Neural Population Control via Deep Image Synthesis

Pouya Bashivan*, Kohitij Kar*, and James J DiCarlo

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

To whom correspondence should be addressed; E-mail: bashivan@mit.edu.
* denotes equal contribution.

Particular deep artificial neural networks (ANNs) are today's most accurate models of the primate brain's ventral visual stream. Here we report that, using an ANN-driven image synthesis method, new luminous power patterns (i.e. images) can be applied to the primate retinae to predictably push the spiking activity of targeted V4 neural sites beyond naturally occurring levels. More importantly, this method, while 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 feed-forward artificial neural network models (ANNs) constitute today's most accurate "understanding" of the initial ∼200ms 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 (area IT) 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 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, showing that these models have 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 the models have only been evaluated in terms of internal functional similarity to the brain (above), 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 if these models would pass a stronger test of functional similarity – does that similarity generalize to entirely novel images? That is, perhaps their reported apparent functional similarity to the brain (*3, 7, 9*), substantially over-estimates their true functional similarity.

Here we conducted a set of non-human primate visual neurophysiology experiments to assess the first potential limitation by asking if 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 if 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"*: synthesize images that "stretch" the maximal firing rate of any single targeted neural site well beyond its naturally occurring maximal rate. ii) *Neural Population State Control*: synthesize images to independently control every neural site in a small recorded population (here, populations of 5-40 neural sites). We here 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 all "clamped" at their baseline activation level. We reasoned that successful experimenter control would demonstrate that at least one ANN model can be used to non-invasively *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 (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 (for details 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), and using 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*) (Conv-3 layer; that 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 a linear mapping function (that model layer selection was also consistent with similarity analysis using 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%, above), 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 removal of that pattern of luminous power will return those V4 neurons to their background firing rates).

While 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 70-170 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* (described 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 retinea are colloquially referred to as "presentation of an image", but we state the precise manipulation here 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 of the two approaches is the assumed mechanisms that intervene between the experimentally-controlled power and the controlled dependent variable (here V4 spiking rate) – steps that the ANN model aims to approximate with stacked synaptic sums, threshold non-linearities, and normalization circuits. In both the control case presented here and the optogenetics control case, those intervening steps are not fully known, but 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 (one day to build all the predictor models, one day to test control), only neural sites that maintained both high SNR and 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 one of the neural site's firing rate 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 only operate on parts of the image that were within the classical receptive field (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 ($\sim$20%) that the control algorithm predicted that it could not ”stretch.” Visual inspection suggests that the five *stretch controller* images generated by the algorithm for each neural site are perceptually more similar to each other compared to those generated for different neural site (see Figures 2 and 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. 2-A), 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. 2 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 bottom 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 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 each and every tested naturalistic image (blue) and above the response to each and every complex curvature stimulus presented within the cRF (purple), (Fig. 2-A).

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 out of 59) statistically beyond its maximal naturally-observed firing rate (unpaired-samples t-test at the level of $p < 0.01$ 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 panel 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 curve 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 curved shapes at a contrast that was matched to the contrast of the synthetic stretch controller images (see supplementary Fig. S4). Interestingly, we found that for each tested neural site, the synthetic controller images generated higher firing rates than the most-effective complex curve shape (Fig. 2-D). Specifically, when we used the maximal response over all the complex curve 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 the synthetic controller images, we carried out extensive simulations to test the predicted effects of a battery of alternative image manipulations. First, to ask if 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 2-5 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 report that the achieved stretch control ability is non-trivial in that, even at high contrast, it cannot be achieved by: simple curvature features, simple transformation on naturalistic images, combining good naturalistic images, or optimizing the spatial extent of the image (see Methods and Fig. S7.)

“One-Hot-Population” Control: Attempt to *only* activate one of many V4 neural sites

Similar to prior single unit visual neurophysiology studies (*16–18*), the stretch control experiment attempted to optimize the response of each V4 neural site one at a time 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 (aka 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 (*One-hot-population* 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 and 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 (that 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 retinea 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; for an example near 0.3 see Fig. 3).

While 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. 3-A). 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 we 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 those two scores as the measure of improved one-hot population control, and we found that the controller typically achieved an improvement of 57% (median over all 33 one-hot-population control tests; see Fig. 3-B and C) and we found that that improved control was statistically significant for 76% of the one-hot population control tests (25 out of 33 tests; unpaired-samples t-test at the level of $p < 0.01$).

We considered the possibility that the improved population control was resulting from the non-overlapping 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 (see Fig. 3-D). 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 (see Fig. 3-D). In total we performed the experiment on 12 target neural sites in two monkeys (4 in monkey-M and 8 in monkey-S) and found that the amplitude of improved control was still 40%. 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 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 the shared cRF of all neural sites in a sub-population with overlapping cRFs (Fig. 3-E). 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 curved 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. 3-E) and the control score was significantly higher than best curvature stimulus for 86% of the neural sites (12 out of 14).

Does the functional fidelity of the ANN brain model generalize to novel images? Besides testing non-invasive causal neural control, these experiments also aimed to ask if ANN models would pass a stronger test of functional similarity to the brain than prior work had shown (*2, 3*). Specifically, does that model-to-brain similarity generalize to entirely novel images? Because the controller images were synthesized anew from random pixel arrangement and they were optimized to drive the firing rates of V4 neural sites both upwards (targets) and downwards (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 farther from the naturalistic images compared to the naturalistic image set to itself (measuring distances in pixels space, recorded V4 neural population space, and model-predicted V4 population space; see Methods and Fig. S6).

To ask 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 than 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). While the model overestimates the neural responses to synthesized stimuli on many occasions and the model-to-brain similarity score is somewhat lower than that obtained for naturalistic images responses (89%), the model still predcits a substantial portion of the variance considering the fact that all parameters were fixed to make these “out-of-naturalistic-domain” image predictions. We believe this is 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 both remarkably similar to the brain’s intermediate ventral stream representation (V4), but also that 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? In sum, we here demonstrate that, using a deep ANN-driven controller method, we can 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 quite well to predict V4 responses to synthetic images – images which are strikingly different than 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*), decades of visual neuroscience 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 2D and 3D (*17, 22*) and tuned to boundary information (*12, 14*), and IT neurons respond to complex object-like patterns (*18*) including faces (*23, 24*) and bodies as special cases (*25*).

While these efforts have been essential to building both a solid foundation and intuitions about the role of neurons in encoding visual information, our results here show how they can be further refined by current and future ANN models of the ventral stream. For instance here we found that synthesis of only few images leads to higher neural response levels that was possible by searching in a relatively large space of natural images (n=640) and complex curved 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 difficult to fully describe by human spoken language (see examples in Fig. S1). This is likely to be especially important in mid and later stages of the visual hierarchy (e.g. in V4 and inferior temporal cortex) where the response complexity and larger receptive fields of neurons makes 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 (understanding and generality)? In our view, the results strongly mitigate both of those critiques, but they do not eliminate them.

On understanding: the ability to use knowledge to gain improved control over things of interest in the world (as we have demonstrated here) is an important test of understanding.

However we acknowledge that this is not the only possible view, and many other notions of "understanding" remain to be explored to see if and how these models add value.

On generality: because we found that even today's ANN models show good generalization to demonstrably novel images, 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 while 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 re-confirms 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 we did not test 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 non-invasive, 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 Figures-3, 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 very interesting scientific and applied questions of if 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 non-perfect accuracy of the current ANN models of the ventral stream, non-perfect 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 maybe only to a few 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 to-be-controlled neural population 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 on 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 $\sim$25 mm^2 of dorsal V4. On each day, recording sites that were visually-driven as measured by response correlation ($r_{pearson} > 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. (Note that, 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 which included 52, 33 and 22 sites in monkey-M (left hemisphere), monkey-N (right hemisphere), and monkey-S (left hemisphere), respectively. Of these sites, 76 of 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, the raw spiking rate of each site from each recording session was normalized (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 three, 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 5 to 7 images, each ON for 100 ms followed by a 100 ms gray blank screen. This was followed by a water reward and an inter-trial interval of 500 ms, followed by the next sequence. Trials were aborted if gaze was not held within $\pm 0.5°$ of the central fixation dot during any point. To estimate the classical receptive field (cRF) of each neural site, we flashed $1° \times 1°$ 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-std 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 sites and every model V4 neural site to each of these images. Each of these images contained a three-dimensional 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 between 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 curve 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 divide the neural response repetitions into two, randomly-selected halves. We then compute the mean firing rate of each neural site in response to each image in each half. The mean responses from the first half are used to find the image that produces the highest score (in that half) and the response to that image is then measured in the second half (and this is the measurement used for further analyses). We repeat 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 compute 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 (a.k.a. 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 here refer to this as model "V4"). We used the same training procedure as was described in (*13*), except we did not split the middle convolutional layers between GPUs.

In addition, the input images were transformed using an eccentricity-dependent function that mimics the known spatial sampling properties of the primate retinae (details 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 $\sim 5-10\%$). The "retinae transformation" was implemented by a fish-eye 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 fish-eye transformation. Parameters of the fish-eye transformation were tuned to mimic the cones density ratio in fovea at 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 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 like principal component regression. Our implementation was a variant of the 2-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 (non-sparse) 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 point-wise 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 the each neural sites 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 $\mathcal{L}_{laplace}$ (defined below). By factorizing the spatial and feature dimensions, this method significantly improves the predictivity of neural responses over the traditional principle 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 which is strongly in line with a 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.

$$\hat{y}_n = \Big(\sum (W_s^{(n)} \cdot X)\Big).W_d^{(n)} + w_b^{(n)} \tag{1}$$

$$\mathcal{L}_2 = \lambda_s \sum_n {W_s^{(n)}}^2 + \lambda_d \sum_n {W_d^{(n)}}^2 \tag{2}$$

$$\mathcal{L}_{laplace} = \lambda_s \sqrt{\sum_n (W_s^{(n)} * L)^2}, \quad L = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{3}$$

$$\mathcal{L}_e = \sqrt{\sum_n (\hat{y_n} - y)^2} \tag{4}$$

$$\mathcal{L} = \mathcal{L}_e + \mathcal{L}_{laplace} + \mathcal{L}_2 \tag{5}$$

We evaluated our model using 2-fold cross-validation and observed that $\sim$89% of the explainable variance could be explained with our model in three monkeys ($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 $\sim$13% over the naive principal component regression applied on features from the model trained without the retinae transformation ($EV_M = 75\%$, $EV_N = 80\%$, $EV_S = 73\%$). Ablation studies on data from each monkey suggested that on average about 3-8% 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 utilizes a backward function $r = g(r\prime)$ that maps the radius of points in the retinae transformed image ($r\prime$) 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*).

$$\rho = \frac{1}{\pi d^2} = e^{-ar'} \tag{6}$$

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

$$d = \frac{1}{\sqrt{\pi}} e^{ar'/2} \tag{7}$$

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 could be written as:

$$\frac{\rho_f}{\rho_p} = e^{ar'_{max}} \tag{8}$$

where ρ_f and ρ_p are the cone densities at the fovea and periphery respectively, and $r\prime_{max}$ is the highest radial distance in the output image (e.g. 150 for an image of size 300). From equation (8) above we can calculate a as a function of ρ_f, ρ_p, and $r\prime_{max}$.

$$a = \frac{\ln \frac{\rho_f}{\rho_p}}{r'_{max}} \tag{9}$$

The $\frac{\rho_f}{\rho_p}$ ratio is known given the size of the visual field in which the stimuli were presented (e.g. 10 for fovea to 4-degrees in this study) and the output image size (e.g. 150 in this study).

We can now formulate the function $g(r\prime)$ as the sum of all the distances d up to radius $r\prime$ weighted by a factor b.

$$g(r\prime) = \frac{b}{\sqrt{\pi}} \sum_{k=0}^{r'-1} d_k = \frac{b}{\sqrt{\pi}} \sum_{k=0}^{r'-1} e^{\frac{ar}{2}} = \frac{b}{\sqrt{\pi}} \frac{1 - e^{ar'/2}}{1 - e^{a/2}} \tag{10}$$

where b is found so that $\frac{r_{max}}{g(r\prime_{max})} = 1$. In our implementation we use Brents 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 the principal component 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. The median normalized EV across all measured sites was used to select the best representation in the ANN model (Fig. S8-A). 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. S8-B). RDMs were computed using the principle components of the features at each layer in response to the naturalistic image set (n=640).

Synthesized "Controller" Images The "response" of artificial neuron in the *mapped V4 encoding model* (above) is a differentiable function of the pixel values $f : \mathcal{I}^{w \times h \times c} \rightarrow \mathbb{R}^n$ that enables us to use the model to analyze the 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 synthesize controller images that attempt to push each individual V4 neural site into its maximal activity state. To do so, we follow an approach first introduced in (*31*) and iteratively change 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 "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 formalize the *One-hot population* goal in the following objective function that we then aim to maximize during the image synthesis procedure:

$$\mathcal{S} = \text{Softmax}_t(y) = \frac{e^{y_t}}{\sum e^{y_i}} \tag{11}$$

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 start from an image that consists of random pixel values drawn from a standard Normal distribution and optimize the objective function for a pre-specified number of steps using gradient ascend algorithm (steps=700). We also use the total variation (defined below) as additional regularization in the optimization loss to reduce the high frequency noise in the generated images:

$$L_{TV} = \sum_{i,j} \left(\|I_{i+1,j} - I_{i,j}\|_2 + \|I_{:,j+1} - I_{i,j}\|_2 \right) \tag{12}$$

During the experiments, monkeys are required to fixate within a 1° circle at the center of the screen. This introduces an uncertainty on the exact gaze location. For this reason, images are synthesized to be robust to small translations of maximum $0.5°$. At every iteration, we translate the image in random directions (i.e. jittering) with a maximum translation length of $0.5°$ in each direction, thereby, generating images that are predicted to elicit similarly high scores regardless of the translations within the range. The total-variation loss and the translation-invariance procedure reduce 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 normalize the computed gradients by its global norm and clip 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 if 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 and the 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 1-30 cycles/degree. We ensured that for all tested neural sites, CE within the cRF for synthetic *Stretch Controller* images were less than or equal to the classic, complex curvature V4 stimuli (Supp 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 2-D Gaussian function (i.e. mu, sigma1, sigma2, rho) 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 2-D 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 picked 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 non-significant 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 used in each one 1) random scaling between 0.5 to 2; 2) random translation between -25 to 25 pixels in each direction; 3) random rotation between 0 to 90 degrees; and 4) 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 number 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 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. Fig. S6 shows the distribution of minimum distances 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 and -E) 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

1. M. Schrimpf, *et al.* (2018), Brain-Score : Which Artificial Neural Network for Object Recognition is most Brain-Like ?, *bioRxiv* pp. 1–9.

2. D. L. K. Yamins, H. Hong, C. Cadieu, J. J. DiCarlo (2013), Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream, *NIPS* .

3. D. L. Yamins, *et al.* (2014), Performance-optimized hierarchical models predict neural responses in higher visual cortex, *PNAS* **111**, 8619.

4. S. A. Cadena, *et al.* (2017), Deep convolutional models improve predictions of macaque V1 responses to natural images, *bioRxiv* p. 201764.

5. R. Rajalingham, K. Schmidt, J. J. DiCarlo (2015), Comparison of Object Recognition Behavior in Human and Monkey, *J. of Neuroscience* **35**, 12127.

6. R. Rajalingham, *et al.* (2018), Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks, *The Journal of neuroscience* **38**, 7255.

7. S. M. Khaligh-Razavi, N. Kriegeskorte (2014), Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation, *PLoS Comp. Bio.* **10**.

8. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva (2016), Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence, *Scientific Reports* **6**, 1.

9. D. L. Yamins, J. J. DiCarlo (2016), Using goal-driven deep learning models to understand sensory cortex, *Nat. Neuroscience* **19**, 356.

10. J. Pearl, *Causality* (Cambridge university press, 2009).

11. M. Jazayeri, A. Afraz (2017), Navigating the Neural Space in Search of the Neural Code, *Neuron* **93**, 1003.

12. A. Pasupathy, C. E. Connor (2001), Shape representation in area V4: position-specific tuning for boundary conformation., *Journal of neurophysiology* **86**, 2505.

13. A. Krizhevsky, I. Sutskever, G. E. Hinton (2012), ImageNet Classification with Deep Convolutional Neural Networks, *NIPS* .

14. A. Pasupathy, C. E. Connor (2002), Population coding of shape in area V4, *Nature Neuroscience* **5**, 1332.

15. J. R. Cavanaugh, W. Bair, J. A. Movshon (2002), Nature and Interaction of Signals From the Receptive Field Center and Surround in Macaque V1 Neurons, *Journal of Neurophysiology* **88**, 2530.

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

17. D. A. Hinkle, C. E. Connor (2002), Three-dimensional orientation tuning in macaque area V4, *Nature Neuroscience* **5**, 665.

18. E. Kobatake, K. Tanaka (1994), Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex., *Journal of neurophysiology* **71**, 856.

19. D. H. Hubel, T. N. Wiesel (1962), Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex, *J. Physiol* **160**, 106.

20. D. H. Hubel, T. N. Wiesel (1968), Receptive Fields and Functional Architecture of monkey striate cortex, *Journal of Physiology* **195**, 215.

21. J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, J. A. Movshon (2013), A functional and perceptual signature of the second visual area in primates, *Nature Neuroscience* **16**, 974.

22. A. Pasupathy, C. E. Connor (1999), Responses to countour features in macaque area V4, *Journal of Neurophysiology* **82**, 2490.

23. R. Desimone, T. D. Albright, C. G. Gross, C. Bruce (1984), Stimulus-selective properties of inferior temporal neurons in the macaque, *The Journal of neuroscience* **4**, 2051.

24. D. Y. Tsao, W. A. Freiwald, R. B. Tootell, M. S. Livingstone (2006), A Cortical Region Consisting Entirely of Face-Selective Cells, *Science* **311**, 670.

25. I. D. Popivanov, J. Jastorff, W. Vanduffel, R. Vogels (2014), Heterogeneous Single-Unit Selectivity in an fMRI-Defined Body-Selective Patch, *Journal of Neuroscience* **34**, 95.

26. J. Kubilius, *et al.* (2018), CORnet: Modeling the Neural Mechanisms of Core Object Recognition, *bioRxiv* pp. 1–9.

27. K. Kar, J. Kubilius, K. M. Schmidt, E. B. Issa, J. J. DiCarlo (2018), Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior, *bioRxiv* p. 354753.

28. N. J. Majaj, H. Hong, E. A. Solomon, J. J. DiCarlo (2015), Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance, *J. of Neuroscience* **35**, 13402.

29. A. B. Watson (2016), A formula for human retinal ganglion cell receptive field density as a function of visual field location, *Journal of Vision* **14**, 1.

30. D. Klindt, A. S. Ecker, T. Euler, M. Bethge (2017), Neural system identification for large populations separating what and where, *NIPS* pp. 3509–3519.

31. D. Erhan, Y. Bengio, A. Courville, P. Vincent (2009), Visualizing higher-layer features of a deep network, *University of Montreal* **1341**, 1.

32. M. D. Zeiler, R. Fergus (2014), Visualizing and understanding convolutional networks, *European conference on computer vision* pp. 818–833.

33. I. J. Goodfellow, J. Shlens, C. Szegedy (2014), Explaining and Harnessing Adversarial Examples, *Internationl Conference on Learning Representations* pp. 1–11.

34. K. Cheng, T. Hasegawa, K. S. Saleem, K. Tanaka (1994), Comparison of neuronal selectivity for stimulus speed, length, and contrast in the prestriate visual cortical areas V4 and MT of the macaque monkey., *Journal of Neurophysiology* **71**, 2269.

35. Jia Deng, *et al.* (2009), ImageNet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp. 248–255.

Acknowledgments We thank Dr. A. Pasupathy for generously providing the complex curvature stimuli, and Kailyn Schmidt and Sachi Sanghavi for technical support. **Funding:** This research was supported by the Intelligence Advanced Research Projects Agency (IARPA), the MIT-IBM Watson AI Lab, US National Eye Institute grants R01-EY014970 (J.J.D.), and the Office of Naval Research MURI-114407 (J.J.D). **Author contributions:** PB, KK and JJD designed research. PB implemented the synthesis algorithm. KK and JJD performed animal surgeries. KK performed neural recordings. PB, and KK analyzed data. PB, KK, and JJD wrote the paper. **Competing interests:** JJD is an Associate Fellow of the Canadian Institute for Advanced Research (CIFAR). JJD 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 is available at `https://s3.amazonaws.com/npc-v4-data/npc_v4_data.h5.zip`.

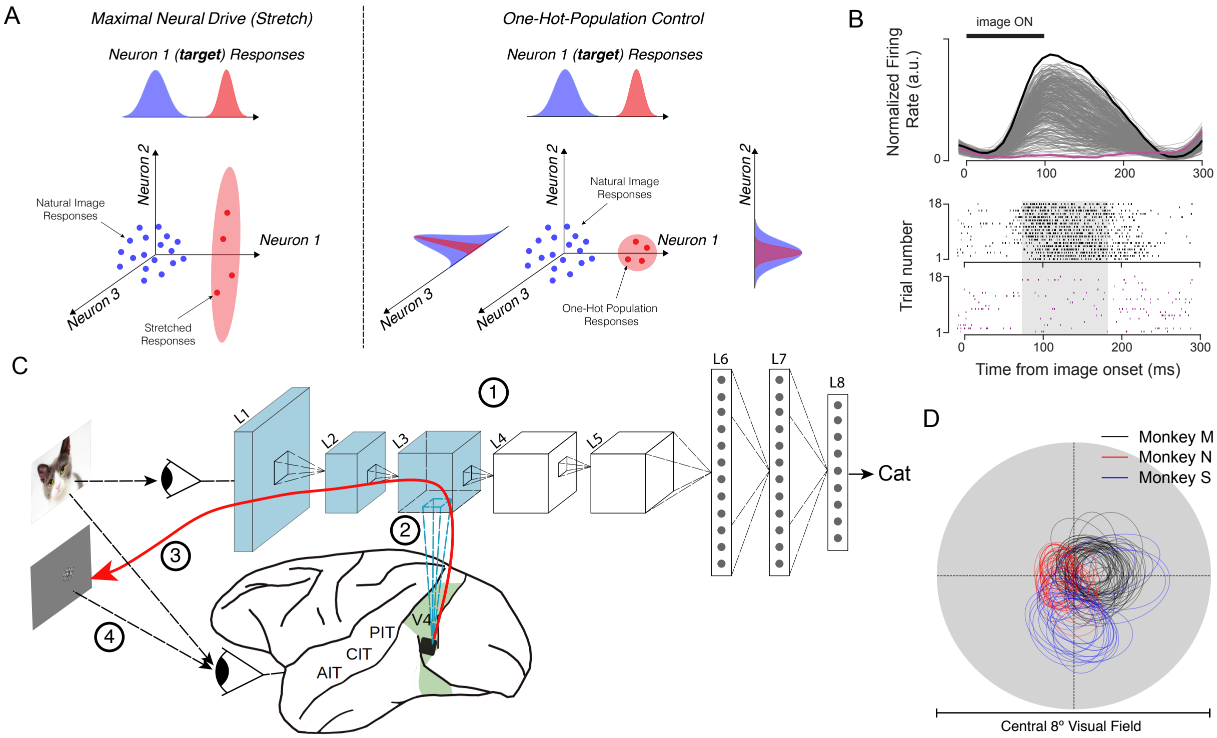

Figure 1: **Overview of the synthesis procedure.** A) Schematic illustration of the two tested control scenarios. Left - the controller algorithm synthesizes novel images that it believes will maximally drive the firing rate of a target neural site (Stretch). In this case, the controller algorithm does 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 - gray lines (overlapping): responses of a single example V4 neural site to 640 naturalistic images (averaged over ∼40 repetitions for each image). Vertical wide black line marks the image presentation period. Bottom - raster plots of highest (black) and lowest (purple) neural response to naturalistic images. 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 those 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. D) Classical receptive fields of neural sites in monkey M (black), Monkey N (red) and Monkey S (blue; see Methods).

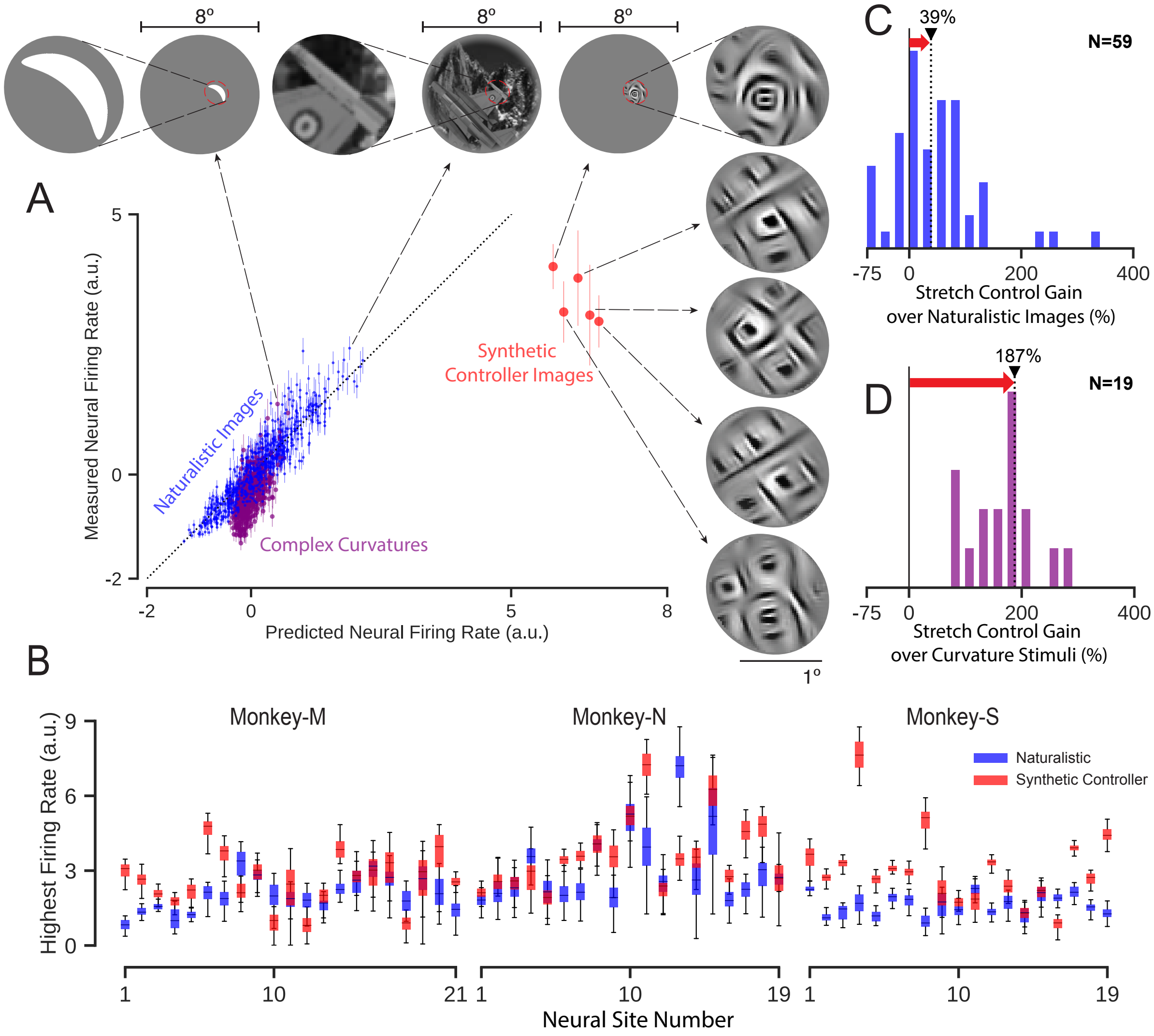

Figure 2: **Maximal drive of individual neural sites (Stretch)**. A) Results for an example successful "stretch" control test. Normalized activity level of the target V4 neural sites is shown for all of the naturalistic images (blue dots), complex curved stimuli (purple dots) and for its five synthetic "stretch" controller images (red dots; see Methods). Best driving images within each category, and the zoomed view of the receptive field are shown on the top. B) Difference in firing rate in response to naturalistic (blue boxes) and synthetic images (red boxes) for each neural site in three monkeys. Controller image synthesis was restricted within the receptive field of the target neural site. 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 curved stimuli. Black triangle with dotted black line marks the median of the scores over all tested neural sites. The 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.

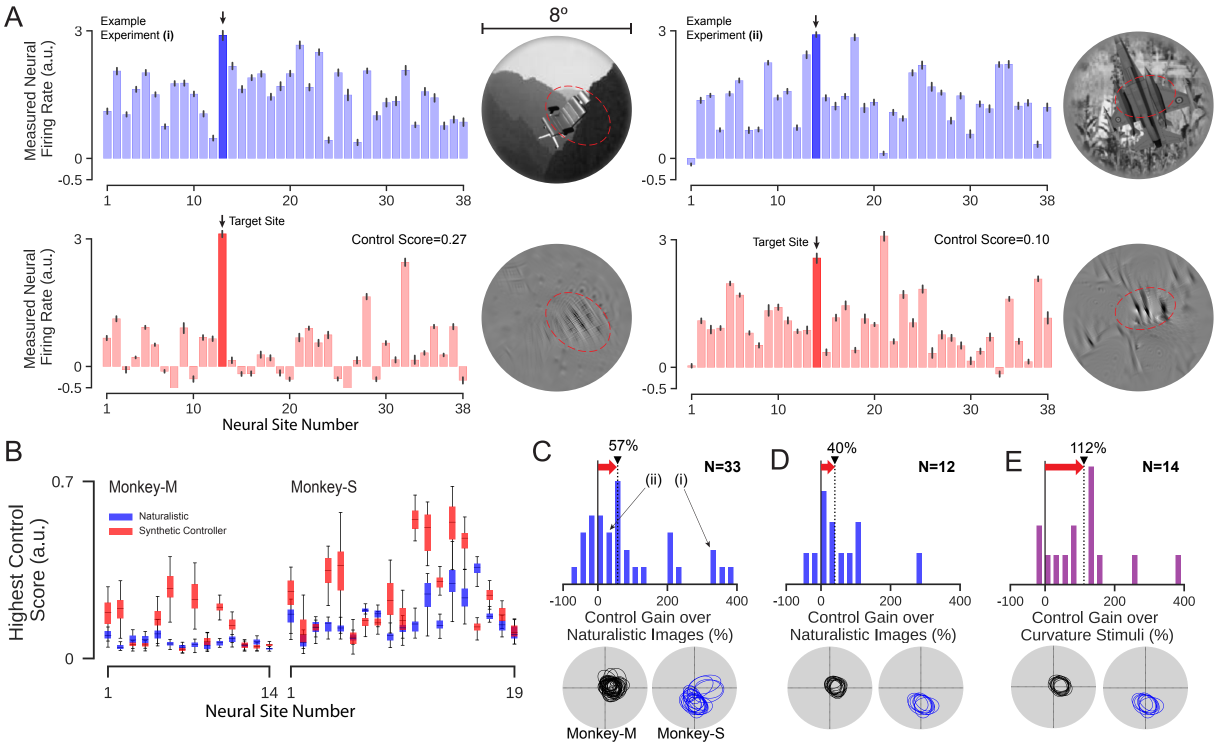

Figure 3: **Neural Population Control.** We synthesized controller images that aimed to set the neural population in a one-hot state (OHP) in which one target neural site is active and all other recorded neural sites are suppressed. A) Two example OHP experiments (left and right). In each case, the neural activity of each of the validated V4 sites (see Methods) in the recorded population are plotted (most have overlapping cRFs), with the target V4 site indicated in dark blue/red). Note that responses are normalized individually on a normalizer image set to make side-by-side comparison of the responses meaningful (see Methods). Upper panel: activity pattern for the best ("best" in the sense of OHP control, see Methods) naturalistic image (shown on the right). Lower panel: activity pattern produced by retinal application of the ANN-model-synthesized controller image (shown on the right). The red dashed line marks the extended receptive field (2-std) of each site. 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). 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 sub-populations selected so that *all* sites have highly overlapping cRFs (see cRFs below). E) OHP control gain where gain is relative to best complex curvature stimulus in the shared cRF (see text) and controller algorithm is also restricted to operate only in that shared cRF (n=14 OHP experiments). Receptive fields of neural sites in each setting (C-E) (black: monkey-M; blue: monkey-S). "N" indicates the number of experiments in each setting. Red arrow highlights the median gain in control (black triangle) achieved in each case.

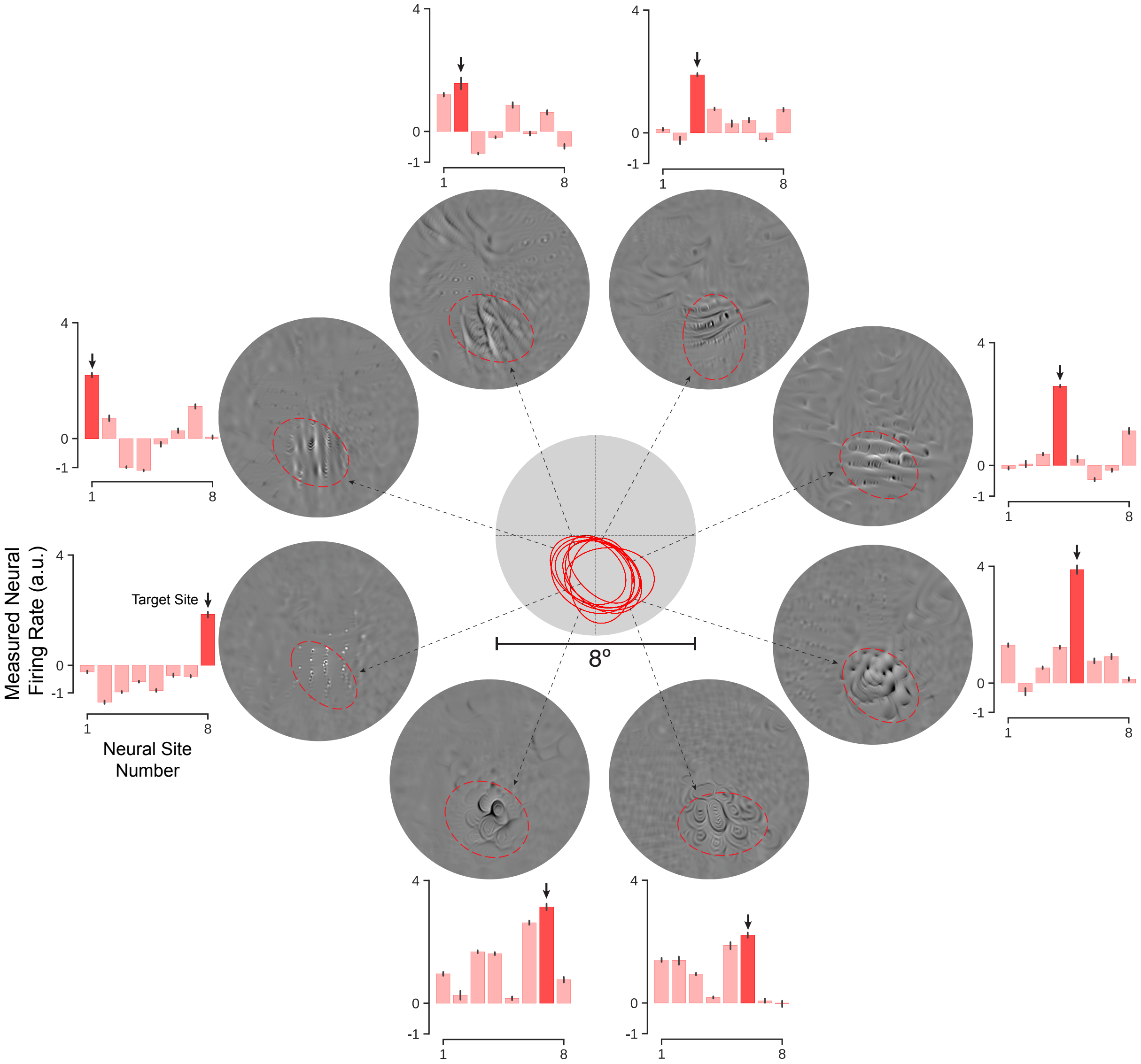

Figure 4: **Example of independent control of each neural site on a subset of V4 neural sites with highly overlapping cRFs.** Controller images were synthesized to try to achieve a one-hot-population over a population of eight neural sites (in each control test, the target neural site is shown as dark red). 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-std) of each site (red).

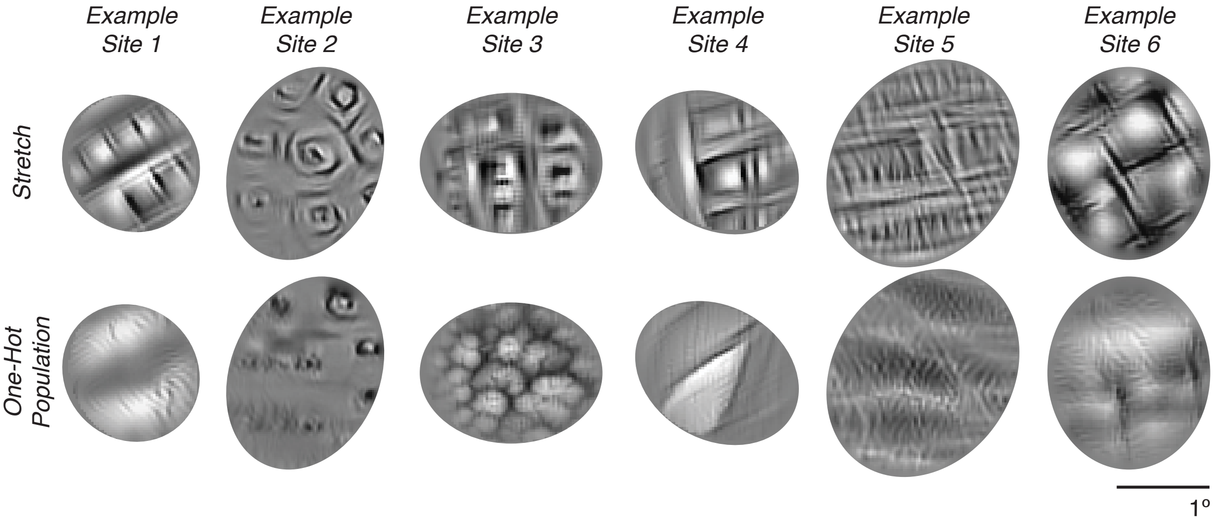

Figure 5: **Example controller images synthesized in "Stretch" and "One-hot population" settings for six example target neural sites.** Controller images were synthesized from the same initial random image, but optimized for each target neural site and for each control goal ("Stretch" or "One-hot population", 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.

Supplementary materials

Figures S1 to S8

Table S1

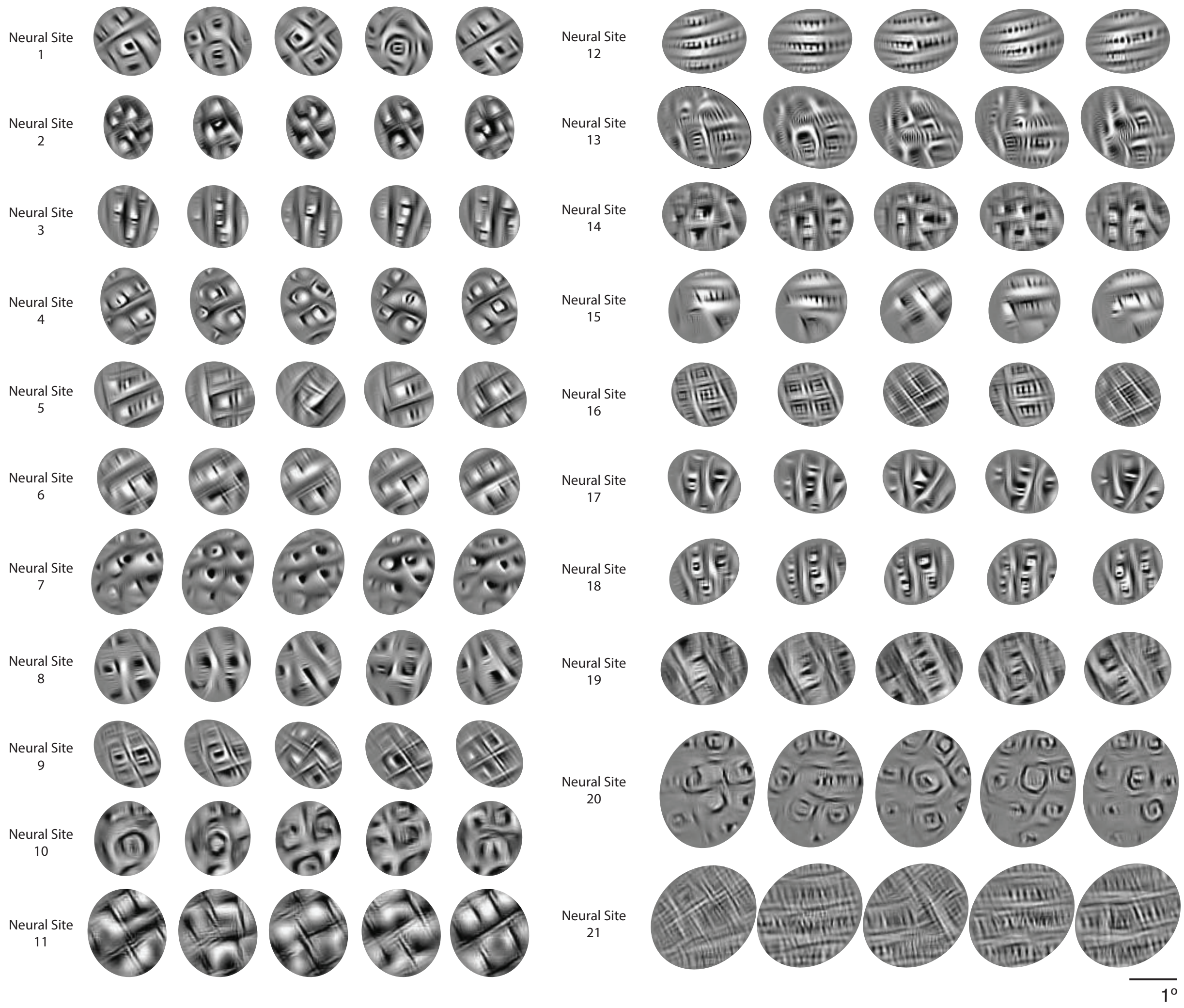

Figure S1: **Stretch synthetic controller images for 21 example V4 neural target sites in Monkey-M.** Each column displays images generated using the same random starting image, but optimized for each target site. Note the perceptual similarity of the controller images synthesized for each site and the dissimilarity between the controller images across sites.

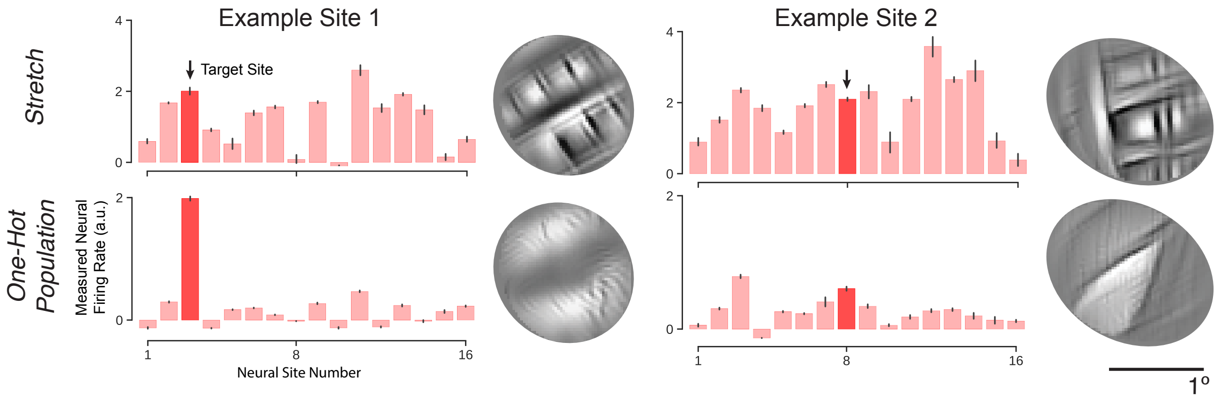

Figure S2: **Comparison of population response in *Stretch* and *One-hot Population* settings.** Population responses in *Stretch* and *One-hot Population* settings are demonstrated for two example neural sites. One-hot population images were generated with an objective function including 16 neural sites with highly overlapping receptive fields. Compared to the *Stretch* controller images, the *one-hot-population* images have fewer identifiable "features". The displayed images were synthesized using the same initial random image.

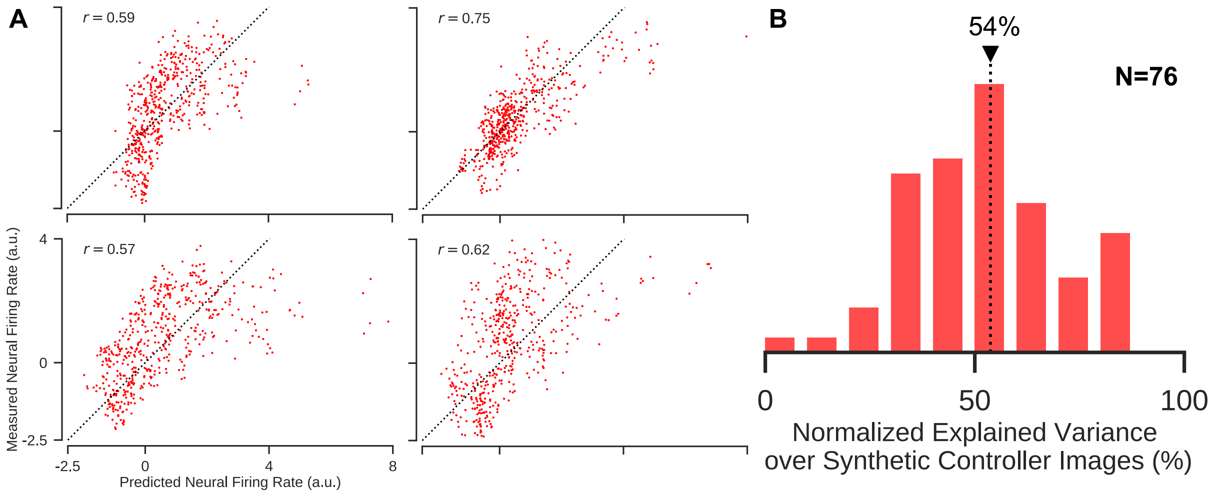

Figure S3: **Predictability of synthetic controller images.** A) Scatter plots of predicted and measured V4 neural responses to synthetic controller images for four example neural sites. For most target neural sites, the predicted and measured neural responses were significantly correlated. Each dot represents the prediction and average measured response to a single image. B) The model accounted for 54% (median across all tested neural sites in three monkeys; N=76) of the explainable variance.

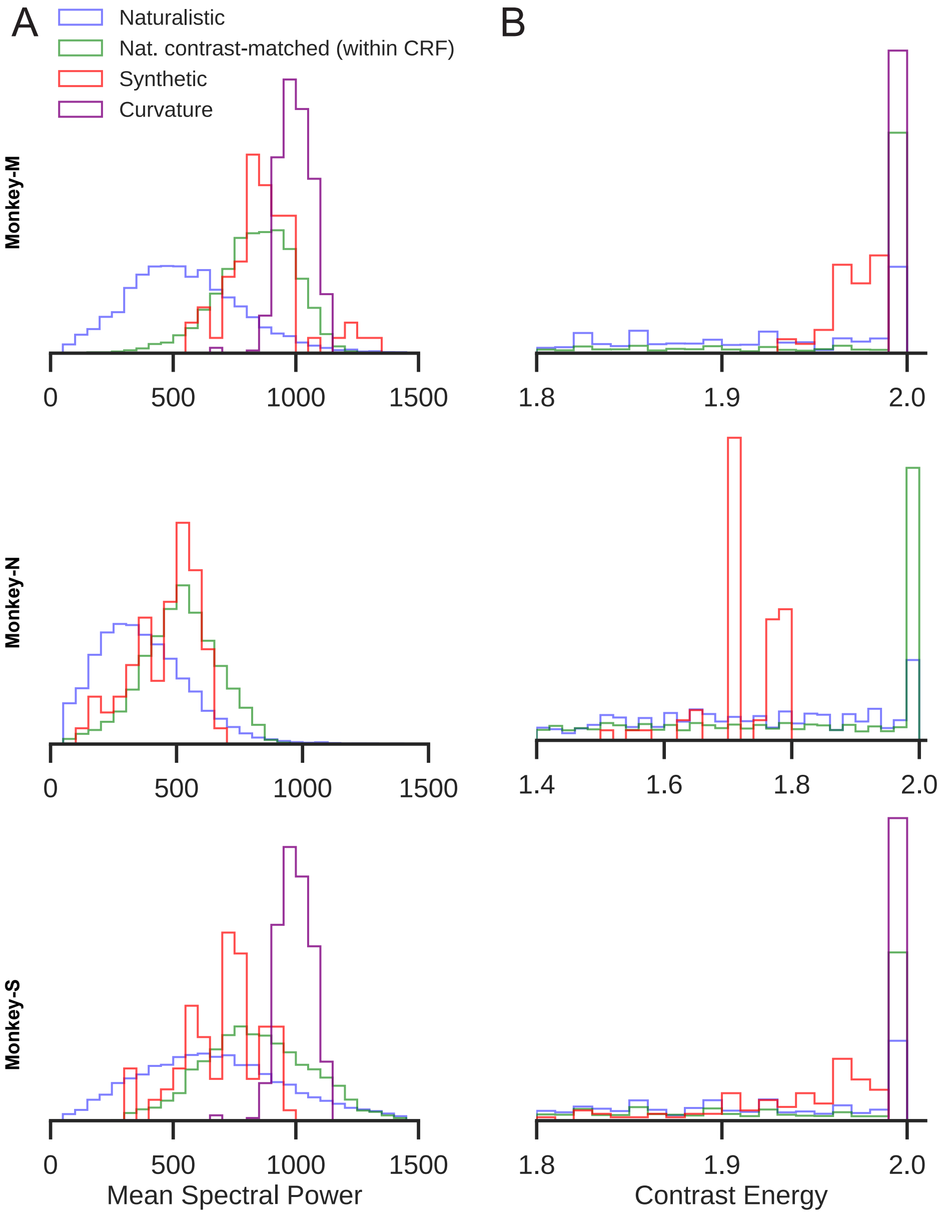

Figure S4: **Comparison of contrast energy between synthetic, naturalistic, and curvature images for each monkey.** A) Distribution of the mean spectral power within target neural sites' classic receptive fields for "Stretch" controller (red), naturalistic (blue), contrast-matched naturalistic (green), and complex curvature (purple) images. Spectral power was computed using 2-D FFT transformation and summed in the frequency range of 1-30 cycles/degree. B) Distribution of contrast energy within target neural sites' classic receptive fields for "Stretch" controller, naturalistic, contrast-matched naturalistic and complex curvature images.

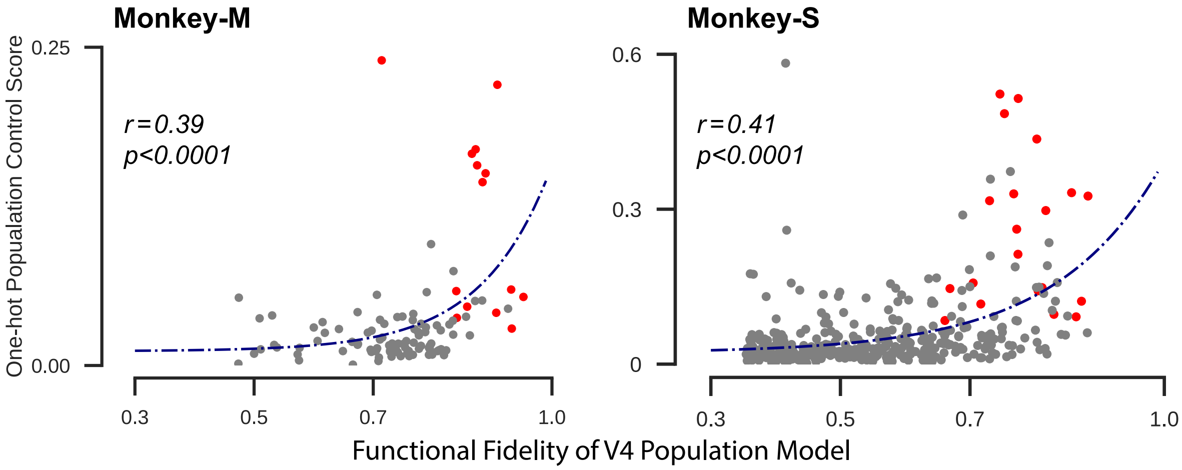

Figure S5: **Higher functional fidelity models increase the ability to control neural responses.** We evaluated the one-hot population control score for each target neural site in each monkey subject for a range of possible models with different prediction accuracy levels. In each monkey session, the functional fidelity of a V4 population model (measured by the mean of: 1) explained variance of target neural site and 2) the mean of the explained variance for all the off-target sites) was plotted against the one-hot population control score achieved with that population model. We found that these were significantly correlated as assessed by Spearman rank order correlation, shown on each panel. For this analysis, for each target neural site, we included not only the original tests, but also tests in which we swapped the predictive model of the target neural site with the model of randomly-chosen off-target site (we do this ”mismatch” test because it is an example of what would have happened in the experiment if the synthesis algorithm had been given the wrong models – it would have produced OHP control stimuli that we already tested – so we can compute the resulting control score without doing new recording experiments). We simply assessed the functional fidelity of V4 population model using the mismatched models and the population control score achieved using the new population model's synthetic control images (again, from population responses to images that we had already tested). Red dots correspond to cases where the target neural site's model and responses were matched (i.e. results of the original OHP tests, see text), and gray dots correspond to the cases where they were mismatched. Dark blue line shows an exponential function fitted to the data points, highlighting the tendency for higher model fidelity to support better control.

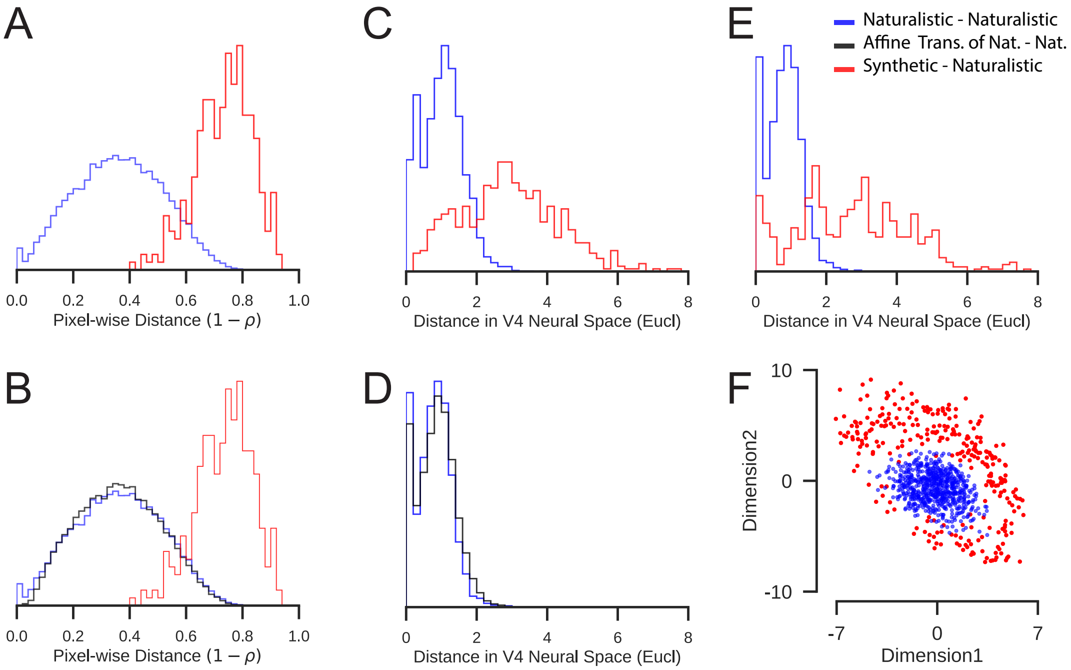

Figure S6: **How ”novel” and ”out-of-domain” are the synthetic images?** A-E) Each distribution plots the minimum distance of each of the images in the test set to the full set of 640 naturalistic images (i.e. minimum over 640). In each case, the reference is the minimum distance of any given naturalistic image to the other naturalistic images (blue distribution in each plot). Note that, in all cases, the synthetic images (red) are farther from the naturalistic images. A) Pixel-space distances (within the receptive field of each neural site). B) Same as panel-A but also showing a new test set: random affine transformations of naturalistic images (black). The random affine transformations naturalistic image set was generated by randomly performing combination of scaling, translation and rotation transformations on random naturalistic images for n=6400. C) Euclidean distances in the predicted V4 population response space (n_M=21, n_N=19, and n_S=19 simulated V4 neural sites). D) Same as Panel-C but also showing a new test set: random affine transformations of naturalistic images (black). E) Euclidean distances in measured neural population responses (n_M=21, n_N=19, and n_S=19 actual V4 neural sites). F) Scatter plot of 640 naturalistic images (blue) and 285 synthetic images (red) where the axes are the first two principle components of the measured V4 population response space (using Multi-Dimensional Scaling; data from Monkey-S).

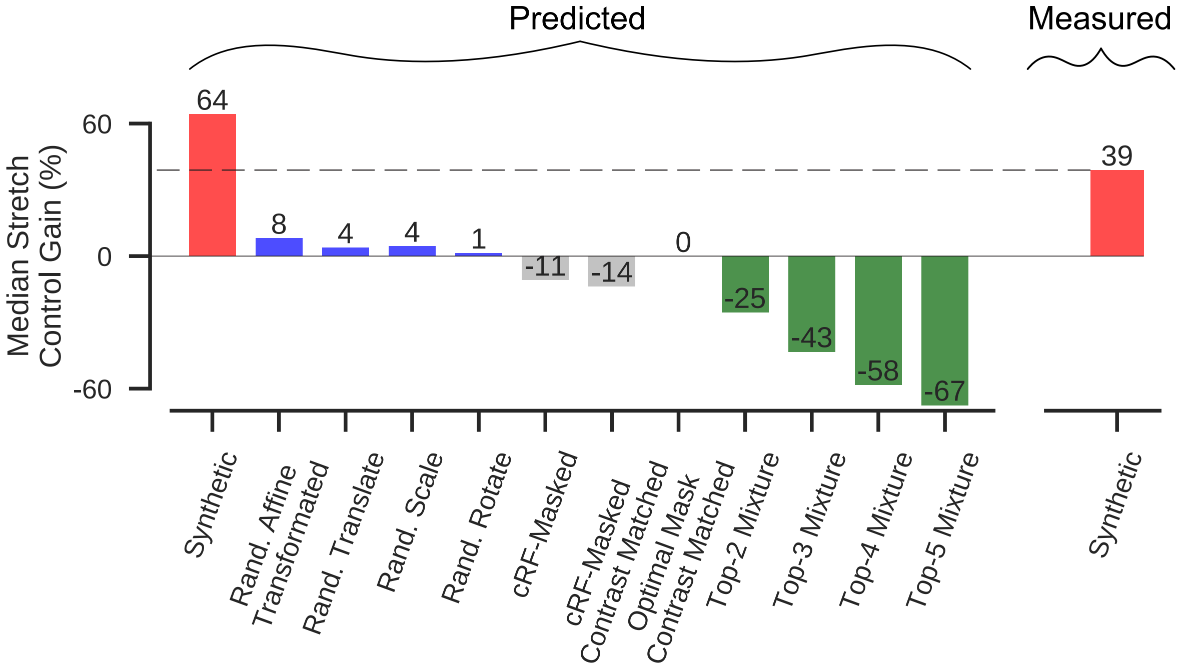

Figure S7: **Comparison of stretch control gain using alternative control methods.** Results on the left panel are computed using the model predictions for each case. Red bar plot on the right indicates the achieved stretch control gain using the synthesis procedure (reported in the main text, see Fig. 2). Control methods are color coded into four categories: ANN-synthesis (red), affine transformations (blue), mask optimization (gray), and image mixing (green). Each bar indicates the median stretch control gain over the naturalistic image set.

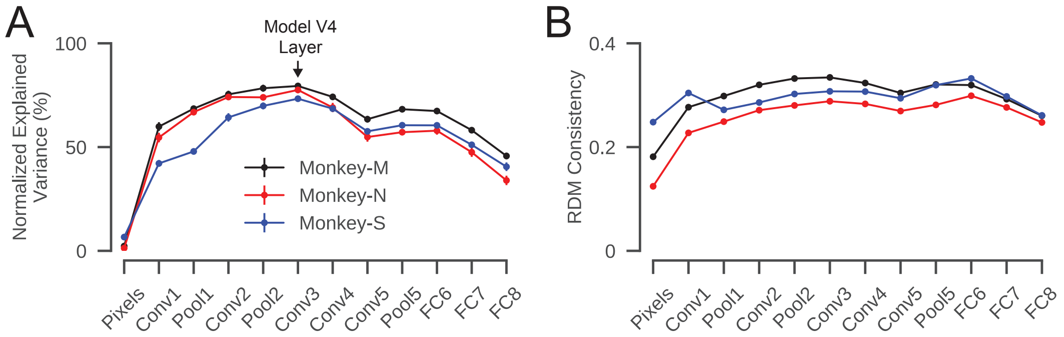

Figure S8: **Similarity of representations to measured neural activities at different stages of processing in the artificial neural network.** A) ANN features at the output of each layer are used to predict the measurements from V4 sites. Amount of explained variance by these features are normalized by the internal consistency of neurons across stimulus presentations. B) Consistency between V4 representation (spanned by the measured neural responses) and representations at each layer of the ANN model is quantified by constructing the image-level representational dissimilarity matrix (RDM) for each one and computing the Pearson correlation between the elements in the upper-triangle of the two matrices.

	Mapping Type	Model Type	Median Normalized EV (%)
Monkey-M	PCR	Conv3	80
	Klindt et al.	Conv3	88
	Klindt et al.	Retinae-Conv3	92
Monkey-N	PCR	Conv3	75
	Klindt et al.	Conv3	84
	Klindt et al.	Retinae-Conv3	92
Monkey-S	PCR	Conv3	72
	Klindt et al.	Conv3	77
	Klindt et al.	Retinae-Conv3	80

Table S1: Median prediction accuracy over all measured neural sites in three monkeys using different mapping methods and model features. Addition of the Retinae transformation and convolutional mapping in Klindt et al. (30) account for 3-8% and 5-9% of the improvement in prediction accuracy compared to the principle component regression method respectively.