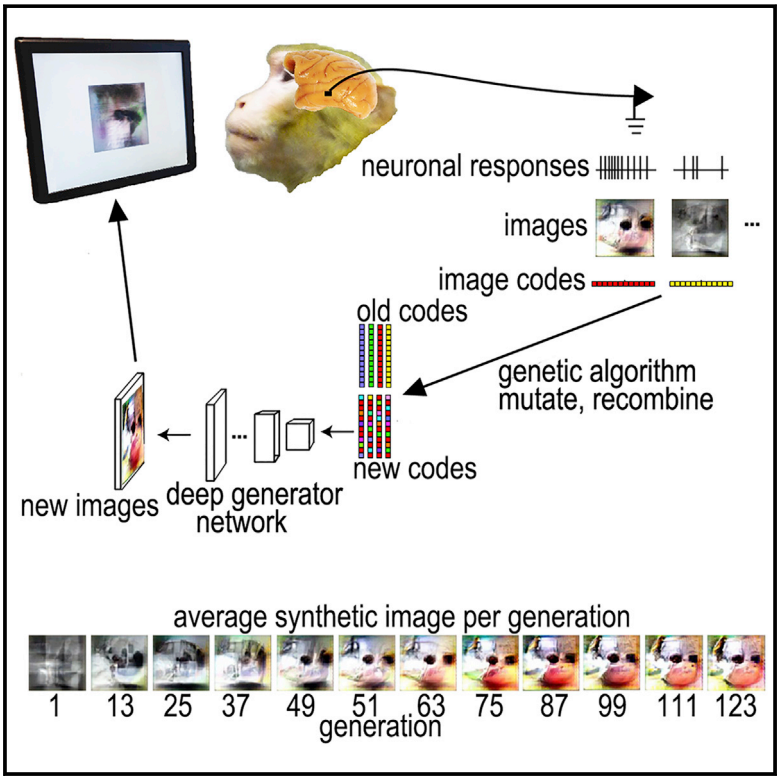
Article

Cell

Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences

Graphical Abstract



Highlights

- A generative deep neural network and a genetic algorithm evolved images guided by neuronal firing
- Evolved images maximized neuronal firing in alert macaque visual cortex
- Evolved images activated neurons more than large numbers of natural images
- Similarity to evolved images predicts response of neurons to novel images

Authors

Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, Margaret S. Livingstone

Correspondence

crponce@wustl.edu (C.R.P.), mlivingstone@hms.harvard.edu (M.S.L.)

In Brief

Neurons guided the evolution of their own best stimuli with a generative deep neural network.

Ponce et al., 2019, Cell *177*, 999–1009
May 2, 2019
https://doi.org/10.1016/j.cell.2019.04.005

CellPress

Article

Cell

Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences

Carlos R. Ponce,[1,4,5,*] Will Xiao,[2,5] Peter F. Schade,[1,5] Till S. Hartmann,[1] Gabriel Kreiman,[3] and Margaret S. Livingstone[1,6,*]
[1]Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA
[2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA
[3]Department of Ophthalmology, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA
[4]Department of Neuroscience, Washington University School of Medicine, St. Louis, MO 63110, USA
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: crponce@wustl.edu (C.R.P.), mlivingstone@hms.harvard.edu (M.S.L.)
https://doi.org/10.1016/j.cell.2019.04.005

SUMMARY

What specific features should visual neurons encode, given the infinity of real-world images and the limited number of neurons available to represent them? We investigated neuronal selectivity in monkey inferotemporal cortex via the vast hypothesis space of a generative deep neural network, avoiding assumptions about features or semantic categories. A genetic algorithm searched this space for stimuli that maximized neuronal firing. This led to the evolution of rich synthetic images of objects with complex combinations of shapes, colors, and textures, sometimes resembling animals or familiar people, other times revealing novel patterns that did not map to any clear semantic category. These results expand our conception of the dictionary of features encoded in the cortex, and the approach can potentially reveal the internal representations of any system whose input can be captured by a generative model.

INTRODUCTION

A transformative revelation in neuroscience was the realization that visual neurons respond preferentially to some stimuli over others (Hubel and Wiesel, 1962). Those findings opened the doors to investigating neural coding for myriad stimulus attributes. A central challenge in elucidating neuronal tuning in visual cortex is the impossibility of testing all stimuli. Even for a small patch of 100 × 100 pixels, there are $\sim 10^{3,010}$ possible binary images, $\sim 10^{24,082}$ grayscale images, or $\sim 10^{72,247}$ 8-bit color images. Using natural images reduces the problem, but it is still impossible to present a neuron with all possible natural stimuli. Investigators circumvent this formidable empirical challenge by using *ad hoc* hand-picked stimuli, inspired by hypotheses that particular cortical areas encode specific visual features (Felleman and Van Essen, 1987; Zeki, 1973, 1974). This approach has led to important insights through the discovery of cortical neurons that respond to stimuli with specific motion directions (Hubel, 1959), color (Michael, 1978), binocular disparity (Barlow et al., 1967), curvature (Pasupathy and Connor, 1999), and even complex natural shapes such as hands or faces (Desimone et al., 1984; Gross et al., 1972).

Despite the successes with hand-picked stimuli, the field might have missed stimulus properties that better reflect the "true" tuning of cortical neurons. A series of interesting alternative approaches have addressed this question. One approach is to start with hand-picked stimuli that elicit strong activation and systematically deform those stimuli (Chang and Tsao, 2017; Freiwald et al., 2009; Kobatake and Tanaka, 1994); this approach has revealed that neurons often tend to respond even better to distorted versions of the original stimuli (Freiwald et al., 2009; Leopold et al., 2006). Another is spike-triggered averaging of noise stimuli (Gaska et al., 1994; Jones and Palmer, 1987), but this has not yielded useful results in higher cortical areas, because it cannot capture non-linearities. An elegant alternative is to use a genetic algorithm whereby the neuron under study can itself guide its own stimulus selection. Connor and colleagues (Carlson et al., 2011; Yamane et al., 2008) pioneered this approach to study selectivity in macaque V4 and IT cortex. Our method extends and complements this approach in order to investigate the tuning properties of inferior temporal cortex (IT) neurons in macaque monkeys.

Here, we use a novel combination of a pre-trained deep generative neural network (Dosovitskiy and Brox, 2016) and a genetic algorithm to allow neuronal responses to guide the evolution of synthetic images. By training on more than one million images from ImageNet (Russakovsky et al., 2015), the generative adversarial network learns to model the statistics of natural images without merely memorizing the training set (Dosovitskiy and Brox, 2016) (Figure S1), thus representing a vast and general image space constrained only by natural image statistics. We reasoned that this would be an efficient space in which to perform the genetic algorithm, because the brain also learns from real-world images, so its preferred images are also likely to follow natural image statistics. Moreover, convolutional neural networks emulate aspects of computation along the primate ventral visual stream (Yamins et al., 2014), and this particular

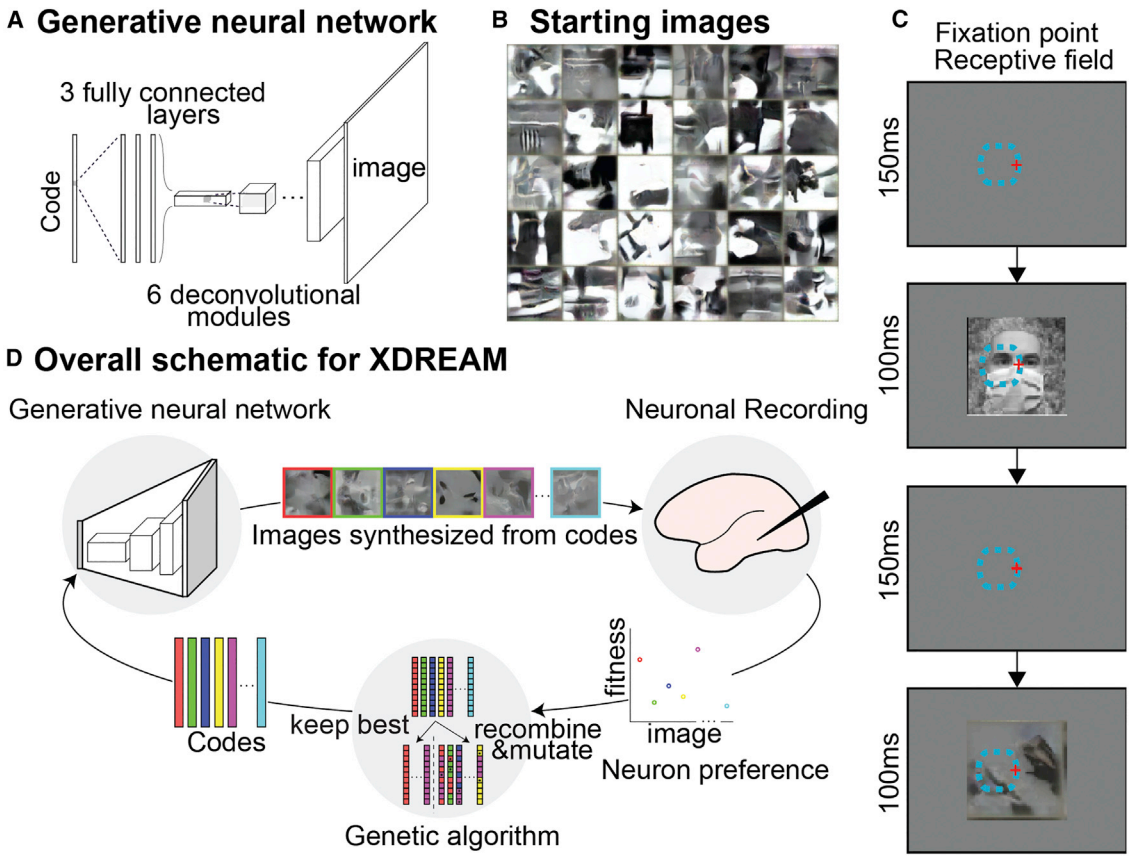


Figure 1. Synthesis of Preferred Stimuli via Neuron-Guided Evolution

(A) Generative adversarial network. Shown is the architecture of the pre-trained deep generative network (Dosovitskiy and Brox, 2016). The network comprised three fully connected layers and six deconvolutional modules.

(B) The initial synthetic images were random achromatic Portilla and Simoncelli (2000) textures; 30 examples are shown here.

(C) Behavioral task. Animals fixated within a 2° diameter window while images were presented for 100 ms followed by a 100 to 200 ms blank period. The red cross represents a fixation point; the dashed line represents population receptive field (RF).

(D) Experimental flow. Image codes were forwarded through the deep generative adversarial network to synthesize images presented to the monkey. Neuronal responses were used to rank image codes, which then underwent selection, recombination, and mutation to generate new image codes (for details see STAR Methods).

generative network has been used to synthesize images that strongly activate units in several convolutional neural networks, including ones not trained on ImageNet (Nguyen et al., 2016; Malakhova, 2018). The network takes 4,096-dimensional vectors (image codes) as input and deterministically transforms them into 256 × 256 RGB images (STAR Methods and Figure 1). In combination, a genetic algorithm used responses of neurons recorded in alert macaques to optimize image codes input to this network. Specifically, each experiment started from an initial population of 40 images created from random achromatic textures (Portilla and Simoncelli, 2000) (Figure 1B). We recorded responses of IT neurons (spike counts 70–200 ms after stimulus onset minus background) while monkeys engaged in a passive fixation task. Images subtended 3° × 3° and covered the unit's receptive field (Figure 1C). Neuronal responses to each synthetic image were used to score the image codes. In each generation, images were generated from the top 10 image codes from the previous generation, unchanged, plus 30 new image codes generated by mutation and recombination of all the codes from the preceding generation selected on the basis of firing rate (Figure 1D). This process was repeated for up to 250 generations over 1–3 h; session duration depended on the monkey's willingness to maintain fixation. To monitor changes in firing rate due to adaptation and to compare synthetic-image responses to natural-image responses, we interleaved reference images that included faces, body parts, places, and simple line drawings.

We term the overall approach XDREAM (EXtending DeepDream with Real-time Evolution for Activity Maximization in real neurons) (Figure 1D). We conducted evolution experiments on IT neurons in six monkeys: two with chronic microelectrode arrays in posterior IT (PIT) (monkeys Ri and Gu), two with chronic arrays in central IT (CIT) (monkeys Jo and Y1), one (monkey Ge) with chronic arrays in both CIT and PIT, and one with a recording chamber over CIT (monkey B3). Lastly, we validated the approach in a seventh monkey with a chronic array in primary visual cortex (V1) (monkey Vi).

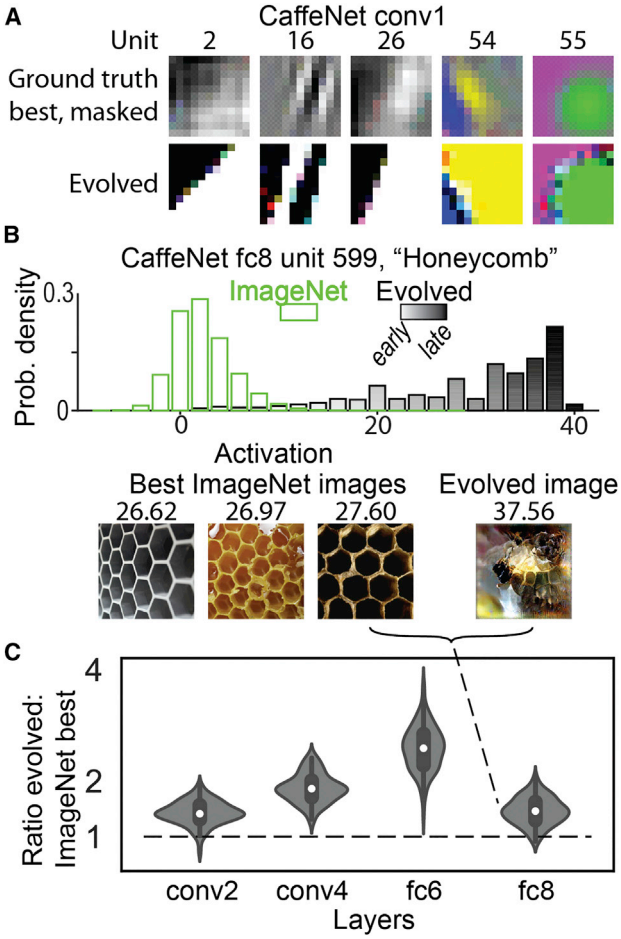


Figure 2. The XDREAM Algorithm Produced Super Stimuli for Units in CaffeNet

(A) Evolved images resembled the ground truth best images in the first layer of CaffeNet. In the ground truth best, transparency indicates the relative contribution of each pixel to the unit's activation. Only the center 11 × 11 pixels of the evolved images are shown, matching the filter size of the units.

(B and C) Most evolved images activated artificial units more strongly than all of > 1.4 million images in the ILSVRC2012 dataset (Russakovsky et al., 2015).

(B) At the top is the distribution of activations to ImageNet images and evolved images for one unit in the classification layer fc8, corresponding to the "Honeycomb" label. Grayscale gradient indicates generation of evolved images. On the bottom are the best 3 ImageNet images and one evolved image, labeled with their respective activations. In this case, the evolved image activated the unit ~1.4 × as strongly as did the best ImageNet image. (C) The violin plot shows distribution of (evolved/best in ImageNet) ratios across 4 layers in CaffeNet, 100 random units per layer. White circles indicate medians of the distributions; thick bars indicate the 25th and 75th quartiles of the distributions.

RESULTS

Evolution of Preferred Stimuli for Units in CaffeNet

We first validated XDREAM on units in an artificial neural network, as models of biological neurons. Our method generated super stimuli for units across layers in CaffeNet, a variant of AlexNet (Figure 2) (Krizhevsky et al., 2012). The evolved images were frequently better stimuli than all of > 1.4 million images, including the training set for CaffeNet, for all 4 layers that we tested (Figure 2C). For units in the first and last layers, the method produced stimuli that matched the ground-truth best stimuli in the first layer and category labels in the last layer. XDREAM is also able to recover the preferred stimuli of units constructed to have a single preferred image (Figure S1). Importantly, well-known methods for feature visualization in artificial neural networks, such as DeepDream, rely on knowledge of network weights (Erhan et al., 2009; Mordvintsev et al., 2015; Nguyen et al., 2016), whereas our approach does not, making it uniquely applicable to neuronal recordings.

Evolution of Preferred Stimuli by One Biological Neuron

We first show an example of an evolution experiment for one PIT single unit (Ri-10) in chronic-array monkey Ri. The synthetic images changed with each generation as the genetic algorithm optimized the images according to the neuron's responses (Figure 3; first half of Video S1). At the beginning of the experiment, this unit responded more strongly to the reference images (Figure S2) than to the synthetic images, but over generations, the synthetic images evolved to become more effective stimuli (Figure 4A). To quantify the change in responses over time, we fit an exponential function to the cell's mean firing rate per generation separately for the synthetic and for the reference images (solid thick lines in Figure 4A). This neuron showed an increase of 51.5 ± 5.0 (95% confidence interval [CI]) spikes per s per generation in response to the synthetic images and a decrease of −15.5 ± 3.5 spikes per s per generation to the reference images—thus the synthetic images became gradually more effective, despite the neuron's slight reduction in firing rate to the reference images, presumably due to adaptation.

We conducted independent evolution experiments with the same single unit on different days, and all final-generation synthetic images featured a brown object against a uniform background, topped by a smaller round pink and/or brown region containing several small dark spots; the object was centered toward the left half of the image, consistent with the recording site being in the right hemisphere (Figure 4B). The synthetic images generated on different days were similar by eye, but not identical, potentially due to invariance of the neuron, response variability, and/or stochastic paths explored by the algorithm in the neuron's response landscape. Regardless, given that this unit was located in PIT, just anterior to the tip of the inferior occipital sulcus and thus early in the visual hierarchy, it was impressive that it repeatedly directed the evolution of images that contained such complex motifs and that evoked such high firing rates. Two days following the evolution experiment in Figure 3, this unit was screened with 2,550 natural images, including animals, bodies, food, faces, and line drawings, plus the top synthetic images from each generation. Among the natural images this neuron responded best to monkey torsos and monkey faces. Of the 10 natural images in this set giving the largest responses, 5 were of the head and torso of a monkey (Figure 4C). The worst natural images were inanimate or rectilinear objects (Figures 4D and 4E).

Evolution of Preferred Stimuli in Other Neurons

We conducted 46 independent evolution experiments on single- and multi-unit sites in IT cortex in six different monkeys. During

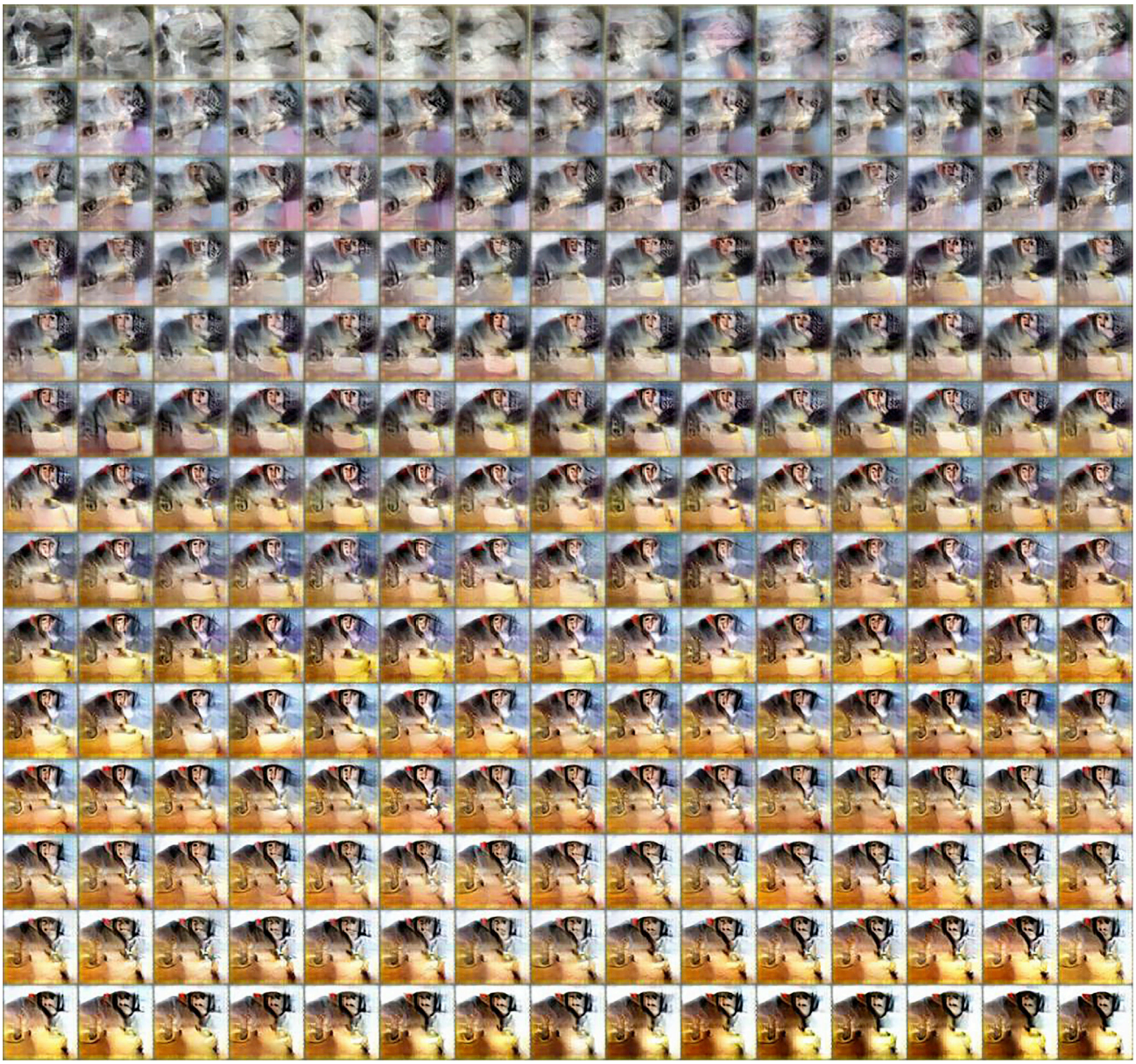


Figure 3. Evolution of Synthetic Images by a Single Monkey-Selective Neuron, Ri-10

Each image is the average of the top 5 synthetic images for each generation (ordered from left to right and top to bottom). The response of this neuron in each of these generations is shown in Figure 4A.

almost all the evolutions, the synthetic images evolved gradually to become increasingly effective stimuli. To quantify the change in stimulus effectiveness over each experiment, we fit an exponential function to the mean firing rate per generation, separately for synthetic and reference images (as in Figure 3A). Synthetic-image firing rate change over the course of each experiment was on average between 25 to 84 spikes per s for the different animals (Figure 5A); synthetic image changes were significantly different from zero in 45 out of 46 individual experiments (95% CI of amplitude estimate not including zero per bootstrap test). In contrast, responses to reference images were stable or decreased slightly across generations (reference firing rate change average for different animals ranged from −11 to 9 spikes per s; this change was significant in 15 out of 46 individual experiments) (Figure 5A and Table S1). Thus, IT neurons could consistently guide the evolution of highly effective images, despite minor adaptation. Moreover, these evolved images were often more powerful stimuli than the best natural images tested, despite the fact that the synthetic images were far from naturalistic. When comparing the cells' maximum responses to natural versus evolved images in every experiment, cells showed significant differences in 25 of 46 experiments ($p < 0.03$,

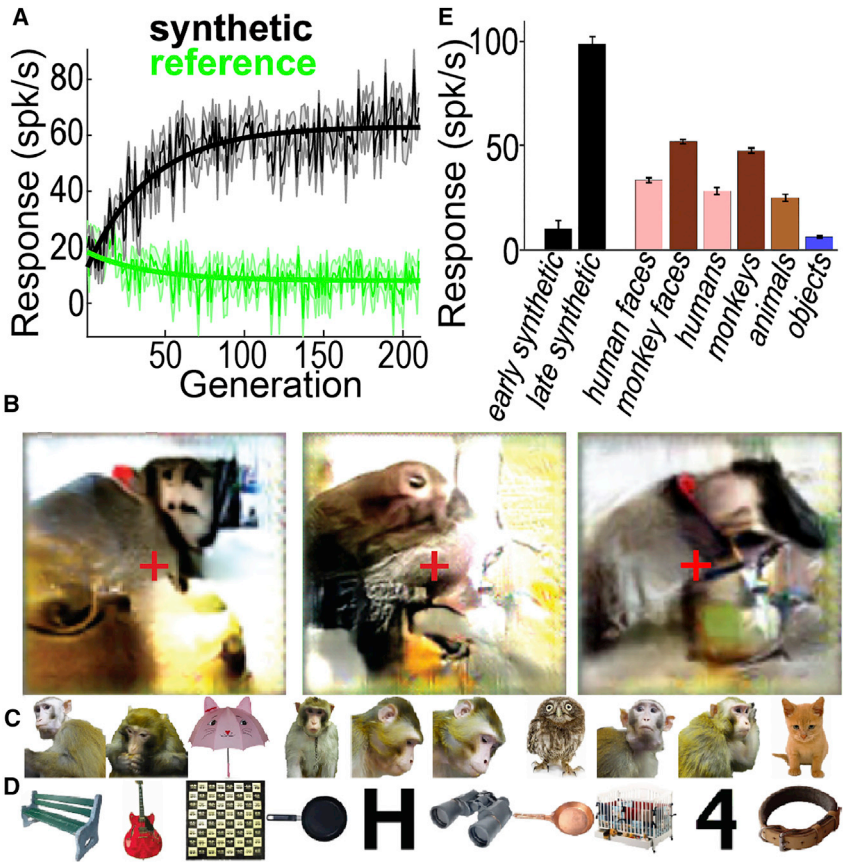


Figure 4. Evolution of Synthetic Images by Maximizing Responses of Single Neuron Ri-10, Same Unit as Figure 3

(A) Mean response to synthetic (black) and reference (green) images for every generation (spikes per s ± SEM). Solid straight lines show an exponential fit to the response over the experiment.

(B) Last-generation images evolved during three independent evolution experiments; the leftmost image corresponds to the evolution in (A); the other two evolutions were carried out on the same single unit on different days. Red crosses indicate fixation. The left half of each image corresponds to the contralateral visual field for this recording site. Each image shown here is the average of the top 5 images from the final generation.

(C–E) Selectivity of this neuron to 2,550 natural images. (C) In (C) are the top 10 images from this image set for this neuron. (D) In (D) are the worst 10 images from this image set for this neuron. The entire rank ordered natural image set is shown in Figure S2. (E) In (E) is the selectivity of this neuron to different image categories (mean ± SEM). The entire image set comprised 2,550 natural images plus selected synthetic images. Early synthetic is defined as the best image from each of the first 10 generations and late from the last 10. Each image response is the average over 10–12 repeated presentations. See Figure S3 for additional independent evolutions from this site.

permutation test after false discovery correction), and, in all but one case, the synthetic images evoked the greater response (Figure 5B; see Table S4 for further quantification of natural and evolved image responses). Figure 5C shows a histogram of response magnitudes for PIT cell Ri-10 to the top synthetic image in each of the 210 generations and responses to each of the 2,550 natural images (data for both synthetic and natural images collected 2 days later). Early generations are indicated by lighter gray and later by darker gray, so it is apparent that later generation synthetic images gave larger responses. We also illustrate one of the few experiments where natural images evoked stronger responses than did synthetic images in Figure 5D (monkey Ge, site 7), which compares the site's responses to synthetic images against 2,550 natural images. This site responded slightly better (by an average of four spikes per s, or 3.7% of its maximum rate) to images of an animal-care person who visits the animals daily, wearing our institution-specific protective mask and gown (see Figure S3 for additional independent evolutions from this site). Even in this case, one clear benefit of XDREAM is that, by coming up with effective stimuli in a manner independent from the hand-picked image set, it reveals specific features of the natural image that drove the neuron's selectivity, stripped of incidental information.

IT neurons guided the evolution of images that varied from experiment to experiment but retained consistent features for any given recording site (see Figure S4 for measure of similarity across and between sites), features that bore some similarity to each neuron's preferences in natural image sets. Figure 6 shows the final-generation evolved images from two independent evolution experiments for IT sites in five monkeys, along with each site's top 10 natural images. In each case a reproducible figure emerged in the part of the synthetic image corresponding to the contralateral visual field. In three monkeys (Ri, Gu, and Ge), response profiles to natural images indicated that the arrays were located in face-preferring regions, whereas in monkey Y1, the array was in a place-preferring region and in monkey Jo, the array was in an object-selective region. Face-selective unit Ri-23 evolved something face-like in the left (contralateral) half of the image; this is most apparent if one covers up the right (ipsilateral) half of the synthetic image. The images evolved by unit Ge-7 bore some resemblance to the unit's top natural image, a familiar person wearing protective clothing (see Figure S3 for additional independent evolutions from this site). Unit Ge-15 consistently evolved a black dot just to the left of fixation on a tan background (see Figure S3 for additional independent evolutions from this site). This unit might be similar to posterior-face-patch neurons described previously that responded optimally to a single eye in the contralateral field (Issa and DiCarlo, 2012) (see Figure S3 for additional independent evolutions from this site). Monkey-face-selective unit Ge-17 evolved a tan area with two large black dots aligned horizontally, and a light area below (see Figure S3 for additional independent evolutions from this site). Unit Jo-6 responded to various body parts and evolved something not inconsistent with a mammalian body; interestingly, a whole "body" in one

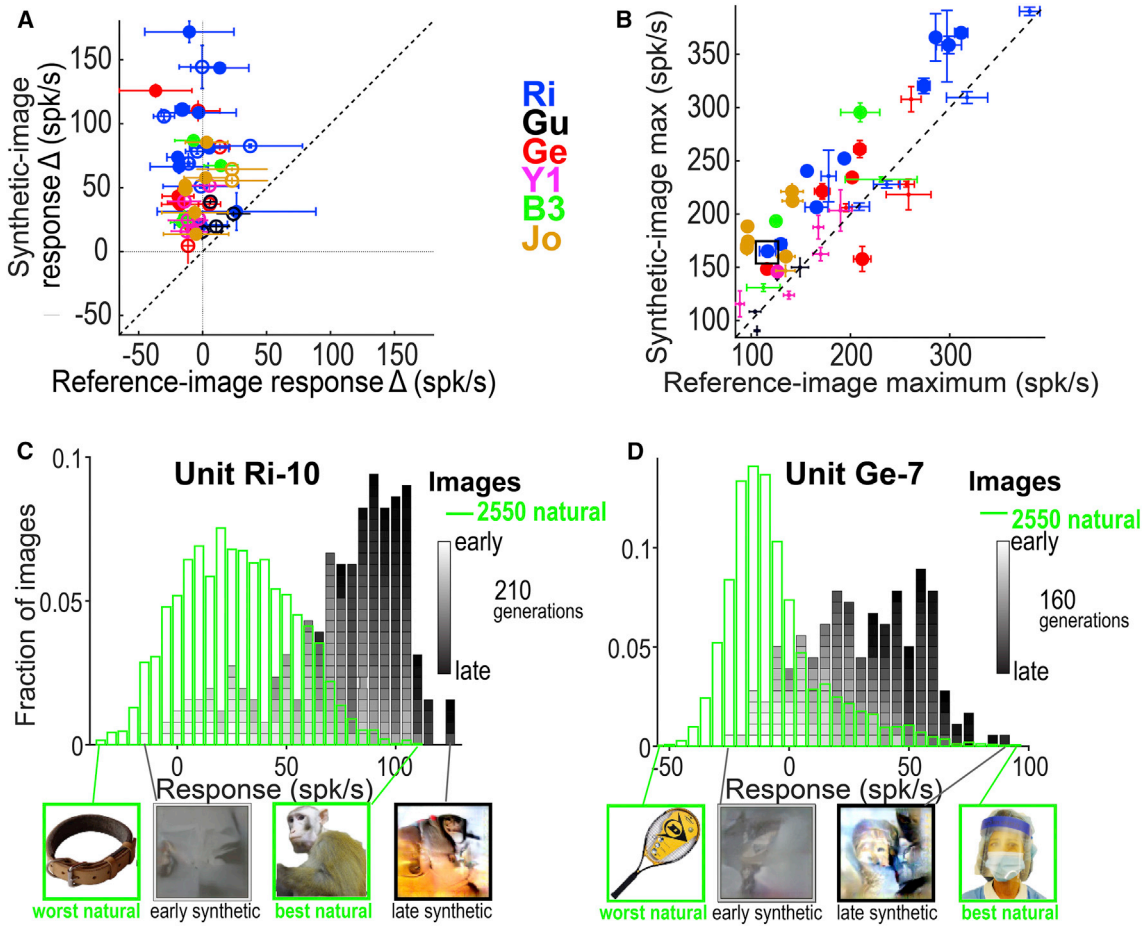


Figure 5. Evolutions for Other IT Cells

(A) Change in response to synthetic versus reference images over generations. Each point shows the mean change in firing rate to reference versus synthetic images in each experiment (change estimated by the amplitude coefficient of an exponential function fitted to the neuron's mean response per generation; error bars represent ± SEM, per bootstrap, 500 iterations of data re-sampling). Solid circles indicate single units; open circles indicate multi-units.

(B) Scatterplot of maximum responses across all images for synthetic versus reference images (measured across all generations, max ± SE per bootstrap). Colors indicate animal. The size of the circle indicates statistical significance (large circle: $p < 0.03$ after false discovery correction). Black square indicates the experiment in Figure 3.

(C) Histogram of response magnitudes to natural (green) and synthetic (gray-to-black) images for unit Ri-10 (same unit as Figures 3 and 4). Below the histogram are shown the best and worst natural and synthetic images.

(D) Color convention of (C) used for unit Ge-7. The evolution for this neuron can be seen in the second half of Video S1.

evolution and a larger partial "body" in the other. Unit Jo-5 evolved a small black square, and unit Jo-4 something black with orange below. Unit Jo-21 consistently evolved a small dark shape in the contralateral half of the image. Scene-selective unit Y1-14 evolved rectilinear shapes in the left (contralateral) field. Additional independent evolutions for some of these and other units are shown in Figure S3.

Predicting Neuronal Responses to a Novel Image from Its Similarity to the Evolved Stimuli

If these evolved images are telling us something important about the tuning properties of IT neurons, then we should be able to use them to predict neurons' responses to novel images. The deep generator network had been trained to synthesize images from their encoding in layer fc6 of AlexNet (4,096 units) (Krizhevsky et al., 2012), so we used the fc6 space to find natural images similar to the evolved images. In particular, we asked whether a neuron's response to a novel image was predicted by the distance in fc6 space between the novel image and the neuron's evolved synthetic image. To do this, we calculated the activation vectors of the evolved synthetic images in AlexNet fc6 and searched for images with similar fc6 activation vectors. We used 2 databases for this: the first comprised ~60,000 images collected in our laboratory over several years, and the second set comprised 100,300 images from the ILSVRC2012 dataset (Russakovsky et al., 2015); we included 100 randomly sampled images from each of its 1,000 categories, and two additional ImageNet categories of faces [ID n09618957], macaques [ID n02487547], and 100 images of local animal care personnel with and without personal protective garb.

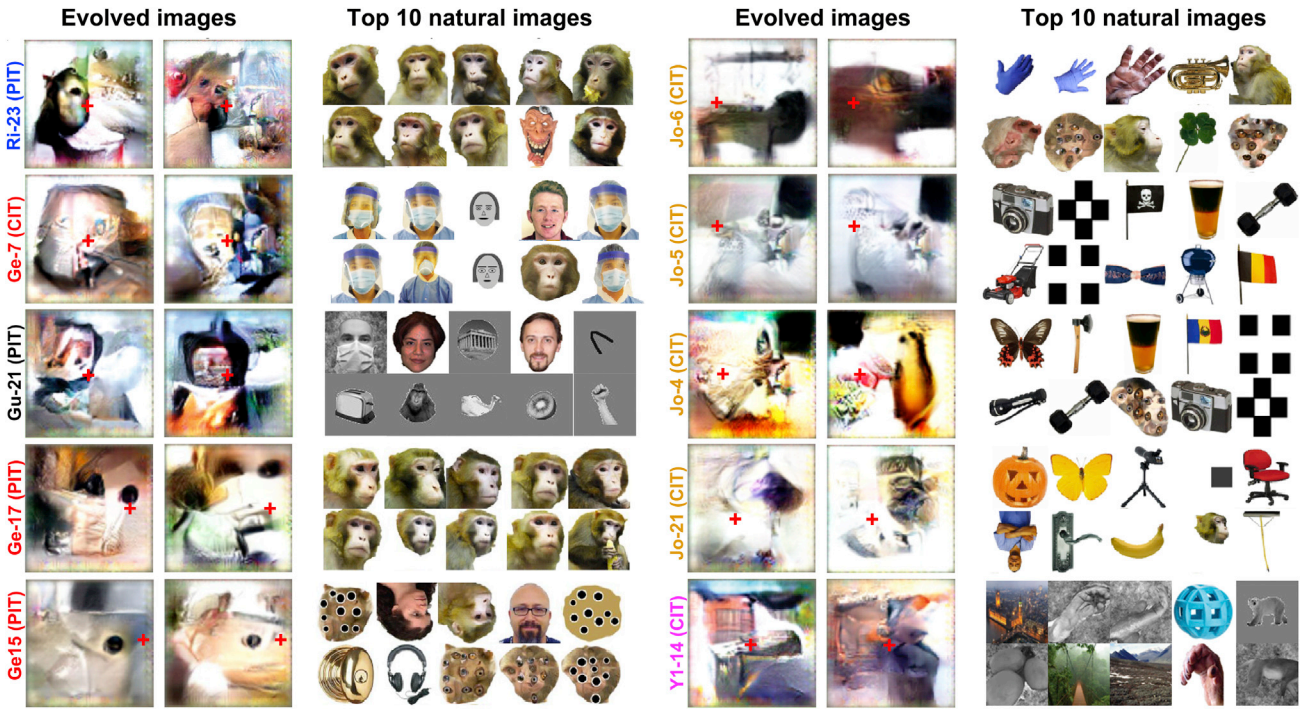


Figure 6. Evolution of Synthetic Images in Other IT Neurons

Each pair of large images shows the last-generation synthetic images from two independent evolution experiments for a single chronic recording site in 5 different animals. To the right of the synthetic images are the top 10 images for each neuron from a natural image set. Red crosses indicate fixation. The arrays were in the left hemisphere of monkey Jo, and in the right hemisphere of all the other animals. The natural images shown interleaved during each evolution were from either a 108-reference image set containing faces, bodies, places, and line segments (used for cells Gu-21 and Y1-14) or the set of 2,550 natural images rank ordered in Figure S2 for unit Ri-10 (used for all the other units in this figure).

First, we focused on the evolution experiment for PIT single unit Ri-17. This cell evolved a discrete shape near the top left of the image frame, comprising a darkly outlined pink convex shape with two dark circles and a dark vertical line between them (Figure S5A). When tested with the 2,550 natural images, this neuron responded best to images of monkeys, dogs, and humans (Figure S5B). We propagated this evolved image through AlexNet, along with the 100,300 ImageNet examples, ranked all the fc6 vectors by their Pearson correlation to the evolved image vector, and identified the closest, middle, and farthest 100 matches. The synthetic image showed an average vector correlation of 0.38 to the closest images, 0.06 to the middle, and −0.14 to the farthest images. The 9 nearest ImageNet images were cats, dogs, and monkeys (Figure S5C). To visualize the common shape motifs of this image cluster, we identified the individual fc6 units most strongly activated by the synthetic image and used activation maximization (*deepDreamImage.m*) to generate examples of preferred shapes for those fc6 units. All the units preferred round tan/pink regions with small dark spots (Figure S5D). To rule out that these matches could be due to an overrepresentation of animals in ImageNet, we also looked at the least correlated matches, which were indeed not animals, but were pictures of places, rectilinear textures, or objects with long, straight contours (Figure S5E).

We applied this image-search approach to all evolution experiments by identifying the top 100 matches to every synthetic image in fc6 space (the Pearson correlation coefficients of these images ranged from 0.30 to 0.61, median 0.36) and visualized the WordNet (PrincetonUniversity, 2010) labels of the matching images via word clouds. In monkey Ri, whose array showed natural-image preferences for faces, the categories that best matched the synthetic images were "macaques," "toy terrier," and "Windsor tie" (the latter containing images of faces and bodies) (Figure S5F); in contrast, in monkey Y1, where most of the neurons in the array had shown natural-image preferences for places, the categories that best matched were "espresso maker," "rock beauty" (a type of fish), and "whiskey jug"; by inspection these images all contained extended contours (Figure S5G). We confirmed this matching trend by quantifying the WordNet hierarchy labels associated with every matched natural image (Table S2).

To find out whether similarity in fc6 space between a neuron's evolved synthetic image and a novel natural image predicted that neuron's response to that novel image, we first performed 3 to 4 independent evolution experiments using the same (single- or multi-) unit in each of three animals. After each evolution, we took the top synthetic image from the final generation and identified the top 10 nearest images in fc6 space, 10 images from the middle of the distance distribution and the farthest (most anticorrelated) 10 images (9 of each shown in Figure 7A). Then, during the same recording session, we presented these images to the same IT neurons and measured the responses

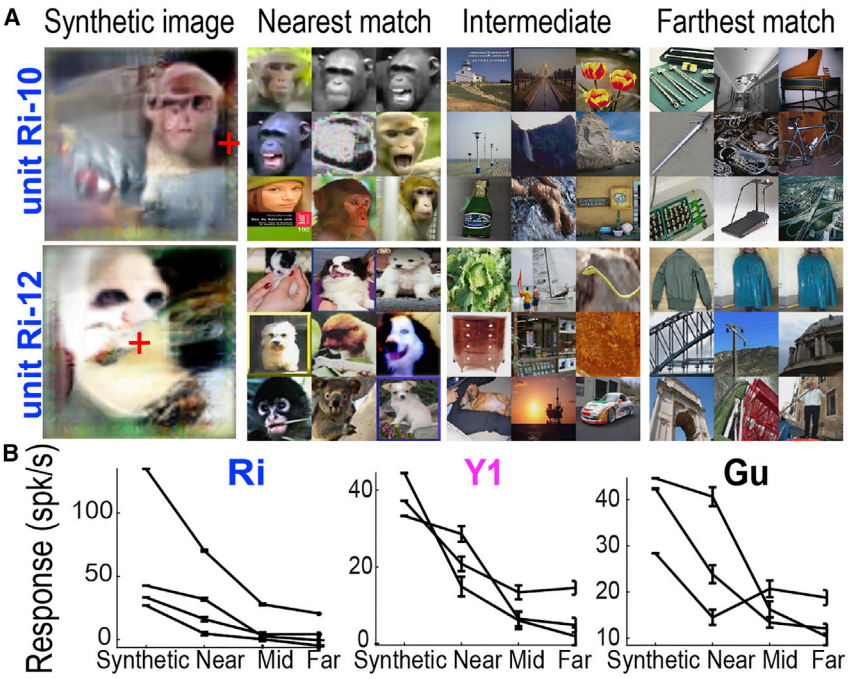


Figure 7. Using Evolved Images to Predict Responses to Novel Images

(A) Final-generation synthetic images from units Ri-10 and Ri-12 and the closest, intermediate, and farthest 9 images from the image set for each. For unit Ri-10 we used the 60,000-image database, and for unit Ri-12 the 100,300-image database.

(B) Responses from each unit to the last-generation evolved images compared with the nearest, intermediate, and farthest images from the evolved images in the fc6 space of AlexNet (mean ± SEM).

to each group (near, middle, and far) as well as to all 40 evolved images of the last generation. Figure 7B shows that synthetic images gave the highest responses, the nearest natural images the next highest responses, and the middle and farthest images the lowest. To quantify this observation, we fit linear regression functions between the ordinal distance from the synthetic image (near, middle, and far) and the unit's mean responses, and found median negative slopes ranging from −5.7 to −21.1 spikes per s across monkeys (Table S3). Thus, distance from the evolved synthetic image in fc6 space predicted responses to novel natural images. This does not indicate that this space is the best model for IT response properties; instead, this shows that it is possible to use the neurons' evolved images to predict responses to other images. Importantly, responses to the synthetic images were the highest of all.

Invariance to Evolved versus Natural Images

IT neurons retain selectivity despite changes in position, size, and rotation (Ito et al., 1995; Kobatake and Tanaka, 1994), although it has been reported that more selective neurons are less transformation-invariant (Zoccolan et al., 2007). The latter observation is also consistent with the alternative interpretation that the more optimal a stimulus is for the neuron, the less invariant the neuron will be, and this is consistent with what we found. To compare the invariance of IT neurons to synthetic and natural images, we presented 3 natural and 3 evolved synthetic images at different positions, sizes, and fronto-parallel rotations in two animals (monkeys Ri and Gu). The natural images were chosen from the nearest, middle, and farthest matches from ImageNet. The synthetic images were chosen from the final generation. Every image was presented at three positions in relation to the fovea: (−2.0°, −2.0°), (−2.0°, 2.0°), and (0.0°, 0.0°); three sizes (widths of 1°, 2°, 4°) and 4 rotations (0°, 22°, 45°, and 80°, counterclockwise from horizontal) (Figure S6A). Invariance was defined as the similarity (correlation coefficient) in the neuron's rank order of preferences for images under different transformation conditions (the more similar the rank order, the higher the invariance). The rank order was better maintained across transformations for the natural images than for the synthetic images (Figures S6B and S6C). Thus, the degree of invariance for these neurons changed depending on the stimulus set, and the neurons were the least invariant for the more optimal synthetic images. This result suggests that the degree of invariance measured for particular neurons might not be a fixed feature of that neuron, but rather might depend on the effectiveness of the stimulus used to test the invariance.

Evolution of Preferred Stimuli for Populations of Neurons

Single- and multi-units in IT successfully guided the evolution of synthetic images that were stronger stimuli for the neuron guiding the evolution than large numbers of natural images. To see if our technique could be used to characterize more coarsely sampled neuronal activity than a single site, we asked whether we could evolve strong stimuli for all 32 sites on an array. Each of the chronically implanted arrays had up to 32 visually responsive sites, spaced 400 μm apart. We conducted a series of evolution experiments in 3 monkeys (Ri, Gu, and Y1) guided by the average population response across the array. Evolution experiments for all 3 monkeys showed increasing responses to synthetic images over generations compared with reference images: the median population response changes to synthetic images for monkeys Ri, Gu, and Y1 were 9.4 spikes per s per generation (2.8–19.6, 25th–75th percentile), 30.7 (17.6–48.3, 25th–75th percentile) and 27.8 (18.8–39.6, 25th–75th percentile). In these population-guided evolutions, 61%, 93%, and 99.5% of individual sites showed increases in firing rate (statistical significance defined by fitting an exponential function to 250 resampled firing rate per generation curves per site; an increase was significant if the 95% CI of the amplitude bootstrap distribution did not include zero). Therefore, larger populations of IT neurons could successfully create images that were on average strong stimuli for that population. When the populations were correlated in their natural-image preferences, the synthetic

images were consistent with those evolved by individual single sites in the array: for example, in monkey Ri, the population-evolved images contained shape motifs commonly found in ImageNet pictures labeled "macaques," "wire-haired fox terrier," and "Walker hound." This suggests that the technique can be used with coarser sampling techniques than single-unit recordings, such as local field potentials, electrocorticography electrodes, or even functional magnetic resonance imaging.

Testing XDREAM Using the Ground Truth of Primary Visual Cortex

We recorded from one single unit and three multiunit sites (six evolution experiments total) in monkey Vi, which had a chronic microelectrode array in V1. The stimuli were centered on each receptive field (measuring ~0.79° square root of the area) but were kept at the same size as in the IT experiments (3° × 3°). In addition to the synthetic images, we interleaved reference images of gratings (3° × 3° area) of different orientations (0°, 45°, 90° and 135°) and spatial frequencies (~0.5, 1, and 2 cycles per degree) at 100% contrast. In all experiments, neurons showed an increase in firing rate to the synthetic images (median 84.0 spikes per s per generation, 77.4–91.2, 25th–75th percentile) (Table S1). Thus, on average, V1 sites, like IT cortex, responded well to late-generation synthetic images (Table S4). To measure the distribution of orientations of the region of the synthetic images that fell within each V1 receptive field (~0.8° × 0.8°), we performed a discrete Fourier transform analysis on the central 0.8° × 0.8° of the synthetic images and correlated the resulting spectrogram to the spectrograms expected from 16 gratings with orientations ranging from 0° to 135°. Across experiments, the mean correlation between the orientation content profile of the patch and the orientation tuning measured from the gratings was 0.59 ± 0.09 (mean ± SEM), compared with 0.01 ± 0.26 for a shuffled distribution (p values $\leq$ 0.006 in 5 out of 6 experiments, permutation test, $N_{iterations}$ = 999). Thus, V1 neurons guided the evolution of images dominated by their independently measured preferred orientation.

DISCUSSION

We introduce XDREAM, a new algorithm for studying the response properties of visual neurons via a vast generative image space. Our approach is an extension of previous work that uses adaptive sampling to reveal neuronal tuning in the ventral stream (Carlson et al., 2011; Hung et al., 2012; Kalfas et al., 2017; Vaziri et al., 2014; Vaziri and Connor, 2016; Yamane et al., 2008). In our approach, adaptive sampling enables the exploration of a stimulus space potentially large enough to span the selectivity of ventral pathway neurons. Starting with random textures and evolving images based on neuronal responses, the algorithm created images that elicited large responses in V1, in different parts of the IT cortex, in single-units, in multi-units, and in average population responses. Remarkably, XDREAM evolved stimuli that evoked higher responses than the best natural images found by an extensive exploration of large image sets.

A powerful approach for modeling neuronal responses has been to use stimuli sampled in a fully defined parametric space. For example, one innovative study (Yamane et al., 2008), which inspired our approach, defined response properties in the IT cortex according to 3D curvature and orientation. In a more recent study, Chang and Tsao (2017) used face images parameterized along appearance and shape axes to describe and predict neuronal responses in face patches. Parametric stimulus spaces lead to quantitative neuronal models that are easier to describe than models built on the learned, vast, and latent space of a deep network. But these approaches are complementary: standard parametric models operate in a circumscribed shape space, so these models might not capture the entire response variability of the neuron. Generative neural networks are powerful enough to approximate a wide range of shape configurations, even shapes diagnostic of monkey faces or bodies. If the response evoked by the best stimulus in a parametric space is less than that evoked by a generative-network stimulus, it would indicate that the parametric model is overly restricted. This is important because a neuron could be tuned to different aspects of a stimulus: a face-preferring neuron might show tuning for high curvature, but curvature tuning alone is insufficient to explain the neuron's selectivity. A generative-network-based approach can serve as an independent, less constrained test of parametric models. Indeed, in some instances, the evolved stimuli contained features of animal faces, bodies, and even animal-care staff known to the monkeys, consistent with theories of tuning to object categories and faces, whereas in other instances, the evolved stimuli were not identifiable objects or object parts, suggesting aspects of neuronal response properties that current theories of the visual cortex have missed. Thus, our approach can also serve as an initial step in discovering novel shape configurations for subsequent parametric quantification. In sum, our approach is well-positioned to corroborate, complement, contrast with, and extend the valuable lessons learned from these previous studies.

What are these powerful stimuli "dreamed of" by the neurons? To start to answer this, we looked for images that were nearest to the evolved images in AlexNet layer fc6 space, because the generative network had been trained on that space and because higher layers of this network can predict neuronal responses (Yamins et al., 2014). Similarity to the evolved images in fc6 space predicted neuronal selectivity. For neurons defined using natural images as monkey- and face-preferring, the closest images to the evolved ones were macaques and dogs or contained faces and torsos of monkeys or other mammals. In contrast, for place-preferring neurons, the nearest images instead included a variety of objects with extended straight contours, such as espresso makers and moving vans.

Although the evolved synthetic images were not life-like, sometimes not even identifiable objects, they nevertheless tell us something quite novel about what information might be encoded by IT neurons. PIT neurons have been reported to be selective for low-level features like radial gratings (Pigarev et al., 2002), color (Zeki, 1977), or single eyes (Issa and DiCarlo, 2012), and face cells in CIT have been shown to be tuned to distinct face-space axes (Chang and Tsao, 2017; Freiwald et al., 2009). Our experiments revealed that both PIT and CIT neurons evolved complex synthetic images, usually consisting of intact objects with multiple, spatially organized, features and colors, not just disjointed object parts. This is consistent with

the finding that even cells tuned to particular feature parameters, like eye size, do not respond well to that parameter in isolation, but rather only in context (Freiwald et al., 2009). It has been proposed that neurons are tuned to abstract parameters, such as axes (Chang and Tsao, 2017; Freiwald et al., 2009; Leopold et al., 2006) that distinguish between things in the environment, rather than being tuned to the things themselves. That is, neuronal responses might carry information about an object in the environment, not a veridical representation of it. The unrealistic nature of our evolved images, plus the fact that these images were more effective than most or all the images in an extensive natural-image database, suggest that these images might lie somewhere on tuning axes that extend beyond anything the animal would normally encounter. Lastly, the neurons that evolved images resembling lab staff wearing face masks indicate a major role for experience in the development of neuronal response properties (Arcaro et al., 2017).

It is unclear whether our approach has uncovered globally optimum stimuli for these cells, but it is equally unclear whether there should be a single global optimum stimulus for a neuron. Different evolutions for the same units yielded synthetic images that shared some features but differed in others (e.g., Figure 4B), as would be expected from a cell that shows invariance to nuisance transformations like position, color, or scale. Even in early visual areas, complex cells respond equally well to the same stimulus in multiple locations within a cell's receptive field (Hubel and Livingstone, 1985; Hubel and Wiesel, 1968), and this complexification; i.e., OR gate-like operation is likely to occur at multiple levels in the hierarchy (Hubel and Livingstone, 1985; Riesenhuber and Poggio, 1999), so it follows that multiple different feature combinations could strongly activate a particular neuron in the IT cortex. However, the facts that our approach can recover known properties of neurons; that it can discover unknown response properties; that the synthetic images can be better than any in a large set of natural stimuli, even better than highly-optimized stimuli based on convolutional neural network models of neurons—currently the best models of visual neurons (Figure S7 and Table S6)—all indicate that we have discovered stimuli that give us a best glimpse at what IT neurons are tuned to. These results complement classical methods for defining neuronal selectivities and demonstrate the potential for uncovering internal representations in any modality that can be captured by generative neural networks.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.04.005.

ACKNOWLEDGMENTS

This work was supported by NIH grants R01EY16187, R01EY25670, R01EY011379, P30EY012196, R01EY026025 and NSF STC award CCF-1231216 to the Center for Brains, Minds and Machines at MIT. We thank Richard Born, Shimon Ullman, and Christof Koch for comments.

AUTHOR CONTRIBUTIONS

G.K. and W.X. conceived of the approach. W.X. implemented the network algorithm and did all the *in silico* experiments. C.R.P. and P.F.S. adapted the algorithm to neurophysiology. P.F.S., M.S.L., T.S.H. and C.R.P. collected and analyzed data. C.R.P., W.X., and M.S.L. wrote the manuscript, all authors revised the manuscript, and M.S.L. acquired funding.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 6, 2018
Revised: November 5, 2018
Accepted: April 2, 2019
Published: May 2, 2019

REFERENCES

Abbasi-Asl, R., Chen, Y., Bloniarz, A., Oliver, M., Willmore, B.D.B., Gallant, J.L., and Yu, B. (2018). The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. https://www.biorxiv.org/content/10.1101/465534v1.

Arcaro, M.J., Schade, P.F., Vincent, J.L., Ponce, C.R., and Livingstone, M.S. (2017). Seeing faces is necessary for face-domain formation. Nat. Neurosci. *20*, 1404–1412.

Barlow, H.B., Blakemore, C., and Pettigrew, J.D. (1967). The neural mechanism of binocular depth discrimination. J. Physiol. *193*, 327–342.

Bashivan, P., Kar, K., and DiCarlo, J.J. (2018). Neural population control via deep image synthesis. https://www.biorxiv.org/content/10.1101/461525v1.

Carlson, E.T., Rasquinha, R.J., Zhang, K., and Connor, C.E. (2011). A sparse object coding scheme in area V4. Curr. Biol. *21*, 288–293.

Chang, L., and Tsao, D.Y. (2017). The code for facial identity in the primate brain. Cell *169*, 1013–1028.e14.

Desimone, R., Albright, T.D., Gross, C.G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. *4*, 2051–2062.

Dosovitskiy, A., and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. advances in neural information processing (NIPS).

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network (University of Montreal).

Felleman, D.J., and Van Essen, D.C. (1987). Receptive field properties of neurons in area V3 of macaque monkey extrastriate cortex. J. Neurophysiol. *57*, 889–920.

Freiwald, W.A., Tsao, D.Y., and Livingstone, M.S. (2009). A face feature space in the macaque temporal lobe. Nat. Neurosci. *12*, 1187–1196.

Gaska, J.P., Jacobson, L.D., Chen, H.W., and Pollen, D.A. (1994). Space-time spectra of complex cell filters in the macaque monkey: a comparison of results obtained with pseudowhite noise and grating stimuli. Vis. Neurosci. *11*, 805–821.

Gross, C.G., Rocha-Miranda, C.E., and Bender, D.B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. J. Neurophysiol. *35*, 96–111.

Hubel, D.H. (1959). Single unit activity in striate cortex of unrestrained cats. J. Physiol. *147*, 226–238.

Hubel, D.H., and Livingstone, M.S. (1985). Complex-unoriented cells in a subregion of primate area 18. Nature *315*, 325–327.

Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. *160*, 106–154.

Hubel, D.H., and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. J. Physiol. *195*, 215–243.

Hung, C.C., Carlson, E.T., and Connor, C.E. (2012). Medial axis shape coding in macaque inferotemporal cortex. Neuron *74*, 1099–1113.

Issa, E.B., and DiCarlo, J.J. (2012). Precedence of the eye region in neural processing of faces. J. Neurosci. *32*, 16666–16682.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. J. Neurophysiol. *73*, 218–226.

Jia, Y., Shelhamer, E., Donohue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, pp. 675–678.

Jones, J.P., and Palmer, L.A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. J. Neurophysiol. *58*, 1187–1211.

Kalfas, I., Kumar, S., and Vogels, R. (2017). Shape selectivity of middle superior temporal sulcus body patch neurons. eNeuro *4*, ENEURO.0113-17.2017.

Kobatake, E., and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J. Neurophysiol. *71*, 856–867.

Konkle, T., Brady, T.F., Alvarez, G.A., and Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. J. Exp. Psychol. Gen. *139*, 558–578.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM. *60*, 84–90.

Leopold, D.A., Bondar, I.V., and Giese, M.A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature *442*, 572–575.

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

Michael, C.R. (1978). Color vision mechanisms in monkey striate cortex: simple cells with dual opponent-color receptive fields. J. Neurophysiol. *41*, 1233–1249.

Mordvintsev, A., Olah, C., and Tyka, M. (2015). Inceptionism: Going deeper into neural networks (Google Research Blog).

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS).

Pasupathy, A., and Connor, C.E. (1999). Responses to contour features in macaque area V4. J. Neurophysiol. *82*, 2490–2502.

Pasupathy, A., and Connor, C.E. (2002). Population coding of shape in area V4. Nat. Neurosci. *5*, 1332–1338.

Pigarev, I.N., Nothdurft, H.C., and Kastner, S. (2002). Neurons with radial receptive fields in monkey area V4A: evidence of a subdivision of prelunate gyrus based on neuronal response properties. Exp. Brain Res. *145*, 199–206.

Portilla, J., and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. Int. J. Comput. Vis. *40*, 49–71.

Princeton University (2010). About WordNet. WordNet (Princeton University), https://wordnet.princeton.edu/.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. *115*, 211–252.

Vaziri, S., and Connor, C.E. (2016). Representation of Gravity-Aligned Scene Structure in Ventral Pathway Visual Cortex. Curr. Biol. *26*, 766–774.

Vaziri, S., Carlson, E.T., Wang, Z., and Connor, C.E. (2014). A channel for 3D environmental shape in anterior inferotemporal cortex. Neuron *84*, 55–62.

Walker, E.Y., Sinz, F.H., Froudarakis, E., Fahey, P.G., Muhammad, T., Ecker, A.S., Cobos, E., Reimer, J., Pitkow, X., and Tolias, A.S. (2018). Inception in visual cortex: in vivo-silico loops reveal most exciting images. https://www.biorxiv.org/content/10.1101/506956v1.

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

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

Zeki, S.M. (1973). Colour coding in rhesus monkey prestriate cortex. Brain Res. *53*, 422–427.

Zeki, S.M. (1974). Functional organization of a visual area in the posterior bank of the superior temporal sulcus of the rhesus monkey. J. Physiol. *236*, 549–573.

Zeki, S.M. (1977). Colour coding in the superior temporal sulcus of rhesus monkey visual cortex. Proc. R. Soc. Lond. B Biol. Sci. *197*, 195–223.

Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J.J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J. Neurosci. *27*, 12292–12307.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MonkeyLogic2		https://www.nimh.nih.gov/labs-at-nimh/research-areas/clinics-and-labs/ln/shn/monkeylogic/index.shtml
Caffe	Jia et al., 2014	N/A
CaffeNet		https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet
Python		https://www.python.org/
MATLAB	Mathworks, Natick, MA	www.mathworks.com
Generative neural network	Dosovitskiy and Brox, 2016	lmb.informatik.uni-freiburg.de/resources/software.php

CONTACT FOR REAGENT AND RESOURCE SHARING

Requests for resources should be directed to and will be fulfilled by the Lead Contact, Margaret Livingstone (mlivingstone@hms.harvard.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

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.

Behavior

Six adult male macaca mulatta (9 −13 kg; 5-14 years old) and one adult male macaca nemestrina (9 kg, 10 years old) were socially housed in standard quad cages on 12/12 hr light/dark cycles.

Recording arrays

Monkeys Ri, Gu, Ge, Jo, and Y1 were implanted with custom floating microelectrode arrays manufactured by MicroProbes for Life Sciences (Gaithersburg, MD); each had 32 platinum/iridium electrodes per ceramic base, electrode lengths of 2-5 mm, impedances between 0.7- 1.0 MΩ. The recording sites in the chronic arrays were stable across days; in particular, for all the experiments where we compare evolutions from the same site recorded on different days, the correlation in category preference was greater than 0.89. Monkey Vi was implanted with a 96-channel Utah array (Blackrock microsystems, Salt Lake City, Utah). Monkey B3 had an acute recording chamber, and neuronal activity was recorded using a 32 channel NeuroNexus Vector array (Ann Arbor, Michigan) that was inserted each recording day.

Surgical procedures

All animals were implanted with custom-made titanium or plastic headposts before fixation training. After several weeks of fixation training, the animals underwent a second surgery for array or chamber implantation. PIT insertion sites were just anterior to the inferior occipital sulcus; CIT sites were on the lower lip of the STS 6-8 mm anterior to the interaural line. All surgeries were done under full surgical anesthesia using sterile technique.

METHOD DETAILS

Behavioral task

The monkeys were trained to perform a fixation task. They fixated on a 0.2°-diameter fixation spot in the middle of the screen. Eye position was monitored using an ISCAN system (Woburn, MA). Animals were rewarded with a drop of water or juice for maintaining fixation within 1.0° of the fixation spot for 2-7 image presentations; the interval was gradually decreased over the experimental session as the monkey's motivation decreased.

Physiological recording

Neural signals were amplified and extracellular action potentials were isolated using the box method of an on-line spike sorting system (Plexon, Dallas, TX). Spikes were sampled at 40 kHz.

Deep Generative Neural Network

The pre-trained generative network (Dosovitskiy and Brox, 2016) was downloaded from the authors' website (https://lmb.informatik.uni-freiburg.de/resources/software.php) and used without further training with the Caffe library (Jia et al., 2014) in Python. To synthesize an image from an input image code, we forward propagated the code through the generative network, clamped the output image pixel values to the valid range between 0 and 1, and visualized them as an 8-bit color image. Some images synthesized by the network contained a patch with a stereotypical shape that occurred in the center right of the image (e.g., second and third image in Figure 4B, and "late synthetic" in Figure 5C and 5D). This was identified as an artifact of the network commonly known as "mode collapse" and it appeared in the same position in a variety of contexts, including a subset of simulated evolutions. This artifact was easily identifiable and it did not affect our interpretations. In the future, more modern GNN training methods should avoid this problem (personal correspondence with Alexey Dosovitskiy).

Initial generation

The initial generation of image codes for all evolution experiments reported here was 40 achromatic textures constructed from a set of Portilla and Simoncelli (2000) textures, derived from randomly sampled photographs of natural objects on a gray background. We started from all-zero codes and optimized for pixelwise loss between the synthesized images and the target images using backpropagation through the network for 125 iterations, with a learning rate linearly decreasing from 8 to 1×10^{-10}. 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.

Genetic algorithm

The algorithm began with an initial population of 40 image codes ("individuals"), each consisting of a 4096-dimensional vector ("genes") and associated with a synthesized image. Images were presented to the subject, and the corresponding spiking response was used to calculate the "fitness" of the image codes by transforming the firing rate into a Z-score within the generation, scaling it by a selectiveness factor of 0.5, and passing it through a softmax function to become a probability. The 10 highest-fitness individuals were passed on to the next generation without recombination or mutation. Another 30 children image codes were produced from recombinations between two parent image codes from the current generation, with the probability for each image code to be a parent being its fitness. The two parents contributed unevenly (75%:25%) to any one child. Individual children genes had a 0.25 probability of being mutated, with mutations drawn from a 0-centered Gaussian with standard deviation 0.75. Hyperparameter values were not extensively optimized. All source code is available upon request.

Evolutions in CaffeNet units

We selected 100 random units each in 4 layers in CaffeNet as targets for evolution. For convolutional layers, only the center unit in each channel was used. Each unit was evolved for 500 generations, 10,000 image presentations total. The best image in the last generation was used in the analysis, although most of the total activation increase was achieved by 200 generations. As a control, we recorded activations of the units to all 1,431,167 images in the ILSVRC2012 dataset, including the training set of CaffeNet. To visualize the ground truth best in CaffeNet layer conv1, $11 \times 11 \times 3$ images were produced from the $11 \times 11 \times 3$ filter weights according to $image(x,y,c) = 0.5 + sign(weight(x,y,c))/2$, because this is a linear transformation. In other words, positive weights corresponded to a pixel value of 1 and negative weights 0, and the ground truth optimum only contained black, white, red, green, blue, magenta, cyan, and yellow pixels. To further visualize the magnitude of the weights, each pixel was made transparent to a gray checkerboard in inverse proportion to its contribution to the overall activation, so that the greyer a pixel, the less it affected the overall activation.

Visual stimuli

We used MonkeyLogic2 (https://www.nimh.nih.gov/labs-at-nimh/research-areas/clinics-and-labs/ln/shn/monkeylogic/index.shtml) as the experimental control software. Images were presented on an LCD monitor 53 cm in front of the monkey at a rate of 100 ms on, 100-200 ms off. Most of the images were from (Konkle et al., 2010); human and monkey images were from our lab, and the rest of the images were from public domain databases.

QUANTIFICATION AND STATISTICAL ANALYSIS

Spike rate changes during evolutions

We defined the neuronal response as the spike rate measured in the 70-200 ms time window after image onset and subtracted the rate in the 0-60 ms window. In the evolution experiments, there were 40 synthetic and 40 reference images per generation, each presented once. To track firing rate change per generation, we averaged responses to all 40 synthetic and 40 reference images separately. To estimate the inter-generational changes in response we fit the mean response/generation separately for synthetic

and reference images with a decaying exponential function $-a \times \exp\left(-\frac{x}{\tau}\right) + c$, which models firing rate as starting at the first block with rate c and asymptotically approaching the rate $(a + c)$ with decay constant τ. We restricted the amplitude change to be within physiologically plausible values (magnitude no more than the absolute maximum rate difference between any two generations in that day). To assess statistical significance, we generated new mean rate per generation curves by resampling responses (with replacement) from each of the 40 synthetic and 40 reference image presentations within one generation, then fit the exponential function each time (N = 500 repetitions). The 95% confidence intervals reported are the 12th and 488th values of the bootstrapped distribution. We used responses from all generations except for one acute experiment in monkey B3, where a single-unit isolation change nullified the initial 15 generations and two chronic experiments in monkey Ge, where spike thresholding adjustments nullified the initial 17 generations in one experiment, and the final 35 generations in the other.

Responses to evolved versus reference images

For every evolution experiment, we measured how the neurons' maximum firing rate to the evolving images compared to the neurons' maximum firing rate to the reference images. To estimate maximum firing rate and its standard error, we re-sampled (with replacement) all single-trial responses to synthetic or reference images, computed the observed maximum per sample, repeated 500 times and then reported the mean and standard deviation across samples. To determine if the max firing rates to synthetic images were different from those to reference images, we used a randomization test: We created a null distribution of differences between two maximum responses by randomly sampling twice from the pooled distribution of synthetic and reference images and measuring the difference in the maximum response between the samples. We repeated this process 499 times, and then measured how often the absolute differences for the mixed-distribution were larger than the observed absolute maximum difference. This two-tailed test indicated if either the reference or synthetic images evoked a significantly larger maximum response.

Relating natural images to the evolved images

Every evolved image was propagated through AlexNet and its fc6-layer activation was compared to those of 100,300 natural images sampled from ImageNet. Every photograph in ImageNet is labeled by categories defined by WordNet, a hierarchically organized lexical database (Princeton University, 2010). After ranking every natural image by its proximity to the evolved image, measured by Pearson correlation coefficient, we used a tree search algorithm to crawl through each labeled image's label hierarchy for the specific search terms–"macaque," "monkey," "face," and "appliance" ("place" was not used because place images often contained people). We measured the frequency of labels associated with all evolved images for every subject. To estimate confidence intervals for every observed frequency, we re-sampled the top matches to each evolved image 200 times (with replacement) and repeated the analysis. To test if the frequency of photographs labeled "monkeys" and "appliance" were statistically different between subjects Ri and Y1, we used a permutation test. The null hypothesis was that these frequency values arose from the same distribution, so we shuffled labels from the Ri and Y1 populations, sampling twice with replacement, and measured the difference, 500 times. We then compared the observed difference in frequency values with the null distribution.

Calculation of stimulus rank order

Background was firing rate 0 to 60 ms after stimulus onset. Spikes evoked in response to each stimulus were averaged over all stimulus presentations from 70 to 200 ms after stimulus onset, minus background. Rank order was determined from the evoked firing rate for each stimulus.

DATA AND SOFTWARE AVAILABILITY

Source code is available at https://github.com/willwx/XDream.

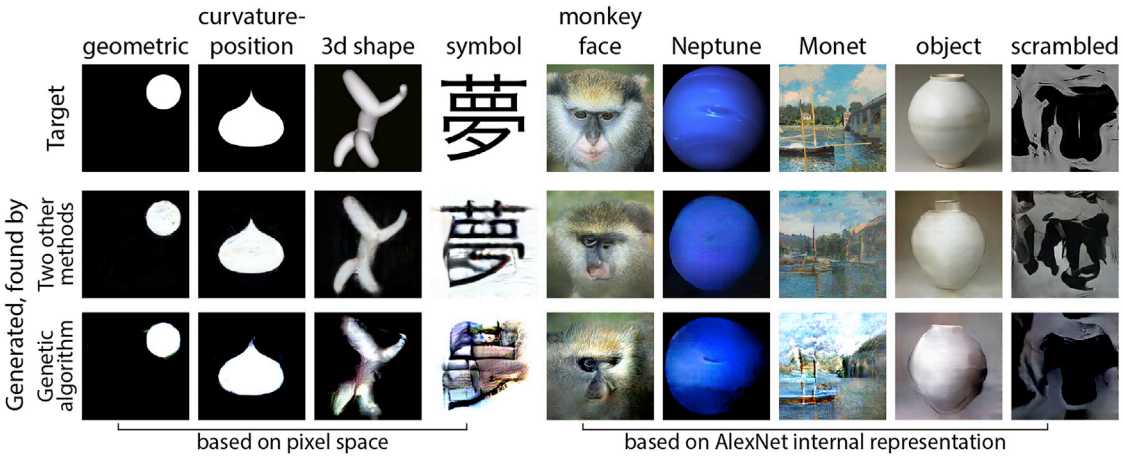


Figure S1. The Deep Generative Adversarial Network Is Expressive and Searchable by A Genetic Algorithm, Related to Figure 1

To qualitatively estimate the expressiveness of the deep generative network, we selected arbitrary images in various styles and categories outside of the training set of the network (first row). To find an image code that would approximately generate each target image (second row), we used either (1) backpropagation to optimize a zero-initialized image code to minimize pixel-space distance (left group; STAR methods, Initial Generation), or (2) the CaffeNet fc6 representations of the target image, as the generator was originally trained to use (right group; Dosovitskiy and Brox, 2016). The existence of codes that produced the images in the second row, regardless of how they were found, demonstrates that the deep generative network is able to encode a variety of images. We then asked whether, given that these images can be approximately encoded by the generator, a genetic algorithm searching in code space ("XDREAM") is able to recover them. To do so, we created dummy "neurons" that calculated the Euclidean distance between the target image and any given image in pixel space (left group) or CaffeNet pool5 space (right group) and used XDREAM to maximize the "neuron responses" (thereby minimizing distance to target), similar to how this network could be used to maximize firing of real neurons in electrophysiology experiments. The genetic algorithm is also able to find codes that produced images (third row) similar to the target images, indicating that not only is the generator expressive, its latent space can be searched with a genetic algorithm. Images that were reproduced from published work with permission are as follows: "Curvature-position" (Pasupathy and Connor, 2002); "3d shape" (Hung et al., 2012); "Monkey face" ILSVRC2012 (Russakovsky et al., 2015). Public domain artwork is as follows: "Monet," *The Bridge at Argenteuil* (National Gallery of Art); "object," "Moon jar" (The Metropolitan Museum of Art). "Neptune" (NASA) is a public domain image.

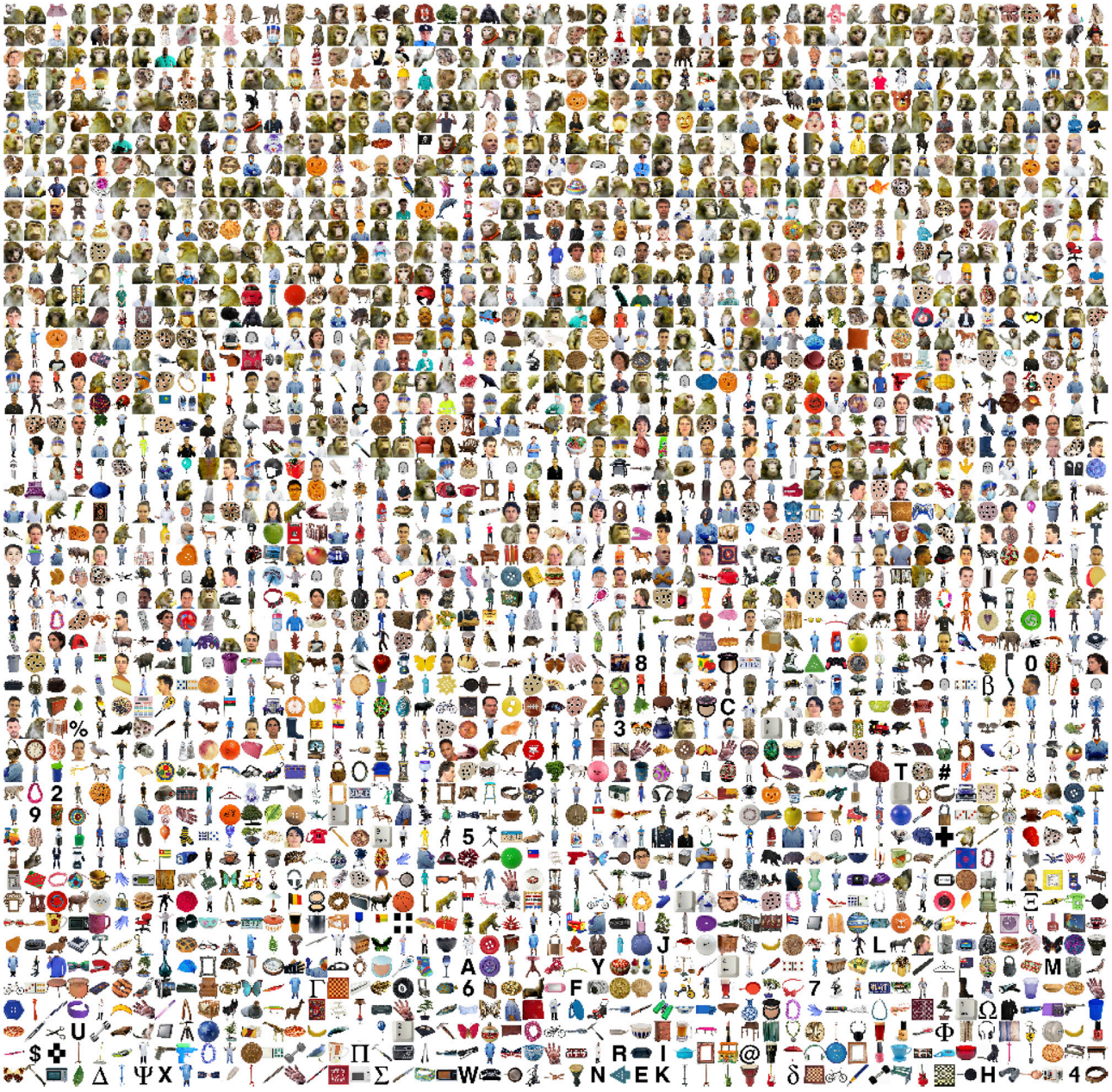


Figure S2. Rank Order from Highest (Top Left) to Lowest (Bottom Right) Responses of Unit Ri-10 to 2550 Images, Related to Figure 4

Responses were the average spikes per image from 70 to 200 ms after stimulus onset, minus baseline (spikes from 1 to 60 ms after stimulus onset). The average response to these images by category is shown in Figure 4E.

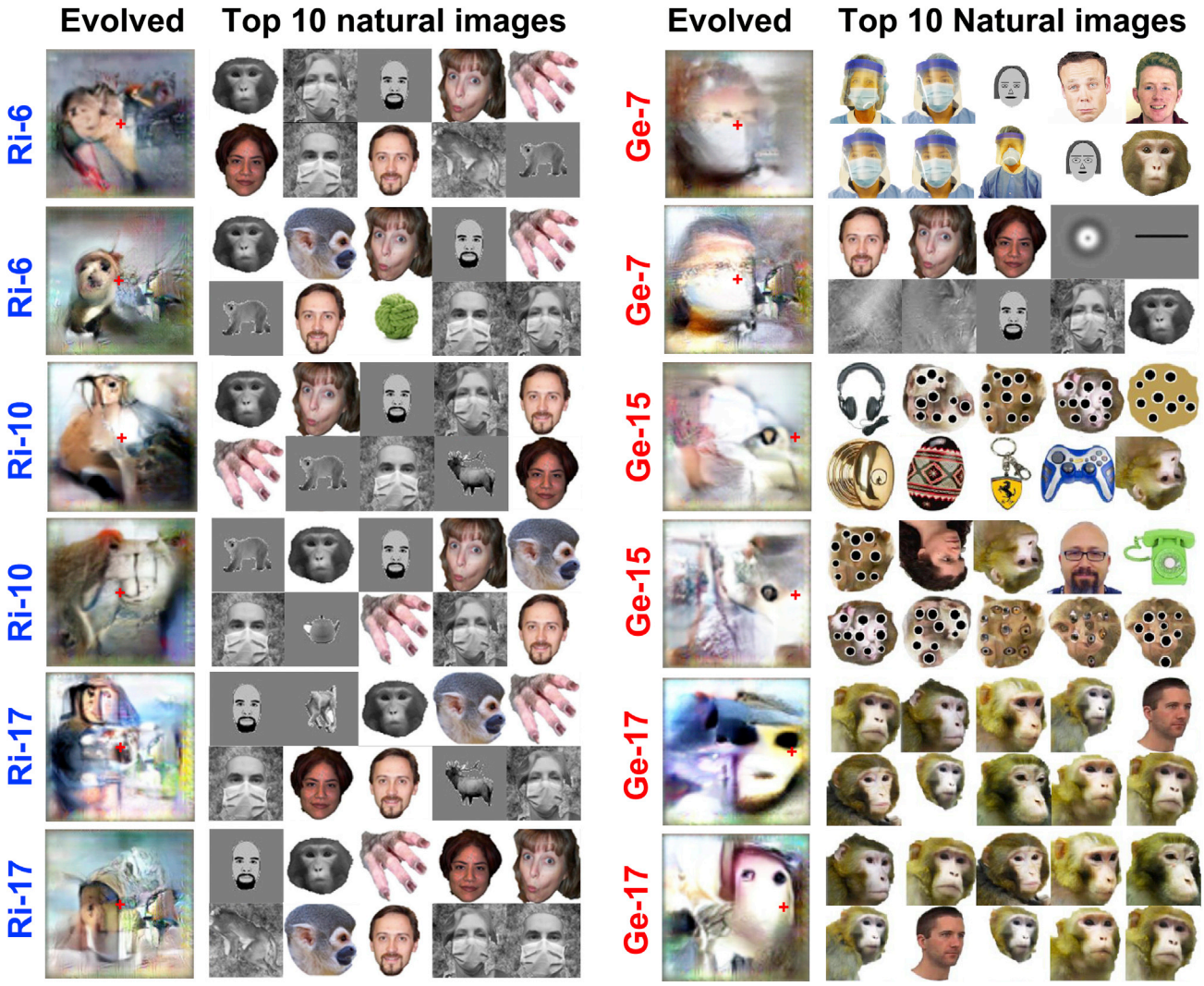


Figure S3. Additional Evolutions of Synthetic Images in IT Neurons, Related to Figure 6

Each large image shows the last-generation synthetic image from one evolution experiment for a single chronic recording site. To the right the synthetic images are shown the top 10 images for that site from a natural image set. Red crosses indicate fixation. The arrays were in the right hemisphere of both animals. As indicated by site number, some of the evolutions shown here are from the same recording sites as shown in Figures 3 & 6, but from independent experiments. For the central-fixation experiments the reader is encouraged to cover the right (ipsilateral) half of the image. For unit Ge-7, first row, unit Ge-15, and unit Ge-17, the natural images were from the set of 2550 natural images shown interleaved with the synthetic images during the evolution experiment; for the other experiments, the natural images were from a 108-image set.

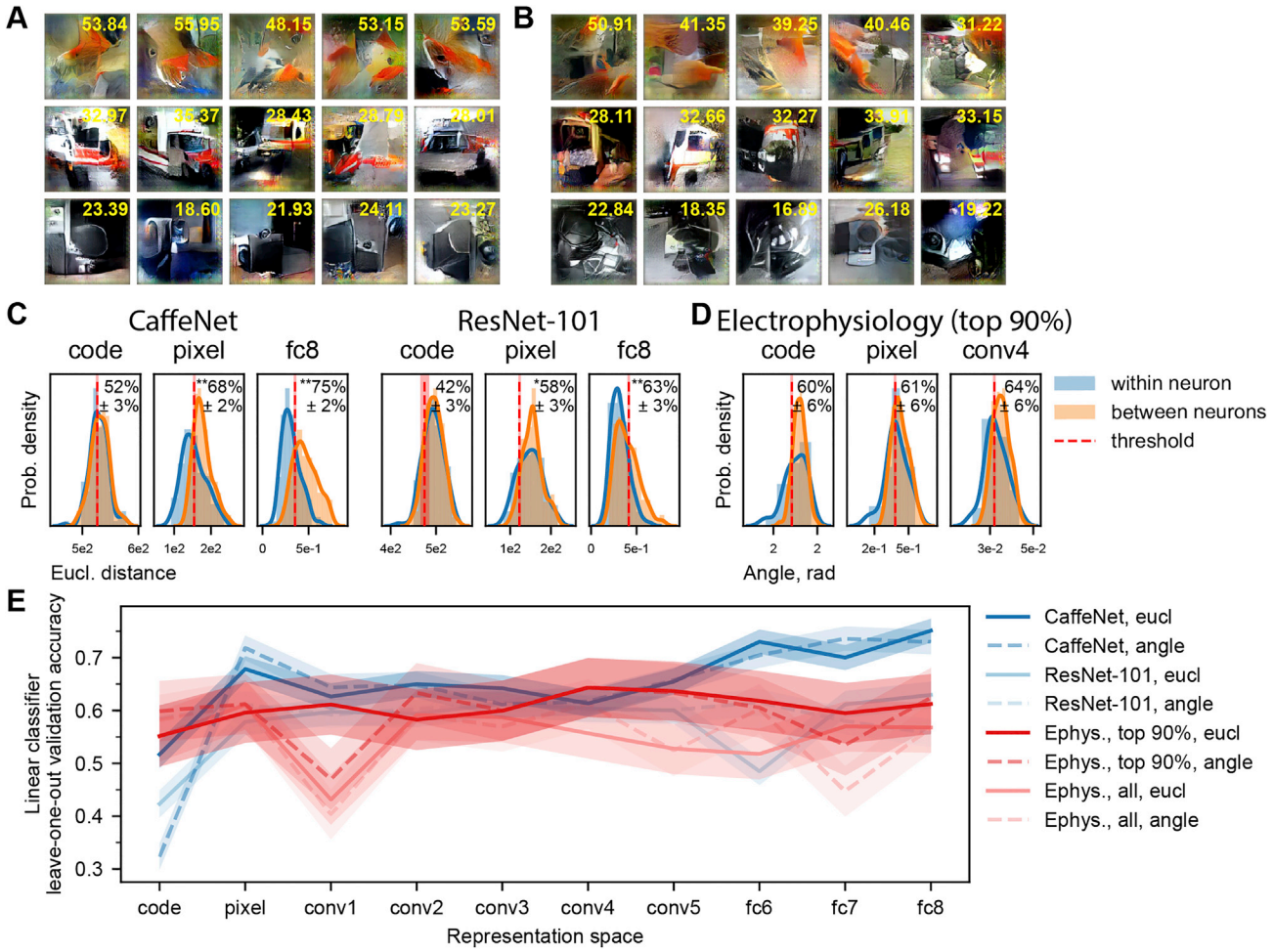


Figure S4. Quantification of Image Similarity across Repeated Experiments using Simple Measures, Related to Figures 4B and 6

(A and B) XDREAM-synthesized stimuli for (A) 3 CaffeNet fc8 units**:** "goldfish," "ambulance," and "loud-speaker;" and (B) the corresponding 3 ResNet-101 fc1000 units, starting from different initial populations (random draws of 40 initial codes from a bank of 1,000). Each row corresponds to a unit and each column a random initialization. Activation for each image is noted on the top right of the image.

(C) Distribution of similarity between images evolved for the same unit versus images evolved for different units, as quantified by Euclidean distance in three spaces (image code, pixel, and CaffeNet layer fc8). To quantify the separation between within- and between-neuron similarities, we estimated the accuracy of simple linear classifiers (thresholds) using leave-one-out cross-validation, separately for each collection of experiments and each measure of similarity. n_{within} = 135; $n_{between}$ = 300. Chance is 50% by equally weighing within-neuron and between-neurons data; accuracy can be < 50% because it is a validation accuracy. Filled bars, histogram. Solid lines, KDE estimate of the distribution, for visual aid only. Shaded red around threshold, standard deviation of threshold across different leave-one-out splits.

(D) Same as (C), but for electrophysiology data with the angle-between-vectors measure in code, pixel, and conv4 spaces. We used the same 45 IT experiments in the main text that showed changes significantly different from zero in neuron response to the evolved images. Here, we excluded 5 experiments with changes within the bottom 10% percentile (analysis with all experiment is shown in the next panel). "Within-neuron" is defined as experiments repeated for the same site in the same animal, and "between-neuron" as experiments in different animals (since sites in the same array can have correlated selectivity). n_{within} = 19; $n_{between}$ = 618. Top-right of each subplot, accuracy ± SEM *p < 0.05; **p < 0.001; both corrected for multiple (20) comparisons.

(E) Accuracy of all 20 measures of similarity (2 distance measures × 10 spaces), separately for the four collections of experiments. Best accuracy for CaffeNet and ResNet-101 data are attained with the Euclidean-fc8 measure. With this measure, accuracies for ResNet-101 data and top-90% electrophysiology data are statistically indistinguishable (p = 0.59). Best accuracy for electrophysiology is attained with the angle-conv4 measure. With this measure, accuracies for electrophysiology data and CaffeNet and ResNet-101 data, respectively. are statistically indistinguishable (p = 0.42 and 0.21 with top-90% electrophysiology data; p = 0.88 and 0.75 with all data). n_{within} = 28; $n_{between}$ = 802 for all 45 electrophysiology experiments. Shaded region indicates SEM.

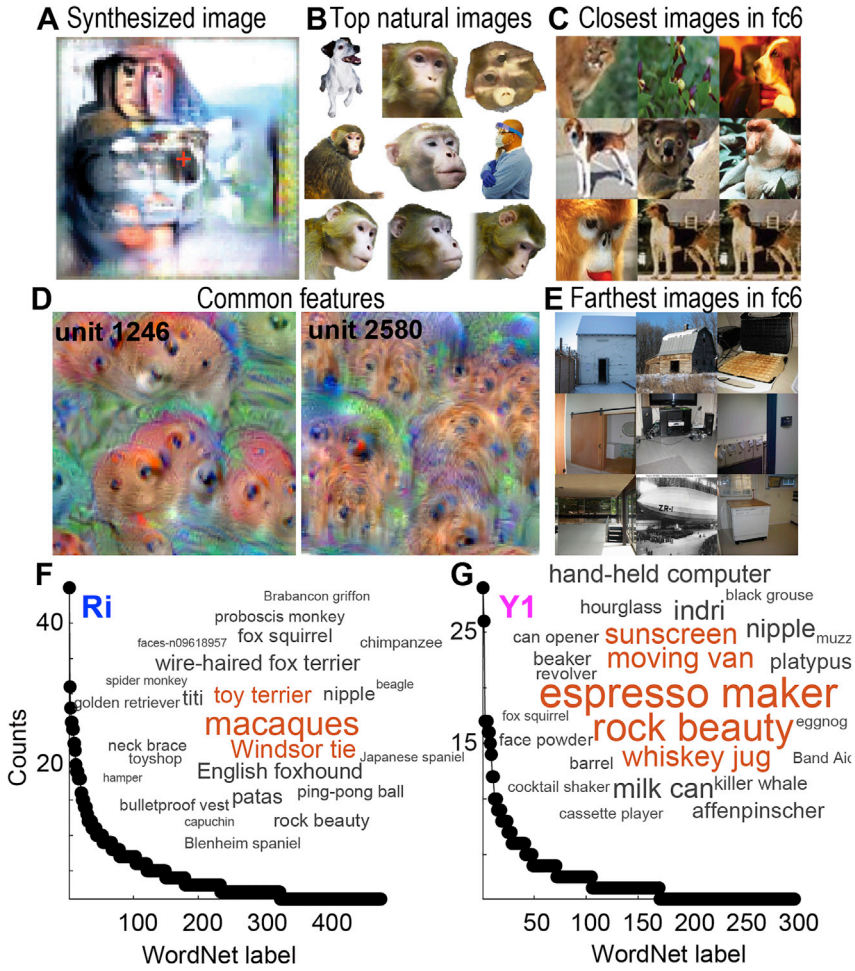


Figure S5. Characterizing Evolved Images using the fc6 Layer of AlexNet, Related to Figure 7

(A) Example image evolved by PIT single unit Ri-17. Its receptive field encompassed the contralateral (left) side of the image; red cross indicates fixation.

(B) Top 9 images for this neuron from the 2,550 image set.

(C) Closest ImageNet pictures based on highest Pearson correlation coefficient in AlexNet layer fc6 representation.

(D) Patterns, as visualized by DeepDream (*deepDreamImage.m*), encoded by the fc6 units that showed highest activations in response to the synthesized image.

(E) Farthest ImageNet pictures.

(F) Word cloud and histogram showing counts of ImageNet labels of the top 150 closest ImageNet pictures to the evolved stimulus, pooled across all experiments for all 14 visually responsive sites in the array in monkey Ri.

(G) Same, but for images evolved by monkey Y1 recording sites, which preferred images of places.

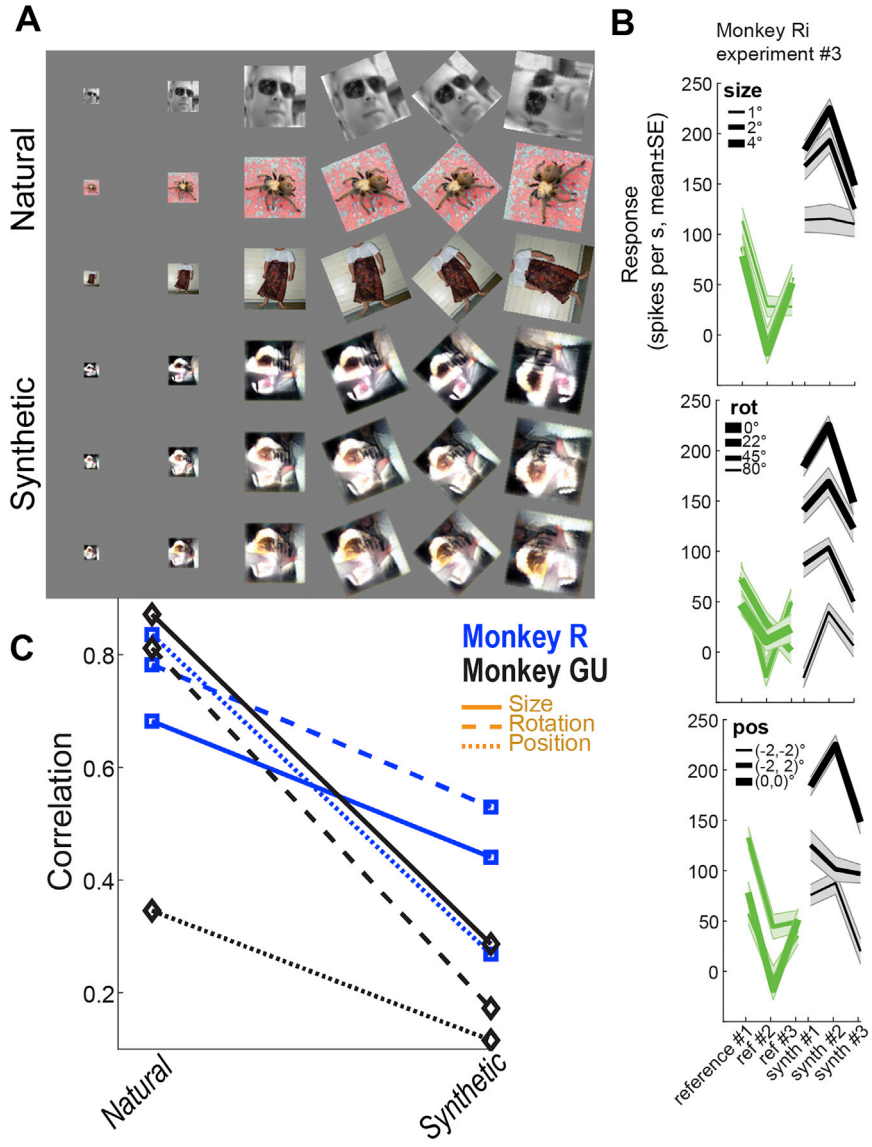


Figure S6. Size, Rotation, and Position Invariance to Natural versus Evolved Images, Related to Figure 7 and Invariance to evolved versus natural images

(A) Transformations applied to natural and synthetic images evolved by PIT units (the natural images were nearest, middle, and farthest fc6 matches to the evolved image, as defined in Figure 7). Images varied in size, rotation and position.

(B) Responses (background subtracted) to six images (3 reference natural images, 3 synthetic images) as a function of (top) size, (center) rotation, and (bottom) position. Each plot shows the mean response (±SEM) as a function of one transformation. Line thickness indicates transformation value.

(C) Correlation of rank orders for natural versus synthetic images across transformations and in two monkeys.

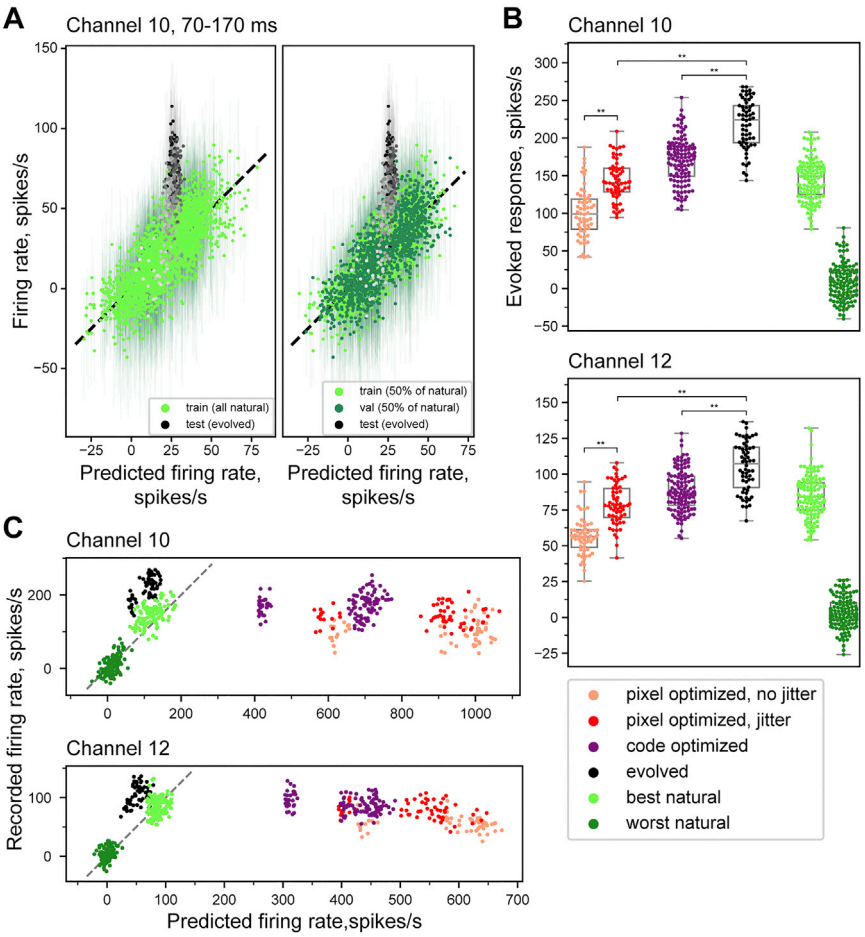


Figure S7. Comparison of XDREAM and Substitute Model-Based Approaches for Making Strong Stimuli, Related to Figure 2 and Discussion

An alternative approach for creating strong stimuli for neurons is to fit a ConvNet-based substitute model on neuronal responses to a reference image set (Yamins et al., 2014), then use gradient-based optimization on the model to predict strong stimuli for the neuron. This approach has been successfully used in V1 and V4 by various groups (Abbasi-Asl et al., 2018; Bashivan et al., 2018; Walker et al., 2018).

To directly compare XDREAM with a substitute model-based approach, we performed two analyses. First, we tested whether a ConvNet model can predict neuron responses to natural and evolved stimuli. We presented ~2,500 natural images and 244 images evolved by a PIT neuron for at least 10 repetitions each, then fit a linear regression from CaffeNet fc6 activations to image-averaged neuron responses. (A) On the left are all natural images are used in fitting the model. On the right, half of the natural images are used to fit the model and the other half used to evaluate it. Although this model was able to predict neuronal firing rate to held-out natural images, it underestimated neuronal firing rates to evolved images, indicating that evolved images contain feature to which the model did not generalize. We repeated the above analyses with another neuron and with other network layers and reached the same conclusion (Table S6). Diagonal, the identity line. Thin lines, standard deviation of neuron responses over repeated image presentations. Grayscale coloring indicates the generation of the evolved stimuli.

Second, we tested whether a ConvNet model can predict better stimuli for the neuron. We performed 4 experiments where we evolved stimuli for two sites in monkey Ri (10 and 12) while simultaneously collecting neuronal responses to natural images. Using these natural image responses, we fit substitute models consisting of AlexNet fc6 units separately for each neuron and each experiment. Based on the models, we generated images predicted to elicit high responses using backpropagation optimization with 3 settings: optimizing directly in pixel space with (red), and without (pink), jitter-robustness (Mordvintsev et al., 2015), and optimizing in the code space of the generative neural network (purple) (Nguyen et al., 2016). We then presented to the monkey the substitute model-optimized images, evolved images (black), and best and worst natural images (green) (1,418 images in total) in a fifth experiment collecting responses from the same two sites. (B) Beeswarm- and box-plot of responses to images by category. Each dot in the beeswarm plot indicates average evoked neuronal response to one image (25–36 trials per image; firing rate between 50–200 ms minus background activity between 1–40 ms). Evolved images were significantly better than images optimized for the substitute model by either optimizing at pixel level or optimizing in the input space of the generative network. Jitter robustness improved substitute-model-based pixel-level optimization as judged by the neuron. (C) Model-predicted responses and actual neuron responses. The model explains responses to natural images reasonably well but diverges from actual neuronal responses in both underpredicting responses to evolved images and overpredicting responses to model-optimized stimuli. **$p < 0.001$. The dashed line indicates identity.