4 million images in ImageNet (**Figure 2.3b**). Intriguingly, the synthetic image contained features reminiscent of both honeycombs and bees but not recognizable as either, whereas the most highly activating natural images only depicted honeycombs. XDream could synthesize super stimuli—images that drove a unit more strongly than its native input—for most units throughout AlexNet processing stages (**Figure 2.3c**) and across a range of classifier networks (**Figure 2.4**). As exemplified by the optimized images (**Figure 2.4b**), units trained to categorize nevertheless responded strongly to visual patterns that fall short of whole objects, even though those patterns may evoke the trained category to a human observer.

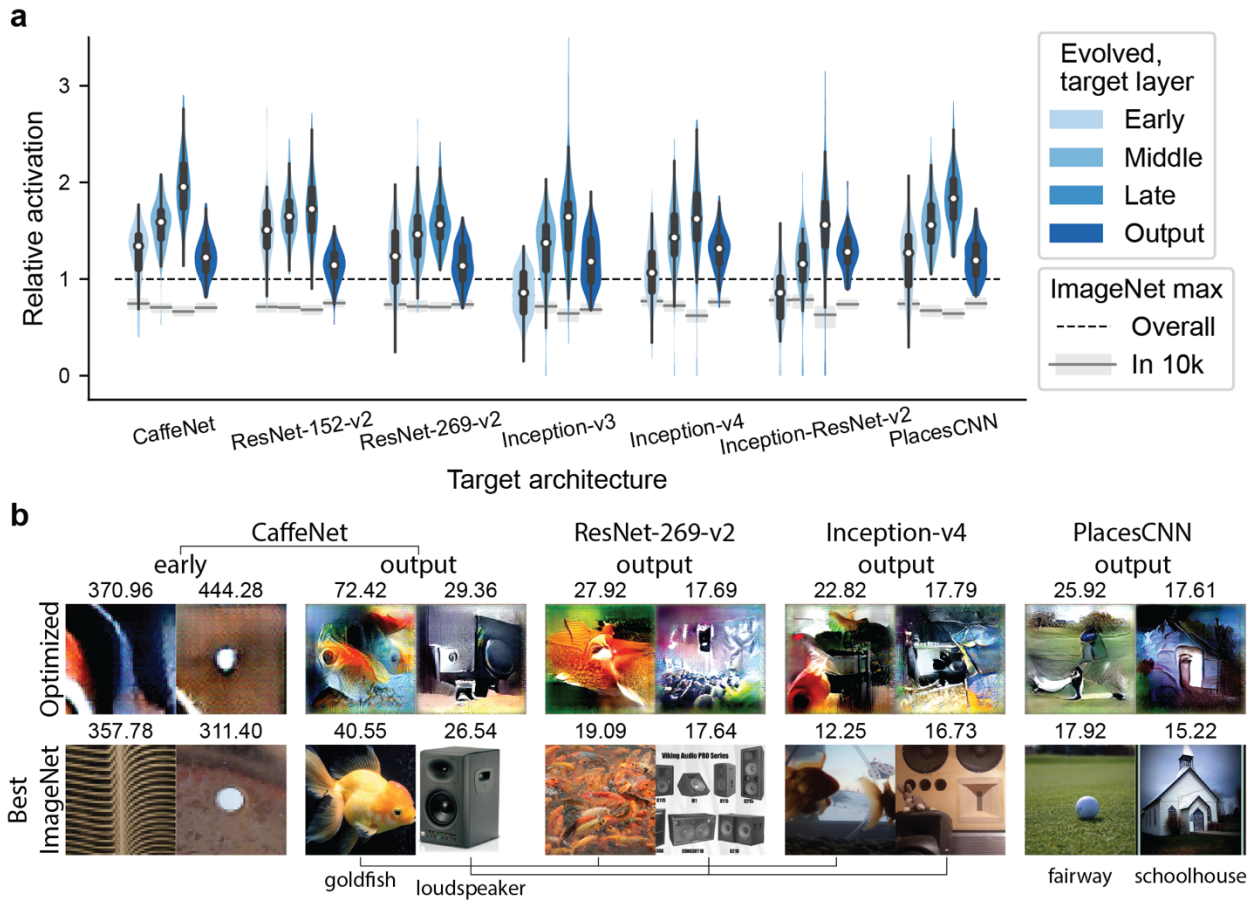

Figure 2.4. *In silico* test of XDream across diverse CNN architectures and layers. **a**, The violin plots show the distribution of relative activation for 100 example units per CNN architecture and layer. The dashed line indicates the best ImageNet image, which had relative activation of 1 by definition. For a fairer comparison, the gray lines and rectangles indicate the distribution (quartiles) of the highest relative activation from 10,000 random ImageNet images, because XDream used 10,000 image presentations to create the evolved stimuli. **b**, The best evolved and ImageNet images, with their activations, for units from different CNN architectures. Adapted from Xiao and Kreiman (2020). CC BY.

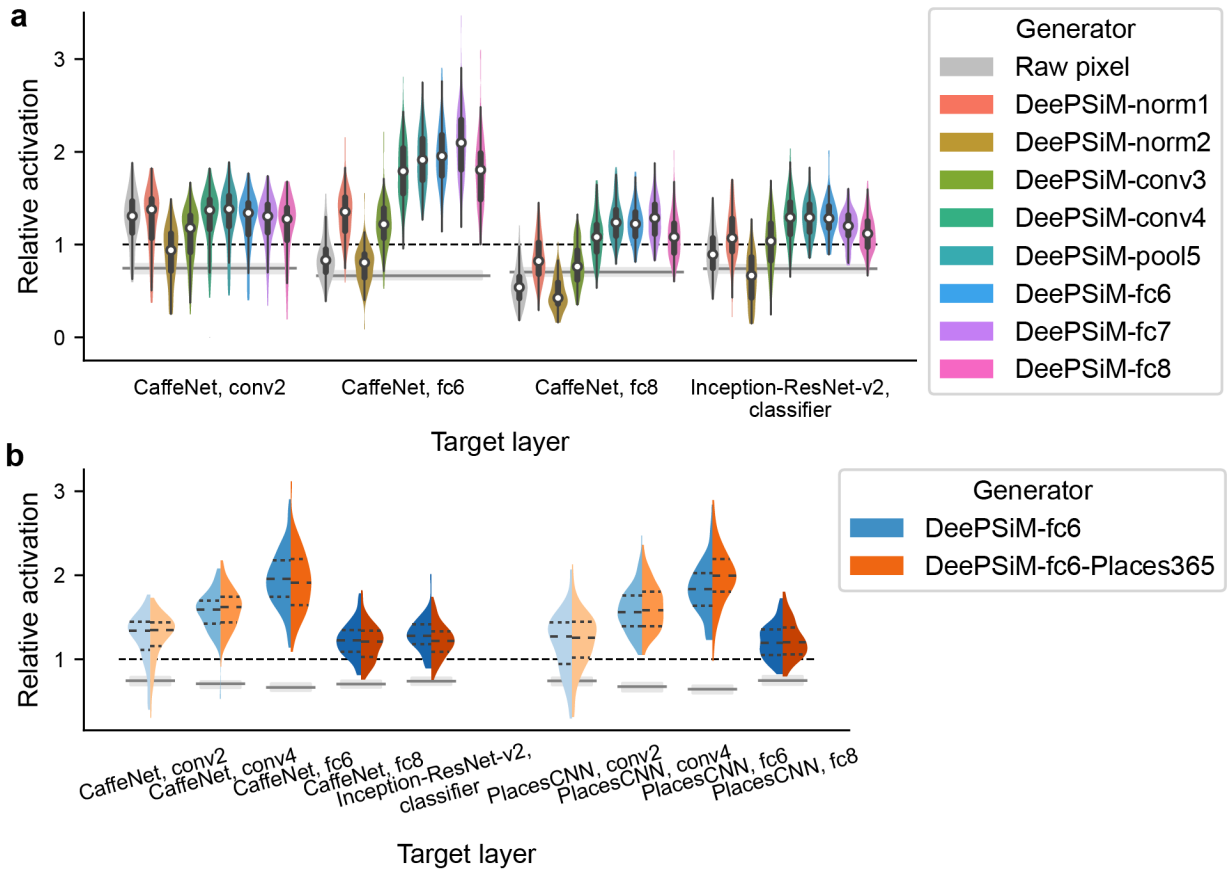

Figure 2.5. *In silico* test of XDream combined with various image generative models. XDream functioned similarly well with various generative models represented by GANs. We compared XDream performance using a range of GANs that **a**, used low- to high-level representations, or **b**, learned on different image sets. Adapted from Xiao and Kreiman (2020). CC BY.

We used *in silico* experiments as a platform for tuning XDream because *in silico* tests were higher throughput than testing with electrophysiological recordings. Since the generative model essentially defined an adaptive stimulus search method, we first tested variants of XDream that used different generative network architectures. These networks learned to use image codes biased toward lower- or higher-level features (Dosovitskiy & Brox, 2016). Many generative networks that used sufficiently high-level features served comparably well as a component of XDream (**Figure 2.5a**). Thus, while we predominantly used the fc6 variant in experiments with real neurons (described below), we also used the pool5 and conv4 variants.

Second, although the generative networks learned from a large and varied natural image dataset, its composition could potentially bias the generative model and, thereby, the search space of XDream. To evaluate this possibility, we compared the same generator architecture trained on two different image datasets—ImageNet and Places (Zhou et al., 2017)—as a component in XDream. Both image datasets primarily comprised photographs, but ImageNet emphasized objects while Places emphasized scene categories. We used the two generative models as a part of XDream to optimize two variants of the same classifier architecture trained on the same two image sets. Both generative network variants performed almost equally well on both classifier networks (**Figure 2.5b**), suggesting that generators did not simply mirror the statistics of a specific image set. Third, unlike model unit activations, neuronal responses were stochastic. We verified that XDream could perform under realistic levels of response stochasticity and degraded performance gracefully (**Figure 2.6**). These tuning experiments paved the way for applying XDream to define feature selectivity in real neuronal recordings.

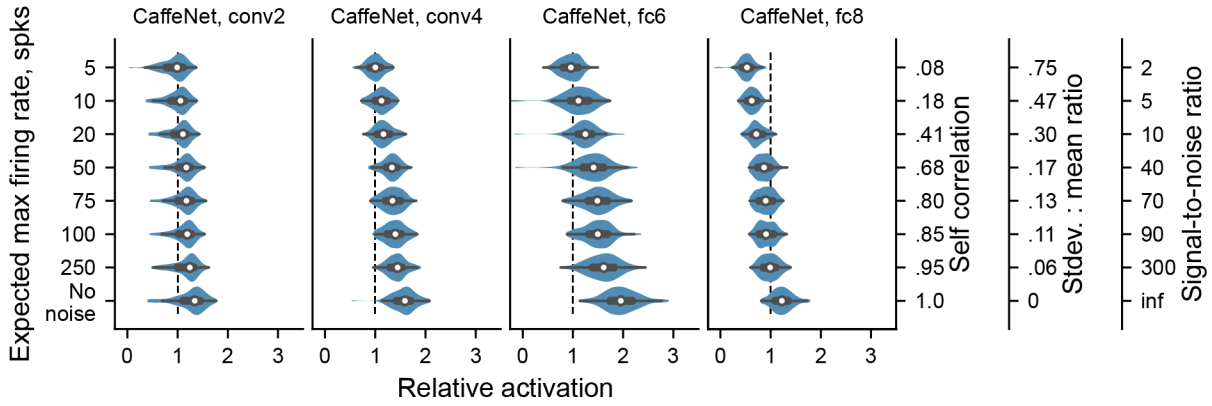

Figure 2.6. Benchmarking XDream with simulated noise.
We benchmarked XDream performance in the face of different levels of response stochasticity within an experimentally relevant range. XDream performance degraded gradually with noise but, especially for low-level target units, remained largely above relative activation = 1. Adapted from Xiao and Kreiman (2020). CC BY.

XDream uncovered strong stimuli for visual neurons

We adapted XDream to use with electrophysiological recording from awake, fixating monkeys. I worked with Carlos Ponce and Peter Schade to integrate XDream into a closed-loop experiment with online data collection. First, XDream produced a generation of images and waited for their neural responses. Then, the MATLAB-based task-control software presented the stimuli, counted the spikes they elicited, and formatted the results for XDream to produce the next generation, closing the loop. The recording equipment software provided online spike sorting and communicated the spike times to the task-control software.

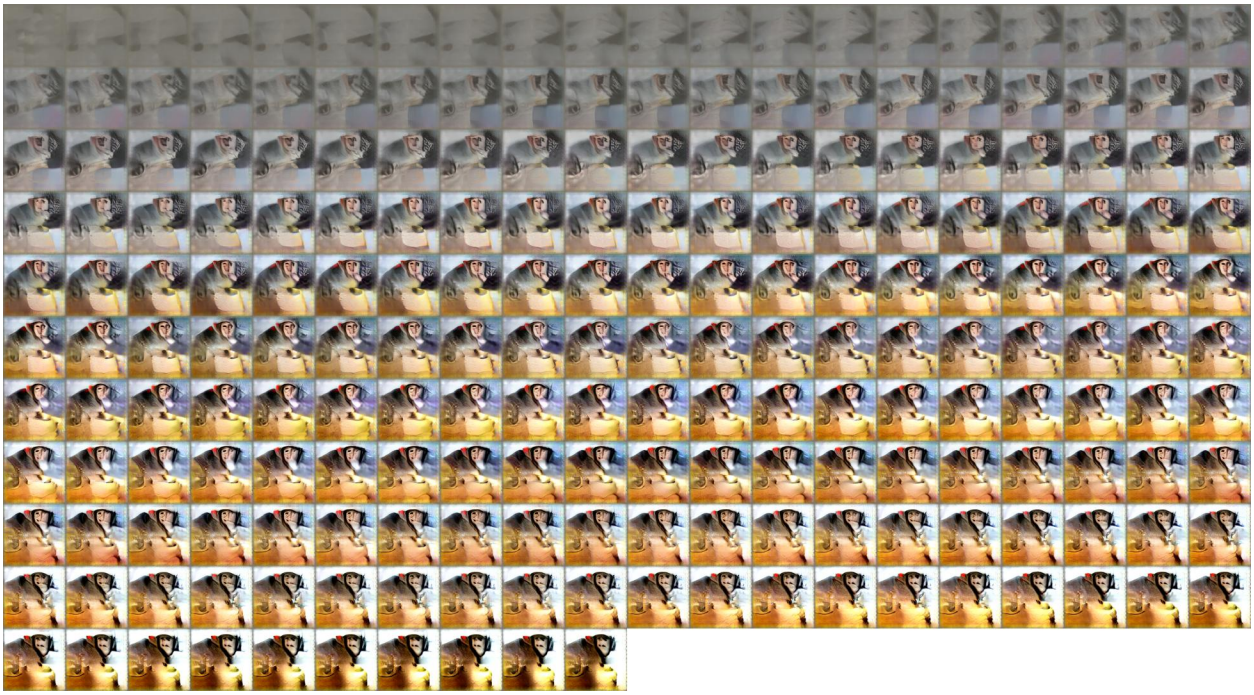

Figure 2.7. Illustration of the progress of an example evolution.
Each image represents one of 210 generations arranged in reading order. The image corresponds to the image code-space average of the population at each generation. Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

In an example experimental session, XDream optimized the responses in an electrode recording multi-unit activity (MUA) from posterior IT. **Figure 2.7** shows an image representing

each generation to illustrate the progressive change in the stimuli over time. XDream started from 40 random grayscale patterns and gradually tweaked them over 210 generations—adding color, outlining a shape, filling in textures, and segregating the foreground from the background. Guiding the evolution, the neuronal responses to the synthetic stimuli steadily increased over generations (**Figure 2.8a**). For reference, we also interleaved synthetic images with an equal number of natural images. Their responses stayed stable, decreasing slightly with time due to repetition suppression. In three sessions over seven days, we repeatedly used XDream to evolve stimuli for this chronically recorded MUA site. Across the sessions, the algorithm converged onto similar stimuli (**Figure 2.8b**). In a fourth session, we compared evolved images from different generations to a large set of 2550 natural images. The top preferred natural images resembled the evolved stimuli more than the non-preferred ones (**Figure 2.8c**). Nonetheless, late-generation evolved images were better stimuli than the best natural ones in this large set (**Figure 2.9a**).

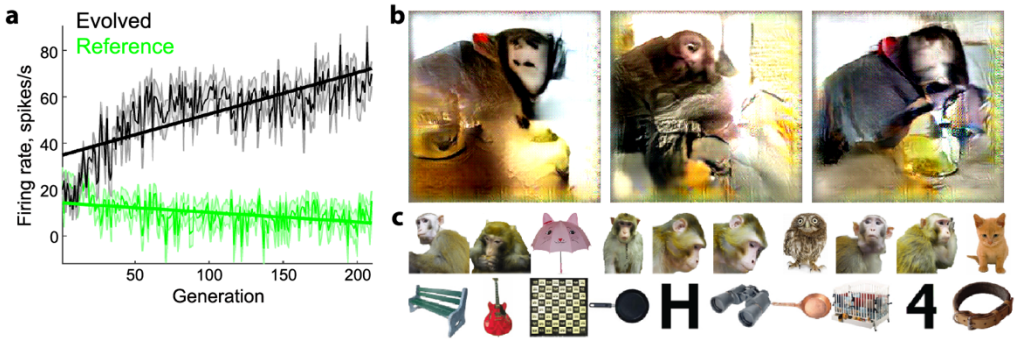

Figure 2.8. Selectivity to evolved and reference stimuli for the example neuron in **Figure 2.7**. **a**, Average firing rates per generation in an example evolution. Each generation had 40 evolving and 40 fixed reference images. Line and shading indicate mean ± s.e.m. **b**, Evolved stimuli for the same multi-unit channel in a chronic recording array over three sessions across seven days. **c**, Selectivity in this channel as illustrated by its top and bottom ten preferred stimuli from a large set of 2550 natural images. Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

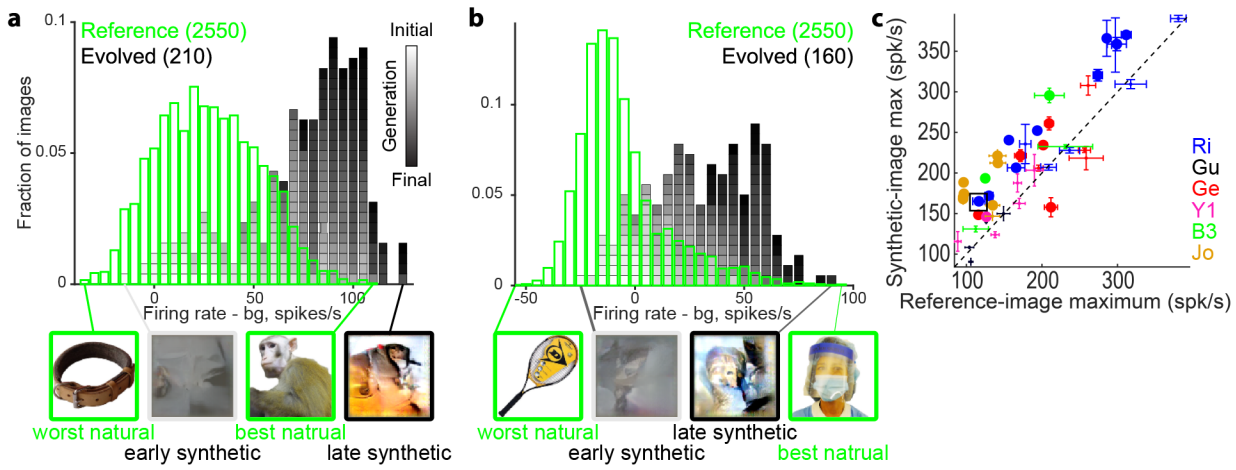

Figure 2.9. Selectivity to evolved images compared with natural image selectivity. **a, b**, After evolving stimuli for two neurons, in separate sessions, we recorded neuronal responses to one evolved image per generation and a large set of 2550 natural images. The histograms show the distribution of neuronal responses separately for natural and synthetic images. The bottom row shows example (best, worst) images for each image type. **a**, The same example neuron from **Figure 2.7**. **b**, A multi-unit recorded in central IT (face patch ML) of a different monkey, Ge. **c**, The maximum synthetic image response compared to the maximum reference image response. Each dot indicates one session; colors indicate different monkeys; the larger dots correspond to statistically significant differences ($p < 0.03$, FDR-corrected). Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

Carlos, Peter, and Margaret Livingstone conducted an initial 46 evolution experiments. Evolved images often elicited higher responses than the top natural image in each experiment (**Figure 2.9c**). In addition to the example unit above, another one, recorded in central IT, was tested with evolved images and the large 2550-image natural set. The best evolved stimulus led to responses almost as high as the best natural one (**Figure 2.9b**).

In other chronic recording sites, repeat evolutions also led to stimuli that share consistent features with each other and the top natural images (**Figure 2.10**). For example, both sites Jo-4 and Jo-5 from central IT evolved stimuli that contained a prominent black square on a light background. This feature echoed the sites' preferred natural pictures, all containing solid black blocks. However, from these pictures alone, one might consider the black blocks as coincidental features in the neuron's preference for man-made objects (Konorski, 1967).

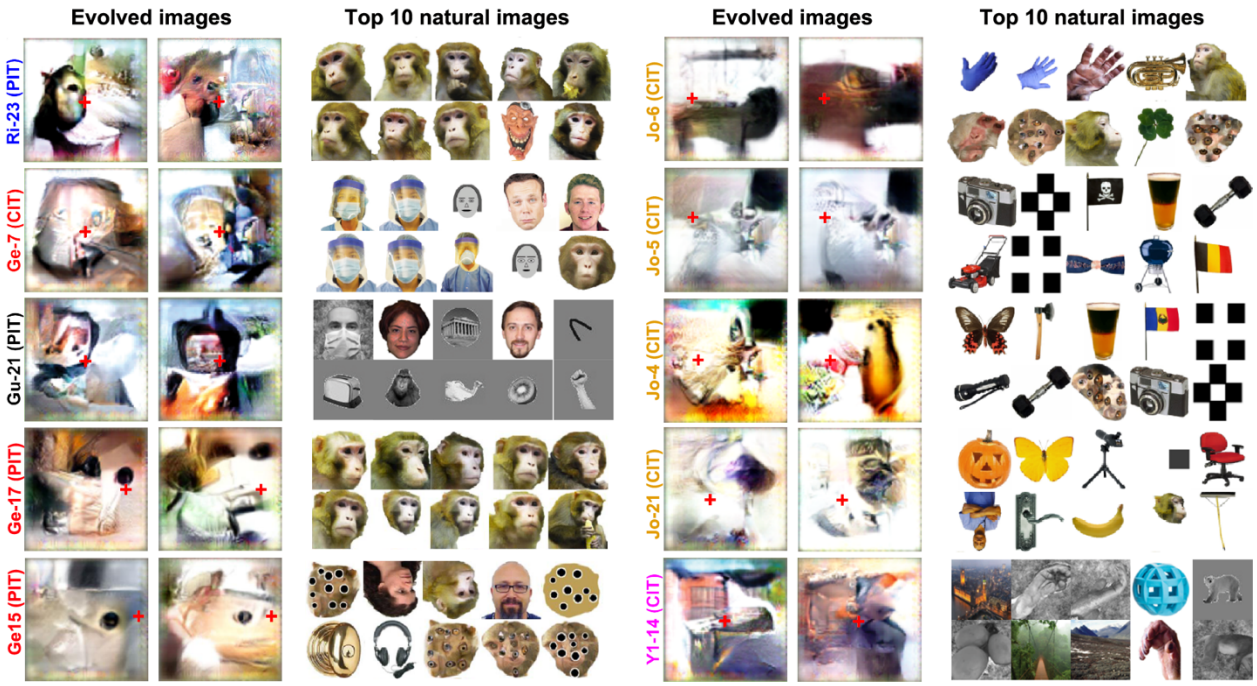

Figure 2.10. Evolved stimuli and preferred natural images for ten example neurons. Each heading on the left indicates one neuron. For each, two evolved stimuli are shown from independent optimizations, followed by the top 10 preferred natural images. Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

Evolved stimuli highly activated face neurons without resembling faces

XDream presented a unique opportunity to test whether IT neurons represented semantic categories. The paragon of category specificity in IT is the so-called 'face neurons,' which respond more strongly to images of faces than nonface objects (Tsao et al., 2006). Face neurons cluster on the cortex into face patches, whose locations are visible on fMRI and stereotyped across individuals. For these and other reasons, investigators have hypothesized that face neurons are specialized for processing faces, a stimulus class with particular ethological relevance. There had been scant evidence that face neurons can be strongly activated by anything but faces because it was hard to identify stimuli that strongly drive these highly selective neurons. XDream provided a tool to systematically ask whether there existed nonface stimuli that could activate face neurons as strongly as realistic faces.

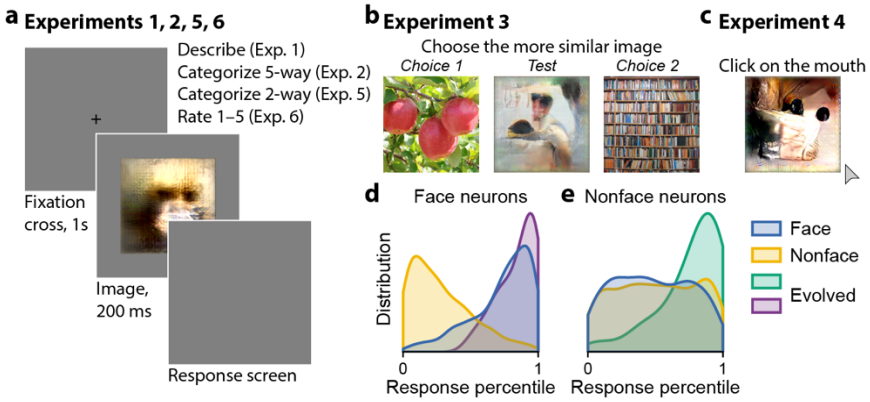

Figure 2.11. Overview of psychophysics experiments to evaluate evolved images. **a**, In four experiments, subjects saw an image briefly, then answered a question about the image. **b**, In experiment 3, subjects chose which one of two images was more visually similar to the test image. **c**, In experiment 4, subjects were instructed to 'click on the mouth' in each image. **d–e**, The distributions show the normalized neuronal responses to face, nonface, and evolved images in the sessions included in the psychophysics experiments. Adapted from Bardon et al. (2022). CC BY-NC-ND.

We used XDream to evolve strongly activating images for IT neurons in hundreds of sessions (**Figure 2.11**). The experiments were done over several years by Margaret Livingstone, Peter Schade, and me. We wanted to evaluate whether the evolved images were consistent with the prevailing view of high-order visual neurons as selective for object categories (Kanwisher, 2010). Cognizant of our potential implicit biases, we recruited naive subjects unaware of the source of the evolved images. We asked the subjects to view, describe, and categorize the evolved images. What words would people use to describe these images? How face-like would these images look? Did face neurons guide the evolution of more face-like stimuli than nonface neurons? A talented undergraduate student, Alex Bardon, helped design six psychophysics experiments (**Figure 2.11a–c**; Methods) progressing from open-ended to more tailored around the face category. Alex conducted the experiments, initially in person but primarily on the crowd-sourcing platform Amazon Mechanical Turk (MTurk), while I helped analyze the data.

We operationally defined a neuron as face- or nonface-selective by calculating a face selectivity index (FSI) using the neuron's mean (background-subtracted) firing rate to faces (r_f) and nonface objects (r_{nf}), FSI = $(r_f - r_{nf}) / (r_f + r_{nf})$. We defined face neurons as central IT (CIT) neurons with FSI greater than 0.5, excluding neurons in the posterior lateral (PL) face patch (Moeller et al., 2008), which also had high FSI but were known to respond to single eyes instead of whole faces (Issa & DiCarlo, 2012). We defined a nonface neuron as any neuron with an FSI less than zero. Based on these criteria, we analyzed 39 evolved images from face neurons and 47 from nonface neurons.

In the first experiment, we asked the subjects to use one word to describe the evolved images. **Figure 2.12a** and **b** show example answers participants gave to a face photo and a face neuron-evolved image. Across face neuron-evolved images, 'monkey' was the most frequent answer (14%), followed by 'dog' and 'art' (both 5%; **Figure 2.12c**). Nonface neuron-evolved images were also primarily described by words related to animals, albeit with lower frequency; the top three responses were 'bird' (7%), 'dog' (5%), and 'animal' (5%). We quantified and compared the descriptions of the face and nonface evolved images, face photos, and nonface pictures. The subjects described face neuron-evolved images more similarly to faces than they described nonface neuron-evolved images or nonface object pictures. These differences reached statistical significance in three of four quantification methods (**Figure 2.12e–h**). However, the participants still described both evolved image groups as less face-like than face photos. Could this result from the limitations of the image generator (**Figure 2.2**)? We tested this possibility by including face photos restricted to the style of the generative network. The subjects described these stylized faces as more face-like than evolved images (**Figure 2.12e–h**), suggesting that the generator was not the limiting factor.

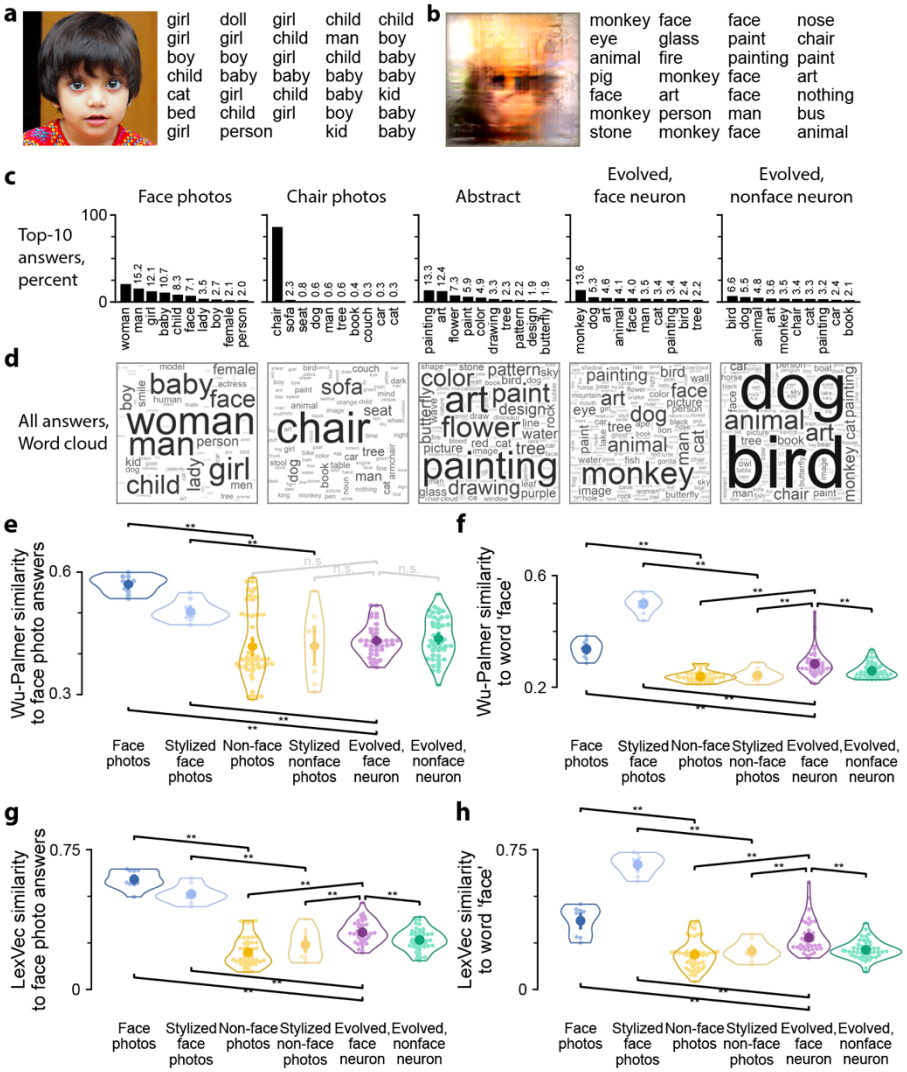

Figure 2.12. One-word image descriptions.
Thirty subjects saw each image and described it with one word. **a, b**, Two example images are shown with their responses across subjects. **c**, The histograms show the ten most frequent answers for each category of images, including example object images and the two sets of evolved images, from face-selective neurons or nonface-selective neurons. **d**, Word cloud visualization of answers given per category. **e–h**, We quantified how much the word descriptions indicated the perception of a face using Wu-Palmer similarity (**e, f**) or LexVec word embeddings (**g, h**). We compared image descriptions to the answers to face images (**e, g**) or the word 'face' (**f, h**). Open contours are violin plots indicating the kernel density estimate of the distribution across images. The lighter dots correspond to images. The darker center dots and lines indicate means and 95%-confidence intervals. Sloped brackets indicate one-tailed permutation tests. **, p < o.o1; n.s., not significant. P-values are FDR-corrected across all seven tests performed per subplot. Adapted from Bardon et al. (2022). CC BY-NC-ND.

The remaining five experiments recapitulated the same trends. In experiment 2, the subjects categorized images into one of five categories randomly drawn per trial from ten total categories. They classified face neuron-evolved images as face 44% of the time when face was an option, significantly more than expected by chance (20%), more than nonface neuron-evolved images (31%), and less than faces (97%) (**Figure 2.13a**). In experiment 3, the subjects chose which one of two images looked more similar to an evolved image. The two choice images depicted objects from two categories, again drawn per trial from ten possibilities. The subjects judged face neuron-evolved images to be more similar to face images than the average alternative category (68%) and did so more than chance (50%) and with nonface neuron-evolved images (56%; **Figure 2.13b**). However, the favorite option for face neuron-evolved images was dog (70%). In experiments 4–6, we tailored the question to evaluate face semblance. In experiment 4, we asked the subjects to click on the mouth in each image and, if unsure, give their best guess. Here, we reasoned that if the participants saw a face, they would click on consistent locations. Therefore, we used entropy to quantify the consistency of click locations as a readout for face perception. In experiment 5, the subjects answered whether each image contained a face or not. In experiment 6, they rated the 'faceness' of each image on a scale from 1 to 5. In all three experiments, the subjects rated face neuron-evolved images as more face-like than nonface neuron-evolved ones but less face-like than actual face photos or face photos in the style of the image generator (**Figure 2.13c–e**).

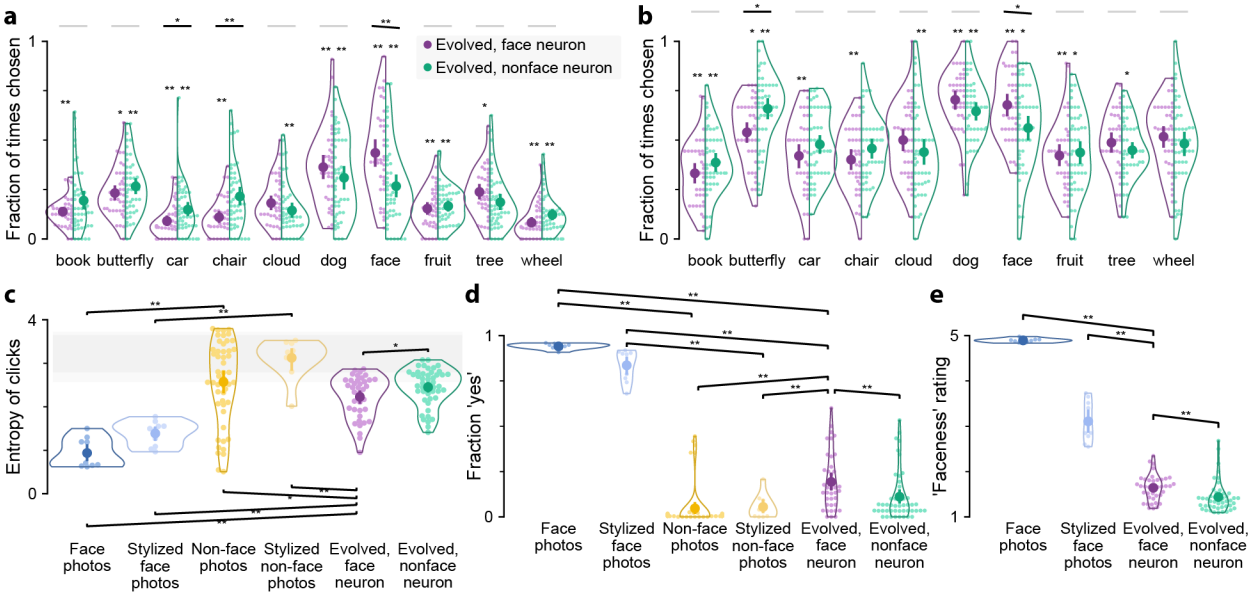

Figure 2.13. Summary of results from the remaining five psychophysics experiments. In panels **a** and **b**, individual asterisks indicate a statistically significant difference from chance, FDR-corrected across all tests performed. The short lines above indicate one- or two-tailed tests comparing face or nonface neuron-evolved images. *, $p < 0.05$; **, $p < 0.01$; gray, not significant. Otherwise, plot conventions follow those in **Figure 2.12**. Adapted from Bardon et al. (2022). CC BY-NC-ND.

Experiment 6 defined a graded quantification of an image's face-likeness to compare to its neuronal responses. We found face neuron responses to be dissociable from the face-likeness of images. On the one hand, face neuron responses were correlated with image faceness more so than nonface neuron responses (**Figure 2.14b–d**). The more face-selective a neuron, the higher this correlation was (**Figure 2.14c**). This result was not foregone. Face-selectivity was conventionally defined using the binary categories of faces (containing only the whole face) or nonface images (containing inanimate objects). Meanwhile, response correlation to faceness accounted for all pictures, including whole-body photos showing a small face, pictures of non-primate animals, and inanimate objects resembling faces (such as jack-o-lanterns); these images received a gradient of faceness ratings (**Figure 2.14a**). Indeed, the correlation between face neuron responses and faceness persisted when we considered nonface objects exclusively (**Figure 2.14e**).

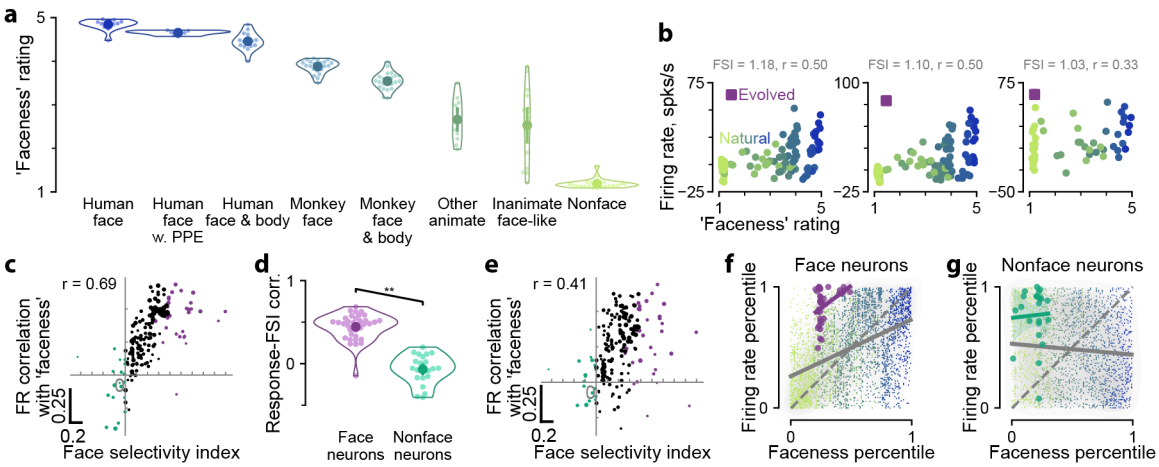

Figure 2.14. Comparing neuron face selectivity, responses, and image face semblance. **a**, We obtained faceness ratings for pictures for which we have collected many neuronal response data. Plot conventions follow those in **Figure 2.12**. **b**, The responses of three example face neurons compared with image faceness ratings. The face selectivity index and response-faceness correlation coefficient, r, are displayed above each plot. **c–d**, More face-selective neurons had higher response-faceness correlation. **c**, All 220 neurons are displayed. The subset of neurons categorized as face and nonface neurons are color-coded and compared in **d**. **e**, The relation between neuronal face-selectivity and response-faceness correlation persisted among exclusively nonface images. **f–g**, A plot like **b** for each neuron normalized and summarized across neurons. Adapted from Bardon et al. (2022). CC BY-NC-ND.

On the other hand, evolved images made counterexamples to face selectivity. They activated face neurons as strongly as most real faces but received faceness ratings lower than all face pictures (face neuron-evolved image ratings: 1.6 ± 0.3, range 1.2–2.3; human face ratings: 4.8 ± 0.1, range 4.5–5.0; monkey face ratings: 3.9 ± 0.1, range 3.5–4.1) (**Figure 2.14f**). Evolved images also strongly activated nonface neurons, although the responses of these neurons did not correlate with faceness (**Figure 2.14g**).

XDream compared to substitute model-based image synthesis

In concurrent studies, several groups developed another family of deep learning-driven methods to define the preferred stimuli of visual neurons (Abbasi-Asl et al., 2018; Bashivan et al., 2019; Malakhova, 2018; Walker et al., 2019). These methods derived from deep net-based encoding models that could predict neuronal responses to novel images (Schrimpf et al., 2020; Yamins et

al., 2014). Encoding models, being image-computable and differentiable, permitted backpropagation to create preferred inputs. Thus, the models *substituted* for neurons during image synthesis to propose preferred stimuli that only needed verification with neuronal recordings. Substitute model-based methods were orthogonal to XDream in essential aspects. We compared XDream and encoding model-based methods to explore whether we could combine the two approaches synergistically.

We started with four pilot sessions in a chronic recording array, where we showed natural images while running two independent evolutions for each of two neurons. Using the natural image responses, we fit encoding models, i.e., linear mapping functions based on neural network representations (Methods). The models accurately predicted held-out reference image responses but underestimated evolved ones (**Figure 2.15a**). Using the substitute models, we optimized images to maximize model-predicted activity with standard feature visualization methods based on backpropagation (Olah et al., 2017). Because of the adversarial examples described above (Szegedy et al., 2013), feature visualization critically depended on regularization, without which the optimization was wont to result in high-frequency visual speckles unlikely to excite neurons. Thus, we compared three conditions of progressively stronger regularization: none, randomly jittering the image, or restricting the image to the same generative model underlying XDream. Substitute model-optimized images evoked increasingly higher responses in the target neurons with higher degrees of regularization (**Figure 2.15b**). When the optimization was constrained to the same generative image space as XDream, the resulting images evoked responses comparable to the highest natural responses. However, the substitute model-optimized images still had lower responses than XDream-evolved images. The substitute models also consistently overestimated neuronal responses to the model-optimized stimuli, mirroring how the models underestimated

evolved image responses. The models predicted the least regularized stimuli to drive the highest activity—sometimes over 1000 spikes/s, a physiologically implausible number and five times beyond the highest observed response. These unregularized images received the lowest actual responses among the synthetic stimuli.

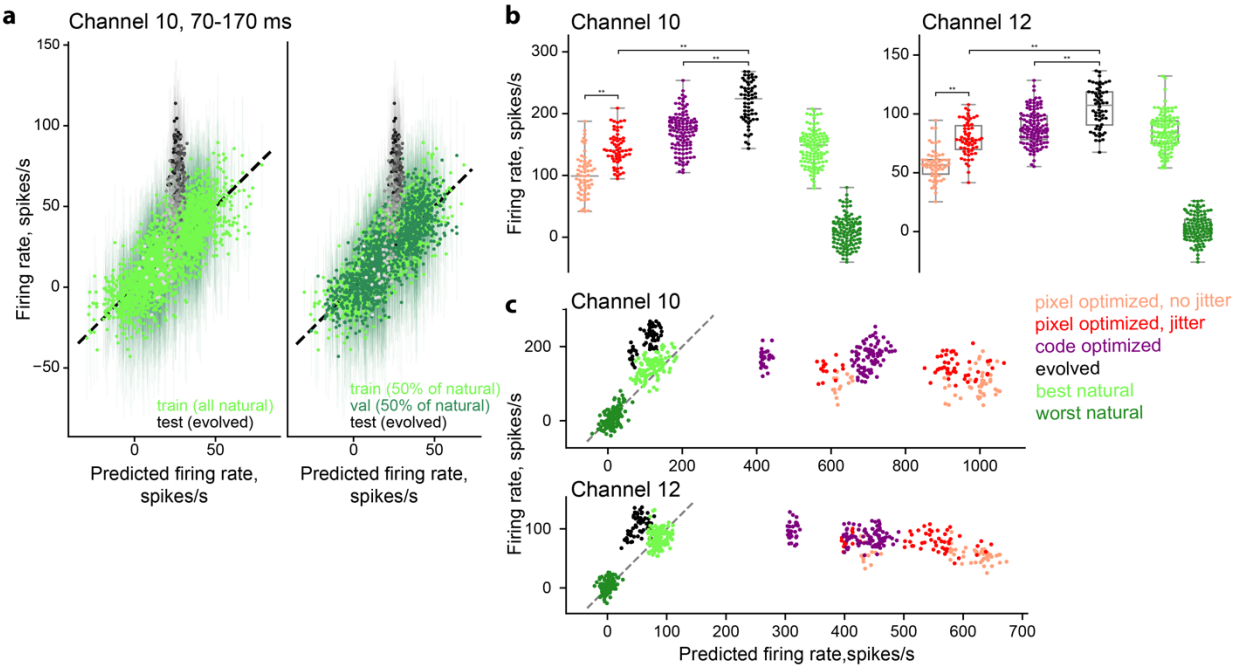

Figure 2.15. Pilot experiments comparing XDream with substitute-model image synthesis. **a**, We validated the substitute model fitting procedure on the data in **Figure 2.9a**. A model fit on the neuronal responses to half of 2550 natural images accurately predicted the held-out half of responses (right). However, the model consistently underpredicted responses to evolved images and did not improve from training on all the natural images (left). **b**, Comparison of neuronal responses to model-synthetic, evolved, and natural images. The subplots correspond to neurons; dots, images. The colors indicate how the image was derived. **, $p < 0.001$. **c**, Comparison of model-predicted and actual neuronal responses. We fit the models on natural images only. Adapted from Ponce et al. (2019) with permission. Copyright 2019 by Elsevier Inc.

The pilot experiments were preliminary, not the least because building substitute models involved numerous design choices and thus required considerable expertise to optimize. I was fortunate to collaborate with Pouya Bashivan, a lead author and modeler of the first study using a substitute model to define neuron stimulus preference (Bashivan et al., 2019). Pouya

implemented an online variant of the substitute model method (Methods) to test in parallel with XDream on the same neurons. We dedicated the same number of trials to guiding an evolution and collecting reference image responses to fit a substitute model. Thus, both methods used an equal amount of information about the target neuron. We collected data under this design from 50 sessions in 2 monkeys with chronic recording arrays in central IT (30 in M1, 20 in M2; **Figure 2.16**).

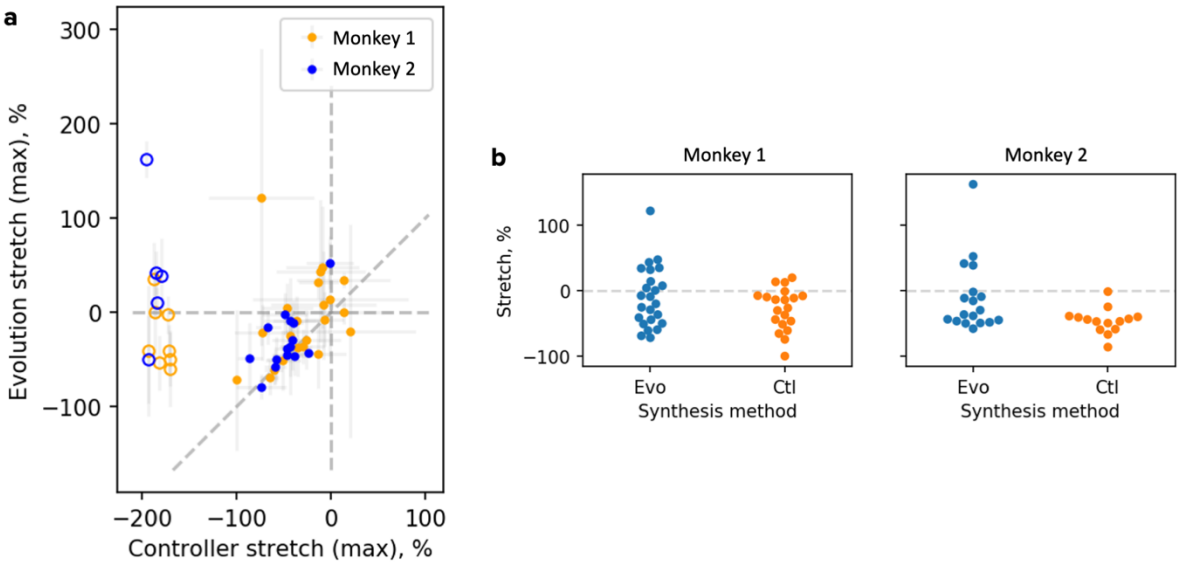

Figure 2.16. Summary of experiments comparing XDream and substitute-model synthesis. 'Ctl' refers to neural population *control* (Bashivan et al., 2019). We tested XDream and neural population control side-by-side. Each dot in the plot indicates one independent synthesis. 'Stretch' is the firing rate normalized by the highest reference rate; see the text.

To summarize the results, we use the 'stretch' metric introduced by Bashivan et al. (2019). The metric was calculated as the response to a synthetic image minus the highest reference image response and divided by it. Thus, a stretch of 0% meant the best synthetic image led to the same response as the best natural image, while 100% indicated the former led to twice the latter response. Overall, the best synthetic stimulus activated the neuron more strongly than the top natural image in 28% of sessions (14 of 50) for XDream and 8.1% of sessions (3 of 37) for

model synthesis (**Figure 2.16a**). Comparing the methods, XDream-evolved images led to higher responses than model-synthetic images in 59% of sessions (22 of 37). In another 15 of the 37 sessions, model-synthetic stimuli led to marginally higher neuronal responses. The results were similar in both monkeys (**Figure 2.16b**). In the 13 sessions remaining from the 50 total, the substitute model prediction accuracy on held-out images did not rise above the threshold required for proposing a synthetic stimulus.

Aside from differences in effectiveness, the two methods produced visually different stimuli that both, in turn, differed from natural images, even in cases when all three types drove comparable responses (**Figure 2.17**). This result suggested that instead of having a compact set of optimal stimuli, individual neurons may respond similarly to a plateau of effective stimuli that vary along null dimensions (Chang & Tsao, 2017).

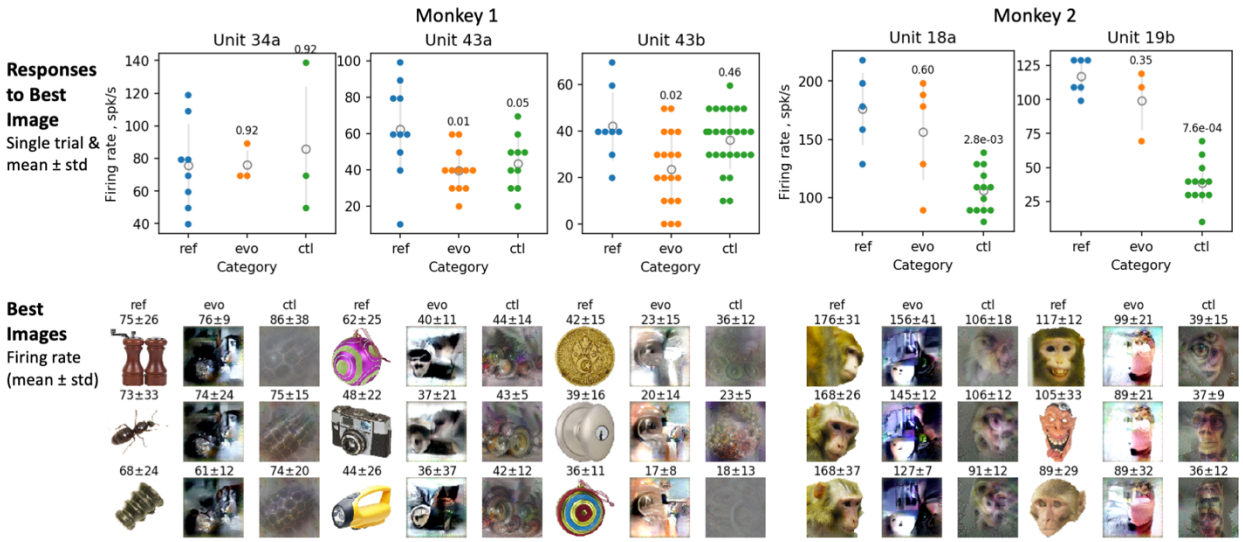

Figure 2.17. Example preferred stimuli found by three methods.
Top, the swarm plots show responses to the highest-activating image found by each method. Each subplot corresponds to one unit; each color, an image. Dots indicate single-trial responses. Numbers denote the p-value of difference from reference image responses using the two-tailed Mann-Whitney U test. Bottom, top-3 images found by each method. A group of three columns corresponds to one experiment. Numbers indicate mean ± stdev. response (spikes/s) across trials.

Discussion

The stimulus preference of neurons has been a classical problem in visual neuroscience. Investigators have lacked a general-purpose way to define effective stimuli outside the narrow boundaries of existing knowledge about neuronal tuning. Instead, researchers have mapped neuronal tuning with theoretically motivated image sets and, occasionally, fortuity. XDream and substitute model-based image synthesis methods represent a qualitative step forward in defining neuronal stimulus preference with less bias in the type of images surveyed and more generality across stages of visual processing.

The simultaneous advent of adaptive search and model-based synthesis was no coincidence. As is the case for much scientific progress, the advancement in defining preferred stimuli for neurons was due to new techniques, namely tools from deep learning. Deep neural networks made surprisingly good models of neuronal responses, which could be queried and optimized more readily than recording from neurons. Deep learning also provided generative image models that could represent diverse images, turbocharging existing adaptive stimulus search methods hamstrung by restrictive image parameterizations.

What conceptual significance preferred stimuli have is contiguous with that of the objective function they optimize, be it the activity of single neurons or some criterion defined by population firing patterns. Face neurons—reminiscent of the grandmother cells conceived by Jerry Lettvin (Gross, 2002)—have been an influential idea for explaining how the primate brain processes a singularly important class of stimuli. Face neurons respond more than twice as strongly to faces than nonface objects (Tsao et al., 2006), thus providing one of the best examples of a sharp category selectivity easily definable in natural language. However, is this descriptive convenience justified by face neurons being truly selective for faces as a semantic category?

Alternatively, whatever their teleological functions, face neurons may functionally respond to visual features enriched in faces but insufficient to constitute a face. XDream-evolved stimuli presented strong evidence that face neurons could be activated as strongly by faces as by images that people perceived to be not quite faces.

Despite the progress in methods for defining stimulus preference, an explanatory gap is already evident in the disparate stimuli that drive a neuron similarly strongly (**Figure 2.17**). The fact that visually dissimilar images can highly activate the same neuron may imply the existence of much more effective stimuli that await improved methods for discovery. A non-exclusive possibility is that the divergent images reveal neurons to be selective for a large and not obviously interrelated set of stimuli. Subsequent studies are already extending XDream to describe the geometry of neuron tuning landscapes, including the sharpness of tuning peaks and features of level sets, i.e., equally preferred stimuli (Wang & Ponce, 2022a, 2022b). Returning to the initial example in the introduction, suppose we could obtain the hypothetical catalog of a neuron's responses to all possible images. Even this whole neuron catalog would not equal a theory of neuronal tuning, let alone function. Understanding requires an accurate predictive model that, arguably, must additionally be succinct and mechanistically grounded. This understanding does not automatically result from finding a set of effective stimuli or fitting a statistically accurate encoding model. In the end, stimulus selectivity is an abstraction and probably only one facet of visual neuron function. Selective responses must ultimately account for the behaviors of the animal whose brain is under study. In Chapter 3, I explore the response properties of visual neurons during more natural behavior.

Methods

Experimental subjects

All procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee and conformed to NIH guidelines provided in the Guide for the Care and Use of Laboratory Animals. Nine adult male macaca mulatta (5–13 kg; 3–17 years old) and two adult male macaca nemestrina (13 and 15 kg, 12 and 14 years old) were socially housed in standard quad cages on 12/12 hour light/dark cycles.

Surgical procedures

Monkeys were implanted with custom-made titanium or plastic headposts before fixation training. After several weeks of fixation training, the animals underwent subsequent surgeries for array or chamber implantation. All surgeries were done under full surgical anesthesia using sterile technique.

Behavioral task and stimulus presentation

Monkeys fixated on a fixation spot in exchange for regular juice rewards. Eye position was monitored using infrared eye trackers (ISCAN, Woburn, MA). During fixation, stimuli were presented rapidly in pseudorandom sequence on an LCD monitor 50–60 cm in front of the monkey. The typical presentation duration and inter-trial intervals were between 100–200 ms, depending on the dynamics of neuronal responses; the typical stimulus size was 4–8 degrees of visual angle. Experiments were controlled using the NIMH MonkeyLogic2 software (Hwang et al., 2019).

Neuronal recording

Single- and multi-unit activity was recorded using chronically implanted intracranial electrode arrays or, in one monkey, using an acute electrode inserted daily in a recording chamber. The chronic arrays included Utah arrays (Blackrock Microsystems, Salt Lake City, Utah), floating

microelectrode arrays (MicroProbes, Gaithersburg, MD), or microwire brush arrays (McMahon et al., 2014; Porada et al., 2000) (MicroProbes). The acute electrode was a 32-channel NeuroNexus vector array (Ann Arbor, Michigan). Neural signals were amplified, digitized, and recorded using OmniPlex (Plexon, Dallas, TX) or Blackrock (Salt Lake City, Utah) data acquisition systems. Neuronal spiking responses were counted in a stimulus onset-aligned response window that typically started at 50–100 ms and ended at 150–250 ms after stimulus onset.

Generative adaptive search (XDream)

The source code of XDream is available at https://github.com/willwx/XDream. XDream comprised a generative neural network and a black-box optimization algorithm in a modular process. Pre-trained generative networks (Dosovitskiy & Brox, 2016) were obtained from the authors' website and used without further training. The generative network was originally trained with the Caffe library (Jia et al., 2014) in Python and subsequently ported to the PyTorch library (Paszke et al., 2019) for ease of use. The generative network was a function mapping from a numerical vector ('image code') to a 256 × 256 × 3 color image. In the majority of experiments, we used the 'fc6' variant of the generative network trained on ImageNet. The input image code to this model was a 4096-dimensional vector. We also used seven other variants in the family of generative models and an instance of the fc6 variant trained on a different image set, Places365 (Zhou et al., 2017).

The optimization algorithm was a custom genetic algorithm. For concreteness, we describe the algorithm with a typical set of parameters. Each generation consisted of a population of 20 image codes. Their corresponding image responses were exponentially transformed into probability weights to determine how frequently each code was selected to produce a progeny in the next generation. Each progeny was produced from two parent image codes by selecting a

random half of the vector elements from either parent. Subsequently, a random half of the vector elements were mutated by adding a random value drawn from a zero-centered Gaussian with a standard deviation of 0.5. In *in silico* experiments, we compared the genetic algorithm to two other black-box optimization algorithms: finite difference gradient descent and natural evolution strategies (Wierstra et al., 2014). We found the genetic algorithm to perform better in noisy conditions that were more relevant for neural recording. Subsequent extensions of XDream used an empirically faster algorithm based on CMA-ES (Hansen et al., 2003; Loshchilov, 2014; Wang & Ponce, 2021).

Initial generation

The initial generation of image codes was 40 achromatic textures constructed from a set of Portilla and Simoncelli textures (Portilla & Simoncelli, 2000) for the initial set of 46 evolution experiments. The textures were derived from randomly sampled photographs of natural objects on a gray background. We optimized image codes to minimize pixelwise loss between the synthesized images and the target images using backpropagation through the generative network. The resulting image codes produced blurred versions of the target images, which was expected from the pixelwise loss function and accepted because the initial images were intended to be quasi-random textures. In *in silico* experiments, the initialization was random vectors drawn from diagonal multivariate Gaussian distributions, with the parameters (mean and standard deviation) matched to the training distribution for the generative networks. Because the *in silico* experiments showed that optimization did not depend strongly on the initialization (Xiao & Kreiman, 2020) we used both initialization schemes above in further neurophysiological experiments.

Reference natural images

Most of the images were from a previous study (Konkle et al., 2010). Human and monkey images were from our lab, and the rest of the images were from public-domain databases.

Models and layers in **in silico** *experiments*

We selected several state-of-the-art convolutional neural networks as target models. The networks were pre-trained on the ImageNet dataset except for PlacesCNN, which was trained on the Places dataset. In each model, we tested what were approximately the early, middle, and late processing stages as well as the output layer. One hundred units were randomly selected from each layer. In convolutional layers, we selected the center spatial position for a feature channel. We used the following networks and layers: CaffeNet and PlacesCNN: conv2, conv4, fc6, fc8; ResNet-152-v2: res15_eletwise, res25_eletwise, res35_eletwise, classifier; ResNet-269-v2: res25_eletwise res45_eletwise, res60_eletwise, classifier; Inception-v3: pool2_3x3_s2 reduction_a_concat reduction_b_concat, classifier; Inception-v4: inception_stem3, reduction_a_concat, reduction_b_concat, classifier; Inception-ResNet-v2: stem_concat, reduction_a_concat, reduction_b_concat, classifier. CaffeNet and PlacesCNN were variants of AlexNet (Jia et al., 2014; Krizhevsky et al., 2017; Zhou et al., 2017). ResNet was developed by He et al (He et al., 2016). Inception-v3, Inception-v4, and Inception-ResNet were developed by Szegedy et al (Szegedy et al., 2017; Szegedy et al., 2016).

Stochastic neuron models

Let y be the activation value of an artificial unit. The activation value corrupted by stochastic noise was drawn from $Poisson(max(0, \gamma))$, where the rate parameter γ was analogous to the number of spikes of a neuron and equaled y times a constant scaling factor. We used a scaling factor because the signal-to-noise ratio of a Poisson process increased with the rate of the Poisson

process, but different network architectures and layers produced activation values of different scales. We used a scaling factor of $20 / y^*$, where y^* was the median max activation to random sets of 2500 ImageNet images, and 20 was a realistic number of spikes a biological neuron might fire to a preferred stimulus within a response time of 200 ms.

Substitute model-based synthesis

In the pilot experiments, we fit a substitute model using AlexNet fc6 layer representations to linearly predict neuronal responses. The linear model was fit using partial least squares with 25 retained components (Yamins et al., 2014). We optimized the input to this substitute model using backpropagation. We tested three levels of regularization in the optimization process. In the 'pixel optimized' condition, the pixel values in the optimized stimuli were directly updated by following the gradient direction maximizing model responses. With the 'jitter' regularization, the input image was shifted by a small, random amount at each iteration of optimization, to combat high spatial frequency features that would strongly excite the model but would be unlikely to affect biological vision (Olah et al., 2017). In the 'code optimized' condition, backpropagation was extended through a GAN (the same underlying XDream) to optimize an image code instead of pixel values.

In further experiments, the model was based on a pretrained ResNet-101 at the layer, 'mdl_block4_unit_3_bottleneck_v2.' The linear mapping model was continually updated online in a closed loop with neuronal recording, using the responses collected so far. The linear mapping procedure followed that described by Klindt et al. (Klindt et al., 2017). Briefly, the linear mapping weights were factorized into a 2D spatial weight and a 1D featurewise weight to apply to the 3D model features per image. These weights were optimized using stochastic gradient descent. The loss function was an L2 loss on the prediction error plus the following regularization

terms: an L2 and a Laplacian smoothness loss (Klindt et al., 2017). Image synthesis was done using the Lucid library (Lucid Authors, 2019) with the following regularizations: jitter, parameterizing the image by its Fourier transform, decorrelated color channels, and power spectrum decay.

Human psychophysics

Behavioral experiments were conducted on Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016). Participants provided informed consent and received monetary compensation for participation in the experiments. Experiments were conducted according to protocols approved by the Institutional Review Board at Boston Children's Hospital. Each subject completed only one experiment. The order of image presentation was randomized in all experiments. Images were presented in color at a size of 256 × 256 pixels. We did not monitor eye movements on the online platform, but the images were flashed for 200 ms, thus minimizing the effects of eye movement during image presentation. There was no time limit to respond in any of the experiments. No feedback was provided.

Semantic similarity

Word similarity was quantified by Wu-Palmer (WP) semantic similarity (Wu & Palmer, 1994) or a metric based on LexVec word embeddings (Salle et al., 2016). Given two words w_1 and w_2, the WP similarity (WPS) was calculated as $WPS = depth(LCS(w_1, w_2)) / (depth(w_1) + depth(w_2))$, where depth corresponded to the number of nodes from the top to arrive at the word in the WordNet hierarchy (Miller, 1995) and LCS corresponded to the least common subsumer, i.e., the most specific shared category. WP similarity ranged from zero (no relation) to one (identity). In the WordNet hierarchy, a word could have multiple hierarchical definitions; the highest WPS over all definitions was used. LexVec word embedding mapped each word to a vector such that the vectorial dot product between two word vectors approximated the log odds (i.e., enhancement in

probability) on one word occurring given that the other word occurred nearby. We used precalculated word embeddings from Salle and Villavicencio (Salle & Villavicencio, 2018) that represented each word as a 300-dimensional vector. LexVec similarity between two words was calculated as the dot product between two word vectors; thus, the value could be interpreted as approximate log odds as defined above. The log odds were lower-bounded by zero, so the minimum dot product value was approximately zero.

Entropy in mouth clicks

Entropy was calculated by putting x and y coordinates of click locations into 121 bins (11 × 11 grid on the image). For each bin i in an image, the click probability p_i was calculated as the number of clicks in the bin divided by the total number of clicks on the image. The entropy for each image was calculated as $Sum_i\ (-p_i \log p_i)$.

Statistical tests

Unless otherwise noted, we used pairwise permutation tests by permuting image assignment to categories for 10,000 permutations. P-values associated with Pearson's r were calculated using the exact distribution for the null hypothesis that the two variables were drawn from a bivariate normal distribution with zero covariance, as implemented in the Python library 'scipy' (Virtanen et al., 2020). One- or two-tailedness of tests and other statistical tests were noted in the text and were implemented using the Python library 'scipy.stats.' P-values for multiple comparisons were corrected to control the false discovery rate at the level of 0.05 using the two-stage Benjamini–Krieger–Yekutieli procedure (Benjamini et al., 2006) as implemented in the Python library 'statsmodels' (Seabold & Perktold, 2010).

3 Responses of ventral visual neurons during natural viewing

Introduction

Vision research typically uses tightly controlled experiments to isolate aspects of visual processing for detailed study. For example, a common way to study visual selectivity is the rapid serial visual presentation (RSVP) task described in Chapter 2. In this task, subjects must hold fixation to view images shown in quick succession. The stimuli are usually simple patterns or objects, and the presentations are typically brief and randomly ordered to focus the investigation on the first pass of feedforward processing isolated from history and context (DiCarlo et al., 2012).

Controlled experiments imply trade-offs between, on the one hand, the simplicity of the visual processes involved, the tractability of forming hypotheses, and the intuitiveness of interpreting results and, on the other hand, the applicability of the results to real-life behavior. Natural vision contrasts starkly with conventional, controlled studies in many ways. Animals use their eyes to extract behaviorally relevant information from scenes dense in information. It is likely that only select aspects of the entire scene image shone onto the retina are analyzed at each moment, contingent on the task at hand or the prevailing mental state. That we do not see—i.e., we cannot report or act on—everything projected into the eyes is evident in perceptual phenomena such as change blindness and inattentional blindness. Natural vision is further active, dynamic, and recurrent. We sample the visual scene over many eye movements in an interactive process unfolding over time. The history and context of viewing affect processing; adaptation alone can impact processing over seconds (Solomon & Kohn, 2014). This embodied nature of vision challenges the premise that there is a well-defined ‘the first hundred milliseconds of visual processing’ to study, an assumption undergirding task paradigms using randomly flashed images.

The considerable differences between natural and laboratory vision matter because they increase the possibility that experimental findings may not translate well to explaining visual processing in its natural context. For example, neuron response properties, such as retinotopic spatial specificity, feature selectivity, and response dynamics, may differ under natural conditions from when studied under controlled viewing. An example of a gap in explanation is the phenomenon of visual stability. Despite frequently moving our eyes, we see the world as stable and unmoving. Humans and monkeys alike can perform behaviors, such as the multi-step saccade task (Hallett & Lightstone, 1976), that need some form of eye movement-invariant information. According to controlled experiments, ventral neurons have retinotopically circumscribed receptive fields that move with the eye and thus could provide no neural basis for stable perception.

Indeed, how we achieve stable vision in the face of constant eye movements is a long-standing question asked since the times of von Helmholtz, Descartes, and Alhazen (Melcher, 2011; Wurtz, 2008). Investigators usually employ another tightly controlled task—a guided saccade task—to study the interaction of visual perception with eye movements. Because saccades take only tens of milliseconds, studies of visual responses during saccades typically use brief, simple stimuli such as light spots, gratings, or letters. These studies have shown that visual stability involves multiple distinct and synergistic processes (Cavanagh et al., 2010; Melcher, 2011; Rolls et al., 2003; Wurtz, 2008). A large part of visual stability is an illusion of absence. Following eye movements, we perceive no gap in vision or whole field motion opposite the direction of eye movement. Backward masking and saccadic suppression can explain why saccades do not momentarily blind or disorient us. Meanwhile, efference copies of the motor signal should explain why we do not see self-motion as world motion (Matin et al., 1982; Stevens et al., 1976; Wurtz, 2008).

The positive content of visual stability—a stable representation of what is where—is harder to explain. One proposed mechanism for stable perception is neurons that predictively remap. Studies have reported frontal, parietal, and superior collicular neurons that have retinotopic visual responses but respond to stimuli displayed in their future receptive fields before an impending eye movement (Duhamel et al., 1992; Umeno & Goldberg, 1997; Walker et al., 1995). Some have posited that predictively remapping neurons contribute to visual stability by monitoring congruity between the pre- and post-saccadic scene (Wurtz, 2008).

Predictive remapping alone is insufficient to account for visual stability. Neurons that eagerly shift their RFs only decouple two changes in time. It is unknown whether or which downstream neurons compare the predictive responses to postdictive ones to maintain a stable representation. Moreover, studies have probed remapped RFs with simple, often transient stimuli. Thus, it is unknown whether remapped RFs exhibit any selectivity necessary for the stable representation of what is where. The processing of detailed visual form is specialized to the ventral visual pathway (Ungerleider & Mishkin, 1982), whose output could contain a stable representation. Ventral visual processing culminates in the inferior temporal cortex (IT). IT is only two to three synapses away from the entorhinal cortex, which contains non-retinotopic gaze direction grid cells (Killian et al., 2012), and the hippocampus, which consolidates episodic memory and contains place cells, both plausibly in a gaze-invariant reference frame.

Previous investigations in ventral visual processing during eye movements have found some neurons that predictively remap in V4 (Neupane et al., 2016; Tolias et al., 2001) and far fewer in V1 and V2 (Nakamura & Colby, 2002). Others examined the feature selectivity of neurons in V1 and IT and found it to remain retinotopic (DiCarlo & Maunsell, 2000; Livingstone et

al., 1996). DiCarlo and Maunsell additionally showed that IT responses had identical dynamics during free and passive vision, providing evidence against predictive remapping in IT.

These studies continued to use simple stimuli in a sparse display. Intuitively, our experience of the world as stable may rely on recurrent processing of a persistent scene rich in framing cues. Visual processing is affected by at least a few seconds of viewing history through adaptation. Active natural vision may recruit additional top-down modulation of visual processing (Gilbert & Li, 2013). A few studies have examined visual processing during free viewing of persisting, natural stimuli (McMahon et al., 2015; Podvalny et al., 2017; Rolls et al., 2003; Sheinberg & Logothetis, 2001) and have provided evidence for non-classical responses. However, challenges inherent in interpretating essentially single-trial neuron activity in the face of unrestricted eye movements, complex stimuli, and intricate feature selectivity hinder the analysis of free-viewing data. Sheinberg and Logothetis (2001) compared object responses during controlled viewing and visual search, when the same objects were embedded into naturalistic backgrounds. They reported similar selectivity in both conditions but, during free viewing, prefixation anticipatory responses and potential dependence on conscious 'noticing,' both absent from DiCarlo and Maunsell's results with simple stimuli. Rolls et al. (2003) used a similar visual search task and found differences in the distance from an effective probe at which IT neurons responded. They interpreted the responsive distance as the receptive field size of IT neurons and concluded that it was larger in a uniform (rather than a natural-image) background and when the effective stimulus was the target (rather than the distractor) of visual search. McMahon et al. (2015) found face neuron responses in anterior IT to be consistent across trials when monkeys freely viewed natural movies, although some of the consistency could be due to stereotyped fixation patterns and position-invariance of the neurons. Podvalny et al. (2017) analyzed

electrocorticography responses to real-life stimuli in an epilepsy patient, tracking vision using a head-mounted camera-eye-tracker. They found the visual cortical activity to depend on whether individual fixations landed near a face and further concluded that responses in the primary but not high-order visual cortex varied with fixation duration. None of the studies combined naturalistic stimuli with detailed characterization of response dynamics, fine-grained feature selectivity, and receptive field location.

I studied the responses of neurons throughout the ventral visual stream as monkeys freely viewed images, aiming to evaluate how traditional response properties manifested during free viewing. To this end, I designed analyses to detect and quantify retinotopic specificity, response latency, spatiotemporal receptive fields, and potential extra-retinotopic responses. Because the methods were general enough to encompass a flexible behavioral paradigm, I benefited from data collected in retrospective and concurrent experiments by my colleague Saloni Sharma and advisor Margaret Livingstone. Their help broadened the scope of the results to many more neurons and areas throughout the ventral stream.

Results

We studied monkeys freely viewing natural images and designed analyses to interpret ventral visual activity during this naturalistic behavior. The analyses were flexible and robust enough to apply to a wide range of experimental conditions. Here we report results summarized over 679 experimental sessions collected over three years, containing 883 hours of recording from 13 monkeys making 4.7 million fixations on thousands of natural images. We report results while accounting for the heterogeneity in the data. While we recorded from chronically implanted multielectrode arrays, for simplicity of interpretation, we treated each unit and session as separate

and instead reported the mean and spread across monkeys using hierarchical permutation. Statistical tests used non-parametric hierarchical permutation tests. (See Methods for details.)

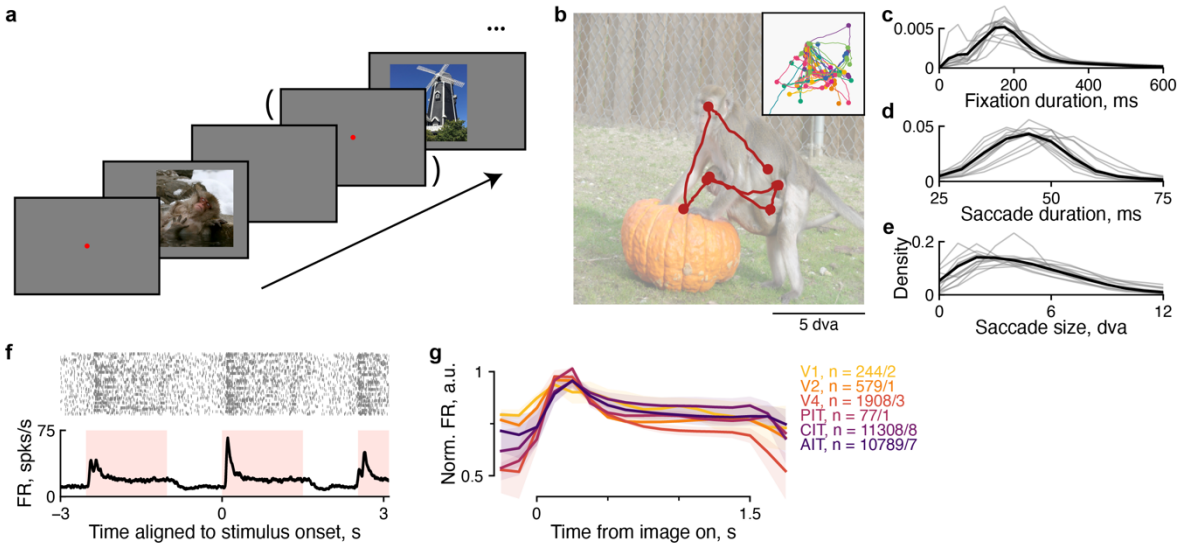

Figure 3.1. Overview of the free viewing experiment.
a, A random image was shown in each trial for the monkey to view freely. In a small subset of trials, a fixation dot was displayed prior to image presentation. **b**, An example eye trajectory is shown for one trial. Inset shows eye trajectories across trials for the same image in one experimental session. Each color indicates a trial; dots indicate fixations. **c–e**, The distribution of fixation durations, **c**, saccade durations, **d**, and saccade sizes, **e**. Each thin line indicates one monkey; the thick lines indicate the average across monkeys. **f**, Example spike rasters and average firing rates aligned to stimulus onset for an example neuron. The pink shading indicates the stimulus presentation cadence in this session. **g**, Average normalized firing rates separately for visual areas. The n values indicate the total number of neurons/monkeys.

In each session, a sequence of natural images was presented and repeated in a pseudorandom block fashion (**Figure 3.1a**). A trial lasted up to 1.5 s in a typical session (410 of 679; range 0.3–60 s in remaining sessions) and was interrupted if the monkey looked away from the image. Images were typically 16 x 16 degrees of visual angle (dva) in size (487 of 679; range 8 × 8–26 × 26 in remaining sessions). Monkeys performed the free-viewing behavior naturally. Across repeats, monkeys examined each picture in varied looking patterns. The basic statistics of the looking behavior were consistent across subjects (**Figure 3.1c–e**) and matched previous studies on similar behaviors (Mitchell et al., 2014; Zhang et al., 2022). An average fixation lasted 248 ± 35

ms, while an average saccade took 48 ± 4 ms and subtended 4.9 ± 0.8 dva (all mean ± stdev. across subjects).

While monkeys freely viewed images, we recorded extracellular single- and multi-unit activity throughout the ventral visual pathway using chronically implanted multielectrode arrays. The recordings spanned six visual areas: V1, V2, V4, and IT in its posterior, central, and anterior subdivisions (PIT, CIT, and AIT). Of these, V4, CIT, AIT, and V1 were represented by at least two monkeys. Overall, ventral visual neurons were more active during stimulus presentation than between trials (**Figure 3.1f–g**).

Face-neuron responses were specific to each fixation

We started to examine how IT responses interacted with eye movements and stimulus content by analyzing a subset of our recordings from face-selective neurons. Face neurons respond more strongly to face images than to nonface object images presented during passive fixation (Tsao et al., 2006). We annotated regions of interest (ROIs) on images to categorize each fixation as face or nonface (**Figure 3.2a**). To allow for the finite receptive field size of IT neurons, we categorized fixations within 2.5 dva of a face ROI as face fixations and the rest as nonface fixations. To identify face neurons functionally, we used 'the zeroth fixation,' the period right after image onset and before the first eye movement on the image. In this period, the appearance of a random image placed either a face or a nonface near where the monkey happened to be fixating. Thus, vision during the zeroth fixation was comparable to conventional passive viewing experiments. We calculated the face selectivity indices (FSI; Methods) of neurons during zeroth fixations using the responses 0–200 ms after image onset. Face neurons were defined as units recorded from four face-patch arrays with FSI of at least 0.2, i.e., at least 50% higher responses to faces than nonface. Analogously, we calculated FSI during free viewing using responses 0–200 ms after the

end of saccades, i.e., the onset of non-zeroth fixations ('fixation 1+'). Neurons were similarly selective to faces during the zeroth and non-zeroth fixations (**Figure 3.2b**).

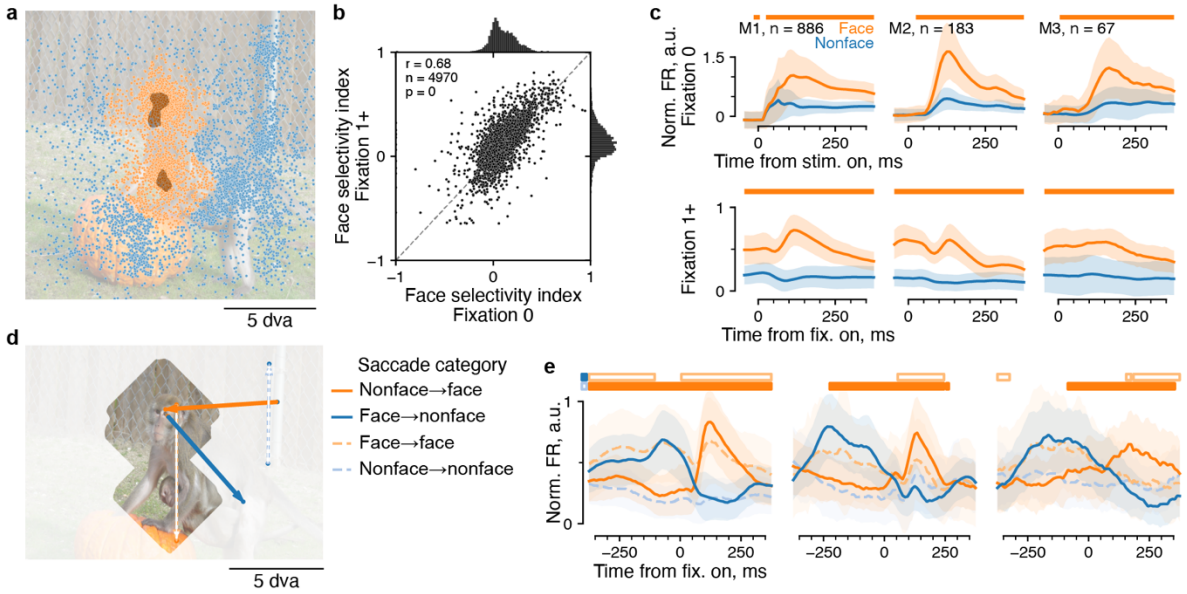

Figure 3.2. Face-selective responses reflected whether each fixation landed near a face. **a**, Face ROIs were used to categorize fixations, illustrated here for the same example image as in **Figure 3.1b**. The dark shaded regions correspond to face ROIs; each dot indicates a fixation; color indicates the categorization of the fixation content (orange, face; blue, nonface). **b**, Face selectivity was quantified by a selectivity index and compared between the zeroth and subsequent fixations. The dashed line is the identity line. **c**, Mean responses per category aligned to the onset of either the zeroth or subsequent fixations. Each column corresponds to one monkey; n indicates the number of neurons; lines and shading indicate mean ± stdev. across neurons; horizontal bars indicate time bins where responses were significantly greater for face than nonface responses ($p < 0.01$, permutation test, FDR-corrected). **d**, Non-zeroth fixations were further divided according to their preceding fixations in two-by-two categories. An example saccade is shown for each category. Open arrows indicate saccades between the same fixation category (blue: nonface-to-nonface; orange: face-to-face). Solid arrows indicate saccades between different categories (blue: face-to-nonface; orange: nonface-to-face). **e**, Mean responses per saccade category for the same neurons as in **c**. Each subplot shows data from one monkey. Horizontal bars indicate time bins where responses were significantly greater for nonface-to-face than for nonface-to-nonface saccades (lower bar), or for face-to-face than for face-to-nonface saccades (upper bar). Other plotting conventions follow those in **c**.

During active vision, face-selective responses began before fixation onset (**Figure 3.2c**). However, the apparent precursory responses could be because a face fixation was often preceded by another face fixation. Thus, we split neighboring fixations into four conditions by the

categories flanking the intervening saccade (**Figure 3.2d**). We further restricted our analysis to large saccades to alleviate any imprecision arising from arbitrary ROI boundaries. Face neuron responses meticulously followed the category of each fixation across a saccade (**Figure 3.2e**). For example, for nonface-to-face saccades, face neuron responses switched from low to high within a few tens of milliseconds after fixation onset. While the fixation category explained broad trends in response magnitude, it also revealed fine differences. For example, face responses following another face fixation were lower than face responses following a nonface fixation (**Figure 3.2e**; compare dashed orange lines to solid lines in the peak after fixation onset), consistent with adaptation effects. Moreover, when comparing nonface-to-face saccades with nonface-to-nonface saccades, responses were significantly higher well before saccade onset in all monkeys (**Figure 3.2e**). We could not rule out the possibility that the unexplained differences were because our face ROIs were conservative, or our ROI dilation (see Methods) did not fully account for neuronal RFs that extended further without an abrupt cut-off. To progress, we looked for an analysis that depended less on a precise, binary RF delineation. Furthermore, we sought a metric to evaluate neuronal responses whose selectivity was less straightforward to delineate with ROIs.

Ventral visual responses were specific to individual fixations

The selectivity of ventral visual neurons did not always correspond to well-defined ROIs, and even when it did, it was problematic to interpret ROIs assuming binary selectivity. Instead, we made use of return fixations to define a self-consistency metric. Primates repeatedly foveate the same location above chance frequency under a variety of task contexts, including visual search and free viewing (Zhang et al., 2022). **Figure 3.3a** shows example pairs of return fixations, within and between trials, a monkey made during one session. Return fixation pairs should lead

to similar responses if neurons encode the retinotopic stimulus content at each fixation. We quantified the self-consistency of responses during return fixations (**Figure 3.3b**) by calculating Pearson's correlation of responses between every pair of return fixations, across pairs; this quantification is analogous to how standard response self-consistency is quantified by the correlation between trials of repeated presentations, across stimuli. To assess temporal specificity, we calculated self-consistency for responses at different times relative to fixation onset. The first two subplots in **Figure 3.3c**, corresponding to the purple bars in **Figure 3.3b**, show the firing rates of an example neuron 200 ms before (subplot 1) and after (subplot 2) the onset of return fixations. The responses were more consistent after (Pearson's $r = 0.55$) than before ($r = 0.30$) fixation onset. For this example neuron, self-consistency was well above zero before fixation onset. We realized that fixations close by in time also tended, on average, to be close by in space. To account for the spatial correlation between subsequent fixations, we contrasted two rules for pairing responses for comparison: pairing responses such that the *current* fixations (response times $t \geq 0$) were return fixations (purple in **Figure 3.3b**), or pairing responses such that the *previous* fixations ($t < 0$) were return fixations (**Figure 3.3b**). The four subplots in **Figure 3.3c** illustrate responses self-consistency for the example neuron at different response windows and with different paring rules. While the responses were similar even before fixation onset (**Figure 3.3c**, first subplot), this similarity was better explained by the previous fixation because the responses were more self-consistent when the previous fixation was paired (**Figure 3.3c**, third subplot) instead of the current one. The converse was true for responses after fixation onset.

Figure 3.3. Response self-consistency indicated specificity to each fixation but no stable representation across fixations.
a, Example pairs of return fixations are shown. A return fixation pair is two nearby fixations on an image, within a trial or across trials. The dots indicate fixations, color coded by trial. Black lines join return fixation pairs within a 1-dva threshold. **b**, The schematic illustrates two pairing rules for calculating response self-consistency based on return fixations. Orange and blue indicate two fixation sequences. The second fixation in each sequence (denoted i and j, respectively) make up a return pair. Neuronal responses aligned to the second fixation onset (purple) are paired based on the rule, 'the current fixation is a return fixation.' Responses aligned to the third fixation onset (green) are paired based on the rule, 'the previous fixation was a return fixation.' **c–e**, Example data illustrate how we quantified self-consistency. **c**, Each dot indicates the firing rates (FR) of an example neuron in a return fixation pair. The four subplots correspond to two response time bins: 200 ms before (columns 1 and 3) or after (columns 2 and 4) fixation onset; and two response pairing rules: paired by the previous or current fixation. Self-consistency was quantified as Pearson's correlation of paired firing rates across pairs. A small amount of random noise was added for visualization purposes only because firing rates were discrete and often overlapped. **d**, Self-consistency for all neurons in this example session. The x- and y-axes correspond to the two response time bins. Color indicates the response pairing rule (previous or current; see **b**). Within a color, each dot indicates one neuron. The example neuron in **c** is indicated by square markers. The dashed line is the identity line. **e**, Self-consistency time courses were calculated using 50 ms sliding response time bins. The lines indicate averages over neurons in the example session. Dashed lines correspond to all return pairs ('default'); solid lines correspond to 'decorrelated' return pairs (see text). **f**, Decorrelated self-consistency time courses summarized over monkeys and neurons. **g**, Self-consistency as a function of the threshold for pairing return fixations, separately for visual areas. **h**, Cumulative distribution of response latency following fixation onset, separately for visual areas. **i**, Comparison of response latency following stimulus onset (x-axis) and fixation onset (x-axis). Each dot indicates one neuron; error bars indicate bootstrap stdev. of the latency estimates; color indicates visual area. Grey diagonal shading indicates 25 ms, the threshold bootstrap stdev. of latency estimates. **j**, Schematics of prediction by the null hypothesis of purely retinotopic responses (H_0, top) and the alternative hypotheses of stable representation (H_1, bottom; see text). Colors indicate various measures of response similarity, or self-consistency. **k**, Mean self-consistency as a function of time from stimulus onset, to compare against the predictions in **j**. In **f**, **g**, **h**, and **k**, lines and shading indicate hierarchical mean ± bootstrap 95%-CI, first over neurons, then over monkeys. The n values indicate the total number of neurons/monkeys.

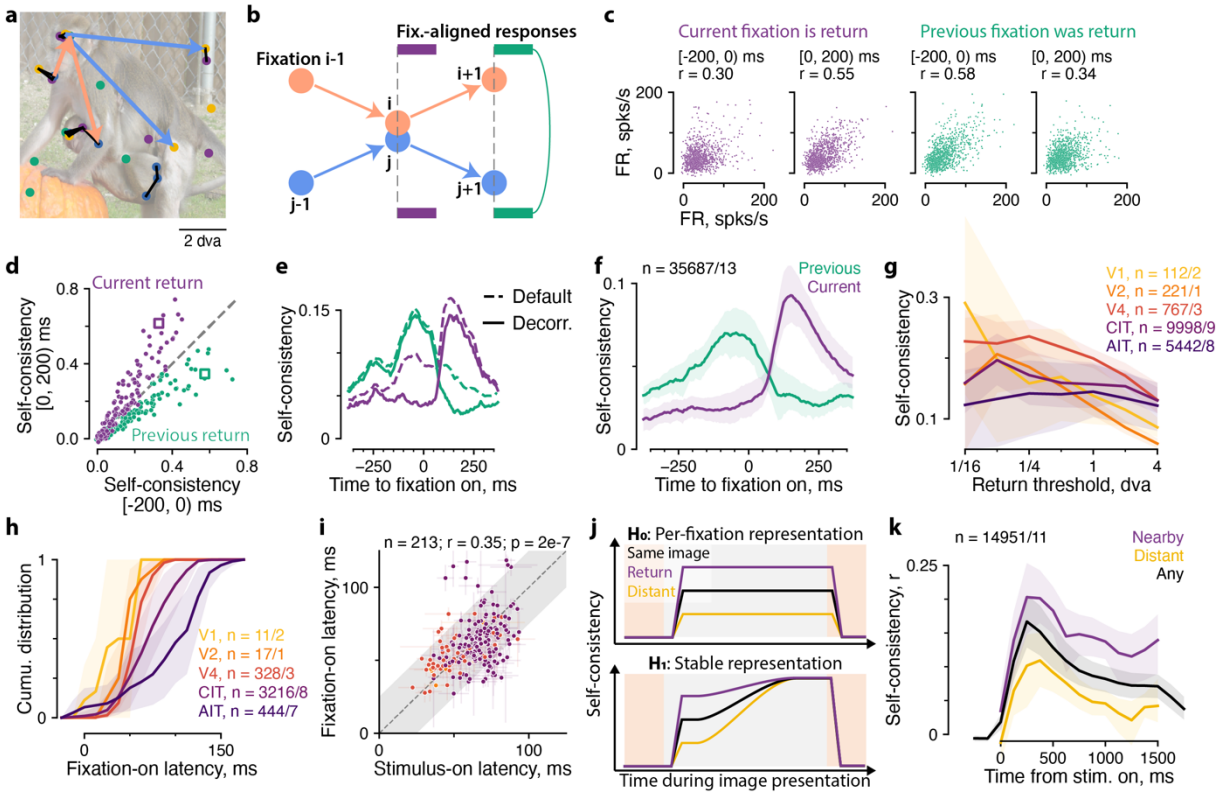

Figure 3.3 (Continued).

Most neurons in this example experimental session showed higher self-consistency in the response time window corresponding to the paired fixation than in the complementary time window (**Figure 3.3d**). This temporal specificity can be revealed in higher resolution by using sliding 50 ms time bins (**Figure 3.3e**, dashed lines). We devised a further control for the baseline self-consistency by *decorrelating* the non-paired fixation. For each pairing rule (previous or current), we sub-selected return fixation pairs such that the fixation during the complementary, non-paired time period must be separated by more than 4 dva. This decorrelation procedure further reduced self-consistency in the unpaired time period (solid lines in **Figure 3.3e**), but barely affected self-consistency values within the paired time period. We used the decorrelated variant of self-consistency in further analyses. The decorrelated self-consistency time course is summarized in **Figure 3.3f**, showing robust specificity to each fixation across monkeys and neurons.

Return fixation self-consistency furnished a measure for the spatial specificity of neuronal responses during free viewing. We asked how self-consistency depended on the distance threshold for defining return fixations. Self-consistency increased with smaller, more stringent thresholds, plateauing at 1 dva for anterior IT (AIT) neurons and continuing to increase in central IT (CIT), V4, V2, and V1 down to 1/4 dva, the resolution of our eye trackers (**Figure 3.3g**). Thus, ventral visual responses were exceedingly retinotopically precise, despite the relatively large sizes of IT receptive fields (Gross & Mishkin, 1977).

The time course of self-consistency also supplied a measure of response latency during free viewing. We estimated the latency of fixation-specific responses as the time point at which responses became better explained by the current than the previous fixation, i.e., the crossing point between the previous- and current-return self-consistency curves. This latency estimate resulted in typical values that, as expected, increased along the processing hierarchy in the ventral

pathway (**Figure 3.3h**). We could further compare, in the same neurons, the fixation-onset latency and the classical, stimulus-onset latency, using zeroth fixation responses to estimate the latter (see Methods). For a small subset of neurons, we could estimate both metrics reliably (bootstrap stdev. < 25 ms; Methods). The two measures of response latency were similar for the same neurons and showed consistent differences across neurons (**Figure 3.3i**).

The results so far indicate that ventral visual responses were yoked to individual fixations during free viewing. It remains possible that some subtle aspect of responses may integrate across fixations to build up a stable representation. We hypothesized that a stable representation would manifest as reduced retinotopic specificity over the course of image presentation after multiple saccades and that this reduced retinotopic specificity would result in a narrowing gap between different measures of response similarity: self-consistency between return fixations, as above; response similarity between any two fixations on the same image; and response similarity between distant fixations > 8 dva apart (**Figure 3.3j**, bottom, alternative hypothesis H_1). In comparison, in the null scenario, responses should remain retinotopic. The null model would predict the same retinotopic gap to persist throughout the trial (**Figure 3.3j**, top, null hypothesis H_0). We evaluated the three measures of response similarity (i.e., self-consistency) and summarized them across all trials lasting 1.5 s (**Figure 3.3k**; Methods). A roughly constant retinotopic gap persisted among the three self-consistency measures, contradicting the alternative hypothesis of a steadily accumulating stable representation. Meanwhile, self-consistency decreased overall throughout the trial, an observation not accounted for by the null hypothesis. We speculate that this decrease is due to the general decrease in firing rate as the monkeys continued to view the same image (**Figure 3.3g**), consistent with adaptation.

Computational models predicted neuronal responses from fixation-aligned stimulus content

The classical response properties exhibited so far encouraged us to test whether the same image-computable models applicable to passive viewing (Schrimpf et al., 2020; Yamins et al., 2014) could also predict visual responses during free viewing. We fit models to predict fixation-aligned responses from a 4 × 4 dva image patch (i.e., $1/16^{th}$ the area of a 16 × 16 dva stimulus) centered on each fixation (**Figure 3.4a**). A pretrained artificial neural network (vision transformer) (Dosovitskiy et al., 2020) converted each image patch into a 1024-D feature vector. From these feature vectors, we fit a linear model (Ridge regression) to predict neuronal responses. Model performance was evaluated using cross-validation across images. The models explained a large fraction of the explainable (i.e., self-consistent) responses, both during the passive viewing-like zeroth fixation (**Figure 3.4b**) and during free viewing (**Figure 3.4c**). The fraction of responses explained peaked at 50% ± 15% during free viewing (median ± stdev. across monkeys). This is in line with the performance of related models in passive viewing studies, despite the fact that we evaluated models on single trial responses, whereas passive viewing studies typically evaluate models with trial-averaged responses over multiple repetitions of the same stimulus.

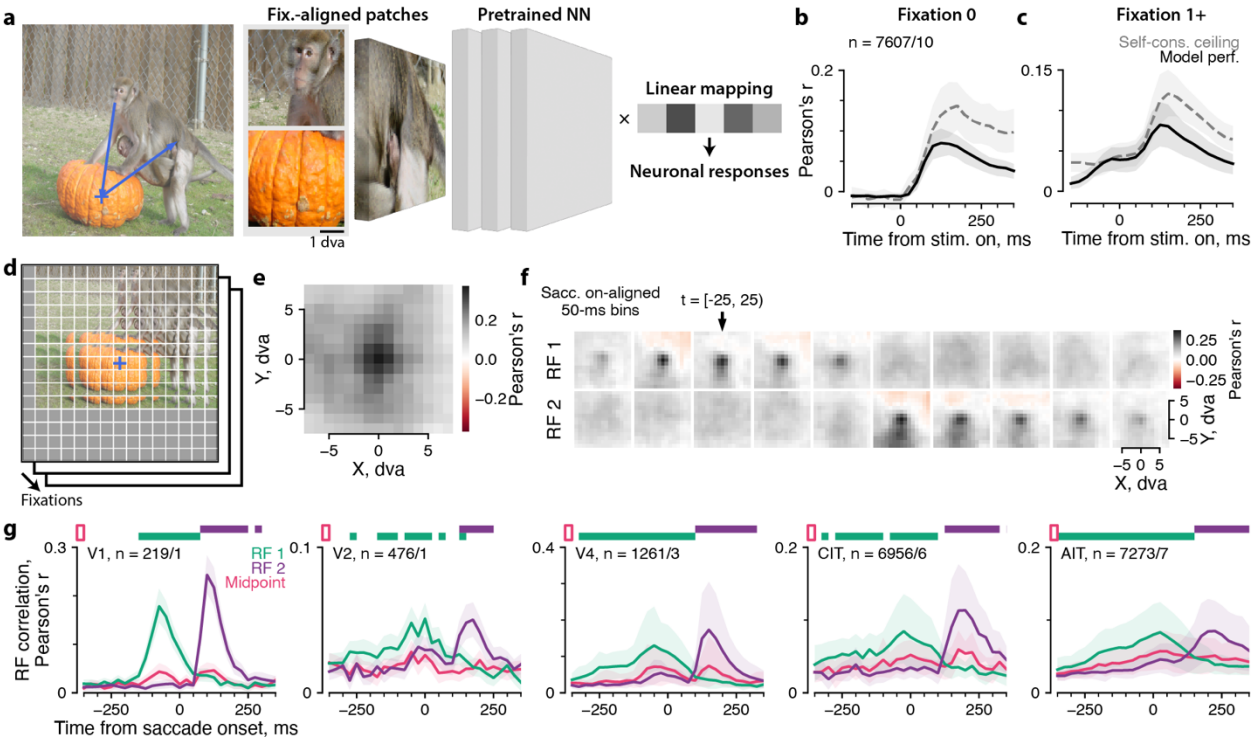

Figure 3.4. Computational models captured neuronal selectivity to individual fixations and revealed the spatiotemporal structure of receptive fields.
a, We adapted computational encoding models of visual neuronal responses to free viewing data. The model comprised a pretrained, fixed neural network (NN) feature extractor (vision transformer; Methods) and a linear mapping fit to neuronal responses. The model inputs were fixation-centered image patches of size 4 dva in panels **b, c** or 2 dva (illustrated) in panels **d–g**. **b, c**, Cross-validated model performance was quantified by Pearson's r and compared to the ceiling of return fixation self-consistency, separately for fixation 0, **b**, or subsequent fixations 1+, **c**. **d**, Illustration of model-based inference of neuronal receptive fields (RFs) during free viewing. At the fixation indicated by a cross (also in **a**), the retinotopic image was partitioned into 2-dva small image patches. The same partitioning was done for different fixations using the same grid of offsets from fixation. The model NN converted each patch to a feature vector. Feature vectors at every offset from fixation were evaluated for their fit to neuronal responses (see text). This process derived a map of local stimulus 'correlation' to neuronal responses. **e**, Example RF inferred from the zeroth-fixation responses of one neuron, represented as a map of model-based, cross-validated reverse correlation and quantified by Pearson's r. **f**, Example spatiotemporal RFs inferred from the responses of the same example neuron in e for rolling time bins aligned to saccade onset. The two rows correspond to RFs centered on the starting (RF 1) or ending fixation (RF 2) of the saccades. **g**, Quantification of RF presence over time and coordinate frames (color indicates RF 1, RF 2, or the counterfactual midpoint control; see text), separately for visual areas. Horizontal bars indicate statistically significant differences from the midpoint control ($p < 0.01$, permutation test, FDR-corrected). In **b** and **g**, lines and shading indicate hierarchical mean ± bootstrap 95%-CI, first over neurons, then over monkeys. The n values indicate the total number of neurons / monkeys.

Stimulus-selectivity implied retinotopic receptive fields that shifted with each fixation

Insofar as the computational models could capture neuronal stimulus selectivity, they provided a means to reveal the structure of neuronal receptive fields (RFs) during natural vision. The RF corresponds to a spatiotemporal section in the visual input that explains a neuron's responses. To localize this explanatory section, we reasoned that stimulus contents within the RF should allow a model to predict neuronal responses, whereas stimulus contents outside the RF should not allow for significant prediction. Therefore, we partitioned the visual scene centered on each fixation into a grid of small, overlapping image patches 2 × 2 dva in size and separated by 1 dva intervals (**Figure 3.4d**). The small patch at a fixed offset from each fixation provided the input to a computational model for predicting neuronal responses. Each different offset corresponded to a different model that competed in predicting the same responses. The resulting map of model performance at different offsets from fixation would indicate the spatial extent of a neuron's RF. Empirically, we found it helpful to further regularize this process by sharing model coefficients across offset locations (see Methods for details), resulting in a metric reminiscent of reverse correlation. Using simulated responses, we validated that this model-based, cross-validated reverse correlation procedure for inferring RF structure could recover the location, size, and shape of ground truth RFs generating simulated responses (**Figure 3.5**).

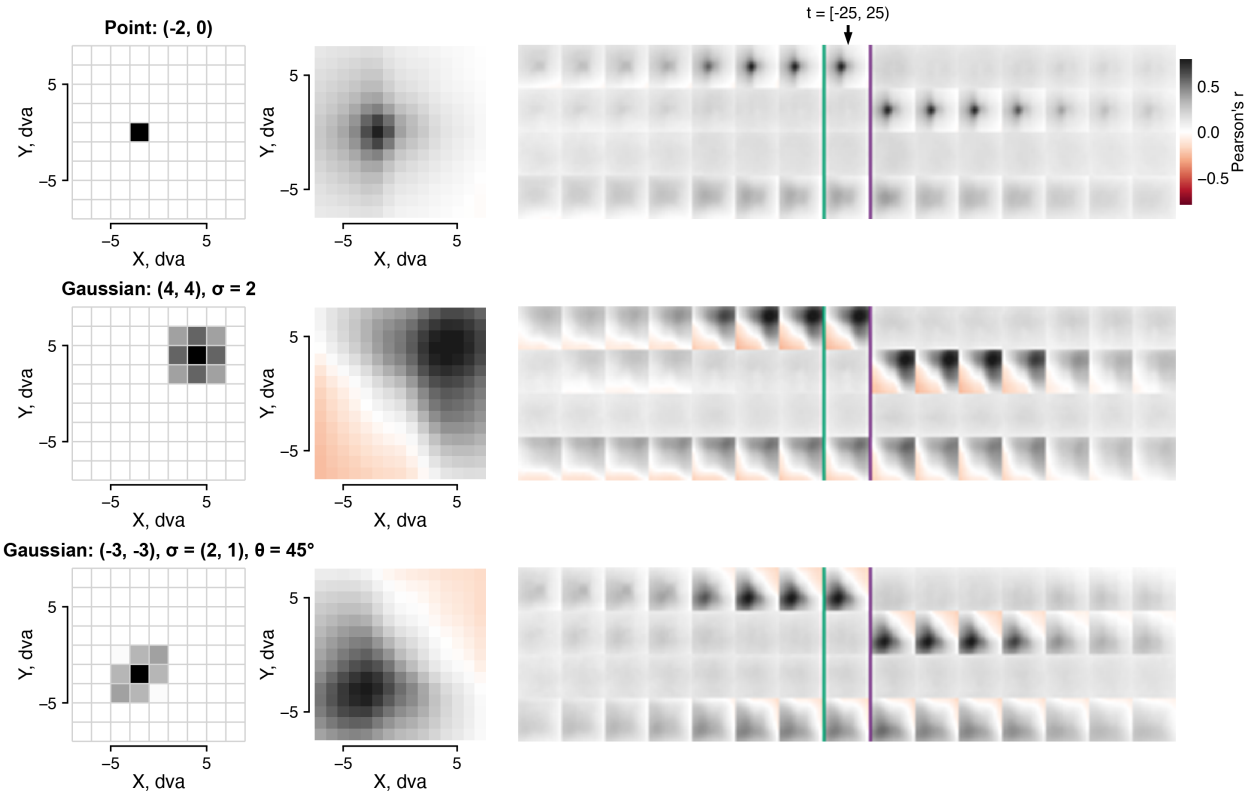

Figure 3.5. Validation of model-inferred RFs using simulated activity.
Stimulus-selective activity was simulated using the behavior data from one example session, then submitted to the model-based RF inference analysis. The simulated responses represented purely retinotopic RFs, a different one per row. The simulation used a weighted combination of the model representation of 2 × 2 image patches. The first column illustrates the image patches and weights underlying each set of simulated responses. Response simulation and RF inference used different models (ResNet-50 and Vision Transformer). The responses were simulated to have zero latency relative to eye position. See Methods for details. The second column corresponds to **Figure 3.4e**. The third column corresponds to **Figure 3.4f**, with two extra rows visualizing two counterfactual controls. Row 3 corresponds to evaluating the RF at a point equidistant to the pre- and post-saccadic fixation point. Row 4 corresponds to the midpoint control described in the main text.

Figure 3.4e shows the RF inferred from responses 0–200 ms following zeroth fixations for an example neuron recorded in central IT. The map showed a focal region of high correlation; the location and extent of this region were consistent with traditionally mapped RFs from dedicated mapping experiments with this chronically recorded site. Across recording arrays, our RF inference method recovered zeroth-fixation RFs that matched traditionally mapped RFs. The model-based RF mapping procedure allowed us to examine the dynamics of RF remapping

directly during free-viewing behavior with naturalistic stimuli. We evaluated RFs for rolling response time bins aligned to saccade onset. The RF was centered alternatively on the fixation point before or after the saccade (FP 1 or FP 2). To aid in disambiguating the RF in the two coordinate frames: RF 1 or RF 2, we included only large saccades (> 5 dva) in this analysis. **Figure 3.4f** shows the two sets of inferred spatiotemporal RFs for the same example neuron in **Figure 3.4e**. When a clear RF was present, it was focal in one of the two fixation-centered coordinate frames. The RF evidently shifted from FP 1 to FP 2 in the response time bin 75–125 ms following saccade onset, consistent with the typical latency in central IT and an average saccade duration of around 50 ms.

To summarize the time course of RF shifts across sessions, we quantified the clear presence of an RF using its consistency across cross-validation splits, regularized via 2D Gaussian fits (**Figure 3.4g**; Methods). We quantified each time bin and RF coordinate frame independently to capture any potential shifts in the RF that could be non-parallel to the saccade (Neupane et al., 2016; Tolias et al., 2001; Zirnsak et al., 2014). The conclusions did not change if we quantified saccadic RFs by their consistency to RFs from the zeroth fixation. Across visual areas, RF presence was clearest in the time period corresponding to the anchoring fixation, i.e., before the saccade for RF 1 (green in **Figure 3.4g**) and after the saccade for RF 2 (purple in **Figure 3.4g**). Evidence for the two RFs crossed over at times broadly consistent with classical response latencies that increased along the processing hierarchy. Nonetheless, in some visual areas, there was an above-zero baseline before the saccade (RF 2) or after it (RF 1), reminiscent of the above-zero baseline in the self-consistency time course (**Figure 3.3**). The baseline could indicate predictive RF remapping; alternatively, the baseline could reflect correlations in stimulus features that the saccade size criterion did not eliminate. To distinguish these two possibilities, we included a

control by inferring counterfactual RFs anchored on the midpoint between FPs 1 and 2. The midpoint control would capture any stimulus features that were shared near FPs 1 and 2 as well as between them. Across visual areas, the evidence for RF 1 and RF 2 significantly exceeded the evidence for the counterfactual midpoint RF only in the time period of the corresponding fixation (**Figure 3.4g**; one-tailed permutation test results indicated as horizontal bars). Thus, our results did not indicate predictive remapping of stimulus-selective RFs along the ventral visual pathway.

Discussion

We studied neurons along the ventral visual cortex, one of two major cortical visual pathways in primates, while monkeys freely viewed natural images. We aimed to design robust analyses that could uncover the general principles governing visual responses during this unconstrained behavior. We found that neurons in the ventral visual pathway maintained their feature selectivity, yoked their responses to eye movements, and responded with classical latencies following fixation shifts. When monkeys viewed images containing faces, face neurons responded more strongly when the monkey fixated near faces than further away, and the responses refreshed after each gaze shift (**Figure 3.2**). Neurons responded consistently during fixations on similar locations (**Figure 3.3a–d**). This consistency evinced high spatial precision (**Figure 3.3g**) and classical response latencies (**Figure 3.3e, f**) that increased along the ventral processing hierarchy (**Figure 3.3h, i**). Throughout the 1.5 seconds an image remained on screen, neuronal responses remained tightly linked to the present fixation and did not become invariant to gaze (**Figure 3.3j, k**). Synthesizing these principles in computational models, we could explain a significant fraction of single-trial, stimulus-selective neuronal responses (**Figure 3.4a, b**). Using these models as a pivot, we factorized out stimulus selectivity to reveal the detailed spatiotemporal structure of neuronal receptive fields (RFs) during free viewing (**Figure 3.4c–g**). The inferred RFs

corroborated response self-consistency to show that ventral neurons responded to a local part of the retinotopic space with classical dynamics and spatial scale.

Our results accord with the textbook model of ventral visual neurons as retinotopic feature detectors. Studies have established this model through controlled experiments that deliberately limit spatial and temporal context and exclude eye movements. This simplifying model also accounted for our data during more natural viewing conditions. Indeed, there was no fundamental mismatch between our results and the conventional understanding of ventral vision as long as we invoked adaptation to explain the general drop in firing rates and selectivity during the trial (**Figure 3.1**g, **Figure 3.3k**). Prior studies reached the same conclusion but used minimalist stimuli (DiCarlo & Maunsell, 2000) or coarser analyses (Livingstone et al., 1996). Others who studied more natural viewing conditions (Gallant et al., 1998; McMahon et al., 2015; Podvalny et al., 2017; Sheinberg & Logothetis, 2001) often hint at non-classical responses but did not decisively challenge or revise the conventional view of visual processing.

We closely examined one candidate non-classical response property, that of predictive remapping. Predictive remapping refers to the anticipatory updating of retinotopic receptive fields before the eye moves. Studies have reported predictive RFs primarily in dorsal cortical and subcortical areas such as the lateral intraparietal area (Duhamel et al., 1992), frontal eye field (Umeno & Goldberg, 1997; Zirnsak et al., 2014), superior colliculus (Walker et al., 1995), area MST (Inaba & Kawano, 2014), and area V3A (Nakamura & Colby, 2002). By comparison, evidence for predictive remapping is limited in the ventral visual cortex. Primarily, studies in V4, a mid-tier region in the ventral stream, have reported transient non-retinotopic visual responses during eye movements consistent with predictive remapping (Neupane et al., 2016; Tolias et al., 2001). These findings appear to conflict with our conclusions that there was no evidence of

anticipatory remapping of feature-selective responses throughout the ventral visual pathway. However, we think our results are reconcilable with previous evidence in V4. Our study differs in several aspects from studies reporting remapping in V4. Neupane et al. (2016) used a flashed probe—a square light spot on a black background. Tolias et al. (2001) used a persistent, cell-specific stimulus such as an oriented bar. Both studies showed the probe on an otherwise empty display and varied the stimulus location but not the identity. Instead, we studied vision in dense, persistent visual scenes that continuously stimulated neurons in their classical RFs. Thus, we quantified how neuronal responses corresponded to stimulus content instead of evaluating the presence or absence of responses as a function of stimulus location. It is unknown whether the predictively shifting RFs in previous studies show the same feature selectivity as the classical RF. Alternatively, shifting RFs could solely indicate the presence of a salient stimulus independent from its content, consistent with their interpretation as a remapped attention pointer (Rolfs, 2015). Attentional remapping could explain why superior collicular remapping diminishes when probed with multiple simultaneous stimuli (Churan et al., 2011), a stimulation condition closer to free viewing of dense natural images. The potential interplay among attention, selectivity, and stimulation conditions suggests caution when extrapolating from highly controlled experiments to explaining vision during natural behavior.

Indeed, there is no reason to expect that seeing involves a detailed, stable map of visual objects. Conscious perception probably does not include a veridical image of the world, with idioms like ‘out of sight, out of mind’ giving voice to this near-truism. Empirical support for the absence of perfect visual stability comes from findings of change or inattentional blindness (Mack & Rock, 1998), memoryless visual search (Horowitz & Wolfe, 1998), and perisaccadic mislocalization (Matin & Pearce, 1965; Ross et al., 2001). Across saccades, people can miss even

relatively large object displacements (Bridgeman et al., 1975). However, behaviors like the multi-step saccade task (Hallett & Lightstone, 1976) require the subject to maintain some information across eye movements. While the nature and content of stable visual perception are unclear, existing evidence suggests that the brain may not stabilize the rich representations of ventral vision but only maintains some heuristics necessary for performing actions.

Our analyses have focused on visually driven, stimulus-selective activity. As quantified by the average self-consistency for responses in 50 ms bins, the visual stimulus only accounted for Pearson's r between 0.1 and 0.2, or 1–4 % of the variance across trials. These numbers may seem small, but to fairly compare them to other studies, it is necessary to control for trial-averaging and response window sizes. More averaging leads to monotonically increasing self-consistency for both averaging across trials and, to an extent, accumulating responses in wider time bins. Preliminary analysis controlling for these variables shows that our results are in line with the self-consistency values reported in another study recording multiunit extracellular activity in macaque V4 during passive viewing (Bashivan et al., 2019) and a published dataset of single-cell calcium imaging responses to natural images in mouse V1 (Stringer et al., 2019). Both studies still fall short of the ideals of perfectly controlled retinal stimulation and veridical recording of single neuron spiking activity. It is possible to do both with existing techniques. Adjusting stimulus position in a closed loop with eye tracking can help closely control retinal stimulation (DiCarlo & Maunsell, 2000); alternatively, downstream analysis can compensate for the eye position (Yates et al., 2021). With high-density CMOS probes such as Neuropixels (Steinmetz et al., 2018), it is possible to detect all action potentials of a well-covered single neuron within a session. Thus, it is now practical to consider what we can hope to explain as the ceiling of neuron response variability for a reliable and objective yardstick of how much progress we have made

with quantitative theories of visual neurons. Studying single-trial responses is essential to reveal potential vision-orthogonal response dimensions, as trial-averaging implicitly removes all variability orthogonal to the stimulus identity that defines repeat trials. Instead of noise, this stimulus-orthogonal variability may encode unexplained variables. Indeed, a recent study found orthogonal dimensions of non-visual activity that multiplexed with visual activity in mouse V1 and correlated with uninstructed facial movements (Stringer et al., 2019). It will be interesting to examine if the monkey visual cortex similarly codes non-visual variables. More tentatively, a fuller consideration of non-visual responses in the visual cortex may lead to a more general understanding of the broader functions of neurons here. This general understanding may help unify existing data attributed to overloaded terms like attention and task demands.

Finally, I wish to underscore the value of studying the brain during natural behavior. Reductionist experiments can isolate and illuminate the mechanisms of cognition as long as the isolated facets reflect how the brain operates during natural behavior. The brain evolved for behavior, with which neuroscience should start and end (Krakauer et al., 2017; Leopold & Park, 2020). Natural behavior is a source for generating hypotheses and should be the final test for principles gleaned from controlled experiments. As Mark Twain famously wrote, 'Truth is stranger than fiction.' The coping strategies evolution discovers can often defy the neat theories preferred by the human mind. Instead, the brain's rich natural activity may provide a starting point for bootstrapping answers to the twin questions: What problems is the brain solving? And, How? One informative analysis might be quantifying the 'reliable information' shared by a neural population (Stringer et al., 2019), definable even without measuring behavior. I speculate that it would be an informative line of inquiry to first identify shared patterns of activity in large populations of (visual) neurons during (visual) behaviors and to subsequently find the grounding of

those activity patterns in external observables and eventually in a computational problem (Marr, 1982). It may be debatable whether free viewing without an explicit task represents natural, everyday vision, but it at least forms a component of broader natural behaviors. I have developed flexible analyses that can be applied to unravel visual response properties across brain areas during natural behaviors. Modern large-scale recording data may help delineate the even broader context of this visual activity.

Methods

Experimental subjects

All procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee and conformed to NIH guidelines provided in the Guide for the Care and Use of Laboratory Animals. Eleven adult *Macaca mulatta* (one female, ten males; 5–13 kg; 2–17 years old) and two adult male *Macaca nemestrina* (13 and 15 kg; 12 and 14 years old) were socially housed in standard quad cages on 12/12 hour light/dark cycles.

Surgical procedures

Animals were implanted with custom-made titanium or plastic headposts before fixation training. After several weeks of fixation training, the animals underwent secondary surgeries for array implantation. All surgeries were done under full surgical anesthesia using sterile technique.

Physiological recording

Animals were implanted with custom floating microelectrode arrays (32 channels, MicroProbes, Gaithersburg, MD or 128 channels, NeuroNexus, Ann Arbor, MI) or microwire bundles (64 channels; MicroProbes). Each animal received 1–5 arrays throughout data collection spanning three years. Neural signals were amplified and sampled at 40 kHz using a data acquisition system

(OmniPlex, Plexon, Dallas, TX). Multi-unit spiking activity was detected using a threshold-crossing criterion. Channels containing separable waveforms were sorted online using a template-matching algorithm.

Behavioral task

Monkeys performed a free viewing task with a range of parameters. Images were typically presented at a size of 16 × 16 degrees of visual angle, but some experiments used other sizes ranging from 8 × 8 to 26 × 26 dva. Most experiments used a 1.5 s presentation duration, but some experiments used other durations ranging from 0.3 to 60 s. Image presentation was pseudorandomly ordered in a block design such that images were repeated when all images had been shown once. Image position was randomly shifted each presentation to encourage free looking because most monkeys have been extensively trained to fixate. Monkeys were rewarded at random intervals with a drop of juice for maintaining their gaze within a window around the image. Task control was handled by a MATLAB-based toolbox, NIMH MonkeyLogic (Hwang et al., 2019). The task-control software monitored and recorded eye position from an infrared eye tracker (ISCAN, Woburn, MA, or EyeLink, Ottawa, Canada). Eye tracking was calibrated before each session using a projective transform.

Data preprocessing

Neural recording was synchronized by TTL events to task control and behavior data including eye tracking. Stimulus onset times were refined using a photodiode signal. We measured and corrected for a fixed latency per tracker in the eye signal. Using a motorized model eye attached to a potentiometer, we compared the eye position signal recorded as a voltage trace alongside neural data to the synchronized behavior data stream to obtain a constant offset. This constant offset

was applied before downstream analysis. Fixations and saccades were detected using ClusterFix (König & Buffalo, 2014) with default parameters.

Neuron selection

Because we recorded from chronic multi-electrode arrays, not all channels contained visually responsive signals. We elected to include only visually selective units in all analyses. Operationally, we defined visually selective units as those that passed a threshold of r = 0.1 on return fixation self-consistency with a 250 ms response window based on either previous or current return; see the section, *Self-consistency metrics* below for how this metric was defined. This criterion included 48 ± 23% of units (mean ± stdev.).

Fixation selection

Fixations were included in analyses based on the following criteria: 1) the fixation lasted at least 100 ms; 2) the fixation landed within the image.

Face-specific analysis

Face regions of interest were either manually drawn for image sets containing both monkey and human faces or detected as bounding boxes using a pre-trained face-detection neural network (RetinaFace) (Deng et al., 2020) for image sets containing human faces only. Fixations were classified as face fixations if they landed within 5 dva of a face ROI and nonface otherwise. To match the face and nonface conditions more closely, we only considered nonface fixations on images that contained a face ROI. Face-selectivity index (FSI) was calculated using the responses 0–200 ms from stimulus onset ('fixation 0') or fixation onset ('fixation 1+') as (a – b) / (a + b), where a and b corresponds to face and nonface fixation responses, respectively. We restricted the face-specific analysis to three monkeys with face-patch arrays and only those units

with an FSI > 1/3 based on fixation 0 responses. Responses were normalized per neuron to the 95th percentile across fixation types (0 or 1+), categories (face or nonface), and time points.

Return fixation self-consistency

In standard controlled viewing experiments, self-consistency is typically defined using the split-half correlation of trial-averaged responses conditioned on stimulus identity. We generalized this metric to free viewing. First, we relaxed trial-averaging by considering 'single-trial', i.e., single-fixation, responses. Second, we considered return fixations to be analogous to repeated presentations. In the 'default' setting, all fixations on the same image in each session were paired based on a proximity threshold, typically 1 dva. Two adjacent fixations tended to be close by (**Figure 3.1**e), leading to nearby locations in the non-paired fixation. To reduce this correlation in the non-paired fixation (i.e., the preceding fixation when the 'current' fixation was paired and the subsequent fixation when the 'previous' fixation was paired), we further sub-selected return fixation pairs where the non-paired fixation was separated by > 4 dva. Neuronal responses were aligned to fixation onset and paired based on the corresponding fixations. Responses were calculated using either the 250 ms before and after fixation onset or using rolling 50 ms time bins with 25 ms steps. Self-consistency was quantified by Pearson's r.

In Figure 3k, we computed variants of self-consistency conditioned on different notions of repeat trials. 'Same image' self-consistency compared all fixation pairs within the same image, across images, regardless of fixation location. 'Distant fixation' self-consistency compared all fixation pairs that were > 8 dva apart. In this analysis, responses were aligned to stimulus rather than fixation onset and computed in 250 ms time bins with 125 ms steps. Every fixation was assigned by its onset time to one time bin; a time bin may contain multiple fixations. Two time bins were compared if they contained any fixations that matched a fixation pairing rule.

Estimates of response latency

Fixation response latency was estimated as the first time point on either side of time 0 that the 'current' return fixation self-consistency exceeded the 'previous,' both using decorrelated return fixation pairs. To estimate stimulus-onset response latency, we elected to use self-consistency instead of the traditional average firing rate to be more comparable to the fixation-onset latency. Self-consistency captured stimulus selectivity, whereas average firing rate may include non-specific onset transients and indeed resulted in systematically lower latency estimates. We calculated return fixation self-consistency using zeroth fixations exclusively. Response latency was estimated as the closest time point on either side of time 0 that the self-consistency time course crossed from below to above a threshold, which was in turn based on the average of the 2.5^{th} and 97.5^{th} percentiles of the self-consistency time course. For all self-consistency metrics, bootstrap estimates were obtained by sampling fixation pairs with replacement. We estimated latency separately for each self-consistency bootstrap sample to obtain a bootstrap distribution of latency estimates. In Figure 3k, we included only units that had both well-estimated latencies based on several quality control criteria: 1) bootstrap stdev. < 25 ms; 2) bootstrap bias < 12.5 ms; 3) peak self-consistency > 0.1; 4) crossing point was unique for 100 ms on either side.

Encoding model of neuronal responses

Image patches were embedded in model features space using a pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2020). We used the model instance, 'vit_large_patch16_384' in the Python library 'pytorch image models' (Wightman, 2019) and features extracted from the layer, 'blocks.10.norm1'. To efficiently fit encoding models, we pre-calculated and cached the model representation of images using a discrete sampling grid, either 4 × 4 dva patches in 1 dva steps for fixation-centered models (Figure 4a–b), or 2 × 2 dva patches in 0.5 dva steps for inferring

receptive fields (Figure 4c–f). Patches extending beyond the image extent were padded with gray. The feature embedding was average over the sequence dimension in the ViT layer to result in a 1024-dimensional feature vector. Each fixation was digitized into the closest bin to obtain the feature vector of the corresponding image patch. Responses were calculated per fixation using a 50 ms window aligned to fixation onset. A Ridge regression model was fit to map between fixation-aligned model features and neuronal responses, with the regularization parameter alpha = 5500. The linear mappings were fitted and evaluated using five-fold cross-validation across images so that no return fixations were present in both the training and testing sets.

Model-based inference of receptive field structure

At each fixation, a grid of 2 × 2 dva image patches was extracted on a fixation-anchored grid of offset locations, from -7 to 7 dva in 1 dva steps. Each image patch was converted into a 1024-D model embedding vector, resulting in a 15 × 15 × 1024 retinotopic stimulus representation akin to a multichannel image. At each of the 15 × 15 = 225 offset locations, a separate encoding model was fit and evaluated across fixations and cross-validated over images as described above. This process resulted in a 15 × 15 × 5 map of model performance per cross-validation (CV) split.

To further regularize this map, we took the 1024-D model weights from the location of peak performance per CV split. The model weights were applied to held-out fixations to project each 15 × 15 × 1024 fixation-centered stimulus representation to a 15 × 15 scalar map, akin to a grayscale image, that was private to each neuron, response window, and CV split. These scalar maps were correlated to fixation-aligned neuronal responses, a process analogous to reverse correlation. The correlation was performed either across CV splits to obtain a single map per condition, or within each split for later comparison across splits as described below. Maps were

clipped at 0 because negative correlation indicated over-fitting; and squared because it resulted in better Gaussian fits (described below).

To quantify the clear and consistent presence of receptive fields (RFs), we fitted a Gaussian distribution to the inferred RF per CV split, then evaluated the goodness-of-fit on the 15 × 15 RF map from other splits. Goodness-of-fit was quantified with Pearson's r and averaged over 5 × 4 pairs of splits (the pairs were directional because only one split contributed to the Gaussian fit).

The process above inferred the spatial structure of an RF and quantified its clear presence separately for each neuron and response time window. This process was repeated over neurons, separately per response window to allow for potential changes in stimulus selectivity and RF structure across time. In Figure 4d, a fixational RF was inferred for the response window 0–200 ms following fixation onset. In Figure 4e–f, time-resolved RFs were inferred for responses aligned to saccade onset in 50 ms rolling response windows from -375–375 ms in 25 ms steps.

For saccade-aligned RFs, the above process was repeated for two retinotopic coordinate frames, anchored on the fixation point either before (F1) or after (F2) the saccade. Saccades were selected that were at least 5 dva in size to reduce stimulus feature correlation. Still, there may be residual spurious correlations due to finite saccade sizes and autocorrelations in natural images. To quantify the empirical baseline RF, we calculated, as a control, a third set of RFs anchored on the midpoint passed by the saccade.

The process above specifies the quantification described in the main text. In the supplementary material, we show implementation detail variants, which do not change the main text conclusions.

Simulation of responses representing ground-truth RFs

Each simulated RF was discretized into one or more offset locations in 2 dva steps, to be indexed into corresponding 2 × 2 image patches aligned to each eye position sample. Offset locations were assigned weights based on a Gaussian decay profile truncated at $\sigma = 2$. Responses were simulated for each eye position sample at its native 1 kHz sampling rate, although downstream analysis would bin responses into 50 ms time bins. A simulated response sample was the weighted sum of the model representations of image patches. No stochasticity was added. The simulated responses were entered into the same analysis pipeline as described above for real data. To prevent trivial generalization from the neural network representations (ViT) underlying the RF inference analysis, the simulated responses were based on model embeddings in ResNet-50 (He et al., 2016) (implementation and ImageNet-pretrained weights from the Python library 'torchvision') at the layer 'layer3.15.bn2'.

Average estimates and statistical tests

In **Figure 3.2c** and **e**, normalized responses were averaged over neurons. In **Figure 3.1c–e** and **g**; **Figure 3.3f–h** and **k**; and **Figure 3.4b** and **f**, the metrics were first averaged over neurons per monkey, then reported as mean ± stdev. over monkeys. The statistical tests in **Figure 3.2c**, **Figure 3.2e**, and **Figure 3.4f** were one-tailed paired Wilcoxon tests across neurons. P-values were corrected to control the false discovery rate at 0.01 using the two-stage Benjamini-Krieger-Yekutieli procedure.

Bibliography

Abbasi-Asl, R., Chen, Y., Bloniarz, A., Oliver, M., Willmore, B. D., Gallant, J. L., & Yu, B. (2018). The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *BioRxiv*, 465534.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103-108.

Bardon, A., Xiao, W., Ponce, C. R., Livingstone, M. S., & Kreiman, G. (2022). Face neurons encode nonsemantic features. *Proceedings of the national academy of sciences*, *119*(16), e2118705119.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491-507.

Bridgeman, B., Hendry, D., & Stark, L. (1975). Failure to detect displacement of the visual world during saccadic eye movements. *Vision research*, *15*(6), 719-722.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Campbell, F., Cooper, G. F., & Enroth-Cugell, C. (1969). The spatial selectivity of the visual cells of the cat. *The Journal of Physiology*, *203*(1), 223.

Carlson, E. T., Rasquinha, R. J., Zhang, K., & Connor, C. E. (2011). A sparse object coding scheme in area V4. *Current Biology*, *21*(4), 288-293.

Cavanagh, P., Hunt, A. R., Afraz, A., & Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in cognitive sciences*, *14*(4), 147-153.

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, *169*(6), 1013-1028. e1014.

Churan, J., Guitton, D., & Pack, C. C. (2011). Context dependence of receptive field remapping in superior colliculus. *Journal of Neurophysiology*, *106*(4), 1862-1874.

Conway, B. R., Hubel, D. H., & Livingstone, M. S. (2002). Color contrast in macaque V1. *Cerebral cortex*, *12*(9), 915-925.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision research*, *22*(5), 545-559.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition, Miami, FL.

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Virtual.

Denton, E. L., Chintala, S., & Fergus, R. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems*, *28*.

Desimone, R., Albright, T., Gross, C., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *The Journal of Neuroscience*, *4*(8), 2051-2062.

DiCarlo, J. J., & Maunsell, J. H. (2000). Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. *Nature neuroscience*, *3*(8), 814-821.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415-434.

Dorkenwald, S., McKellar, C. E., Macrina, T., Kemnitz, N., Lee, K., Lu, R., Wu, J., Popovych, S., Mitchell, E., & Nehoran, B. (2022). FlyWire: online community for whole-brain connectomics. *Nature Methods*, *19*(1), 119-128.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, *29*.

Duhamel, J.-R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, *255*(5040), 90-92.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845-851.

Gallant, J. L., Connor, C. E., & Van Essen, D. C. (1998). Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport*, *9*(7), 1673-1678.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350-363.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Gross, C. G. (1994). How inferior temporal cortex became a visual area. *Cerebral cortex*, *4*(5), 455-469.

Gross, C. G. (2002). Genealogy of the "grandmother cell". *The Neuroscientist*, *8*(5), 512-518.

Gross, C. G., Bender, D. B., & Gerstein, G. L. (1979). Activity of inferior temporal neurons in behaving monkeys. *Neuropsychologia*, *17*(2), 215-229.

Gross, C. G., & Mishkin, M. (1977). The neural basis of stimulus equivalence across retinal translation. *Lateralization in the nervous system*, 109-122.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829-842.

Hallett, P. E., & Lightstone, A. (1976). Saccadic eye movements to flashed targets. *Vision research*, *16*(1), 107-114.

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary computation*, *11*(1), 1-18.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *394*(6693), 575-577.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106-154.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215-243.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science*, *310*(5749), 863-866.

Hwang, J., Mitz, A. R., & Murray, E. A. (2019). NIMH MonkeyLogic: Behavioral control and data acquisition in MATLAB. *Journal of neuroscience methods*, *323*, 13-21.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, *32*.

Inaba, N., & Kawano, K. (2014). Neurons in cortical area MST remap the memory trace of visual motion across saccadic eye movements. *Proceedings of the national academy of sciences*, *111*(21), 7825-7830.

Issa, E. B., & DiCarlo, J. J. (2012). Precedence of the eye region in neural processing of faces. *Journal of Neuroscience*, *32*(47), 16666-16682.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM international conference on Multimedia,

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, *107*(25), 11163-11170.

Killian, N. J., Jutras, M. J., & Buffalo, E. A. (2012). A map of visual space in the primate entorhinal cortex. *Nature*, *491*(7426), 761-764.

Klindt, D., Ecker, A. S., Euler, T., & Bethge, M. (2017). Neural system identification for large populations separating "what" and "where". *Advances in neural information processing systems*, *30*.

Klüver, H. (1951). *Functional Differences Between the Occipital and Temporal Lobes: With Special Reference to the Interrelations of Behavior and Extracerebral Mechanisms*. John Wiley & Sons.

König, S. D., & Buffalo, E. A. (2014). A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *Journal of neuroscience methods*, *227*, 121-131.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of experimental Psychology: general*, *139*(3), 558.

Konorski, J. (1967). *Integrative activity of the brain; an interdisciplinary approach*. University of Chicago Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Leopold, D. A., & Park, S. H. (2020). Studying the visual brain in its natural rhythm. *Neuroimage*, *216*, 116790.

Livingstone, M., Freeman, D., & Hubel, D. (1996). Visual responses in V1 of freely viewing monkeys. Cold Spring Harbor Symposia on Quantitative Biology,

Loshchilov, I. (2014). A computationally efficient limited memory CMA-ES for large scale optimization. Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation,

Lucid Authors, T. (2019, March 19, 2021). *Lucid*. GitHub. Retrieved July 31, 2020 from https://github.com/tensorflow/lucid

Mack, A., & Rock, I. (1998). Inattentional blindness: Perception without attention. *Visual attention*, *8*, 55-76.

Malakhova, K. (2018). Visualization of information encoded by neurons in the higher-level areas of the visual system. *Journal of Optical Technology*, *85*(8), 494-498.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

Matin, L., & Pearce, D. G. (1965). Visual perception of direction for stimuli flashed during voluntary saccadic eye movements. *Science*, *148*(3676), 1485-1488.

Matin, L., Picoult, E., Stevens, J. K., Edwards Jr, M. W., Young, D., & MacArthur, R. (1982). Oculoparalytic illusion: Visual-field dependent spatial mislocalizations by humans partially paralyzed with curare. *Science*, *216*(4542), 198-201.

McMahon, D. B., Bondar, I. V., Afuwape, O. A., Ide, D. C., & Leopold, D. A. (2014). One month in the life of a neuron: longitudinal single-unit electrophysiology in the monkey visual system. *Journal of Neurophysiology*, *112*(7), 1748-1762.

McMahon, D. B., Russ, B. E., Elnaiem, H. D., Kurnikova, A. I., & Leopold, D. A. (2015). Single-unit activity during natural vision: diversity, consistency, and spatial sensitivity among AF face patch neurons. *Journal of Neuroscience*, *35*(14), 5537-5548.

Melcher, D. (2011). Visual stability. *366*(1564), 468-475.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Mishkin, M., & Pribram, K. H. (1954). Visual discrimination performance following partial ablations of the temporal lobe: I. Ventral vs. lateral. *Journal of comparative and physiological psychology*, *47*(1), 14.

Mitchell, J. F., Reynolds, J. H., & Miller, C. T. (2014). Active vision in marmosets: a model system for visual neuroscience. *Journal of Neuroscience*, *34*(4), 1183-1194.

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*, *320*(5881), 1355-1359.

Mordvintsev, A., Olah, C., & Tyka, M. (2015, July 13). *Inceptionism: Going deeper into neural networks*. Retrieved Janurary 15, 2023 from https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Nakamura, K., & Colby, C. L. (2002). Updating of the visual representation in monkey striate and extrastriate cortex during saccades. *Proceedings of the national academy of sciences*, *99*(6), 4026-4031.

Neupane, S., Guitton, D., & Pack, C. C. (2016). Two distinct types of remapping in primate cortical area V4. *Nature communications*, *7*(1), 1-11.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, *29*.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding V1? *Neural computation*, *17*(8), 1665-1699.

Pasupathy, A., & Connor, C. E. (1999). Responses to Contour Features in Macaque Area V4. *Journal of Neurophysiology*, *82*(5), 2490-2502.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Podvalny, E., Yeagle, E., Mégevand, P., Sarid, N., Harel, M., Chechik, G., Mehta, A. D., & Malach, R. (2017). Invariant temporal dynamics underlie perceptual stability in human visual cortex. *Current Biology*, *27*(2), 155-165.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999-1009. e1010.

Porada, I., Bondar, I., Spatz, W., & Krüger, J. (2000). Rabbit and monkey visual cortex: more than a year of recording with up to 64 microelectrodes. *Journal of neuroscience methods*, *95*(1), 13-28.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*(1), 49-70.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019-1025.

Rolfs, M. (2015). Attention in active vision: A perspective on perceptual continuity across saccades. *Perception*, *44*(8-9), 900-919.

Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *Journal of Neuroscience*, *23*(1), 339-348.

Ross, J., Morrone, M. C., Goldberg, M. E., & Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends in neurosciences*, *24*(2), 113-121.

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature neuroscience*, *8*(12), 1647-1650.

Salle, A., Idiart, M., & Villavicencio, A. (2016). Matrix factorization using window sampling and negative sampling for improved word representations. *arXiv preprint arXiv:1606.00819*.

Salle, A., & Villavicencio, A. (2018). Incorporating subword information into matrix factorization word embeddings. *arXiv preprint arXiv:1805.03710*.

Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S.-y., Hayworth, K. J., Huang, G. B., Shinomiya, K., Maitlin-Shepard, J., & Berg, S. (2020). A connectome and analysis of the adult Drosophila central brain. *Elife*, *9*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., & Geiger, F. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference, Austin, TX.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, *21*(4), 1340-1350.

Solomon, S. G., & Kohn, A. (2014). Moving sensory adaptation beyond suppressive effects in single neurons. *Current Biology*, *24*(20), R1012-R1022.

Steinmetz, N. A., Koch, C., Harris, K. D., & Carandini, M. (2018). Challenges and opportunities for large-scale electrophysiology with Neuropixels probes. *Current opinion in neurobiology*, *50*, 92-100.

Stevens, J. K., Emerson, R. C., Gerstein, G. L., Kallos, T., Neufeld, G. R., Nichols, C. W., & Rosenquist, A. C. (1976). Paralysis of the awake human: visual perceptions. *Vision research*, *16*(1), 93-IN99.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, *364*(6437), eaav7893.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence, San Francisco, CA.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tolias, A. S., Moore, T., Smirnakis, S. M., Tehovnik, E. J., Siapas, A. G., & Schiller, P. H. (2001). Eye movements modulate visual receptive fields of V4 neurons. *Neuron*, *29*(3), 757-767.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670-674.

Umeno, M. M., & Goldberg, M. E. (1997). Spatial processing in the monkey frontal eye field. I. Predictive visual responses. *Journal of Neurophysiology*, *78*(3), 1373-1383.

Ungerleider, L. G., & Mishkin, M. (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of VIsual Behavior* (pp. 549-586). The MIT Press.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., & Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261-272.

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, *22*(12), 2060-2065.

Walker, M. F., Fitzgibbon, E. J., & Goldberg, M. E. (1995). Neurons in the monkey superior colliculus predict the visual result of impending saccadic eye movements. *Journal of Neurophysiology*, *73*(5), 1988-2003.

Wang, B., & Ponce, C. R. (2021). The geometry of deep generative image models and its applications. *arXiv preprint arXiv:2101.06006*.

Wang, B., & Ponce, C. R. (2022a). On the Level Sets and Invariance of Neural Tuning Landscapes. *arXiv preprint arXiv:2212.13285*.

Wang, B., & Ponce, C. R. (2022b). Tuning landscapes of the ventral stream. *Cell Reports*, *41*(6), 111595.

White, J. G., Southgate, E., Thomson, J. N., & Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci*, *314*(1165), 1-340.

Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., & Schmidhuber, J. (2014). Natural evolution strategies. *The Journal of Machine Learning Research*, *15*(1), 949-980.

Wightman, R. (2019). *PyTorch Image Models*. Retrieved November 14, 2022 from https://github.com/rwightman/pytorch-image-models

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Wurtz, R. H. (2008). Neuronal mechanisms of visual stability. *Vision research*, *48*(20), 2070-2089.

Xiao, W., Chen, H., Liao, Q., & Poggio, T. (2018). Biologically-plausible learning algorithms can scale to large datasets. *arXiv preprint arXiv:1811.03567*.

Xiao, W., & Kreiman, G. (2020). XDream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLoS computational biology*, *16*(6), e1007973.

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience*, *11*(11), 1352-1360.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619-8624.

Yates, J. L., Coop, S. H., Sarch, G. H., Wu, R.-J., Butts, D. A., Rucci, M., & Mitchell, J. F. (2021). Beyond Fixation: detailed characterization of neural selectivity in free-viewing primates. *BioRxiv*, 2021.2011.2006.467566.

Yuan, L., Xiao, W., Kreiman, G., Tay, F. E., Feng, J., & Livingstone, M. S. (2020). Adversarial images for the primate brain. *arXiv preprint arXiv:2011.05623*.

Zhang, M., Armendariz, M., Xiao, W., Rose, O., Bendtz, K., Livingstone, M., Ponce, C., & Kreiman, G. (2022). Look twice: A generalist computational model predicts return fixations across tasks and species. *PLoS computational biology*, *18*(11), e1010654.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1452-1464.

Zirnsak, M., Steinmetz, N. A., Noudoost, B., Xu, K. Z., & Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. *Nature*, *507*(7493), 504-507.