RESEARCH ARTICLE SUMMARY

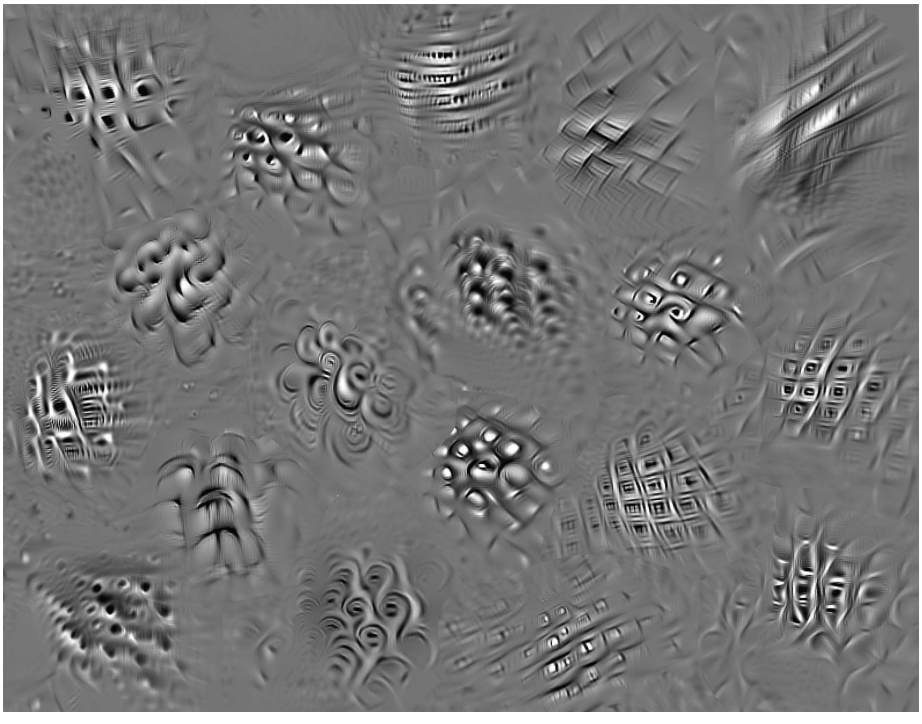
NEUROSCIENCE

Neural population control via deep image synthesis

Pouya Bashivan*, Kohitij Kar*, James J. DiCarlo†

INTRODUCTION: The pattern of light that strikes the eyes is processed and re-represented via patterns of neural activity in a "deep" series of six interconnected cortical brain areas called the ventral visual stream. Visual neuroscience research has revealed that these patterns of neural activity underlie our ability to recognize objects and their relationships in the world. Recent advances have enabled neuroscientists to build ever more precise models of this complex visual processing. Currently, the best such models are particular deep artificial neural network (ANN) models in which each brain area has a corresponding model layer and each brain neuron has a corresponding model neuron. Such models are quite good at predicting the responses of brain neurons, but their contribution to an understanding of primate visual processing remains controversial.

RATIONALE: These ANN models have at least two potential limitations. First, because they aim to be high-fidelity computerized copies of the brain, the total set of computations performed by these models is difficult for humans to comprehend in detail. In that sense, each model seems like a "black box," and it is unclear what form of understanding has been achieved. Second, the generalization ability of these models has been questioned because they have only been tested on visual stimuli that are similar to those used to "teach" the models. Our goal was to assess both of these potential limitations through nonhuman primate neurophysiology experiments in a mid-level visual brain area. We sought to answer two questions: (i) Despite these ANN models' opacity to simple "understanding," is the knowledge embedded in them already useful for a potential application (i.e., neural activity control)? (ii) Do these models accurately predict brain responses to novel images?



Collection of images synthesized by a deep neural network model to control the activity of neural populations in primate cortical area V4. We used a deep artificial neural network to control the activity pattern of a population of neurons in cortical area V4 of macaque monkeys by synthesizing visual stimuli that, when applied to the subject's retinae, successfully induced the experimenter-desired neural response patterns.

RESULTS: We conducted several closed-loop neurophysiology experiments: After matching model neurons to each of the recorded brain neural sites, we used the model to synthesize entirely novel "controller" images based on the model's implicit knowledge of how the ventral visual stream works. We then presented those images to each subject to test the model's ability to control the subject's neurons. In one test, we asked the model to try to control each brain neuron so strongly as to activate it beyond its typically observed maximal activation level. We found that the model-generated synthetic stimuli successfully drove 68% of neural sites beyond their naturally observed activation levels (chance level is 1%). In an even more stringent test, the model revealed that it is capable of selectively controlling an entire neural subpopulation, activating a particular neuron while simultaneously inactivating the other recorded neurons (76% success rate; chance is 1%).

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aav9436

Next, we used these non-natural synthetic controller images to ask whether the model's ability to predict the brain responses would hold up for these highly novel images. We found that the model was indeed quite accurate, predicting 54% of the image-evoked patterns of brain response (chance level is 0%), but it is clearly not yet perfect.

CONCLUSION: Even though the nonlinear computations of deep ANN models of visual processing are difficult to accurately summarize in a few words, they nonetheless provide a shareable way to embed collective knowledge of visual processing, and they can be refined by new knowledge. Our results demonstrate that the currently embedded knowledge already has potential application value (neural control) and that these models can partially generalize outside the world in which they "grew up." Our results also show that these models are not yet perfect and that more accurate ANN models would produce even more precise neural control. Such noninvasive neural control is not only a potentially powerful tool in the hands of neuroscientists but also could lead to a new class of therapeutic applications. ■

Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, and Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: dicarlo@mit.edu
Cite this article as P. Bashivan *et al.*, *Science* **364**, eaav9436 (2019). DOI: 10.1126/science.aav9436

RESEARCH ARTICLE

NEUROSCIENCE

Neural population control via deep image synthesis

Pouya Bashivan*, Kohitij Kar*, James J. DiCarlo†

Particular deep artificial neural networks (ANNs) are today's most accurate models of the primate brain's ventral visual stream. Using an ANN-driven image synthesis method, we found that luminous power patterns (i.e., images) can be applied to primate retinae to predictably push the spiking activity of targeted V4 neural sites beyond naturally occurring levels. This method, although not yet perfect, achieves unprecedented independent control of the activity state of entire populations of V4 neural sites, even those with overlapping receptive fields. These results show how the knowledge embedded in today's ANN models might be used to noninvasively set desired internal brain states at neuron-level resolution, and suggest that more accurate ANN models would produce even more accurate control.

Particular deep feedforward artificial neural network models (ANNs) constitute today's most accurate "understanding" of the initial ~200 ms of processing in the primate ventral visual stream and the core object recognition behavior it supports [see (*1*) for the currently leading models]. In particular, visually evoked internal "neural" representations of these specific ANNs are remarkably similar to the visually evoked neural representations in mid-level (area V4) and high-level (inferior temporal) cortical stages of the ventral stream (*2*, *3*)—a finding that has been extended to neural representations in visual area V1 (*4*), to patterns of behavioral performance in core object recognition tasks (*5*, *6*), and to both magnetoencephalography and functional magnetic resonance imaging (fMRI) measurements from the human ventral visual stream (*7*, *8*). Notably, these prior findings of model-to-brain similarity were not curve fits to brain data; they were predictions evaluated using images not previously seen by the ANN models. This has been construed as evidence that these models demonstrate some generalization of their ability to capture key functional properties of the ventral visual stream.

However, at least two important potential limitations of this claim have been raised. First, because the visual processing that is executed by the models is not simple to describe, and because the models have only been evaluated in terms of internal functional similarity to the brain, perhaps they are more like a copy of, rather than a useful "understanding" of, the ventral stream. Second, because the images to assess similarity were sampled from the same distribution as that used to set the model's internal parameters (photograph and rendered object databases), it is unclear whether these models would pass a stronger test of functional similarity—specifically, whether that similarity would generalize to entirely novel images. Perhaps the models' reported apparent functional similarity to the brain (*3*, *7*, *9*) substantially overestimates their true functional similarity.

We conducted a set of nonhuman primate visual neurophysiology experiments to assess the first potential limitation by asking whether the detailed knowledge that the models contain is useful for one potential application (neural activity control) and to assess the second potential limitation by asking whether the functional similarity of the model to the brain generalizes to entirely novel images. Specifically, we used one of the leading deep ANN ventral stream models (i.e., a specific model with a fully fixed set of parameters) to synthesize new patterns of luminous power ("controller images") that, when applied to the retinae, were intended to control the neural firing activity of particular, experimenter-chosen neural sites in cortical visual area V4 of macaques in two settings: (i) neural "stretch," in which synthesized images stretch the maximal firing rate of any single targeted neural site well beyond its naturally occurring maximal rate, and (ii) neural population state control, in which synthesized images independently control every neural site in a small recorded population (here, populations of 5 to 40 neural sites). We tested that population control by aiming to use such model-designed retinal inputs to drive the V4 population into an experimenter-chosen "one-hot" state in which one neural site is pushed to be highly active while all other nearby sites are simultaneously "clamped" at their baseline activation level. We reasoned that successful experimenter control would demonstrate that at least one ANN model can be used to noninvasively control the brain—a practical test of useful, causal "understanding" (*10*, *11*).

We used chronic implanted microelectrode arrays to record the responses of 107 neural multi-unit and single-unit sites from visual area V4 in three awake, fixating rhesus macaques designated as monkeys M, N, and S (n_M = 52, n_N = 33, n_S = 22). We first determined the classical receptive field (cRF) of each site with briefly presented small squares (see methods). We then tested each site using a set of 640 naturalistic images (always presented to cover the central 8° of the visual field that overlapped with the estimated cRFs of all the recorded V4 sites), as well as a set of 370 complex-curvature stimuli previously determined to be good drivers of V4 neurons (*12*) (location-tuned for the cRFs of the neural sites). Using each site's visually evoked responses (see methods) to 90% of the naturalistic images (n = 576), we created a mapping from a single "V4" layer of a deep ANN model (*13*) (the Conv-3 layer, which we had established in prior work) to the neural responses. We selected the model layer that maximally predicted the area V4 responses to the set of naturalistic images using linear mapping with two-fold cross-validation (this model layer selection was also consistent with similarity analysis using a representational dissimilarity matrix; see methods and fig. S8). The predictive accuracy of this model-to-brain mapping has previously been used as a measure of the functional fidelity of the brain model to the brain (*1*, *3*). Indeed, using the V4 responses to the held-out 10% of the naturalistic images as tests, we replicated and extended that prior work. We found that the neural predictor models correctly predicted 89% of the explainable (i.e., image-driven) variance in the V4 neural responses (median over the 107 sites, each site computed as the mean over two mapping/testing splits of the data; see methods).

Besides generating a model V4–to–brain V4 similarity score (89% in this case), this mapping procedure produces a potentially powerful tool: an image-computable predictor model of the visually evoked firing rate of each of the V4 neural sites. If truly accurate, this predictor model is not simply a data-fitting device and not just a similarity scoring method; instead, it must implicitly capture a great deal of visual knowledge that may be difficult to express in human language but is hypothesized (by the model) to be used by the brain to achieve successful visual behavior. To extract and deploy that knowledge, we used a model-driven image synthesis algorithm (see Fig. 1 and methods) to generate controller images that were customized for each neural site (i.e., according to its predictor model) so that each image should predictably and reproducibly control the firing rates of V4 neurons in a particular, experimenter-chosen way. That is, we aimed to test the hypothesis that experimenter-delivered application of a particular pattern of luminous power on the retinae will reliably and reproducibly cause V4 neurons to move to a particular, experimenter-specified activity state (and that the removal of this pattern of luminous power will return those V4 neurons to their background firing rates).

Department of Brain and Cognitive Sciences, McGovern Institute for Brain Research, and Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: dicarlo@mit.edu

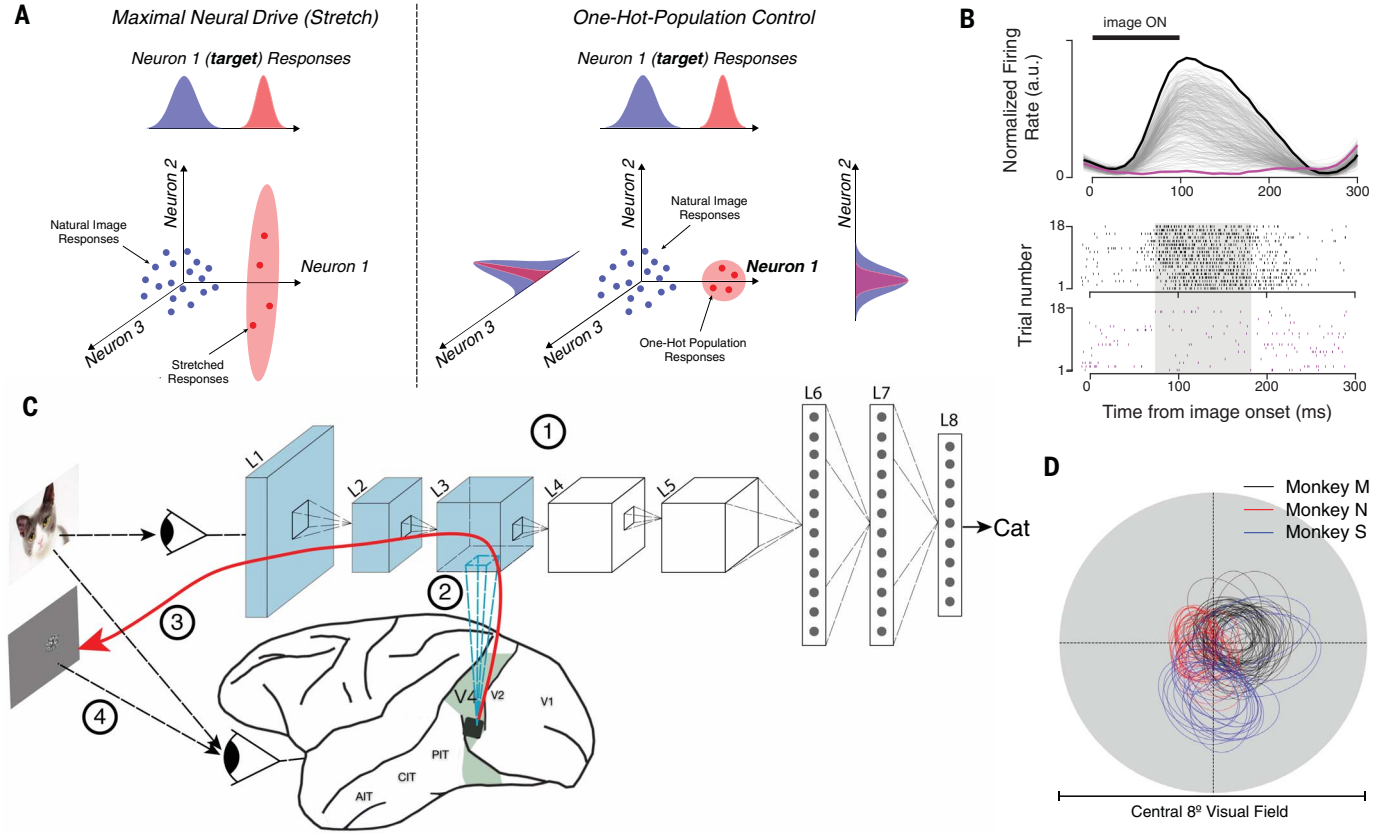


Fig. 1. Overview of the synthesis procedure. (**A**) Schematic illustration of the two tested control scenarios. Left: The controller algorithm synthesizes novel images that it believes will maximally drive the firing rate of a target neural site (stretch). In this case, the controller algorithm does not attempt to regulate the activity of other measured neurons (e.g., they might also increase as shown). Right: The controller algorithm synthesizes images that it believes will maximally drive the firing rate of a target neural site while suppressing the activity of other measured neural sites (one-hot population). (**B**) Top: Responses of a single example V4 neural site to 640 naturalistic images (averaged over ~40 repetitions for each image) are represented by overlapping gray lines; black line at upper left denotes the image presentation period. Bottom: Raster plots of highest and lowest neural responses to naturalistic images, corresponding to the black and purple lines in the top panel, respectively. The shaded area indicates the time window over which the activity level of each V4 neural site is computed (i.e., one value per image for each neural site). (**C**) The neural control experiments are done in four steps: (1) Parameters of the neural network are optimized by training on a large set of labeled natural images [Imagenet (*35*)] and then held constant thereafter. (2) ANN "neurons" are mapped to each recorded V4 neural site. The mapping function constitutes an image-computable predictive model of the activity of each of these V4 sites. (3) The resulting differentiable model is then used to synthesize "controller" images for either single-site or population control. (4) The luminous power patterns specified by these images are then applied by the experimenter to the subject's retinae, and the degree of control of the neural sites is measured. AIT, anterior inferior temporal cortex; CIT, central inferior temporal cortex; PIT, posterior inferior temporal cortex. (**D**) Classical receptive fields of neural sites in monkey M (black), monkey N (red), and monkey S (blue; see methods).

Although there are an extremely large number of possible neural activity states that an experimenter might ask a controller method to try to achieve, we restricted our experiments to the V4 spiking activity 70 to 170 ms after retinal power input (the time frame where the ANN models are presumed to be most accurate), and we have thus far tested two control settings: stretch control and one-hot-population control (see below). To test and quantify the goodness of control, we applied patterns of luminous power specified by the synthesized controller images to the retinae of the animal subjects while we recorded the responses of the same V4 neural sites (see methods).

Each experimental manipulation of the pattern of luminous power on the retinae is colloquially referred to as "presentation of an image." However, here we state the precise manipulation of applied power that is under experimenter control and fully randomized with other applied luminous power patterns (other images) to emphasize that this is logically identical to more direct energy application (e.g., optogenetic experiments) in that the goodness of experimental control is inferred from the correlation between power manipulation and the neural response in exactly the same way in both cases [see (*11*) for review]. The only difference between the two approaches is the assumed mechanisms that intervene between the experimentally controlled power and the controlled dependent variable (here, V4 spiking rate). These are steps that the ANN model aims to approximate with stacked synaptic sums, threshold nonlinearities, and normalization circuits. In both the control cases presented here and the optogenetics control case, these intervening steps are not fully known but are approximated by a model of some type; that is, neither experiment is "only correlational" because causality is inferred from experimenter-delivered, experimenter-randomized application of power to the system.

Because each experiment was performed over separate days of recording (1 day to build all the predictor models, 1 day to test control), only neural sites that maintained both a high signal-to-noise ratio and a consistent rank order of responses to a standard set of 25 naturalistic images across the two experimental days were considered further (n_M = 38, n_N = 19, and n_S = 19 for stretch experiments; n_M = 38 and n_S = 19 for one-hot-population experiments; see methods).

Stretch control: Attempt to maximize the activity of individual V4 neural sites

We first defined each V4 site's "naturally observed maximal firing rate" as that which was

found by testing its response to the best of the 640 naturalistic test images (cross-validated over repeated presentations; see methods). We then generated synthetic controller images for which the synthesis algorithm was instructed to drive the firing rate of one of the neural sites as high as possible beyond that rate, regardless of the other V4 neural sites. For our first stretch control experiment, we restricted the synthesis algorithm to operate only on parts of the image that were within the cRF of each neural site. For each target neural site (n_M = 21, n_N = 19, and n_S = 19), we ran the synthesis algorithm from five different random image initializations. For 79% of neural sites, the synthesis algorithm successfully found at least one image that it predicted to be at least 10% above the site's naturally observed maximal firing rate (see methods). However, in the interest of presenting an unbiased estimate of the stretch control goodness for randomly sampled V4 neural sites, we included all sites in our analyses, even those (~20%) that the control algorithm predicted that it could not stretch. Visual inspection suggested that the five stretch controller images generated by the algorithm for each neural site are perceptually more similar to each other than to those generated for different neural sites (see Fig. 2 and fig. S1), but we did not psychophysically quantify that similarity.

An example of the results of applying the stretch control images to the retinae of one monkey to target one of its V4 sites is shown in Fig. 2A, along with the ANN model–predicted responses of this site for all tested images. A closer visual inspection of this neural site's "best" natural and complex-curvature images within the site's cRF (Fig. 2A, top) suggests that it might be especially sensitive to the presence of an angled convex curvature in the middle and a set of concentric circles at the lower left side. This is consistent with extensive systematic work in V4 using such stimuli (*12*, *14*), and it suggests that we had successfully located the cRF and tuned our stimulus presentation to maximize the firing rate by the standards of such prior work. Interestingly, however, we found that all five synthetic stretch control images (red) drove the neural responses above the response to every tested naturalistic image (blue) and above the response to every complex-curvature stimulus presented within the cRF (purple) (Fig. 2A).

To quantify the goodness of this stretch control, we measured the neural response to the best of the five synthetic images (again, cross-validated over repeated presentations; see methods) and compared it with the naturally observed maximal firing rate (defined above). We found that the stretch controller images successfully drove 68% of the V4 neural sites (40 of 59) statistically beyond its maximal naturally observed firing rate ($P < 0.01$, unpaired-samples t test between distribution of highest firing rates for naturalistic and synthetic images; distribution generated from 50 random cross-validation samples; see methods). Measured as an amplitude, we found that the stretch controller images typically produced a firing rate that was 39% higher than the maximal naturalistic firing rate (median over all tested sites; Fig. 2, B and C).

Because our fixed set of naturalistic images was not optimized to maximally drive each V4 neural site, we considered the possibility that our stretch controller was simply rediscovering image pixel arrangements that are already known from prior systematic work to be good drivers of V4 neurons (*12*, *14*). To test this hypothesis, we tested 19 of the V4 sites (n_M = 11, n_S = 8) by presenting, inside the cRF of each neural site, each of 370 complex-curvature shapes (*14*)—a stimulus set that has been previously shown to contain image features that are good at driving V4 neurons when placed within the cRF. Because we were also concerned that the fixed set of naturalistic images did not maximize the local image contrast within each V4 neuron's cRF, we presented the complex-curvature shapes at a contrast that was matched to the contrast of the synthetic stretch controller images (fig. S4). Interestingly, we found that for each tested neural site, the synthetic controller images generated higher firing rates than the most effective complex-curvature shape (Fig. 2D). Specifically, when we used the maximal response over all the complex-curvature shapes as the reference (again, cross-validated over repeated presentations), we found that the median stretch amplitude was even larger (187%) than when the maximal naturalistic image was used as the reference (73% for the same 19 sites). In sum, the ANN-driven stretch controller had discovered pixel arrangements that were better drivers of V4 neural sites than prior systematic attempts to do so.

To further test the possibility that the relatively simple image transformations might also achieve neural response levels that were as high as those elicited by the synthetic controller images, we carried out extensive simulations to test the predicted effects of a battery of alternative image manipulations. First, to investigate whether the response might be increased simply by reducing surround suppression effects (*15*), we assessed each site's predicted response to its best naturalistic image, spatially cropped to match the site's cRF. We also adjusted the contrast of that cropped image to match the average contrast of the synthetic images for the site (also measured within the site's cRF). Over all tested sites, the predicted median stretch control gain achieved using these newly generated images was 14% lower than the original naturalistic set (n = 59 sites; see fig. S7). To explore this further, we optimized the size and location of the cropped region of the natural image (see methods). The stretch control gain achieved with this procedure was 0.1% lower than that obtained for the original naturalistic images. Second, we tested response-optimized affine transformations of the best naturalistic images (position, scale, rotations). Third, to place some energy from multiple features of natural images in the cRF, we tested contrast blends of the best two to five images for each site (see methods). The predicted stretch control gain of each of these manipulations was still far below that achieved with the synthetic controller images. In summary, we found that the achieved stretch control ability is nontrivial, in that even at high contrast, it cannot be achieved by complex-curvature features, simple transformation on naturalistic images, combination of good naturalistic images, or optimization of the spatial extent of the image (see methods and fig. S7).

One-hot-population control: Attempt to activate only one of many V4 neural sites

Similar to prior single-unit visual neurophysiology studies (*16*–*18*), the stretch control experiments attempted to optimize the response of each V4 neural site individually, without regard to the rest of the neural population. But the ANN model potentially enables much richer forms of population control in which each neural site might be independently controlled. As a first test of this, we asked the synthesis algorithm to try to generate controller images with the goal of driving the response of only one "target" neural site high while simultaneously keeping the responses of all other recorded neural sites low (i.e., a one-hot-population activity state; see methods).

We attempted this one-hot-population control on neural populations in which all sites were simultaneously recorded (experiment 1, n = 38 in monkey M; experiment 2, n = 19 in monkey S). Specifically, we randomly chose a subset of neural sites as "target" sites (14 in monkey M, 19 in monkey S) and we asked the synthesis algorithm to generate five one-hot-population controller images for each of these sites (i.e., 33 tests in which each test is an attempt to maximize the activity of one site while suppressing the activity of all other measured sites from the same monkey). For these control tests, we allowed the controller algorithm to optimize pixels over the entire 8° diameter image (which included the cRFs of all the recorded neural sites; see Fig. 3), and we then applied the one-hot-population controller images to the monkey retinae to assess the goodness of control. The synthesis procedure predicted a softmax score of at least 0.5 for 77% of population experiments (as a reference, the maximum softmax score is 1 and is obtained when only the target neural site is active and all off-target neural sites are completely inactive; see Fig. 3A for an example near 0.3).

Although the one-hot-population controller images did not achieve perfect one-hot-population control, we found that the controller images were typically able to achieve enhancements in the activity of the target site without generating much increase in off-target sites (relative to naturalistic images; see examples in Fig. 3A). To quantify the goodness of one-hot-population control in each of the 33 tests, we computed a one-hot-population score on the responses of the activity profile of each population (softmax score; see methods) and referenced that score to the one-hot-population control score that could be achieved using only the naturalistic images (i.e., without the benefit of the ANN model and synthesis algorithm). We took the ratio of these two scores as the measure of improved one-hot-population control, and we found that the

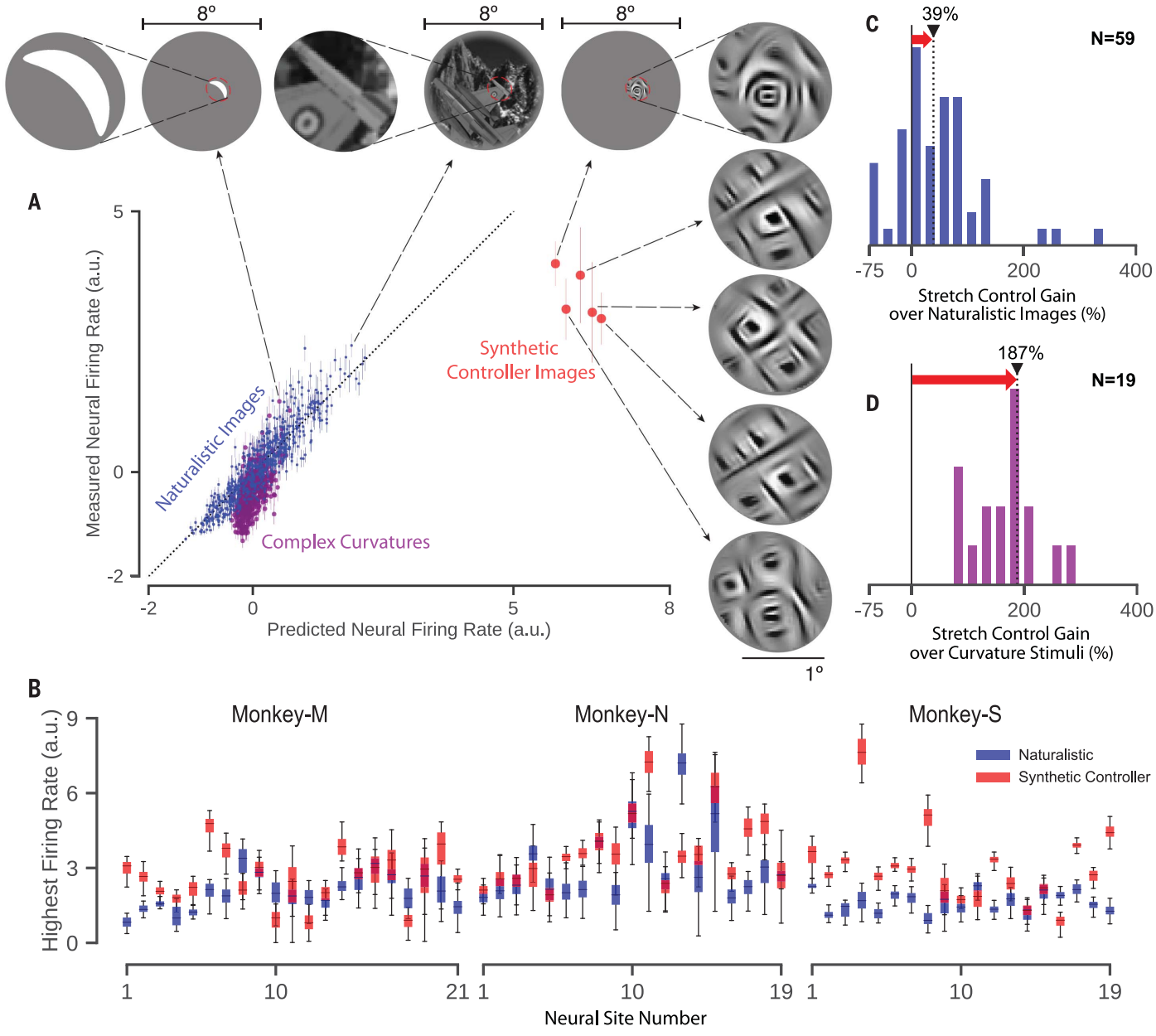


Fig. 2. Maximal drive of individual neural sites (stretch). (**A**) Results for an example successful stretch control test. Normalized activity level of the target V4 neural sites is shown for all of the naturalistic images (blue dots), complex-curvature stimuli (purple dots), and five synthetic stretch controller images (red dots; see methods). Best driving images within each category and a zoomed view of the receptive field are shown at the top. (**B**) Difference in firing rate in response to naturalistic (blue) and synthetic images (red) for each neural site in three monkeys. Controller image synthesis was restricted within the receptive field of the target neural site. Error bars denote range of the data. (**C**) Histogram of increase in the firing rate over naturalistic images for cRF-restricted synthetic images. (**D**) Histogram of increase in the firing rate over complex-curvature stimuli. Black triangle with dotted black line marks the median of the scores over all tested neural sites. Red arrow highlights the gain in firing rate in each experiment achieved by the controller images. *N* indicates the number of neural sites included in each experiment.

controller typically achieved an improvement of 57% (median over all 33 one-hot-population control tests; Fig. 3, B and C) and that this improved control was statistically significant for 76% of the one-hot-population control tests (25 of 33 tests; $P < 0.01$, unpaired-samples *t* test).

We considered the possibility that the improved population control was a result of the nonoverlapping cRFs that would allow neural sites to be independently controlled simply by restricting image contrast energy to each site's cRF. To test this possibility, we analyzed a subsample of the measured neural population in which all sites had strongly overlapping cRFs (Fig. 3D). We considered a neural population of size 10 in monkey M and of size 8 in monkey S for this experiment with largely overlapping cRFs (Fig. 3D). In total, we performed the experiment on 12 target neural sites in two monkeys (four in monkey M and eight in monkey S) and found that the amplitude of improved control was still 40% (Fig. 3D). Thus, a large portion of the improved control is the result of specific spatial arrangements of luminous power within the retinal input region shared by multiple V4 neural sites that the ANN model has implicitly captured and predicted and that the synthesis algorithm has successfully recovered (Fig. 4).

As another test of one-hot-population control, we conducted an additional set of experiments in which we restricted the one-hot control synthesis algorithm to operate only on image pixels within

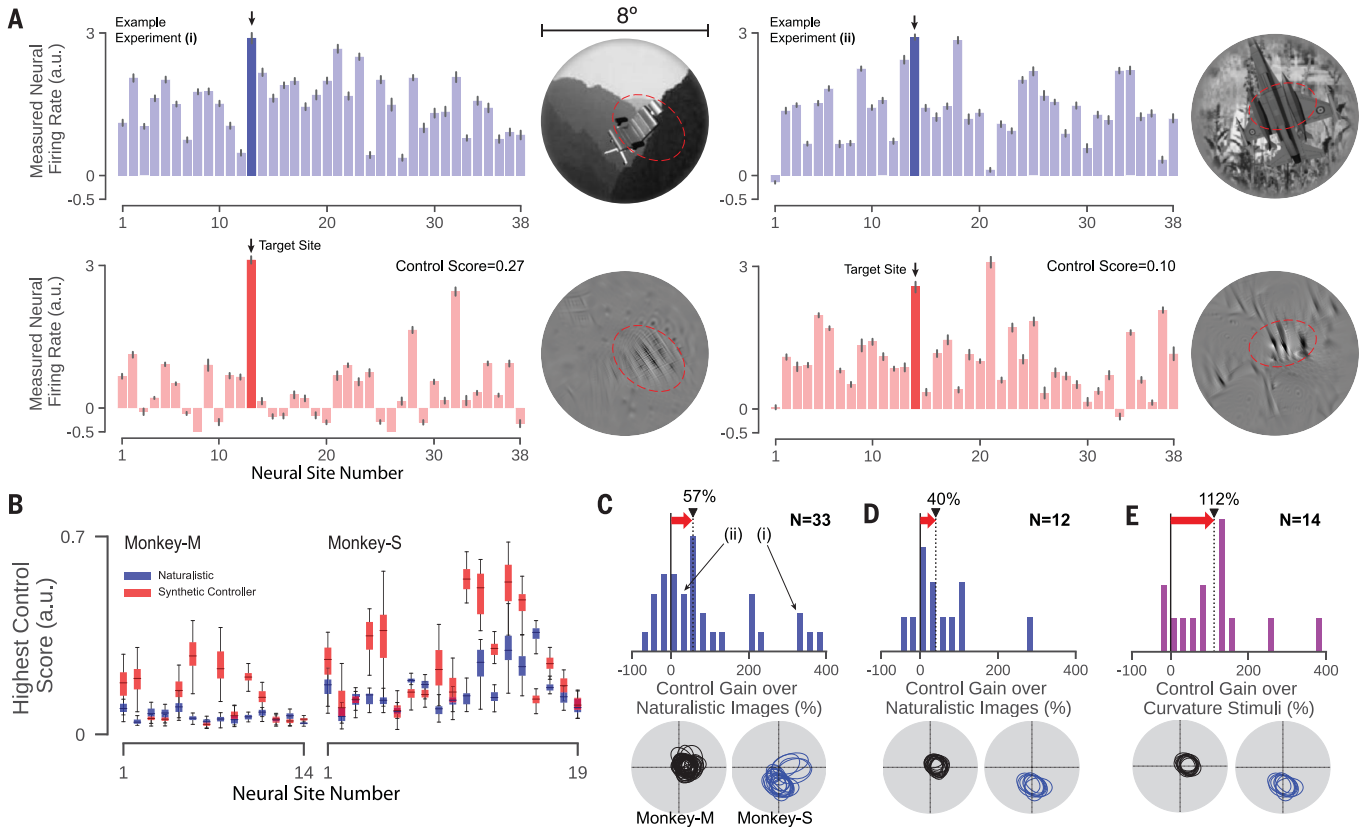


Fig. 3. Neural population control. We synthesized controller images that aimed to set the neural population in a one-hot state (OHP) in which one target neural site is active and all other recorded neural sites are suppressed. (**A**) Two example OHP experiments (left and right). In each case, the neural activity of each of the validated V4 sites (see methods) in the recorded population is plotted (most have overlapping cRFs), with the target V4 site (dark blue or red) indicated by an arrow. Note that responses are normalized individually on a normalizer image set to make side-by-side comparison of the responses meaningful (see methods). Top row: Activity pattern for the best ("best" in the sense of OHP control; see methods) naturalistic image (shown at right). Bottom row: Activity pattern produced by retinal application of the ANN model–synthesized controller image (shown at right). The red dashed oval marks the extended receptive field (2 SD) of each site. Error bars denote 95% confidence interval. (**B**) Distribution of control scores for best synthetic and naturalistic images for all 33 OHP full-image controller experiments (n_M = 14, n_S = 19). Control scores are computed using cross-validation (see methods). Error bars denote range of the data. (**C**) Histogram of OHP control gain (i.e., improvement over naturalistic images) for results in (B); (i) and (ii) indicate the scores corresponding to example experiments shown in (A). (**D**) Same experimental data as (C) except analyzed for subpopulations selected so that all sites have highly overlapping cRFs [see cRFs below (C) to (E); black, monkey M; blue: monkey S]. (**E**) OHP control gain, where gain is relative to the best complex-curvature stimulus in the shared cRF (see text) and the controller algorithm is also restricted to operate only in that shared cRF (n = 14 OHP experiments). In (C) to (E), N indicates the number of experiments in each setting; red arrow highlights the median gain in control (black triangle) achieved in each case.

the shared cRF of all neural sites in a subpopulation with overlapping cRFs (Fig. 3E). We compared this within-cRF synthetic one-hot-population control with the within-cRF one-hot-population control that could be achieved with the complex-curvature shapes (because the prior experiments with these stimuli were also designed to manipulate V4 responses only using pixels inside the cRF). We found that for the same set of neural sites, the synthetic controller images produced a very large one-hot-population control gain (median 112%; Fig. 3E) and the control score was significantly higher than the best complex-curvature stimulus for 86% of the neural sites (12 of 14).

Does the functional fidelity of the ANN brain model generalize to novel images?

Besides testing noninvasive causal neural control, these experiments also aimed to ask whether ANN models would pass a stronger test of functional similarity to the brain than prior work had shown (*2*, *3*)—specifically, whether this model-to-brain similarity would generalize to entirely novel images. Because the controller images were synthesized anew from random pixel arrangement and were optimized to drive the firing rates of V4 neural sites both upward (targets) and downward (one-hot-population off-targets), we considered them to be a potentially novel set of neural-modulating images that is far removed from the naturalistic images. We quantified and confirmed this notion of novelty by demonstrating that synthetic images were indeed statistically less similar to any of the naturalistic images than the naturalistic images were to themselves (measuring distances in pixel space, recorded V4 neural population space, and model-predicted V4 population space; see methods and fig. S6).

To determine how well the V4 predictor model generalizes to these novel synthetic images, for each neural site we compared the predicted response to every tested synthetic image with the actual neural response, using the same similarity measure as prior work (*2*, *3*), but now with zero parameters to fit. That is, a good model-to-brain similarity score required that the ANN predictor model for each V4 neural site accurately predict the response of that neural site for all of many synthetic images that are each very different from those that we used to train the ANN (photographs) and also very different from the images used to map ANN "V4" sites to individual V4 neural sites (naturalistic images).

Consistent with the control results (above), we found that the ANN model accounted for 54% of the explainable variance for the set of synthetic images (median over 76 neural sites in three monkeys; fig. S3). Although the model overestimated the neural responses to synthesized stimuli on many occasions and the model-to-brain

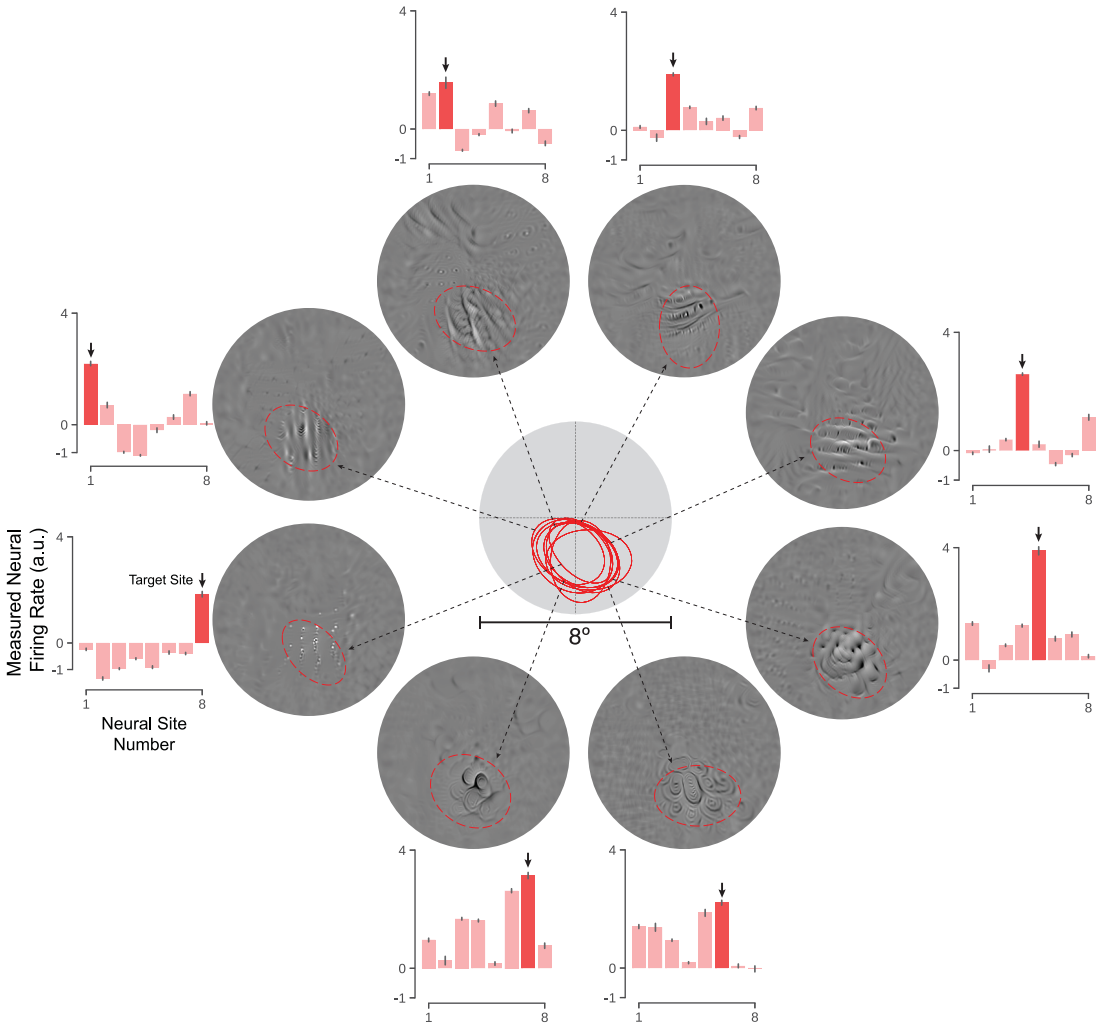


Fig. 4. Example of independent control of each neural site on a subset of V4 neural sites with highly overlapping cRFs. Controller images were synthesized to try to achieve a one-hot population over a population of eight neural sites (in each control test, the target neural site is shown in dark red and designated by an arrow). Despite highly overlapping receptive fields (center), most of the neural sites could be individually controlled to a reasonable degree. Controller images are shown along with the extended cRF (2 SD) of each site (red dashed ovals). Error bars denote 95% confidence interval.

similarity score was somewhat lower than that obtained for responses to naturalistic images (89%), the model still predicted a substantial portion of the variance, given that all parameters were fixed to make these "out of naturalistic domain" image predictions. We believe this to be the strongest test of generalization of today's ANN models of the ventral stream thus far, and it again shows that the model's internal neural representation is remarkably similar to the brain's intermediate ventral stream representation (V4), although it is still not a perfect model of the representation. We also note that because the synthetic images were generated by the model, we cannot assess the accuracy of predictions for images that are entirely "out of model domain."

How do we interpret these results?

Our results show that a deep ANN-driven controller method can be used to push the firing rates of most V4 neural sites beyond naturally occurring levels and that V4 neural sites with overlapping receptive fields can be partly—but not yet perfectly—independently controlled. In both cases, we show that the goodness of this control is unprecedented in that it is superior to that which can be obtained without the ANN. Finally, we find that with no parameter tuning at all, the ANN model generalizes moderately well (54%) to predict V4 responses to synthetic images that are strikingly different from the real-world photographs used to tune the ANN synaptic connectivity and map the ANN's "V4" to each V4 neural site. We believe that these results are the strongest test thus far of today's deep ANN models of the ventral stream.

Beginning with the work of Hubel and Wiesel (*19*, *20*), visual neuroscience research has closely equated an understanding of how the brain represents the external visual world with an understanding of what stimuli cause each neuron to respond the most. Indeed, textbooks and important recent results tell us that V1 neurons are tuned to oriented bars (*20*), V2 neurons are tuned to correlated combinations of V1 neurons found in natural images (*21*), V4 neurons are tuned to complex-curvature shapes in both two and three dimensions (*17*, *22*) and tuned to boundary information (*12*, *14*), and inferior temporal (IT) neurons respond to complex object-like patterns (*18*) including faces (*23*, *24*) and bodies as special cases (*25*).

Whereas these efforts have been essential to building both a solid foundation and intuitions about the role of neurons in encoding visual information, our results show how they can be further refined by current and future ANN models of the ventral stream. For instance, we found that synthesis of only a few images leads to higher neural response levels than was possible by searching in a relatively large space of natural images (n = 640) and complex-curvature stimuli (n = 370) derived from those prior intuitions. This shows that even today's ANN models—which are clearly not yet perfect (*1*, *6*)—already give us new ability to find manifolds of more optimal stimuli for each neural site at a much finer degree of granularity and to discover such stimuli unconstrained by human intuition and the limits of human language (see examples in fig. S1). This is likely to be especially important in middle and later stages of the visual hierarchy (e.g., in V4 and IT cortex), where the response complexity and larger receptive fields of neurons make manual search intractable.

In light of these results, what can we now say about the two important critiques of today's ANN models raised at the outset of this study

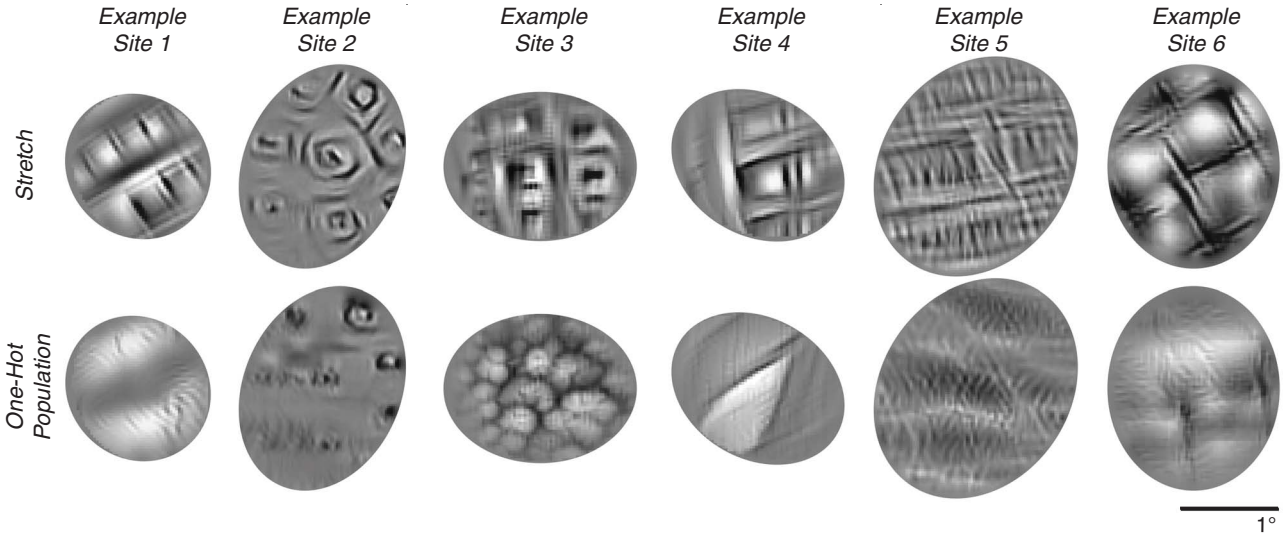


Fig. 5. Example controller images synthesized in stretch and one-hot population settings for six example target neural sites. Controller images were synthesized from the same initial random image but optimized for each target neural site and for each control goal (stretch or one-hot population; see text). Visual inspection suggests that for each target site, the one-hot-population control images contain only some aspects of the image features in the stretch images.

(understanding and generality)? In our view, the results strongly mitigate both of those critiques, but they do not eliminate them. An important test of understanding is the ability to use knowledge to gain improved control over things of interest in the world, as we have demonstrated; however, we acknowledge that this is not the only possible view, and many other notions of "understanding" remain to be explored to see whether and how these models add value. With respect to generality, we found that even today's ANN models show good generalization to demonstrably novel images, so we believe these results close the door on critiques that argue that current ANN models are extremely narrow in the scope of images they can accurately cover. However, we note that although 54% of the explainable variance in the generalization test was successfully predicted, this is somewhat lower than the 89% explainable variance that is found for images that are "closer" to (but not identical to) the mapping images. This not only reconfirms that these brain models are not yet perfect, but also suggests that a single metric of model similarity to each brain area is insufficient to characterize and distinguish among alternative models [e.g., (*1*)]. Instead, multiple similarity tests at different generalization "distances" could be useful, as we can imagine future models that show less decline in successfully predicted variance as one moves from testing images "near" the training and mapping distributions (typically photographs and naturalistic images) to "far" (such as the synthetic images like those used here) to "extremely far," such as images that cannot even be synthesized under the guidance of current models and thus remained untested here.

From an applications standpoint, the results presented here show how today's ANN models of the ventral stream can already be used to achieve improved noninvasive population control (e.g., Fig. 4). However, the control results are clearly not yet perfect. For example, in the one-hot population control setting, we were not able to fully suppress each and every one of the responses of the "off-target" neural sites while keeping the target neural site active (see examples in Figs. 3 and 4). Post hoc analysis showed that we could partially anticipate which off-target sites would be most difficult to suppress: They were typically (and not surprisingly) the sites that had high patterns of response similarity with the target site (r = 0.49, $P < 10^{-4}$; correlation between response similarity with the target neural site over naturalistic images and the off-target activity level in the full image one-hot population experiments; n = 37 off-target sites). Such results raise interesting scientific and applied questions of whether and when perfect independent control is possible at neuron-level resolution. Are our current limitations on control due to anatomical connectivity that restricts the potential population control, the nonperfect accuracy of the current ANN models of the ventral stream, nonperfect mapping of the model neurons to the individual neural site in the brain, the fact that we are attempting to control multi-unit activity, inadequacy of the controller image synthesis algorithm, or some combination of all of these and other factors?

Consider the synthesis algorithm: Intuitively, each particular neural site might be sensitive to many image features, but perhaps each site is only sensitive to a few features that the other neural sites are not sensitive to. This intuition is consistent with the observation that, using the current ANN model, it was more difficult for our synthesis algorithm to find good controller images in the one-hot-population setting than in the stretch setting (the one-hot-population optimization typically took more than twice as many steps to find a synthetic image that is predicted to drive the target neural site response to the same level as in the stretch setting), and visual inspection of the images suggests that the one-hot-population images have fewer identifiable "features" (Fig. 5 and fig. S2). As the size of the neural population to be controlled is increased, it would likely become increasingly difficult to achieve fully independent control, but this is an open experimental question.

Consider the current ANN models: Our data suggest that future improved ANN models are likely to enable even better control. For example, better ANN V4 population predictor models generally produced better one-hot-population control of that V4 population (fig. S5). One thing is clear already: Improved ANN models of the ventral visual stream have led to control of high-level neural population that was previously out of reach. With continuing improvement of the fidelity of ANN models of the ventral stream (*1*, *26*, *27*), the results presented here have likely only scratched the surface of what is possible with such implemented characterizations of the brain's neural networks.

Methods

Electrophysiological recordings in macaques

We sampled and recorded neural sites across the macaque V4 cortex in the left, right, and left hemisphere of three awake, behaving macaques, respectively. In each monkey, we implanted one chronic 96-electrode microelectrode array (Utah array), immediately anterior to the lunate sulcus (LS) and posterior to the inferior occipital sulcus (IOS), with the goal of targeting the central visual representation (<5° eccentricity, contralateral lower visual field). Each array sampled from ~25 mm^2 of dorsal V4. On each day, recording sites that were visually driven as measured by response correlation (Pearson r > 0.8) across

split-half trials of a fixed set of 25 out-of-set naturalistic images shown for every recording session (termed the normalizer image set) were deemed "reliable."

We do not assume that each V4 electrode was recording only the spikes of a single neuron. Hence, we use the term "neural site" throughout the manuscript. But we did require that the spiking responses obtained at each V4 site maintained stability in its image-wise "fingerprint" between the day(s) that the mapping images were tested (i.e., the response data used to build the ANN-driven predictive model of each site; see text) and the days that the controller images or the complex-curvature images were tested (see below). Specifically, to be "stable," we required an image-wise Pearson correlation of at least 0.8 in its responses to the normalizer set across recording days.

Neural sites that were reliable on the experimental mapping day and the experimental test days, and were stable across all those days, were termed "validated." All validated sites were included in all presented results. (To avoid any possible selection biases, this selection of validated sites was done on data that were completely independent from the main experimental result data.) In total, we recorded from 107 validated V4 sites during the ANN-mapping day, including 52, 33, and 22 sites in monkey M (left hemisphere), monkey N (right hemisphere), and monkey S (left hemisphere), respectively. Of these sites, 76 were validated for the stretch control experiments (n_M = 38, n_N = 19, n_S = 19) and 57 were validated for the one-hot-population control experiments (n_M = 38, n_S = 19).

To allow meaningful comparisons across recording days and across V4 sites, we normalized the raw spiking rate of each site from each recording session (within just that session) by subtracting its mean response to the 25 normalizer images and then dividing by the standard deviation of its response over those normalizer images (these are the arbitrary units shown as firing rates in Figs. 2A, 3A, and 4). The normalizer image set was always randomly interleaved with the main experimental stimulus set(s) run on each day.

Control experiments consisted of three steps. In the first step, we recorded neural responses to our set of naturalistic images that were used to construct the mapping function between the ANN activations and the recorded V4 sites. In a second, offline step, we used these mapping functions (i.e., a predictive model of the neural sites) to synthesize the controller images. Finally, in step 3, we closed the loop by recording the neural responses to the synthesized images. The time between step 1 and step 3 ranged from several days to 3 weeks.

Fixation task

All images were presented while monkeys fixated a white square dot (0.2°) for 300 ms to initiate a trial. We then presented a sequence of five to seven images, each ON for 100 ms followed by a 100-ms gray blank screen. This was followed by a water reward and an intertrial interval of 500 ms, followed by the next sequence. Trials were aborted if gaze was not held within ±0.5° of the central fixation dot during any point. To estimate the cRF of each neural site, we flashed 1° × 1° white squares across the central 8° of the monkeys' visual field, measured the corresponding neural responses, and then fitted a 2D Gaussian to the data. We defined 1 SD as the cRF of each site.

Naturalistic image set

We used a large set (N = 640) of naturalistic images to measure the response of each recorded V4 neural site and every model "V4" neural site to each of these images. Each of these images contained a 3D-rendered object instantiated at a random view overlaid on an unrelated natural image background; see (*28*) for details.

Complex-curvature stimuli

We used a set of images consisting of closed shapes constructed by combining concave and convex curves (*12*). These stimuli are constructed by parametrically defining the number and configuration of the convex projections that constituted the shapes. Previous experiments with these shapes showed that curvature and polar angle were quite good at describing the shape tuning (*12*). The number of projections varied from 3 to 5, and the angular separation between projections was in 45° increments. These shapes were previously shown to contain good drivers of V4 neurons of macaque monkeys (*12*, *14*). The complex-curvature images were generated using the code generously supplied by the authors of that prior work (http://depts.washington.edu/shapelab/resources/stimsonly.php). The stimuli were presented at the center of the receptive field of the neural sites (detailed below).

Cross-validation procedure for evaluating control scores

To evaluate the scores from the neural responses to an image set, we divided the neural response repetitions into two randomly selected halves. We then computed the mean firing rate of each neural site in response to each image in each half. The mean responses from the first half were used to find the image that produces the highest score (in that half) and the response to that image was then measured in the second half (this is the measurement used for further analyses). We repeated this procedure 50 times for each neural site (i.e., 50 random half splits). For stretch and one-hot-population experiments, the score functions were the "neural firing rate" and "softmax score," respectively. We computed each score for the synthetic controller images and for the reference images (either the naturalistic or the complex-curvature sets; see text). The synthetic "gain" in the control score is calculated as the difference between the synthetic controller score and the reference score, divided by the reference score.

V4 encoding model

To use the ANN model to predict each recorded neural site (or neural population), the internal "V4-like" representation of the model must first be mapped to the specific set of recorded neural sites. The assumptions behind this mapping are discussed elsewhere (*9*), but the key idea is that any good model of a ventral stream area must contain a set of artificial neurons ("features") that together span the same visual encoding space as the brain's population of neurons in that area (i.e., the model "layer" must match the brain area up to a linear mapping). To build this predictive map from model to brain, we started with a specific deep ANN model with locked parameters. Here we used a variant of Alexnet architecture trained on Imagenet (*13*), as we have previously found the feature space at the output of Conv-3 layer of Alexnet to be a good predictor of V4 neural responses (we refer to this as model "V4"). We used the same training procedure as described in (*13*), except that we did not split the middle convolutional layers between graphics processing units (GPUs).

In addition, the input images were transformed using an eccentricity-dependent function that mimics the known spatial sampling properties of the primate retinae (see below). We termed this the "retinae transformation." We had previously found that training deep convolutional ANN models with retinae-transformed images improves the neural prediction accuracy of V4 neural sites (an increase in explained variance by ~5 to 10%). The retinae transformation was implemented by a fisheye transformation that mimics the eccentricity-dependent sampling performed in primate retinae (code available at https://github.com/dicarlolab/retinawarp). All input images to the neural network were preprocessed by randomly cropping followed by applying the fisheye transformation. Parameters of the fisheye transformation were tuned to mimic the cones density ratio between the fovea and 4° peripheral vision (*29*).

We used the responses of the recorded V4 neural sites in each monkey and the responses of all the model "V4" neurons to build a mapping from the model to the recorded population of V4 neural sites (Fig. 1). We used a convolutional mapping function that significantly reduces the neural prediction error compared to other methods such as principal component regression. Our implementation was a variant of the two-stage convolutional mapping function proposed in (*30*) in which we substituted the group sparsity regularization term with an L2 loss term to allow for smooth (nonsparse) feature mixing. The first stage of the mapping function consists of a learnable spatial mask (W_s) that is parameterized separately for each neural site (n) and is used to estimate the receptive field of each neuron. The second stage consists of a mixing pointwise convolution (W_d) that computes a weighted sum of all feature maps at a particular layer of the ANN model (Conv3 layer in our case). The mixing stage finds the best combination of model features that are predictive of each neural site's response. The final output is then averaged over all spatial locations to form a scalar prediction of the neural response. Parameters are jointly optimized to minimize the prediction error $\mathcal{L}_e$ on

the training set regularized by combination of $\mathcal{L}_2$ and smoothing Laplacian losses $\mathcal{L}_{\text{Laplace}}$ (defined below). By factorizing the spatial and feature dimensions, this method significantly improves the predictivity of neural responses over the traditional principal component regression. We interpret this improved predictive power as resulting from the fact that it imposes a prior on the model-to-brain mapping procedure that is strongly in line with an empirical fact: that each neuron in area V4 has a receptive field. That neuron is thus best explained by linear combinations of simulated neurons that have similar receptive fields.

$$\hat{y}_n = \left[\sum \left(W_{\text{s}}^{(n)} \cdot X\right)\right] * W_{\text{d}}^{(n)} + W_{\text{b}}^{(n)} \quad (1)$$

$$\mathcal{L}_2 = \lambda_{\text{s}} \sum_n {W_{\text{s}}^{(n)}}^2 + \lambda_{\text{d}} \sum_n {W_{\text{d}}^{(n)}}^2 \quad (2)$$

$$\mathcal{L}_{\text{Laplace}} = \lambda_{\text{s}} \sqrt{\sum_n \left(W_{\text{s}}^{(n)} * L\right)^2}, \; L = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \quad (3)$$

$$\mathcal{L}_{\text{e}} = \sqrt{\sum_n (\hat{y}_n - y)^2} \quad (4)$$

$$\mathcal{L} = \mathcal{L}_{\text{e}} + \mathcal{L}_{\text{Laplace}} + \mathcal{L}_2 \quad (5)$$

We evaluated our model using two-fold cross-validation and observed that ~89% of the explainable variance could be explained with our model in three monkeys (EV_{M} = 92%, EV_{N} = 92%, EV_{S} = 80%). The addition of the retinae transformation together with the convolutional mapping function increased the explained variance by ~13% over the naive principal component regression applied on features from the model trained without the retinae transformation (EV_{M} = 75%, EV_{N} = 80%, EV_{S} = 73%). Ablation studies on data from each monkey suggested that on average about 3 to 8% of the improvements were due to the addition of the retinae transformation (see table S1). For constructing the final mapping function, adopted for image synthesis, we optimized the mapping function parameters on 90% of the data, selected randomly.

The resulting predictive model of V4 (ANN features plus linear mapping) is referred to as the mapped V4 encoding model and, by construction, it produces the same number of artificial V4 "neurons" as the number of recorded V4 neural sites (52, 33, and 22 neural sites in monkeys M, N, and S, respectively).

Retinae transformation

To retain the resolution of the retinae-transformed images as high as possible, we did not subsample the input image with a fixed sampling pattern. Instead, our implementation of the retinae sampling uses a backward function $r = g(r')$ that maps the radius of points in the retinae-transformed image (r') to those in the input image (r). In this way, for every pixel in the output image, we can find the corresponding pixel in the input image using the pixel-mapping function g. To formulate the pixel-mapping function g, we take advantage of the known rate of change of cones density (ρ) in the primate retinae that exponentially decreases with eccentricity (*29*):

$$\rho = \frac{1}{\pi d^2} = \exp(-ar') \quad (6)$$

where d is the distance between nearby cones and r' is the radial distance from the fovea in the transformed image. From this, we can write d as a function of r':

$$d = \frac{1}{\sqrt{\pi}} \exp\left(\frac{ar'}{2}\right) \quad (7)$$

The ratio between the cones density in the fovea and the outmost periphery, given the specific visual field size in which the stimulus has been presented in the experiment, can be written as

$$\frac{\rho_{\text{f}}}{\rho_{\text{p}}} = \exp(ar'_{\max}) \quad (8)$$

where ρ_{f} and ρ_{p} are the cone densities at the fovea and periphery, respectively, and $r'_{\max}$ is the highest radial distance in the output image (e.g., 150 for an image of size 300). From Eq. 8, we can calculate a as a function of ρ_{f}, ρ_{p}, and $r'_{\max}$:

$$a = \frac{\ln(\rho_{\text{f}}/\rho_{\text{p}})}{r'_{\max}} \quad (9)$$

The $\rho_{\text{f}}/\rho_{\text{p}}$ ratio is known given the size of the visual field in which the stimuli were presented (e.g., 10 for fovea to 4° in this study) and the output image size (e.g., 300 in this study). We can now formulate the function $g(r')$ as the sum of all the distances d up to radius r' weighted by a factor b:

$$g(r') = \frac{b}{\sqrt{\pi}} \sum_{k=0}^{r'-1} d_k = \frac{b}{\sqrt{\pi}} \sum_{k=0}^{r'-1} \exp\left(\frac{ar}{2}\right)$$

$$= \frac{b}{\sqrt{\pi}} \frac{1 - \exp\left(\frac{ar'}{2}\right)}{1 - \exp\left(\frac{a}{2}\right)} \quad (10)$$

where b is found so that $r_{\max}/g(r'_{\max}) = 1$. In our implementation we use Brent's method to find the optimal b value.

Finding the best representation in the ANN model

We used linear mapping from model features to neural measurements to compare the representation at each stage of processing in the ANN model. For features in each layer of the ANN model, we applied principal components analysis and extracted the top 640 dimensions. We then fitted a linear transformation to the data using Ridge regression method and computed the amount of explained variance (EV) by the mapping function. For each neural site we normalized the EV by the internal consistency of measurements across repetitions using two-fold cross-validation. The median normalized EV across all measured sites was used to select the best representation in the ANN model (fig. S8A). We also quantified the similarity of representations at each layer of the ANN model and the neural measurements using the image-level representational dissimilarity matrix (RDM) that followed the same pattern as that which was obtained from linear mapping method (fig. S8B). RDMs were computed using the principal components of the features at each layer in response to the naturalistic image set (n = 640).

Synthesized "controller" images

The "response" of an artificial neuron in the mapped V4 encoding model (above) is a differentiable function of the pixel values $f: \mathcal{I}^{w \times h \times c} \rightarrow \mathbb{R}$ that enables us to use the model to analyze the predicted sensitivity of neurons to patterns in the pixels space. We formulate the synthesis operation as an optimization procedure during which images are synthesized to control the neural firing patterns in the following two settings:

1) Stretch: We synthesized controller images that attempt to push each individual V4 neural site into its maximal activity state. To do so, we followed an approach first introduced in (*31*) and iteratively changed the pixel values in the direction of the gradient that maximizes the firing rate of the corresponding model V4 neural site. We repeated the procedure for each neural site using five different random starting images, thereby generating five stretch controller images for each V4 neural site.

2) One-hot population: Similar to the stretch scenario, except that here we chose the optimization to change the pixel values in a way that (i) attempts to maximize firing rate of the target V4 neural site, and (ii) attempts to maximally suppress the firing rates of all other recorded V4 neural sites. We formalized the one-hot population goal in the following objective function that we then aimed to maximize during the image synthesis procedure:

$$\mathcal{S} = \text{softmax}_t(y) = \frac{\exp(y_t)}{\sum_i \exp(y_i)} \quad (11)$$

where t is the index of the target neural site, and y_i is the response of the model V4 neuron i to the synthetic image.

For each optimization run, we started from an image that consists of random pixel values drawn from a standard normal distribution and optimized the objective function for a prespecified number of steps using a gradient ascent algorithm (steps = 700). We also used the total variation (defined below) as additional regularization in the optimization loss to reduce the high-frequency noise in the generated images:

$$L_{\text{TV}} = \sum_{i,j} (\|I_{i+1,j} - I_{i,j}\|_2 + \|I_{i,j+1} - I_{i,j}\|_2) \quad (12)$$

During the experiments, monkeys were required to fixate within a 1° circle at the center of the screen. This introduced an uncertainty on the exact gaze location. For this reason, images were

synthesized to be robust to small translations of maximum 0.5°. At every iteration, we translated the image in random directions (i.e., jittering) with a maximum translation length of 0.5° in each direction, thereby generating images that were predicted to elicit similarly high scores regardless of the translations within the range. The total variation loss and the translation invariance procedure reduced the amount of high-frequency noise patterns in the generated images commonly known as adversarial examples (*32*, *33*). In addition, at every iteration during the synthesis procedure, we normalized the computed gradients by its global norm and clipped the pixel values at −1 and 1.

Contrast energy

It has been shown that neurons in area V4 respond more strongly to higher contrast stimuli (*34*). To ask whether contrast energy (CE) was the main factor in stretching the V4 neural firing rates, we computed the contrast energy within the receptive field of the neural sites for all the synthetic, naturalistic, and classic V4 stimuli. Contrast energy was calculated as the ratio between the maximum and background luminances. For all images, the average luminance was used as the background value. Because the synthetic images consisted of complex visual patterns, we also computed the contrast energy using an alternative method based on spectral energy within the receptive field. We calculated the average power in the cRF in the frequency range of 1 to 30 cycles per degree. We ensured that for all tested neural sites, CE values within the cRF for synthetic stretch controller images were less than or equal to the classic, complex-curvature V4 stimuli (fig. S4).

cRF-cropped contrast-matched naturalistic stimuli

For each neural site, we first produced a new naturalistic image set by cropping the older naturalistic image set at the estimated cRF of the respective site. We then matched the contrast of these naturalistic images (within the cRF of that neuron) to the average contrast across all five synthesized images (generated for the same neural site). We then computed the predicted neural responses to all these new cRF-masked, contrast-matched naturalistic images and evaluated the stretch control gain achieved with this set over the original naturalistic images. The stretch control gain using these images showed a 14% decrease in the median gain over all target neurons. This meant that the original naturalistic image set without the cRF masking and contrast matching contained better drivers of the neural sites measured in our experiments. We noticed that masking the images with the estimated cRF was responsible for most of the drop in the observed stretch control gain (11%; see fig. S7). We also noted that the contrast energy within the cRF was higher for best naturalistic images compared to synthetic images for most sites (median ratio of synthetics contrast to best naturalistic images was 0.76 over all tested sites).

Monte Carlo mask optimization

We estimated the optimal mask parameters formulated as a 2D Gaussian function (i.e., μ, σ_1, σ_2, ρ) for each neural site via Monte Carlo simulations (n = 500). We sampled each parameter from the corresponding distribution derived from the measured neural sites in each monkey. For each Monte Carlo simulation, we sampled the mask parameters from the above-mentioned distributions and constructed a 2D mask. We then masked the naturalistic images with the sampled mask (cropped at 1 SD) and matched image contrasts to the average contrast of synthetic images produced for each neural site within the mask. For each neural site, we chose the optimal mask parameters that elicited the maximum average firing rate (predicted) across all images in the naturalistic set. The maximum predicted output for each neural site in response to these images was used to evaluate the stretch control gain that showed a nonsignificant gain over the naturalistic images.

Affine transformations of the naturalistic image set

There might be simple image transformations that could achieve the same level of control as that obtained by the synthetic images. To test this, we conducted an additional analysis in which we randomly transformed the best naturalistic image for each neural site using various affine transformations (i.e., translation, scale, and rotation; n = 100) and calculated the predicted responses to those images. We considered four experiments with the following transformations: (i) random scaling between 0.5 and 2, (ii) random translation between −25 and 25 pixels in each direction, (iii) random rotation between 0° and 90°, and (iv) mixture of all three transformations. For each experiment, we evaluated the stretch control gain over the naturalistic image set achieved with these new images that showed significantly lower gains for all of the alternative methods compared to our proposed model-based method (see fig. S7).

Combining best driver images

Images that are good drivers of the measured neurons could be combined together to form new mixed images that might drive the neurons even higher. To test this hypothesis, we combined the top naturalistic images for each neuron by taking the average pixel value over all select images and matched the contrast (within cRF of each neural site) of the mixed image to the average contrast across synthetic images generated for each neuron. We tried various numbers of top images to create the mixed image (i.e., top 2, 3, 4, and 5). We computed the predicted stretch control gain using these mixed images over the naturalistic image set and found that these images were considerably weaker drivers of the same neurons (see fig. S7).

Quantifying the novelty of synthetic images

We hypothesized that if the synthetic stimuli are indeed novel, they should be less similar (i.e., correlated) to any of the naturalistic images than the naturalistic images are to themselves. We computed the distances between synthetic and naturalistic images in pixel space as well as in the space of neural responses. To test this, we measured the minimum Euclidean distance (in the space of measured neural responses) between each synthetic image and all naturalistic images and compared them with minimum distances obtained for naturalistic images. Figure S6 shows the distribution of minimum distances of synthetic and naturalistic images to any naturalistic images and illustrates the point that the responses to synthetic images are significantly farther from the distribution of responses to naturalistic images than expected from sampling within the naturalistic space (fig. S6, A, C, E, and F) or by applying simple image transformations on images sampled from that space (fig. S6, B and D). Therefore, we can quantifiably call these images "out-of-domain" [Wilcoxon rank-sum test; $Z(3798) = 30.8$; $P < 0.0001$]. We also computed the distances between synthetic and naturalistic images in the pixel space using the correlation distance ($1 - \rho$) that showed a similar distinction between the two [Wilcoxon rank-sum test; $Z(37120) = 29.3$; $P < 0.0001$].

REFERENCES AND NOTES

1. M. Schrimpf *et al*., Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv [preprint]. 5 September 2018. pmid: 407007
2. D. L. K. Yamins, H. Hong, C. Cadieu, J. J. DiCarlo, in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges *et al*., Eds. (Neural Information Processing Systems Foundation, 2013); https://papers.nips.cc/paper/4991-hierarchical-modular-optimization-of-convolutional-networks-achieves-representations-similar-to-macaque-it-and-human-ventral-stream.
3. D. L. Yamins *et al*., Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014). doi: 10.1073/pnas.1403112111; pmid: 24812127
4. S. A. Cadena *et al*., Deep convolutional models improve predictions of macaque V1 responses to natural images. bioRxiv [preprint]. 11 October 2017. pmid: 201764
5. R. Rajalingham, K. Schmidt, J. J. DiCarlo, Comparison of Object Recognition Behavior in Human and Monkey. *J. Neurosci.* **35**, 12127–12136 (2015). doi: 10.1523/JNEUROSCI.0573-15.2015; pmid: 26338324
6. R. Rajalingham *et al*., Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018). doi: 10.1523/JNEUROSCI.0388-18.2018; pmid: 30006365
7. S. M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput. Biol.* **10**, e1003915 (2014). pmid: 25375136
8. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016). doi: 10.1038/srep27755; pmid: 27282108
9. D. L. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016). doi: 10.1038/nn.4244; pmid: 26906502
10. J. Pearl, *Causality* (Cambridge Univ. Press, 2009).
11. M. Jazayeri, A. Afraz, Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017). doi: 10.1016/j.neuron.2017.02.019; pmid: 28279349
12. A. Pasupathy, C. E. Connor, Shape representation in area V4: Position-specific tuning for boundary conformation. *J. Neurophysiol.* **86**, 2505–2519 (2001). doi: 10.1152/jn.2001.86.5.2505; pmid: 11698538

13. A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Advances in Neural Information Processing Systems 25*, F. Pereira *et al*., Eds. (Neural Information Processing Systems Foundation, 2012); https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.
14. A. Pasupathy, C. E. Connor, Population coding of shape in area V4. *Nat. Neurosci.* **5**, 1332–1338 (2002). doi: 10.1038/972; pmid: 12426571
15. J. R. Cavanaugh, W. Bair, J. A. Movshon, Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophysiol.* **88**, 2530–2546 (2002). doi: 10.1152/jn.00692.2001; pmid: 12424292
16. E. T. Carlson, R. J. Rasquinha, K. Zhang, C. E. Connor, A sparse object coding scheme in area V4. *Curr. Biol.* **21**, 288–293 (2011). doi: 10.1016/j.cub.2011.01.013; pmid: 21315595
17. D. A. Hinkle, C. E. Connor, Three-dimensional orientation tuning in macaque area V4. *Nat. Neurosci.* **5**, 665–670 (2002). doi: 10.1038/nn875; pmid: 12068303
18. E. Kobatake, K. Tanaka, Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* **71**, 856–867 (1994). doi: 10.1152/jn.1994.71.3.856; pmid: 8201425
19. D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962). doi: 10.1113/jphysiol.1962.sp006837; pmid: 14449617
20. D. H. Hubel, T. N. Wiesel, Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243 (1968). doi: 10.1113/jphysiol.1968.sp008455; pmid: 4966457
21. J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, J. A. Movshon, A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* **16**, 974–981 (2013). doi: 10.1038/nn.3402; pmid: 23685719
22. A. Pasupathy, C. E. Connor, Responses to contour features in macaque area V4. *J. Neurophysiol.* **82**, 2490–2502 (1999). doi: 10.1152/jn.1999.82.5.2490; pmid: 10561421
23. R. Desimone, T. D. Albright, C. G. Gross, C. Bruce, Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* **4**, 2051–2062 (1984). doi: 10.1523/JNEUROSCI.04-08-02051.1984; pmid: 6470767
24. D. Y. Tsao, W. A. Freiwald, R. B. Tootell, M. S. Livingstone, A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006). doi: 10.1126/science.1119983; pmid: 16456083
25. I. D. Popivanov, J. Jastorff, W. Vanduffel, R. Vogels, Heterogeneous single-unit selectivity in an fMRI-defined body-selective patch. *J. Neurosci.* **34**, 95–111 (2014). doi: 10.1523/JNEUROSCI.2748-13.2014; pmid: 24381271
26. J. Kubilius *et al*., CORnet: Modeling the Neural Mechanisms of Core Object Recognition. bioRxiv [preprint]. 4 September 2018. pmid: 408385
27. K. Kar, J. Kubilius, K. M. Schmidt, E. B. Issa, J. J. DiCarlo, Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* 10.1038/s41593-019-0392-5 (2019). doi: 10.1038/s41593-019-0392-5.
28. N. J. Majaj, H. Hong, E. A. Solomon, J. J. DiCarlo, Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *J. Neurosci.* **35**, 13402–13418 (2015). doi: 10.1523/JNEUROSCI.5181-14.2015; pmid: 26424887
29. A. B. Watson, A formula for human retinal ganglion cell receptive field density as a function of visual field location. *J. Vis.* **14**, 1 (2014). pmid: 24982468
30. D. Klindt, A. S. Ecker, T. Euler, M. Bethge, in *Advances in Neural Information Processing Systems 31*, I. Guyon *et al*., Eds. (Neural Information Processing Systems Foundation, 2017); https://papers.nips.cc/paper/6942-neural-system-identification-for-large-populations-separating-what-and-where.pdf.
31. D. Erhan, Y. Bengio, A. Courville, P. Vincent, Visualizing higher-layer features of a deep network (Departement d'Informatique et Recherche Opérationnelle, Université de Montréal, 2009); https://pdfs.semanticscholar.org/65d9/94fb778a8d9e0f632659fb33a082949a50d3.pdf.
32. M. D. Zeiler, R. Fergus, in *European Conference on Computer Vision 2014* (Springer, 2014), pp. 818–833; https://cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf.
33. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. arXiv 1412.6572 [stat.ML] (20 March 2015).
34. K. Cheng, T. Hasegawa, K. S. Saleem, K. Tanaka, Comparison of neuronal selectivity for stimulus speed, length, and contrast in the prestriate visual cortical areas V4 and MT of the macaque monkey. *J. Neurophysiol.* **71**, 2269–2280 (1994). doi: 10.1152/jn.1994.71.6.2269; pmid: 7931516
35. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255; www.image-net.org/papers/imagenet_cvpr09.pdf.

ACKNOWLEDGMENTS

We thank A. Pasupathy for generously providing the complex-curvature stimuli, and K. Schmidt, C. Shay, and S. Sanghavi for technical support. **Funding:** Supported by the Intelligence Advanced Research Projects Agency, U.S. National Eye Institute grant R01-EY014970 (J.J.D.), Office of Naval Research grant MURI-114407 (J.J.D.), and Simons Foundation grant SCGB-542965 (J.J.D.). **Author contributions:** P.B., K.K., and J.J.D. designed research; P.B. implemented the synthesis algorithm; K.K. and J.J.D. performed animal surgeries; K.K. performed neural recordings; P.B. and K.K. analyzed data; and P.B., K.K., and J.J.D. wrote the paper. **Competing interests:** J.J.D. is an associate fellow of the Canadian Institute for Advanced Research (CIFAR). J.J.D. has served as a scientific advisor for, and has a financial interest in, Bay Labs Inc. **Data and materials availability:** The methods are clearly described, and the primary data are available at https://github.com/dicarlolab/npc.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/364/6439/eaav9436/suppl/DC1
Figs. S1 to S8
Table S1

4 November 2018; accepted 5 March 2019
10.1126/science.aav9436

Science

Neural population control via deep image synthesis

Pouya Bashivan, Kohitij Kar and James J. DiCarlo

Science **364** (6439), eaav9436.
DOI: 10.1126/science.aav9436

Predicting behavior of visual neurons

To what extent are predictive deep learning models of neural responses useful for generating experimental hypotheses? Bashivan *et al.* took an artificial neural network built to model the behavior of the target visual system and used it to construct images predicted to either broadly activate large populations of neurons or selectively activate one population while keeping the others unchanged. They then analyzed the effectiveness of these images in producing the desired effects in the macaque visual cortex. The manipulations showed very strong effects and achieved considerable and highly selective influence over the neuronal populations. Using novel and non-naturalistic images, the neural network was shown to reproduce the overall behavior of the animals' neural responses.

Science, this issue p. eaav9436