Adversarial images for the primate brain

Li Yuan,[1,2,6], Will Xiao,[1,3,6,*] Gabriel Kreiman,[4] Francis E.H. Tay,[5] Jiashi Feng,[2]
Margaret S. Livingstone[1,*]

[1]Department of Neurobiology, Harvard Medical School, Boston, MA 02115, U.S.A.

[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583

[3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02134, U.S.A.

[4]Department of Ophthalmology, Boston Children's Hospital, Boston, MA 02115, U.S.A.

[5]Department of Mechanical Engineering, National University of Singapore, Singapore 117583

[6]These authors contributed equally: Li Yuan, Will Xiao

[*]To whom correspondence should be addressed;

E-mail: xiaow@fas.harvard.edu; margaret_livingstone@hms.harvard.edu

Abstract

Deep artificial neural networks have been proposed as a model of primate vision. However, these networks are vulnerable to adversarial attacks, whereby introducing minimal noise can fool networks into misclassifying images. Primate vision is thought to be robust to such adversarial images. We evaluated this assumption by designing adversarial images to fool primate vision. To do so, we first trained a model to predict responses of face-selective neurons in macaque inferior temporal cortex. Next, we modified images, such as human faces, to match their model-predicted neuronal responses to a target category, such as monkey faces. These adversarial images elicited neuronal responses similar to the target category. Remarkably, the same images fooled monkeys and humans at the behavioral level. These results challenge fundamental assumptions about the similarity between computer and primate vision and show that a model of neuronal activity can selectively direct primate visual behavior.

Introduction

Artificial neural networks (ANNs) can now match, and often even exceed, human performance in tasks such as visual categorization[1,2]. Together with this rapid development came the surprising finding that a family of ANNs loosely inspired by biological visual systems—deep convolutional neural networks—now make the best model of visual cortical neuron responses[3,4]. In this light, a perhaps equally surprising finding is that ANNs are susceptible to adversarial attacks that do not fool biological visual systems. That is, adding carefully crafted, minute noise can cause ANNs to misclassify images with high confidence[5,6]. Often, such adversarial noise is too small to be displayed on a standard monitor, let alone to be perceived by human observers. Thus, an intuitive assumption is that humans are immune to adversarial images, with the corollary that ANNs 'reason' in a very different way than brains.

To better understand how artificial vision resembles and differs from biological vision, here we ask whether there are adversarial images that can fool the primate brain. Adversarial images have sometimes been defined relative to ground truth labels assigned by humans; by this definition, adversarial images cannot exist for humans. Adversarial images constitute as a measure of robustness of perception, and thus we define adversarial images based on the amount of change from an unambiguously classified original (*clean*) image.

There have been hints indicating that primate vision exhibits some sensitivity to adversarial images. Elsayed et al.[7] showed that humans doing visual categorization under tight time constraints can be biased by adversarial images. Zhou et al.[8] showed that humans can decipher the attack target in adversarial images crafted for CNNs (without necessarily making mistakes if asked to categorize the images). Berardino et al.[9] identified 'eigen-distortions,' which are directions of small pixel value change that most readily make images appear different, although this study did not investigate changing the categorization of images. None of these studies exam-

ined neuronal responses to the altered images, or crafted adversarial images based on neuronal responses.

We developed an adversarial attack method for primate visual recognition, a method we term *gray box attack* because it utilizes partial knowledge about the attacked system. Adversarial attack in ANNs typically relies on full knowledge of the internal architecture and weights in the system. In the case of primate vision, full access to the relevant circuitry is currently lacking. In our method, the primate visual system is regarded as a gray box, for which we know only the input (image stimuli) and a small sample of the internal representations (recorded responses from visual neurons) (Figure 1A). Using this information, we fit a *substitute model* that takes the same input and predicts corresponding neuronal responses. The substitute model comprises a linear mapping module[10] attached to the final convolutional layer of ResNet-101[11]. This model family has been shown to be able to predict a large fraction of visual responses[3,12]. We hypothesize that if the substitute model provides a reasonable approximation to visual neuron activity, we could use the model to design images to mislead the neurons. Moreover, to the extent that the modeled neural population underlies behavior, we should be able to fool the subject in a behavior task.

We focused on the well-characterized face-processing system in primate inferior temporal (IT) cortex[13] and created adversarial human faces to look like monkey faces, adversarial monkey faces to look like human faces, and adversarial non-face images to look like human faces. To anticipate the results, face-selective neurons in macaque IT miscategorized gray box adversarial images as the target category. Further, these images misled monkeys and humans during visual categorization.

Results

We created a substitute model of neuronal responses recorded from macaque inferior temporal (IT) cortex. To build the substitute model, we used the ResNet-101 neural network[11] pre-trained on object categorization on ImageNet[14]. The network was fine-tuned using monkey and human faces, two additional categories not present in ImageNet, thus creating a total of 1,002 categories. A linear mapping module was then fit from the last convolutional layer features of the network to predict responses of 22 neurons (recorded from monkey P) to a few thousand pictures of objects[15] (Figure 1A). The substitute model achieved an average correlation of 0.70 between predicted and actual responses on held-out images not used during fitting.

Adversarial Images Altered Responses of Visually Selective Neurons to Target Patterns

First, we created adversarial human face images that were predicted to elicit monkey face-like neuronal responses (human→monkey attack). The original images were 40 human face images from the Chicago dataset[16]. Using the substitute model, we modified each human face image to make its model-predicted neuronal response more similar to measured neuronal responses to monkey faces. This objective function was optimized using the multi-step Fast Gradient Sign Method (FGSM)[6]. We defined 10 *noise levels* corresponding to adding different amounts of modification (Figure 1C) and generated 40 images at each noise level, resulting in 400 adversarial images in total. The average pixel-level distance between 600 clean human and monkey faces had a mean-squared error (mse) of 7000 (minimum = 2500). The highest level of noise we used was mse = 100, which was much lower than the minimum clean image distance. Therefore, the original image identity was largely preserved. Notably, the adversarial images were generated only to alter model-predicted neuronal responses, not to change categorization by ResNet. Nevertheless, all adversarial human faces were categorized as monkey faces by our

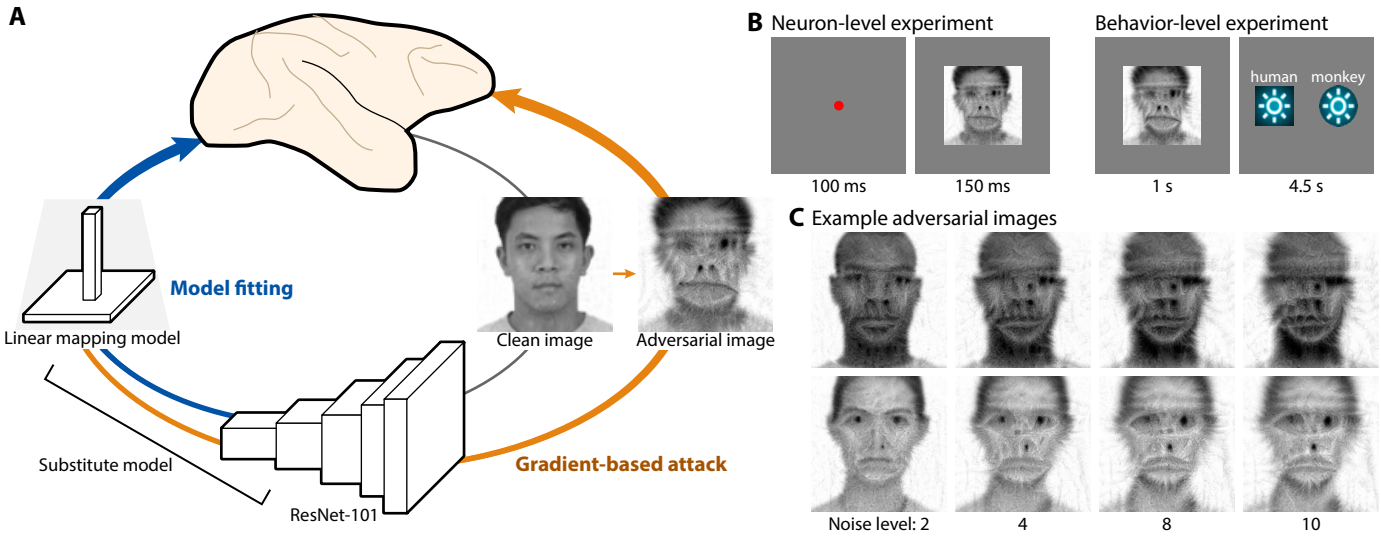

Figure 1: Overview of adversarial attack. (A) A substitute model was fit on IT neuron responses. The substitute model consisted of a pre-trained ResNet-101 (excluding the final fully-connected layer) and a linear mapping model. Adversarial images were generated by gradient-based optimization of the image to create the desired neuronal response pattern as predicted by the substitute model. (B, left) The adversarial images were tested in monkeys in neuron-level experiments. Monkeys fixated on a red fixation point while images were presented in random order and neuronal responses were recorded. (B, right) The images were tested in behavioral experiments with monkeys and human subjects. Each image was presented for 1000 ms. For monkeys, two choice buttons were presented (text for illustration only). Monkeys were rewarded for touching the correct button for training images and a random button for test images. Humans were instructed to press a key to indicate the correct option. (C) Example human→monkey attack images, based on two original human faces, are shown for different noise levels.

fine-tuned ResNet in a 1002-way categorization.

Next, we presented the adversarial images, together with 300 unmodified (*clean*) human faces and 300 clean monkey faces (Figures S4), to three monkeys (P, R, and B1). The clean images did not include originals of the attack images. Neuronal responses were recorded from the medial-lateral (ML) face patch (monkeys P and R) and anterior-medial (AM) face patch (monkey B1) in inferior temporal cortex, areas strongly selective for faces[17] (Figure S1).

We visualized the neuron population representation of clean and adversarial images (level 10) using Uniform Manifold Approximation and Projection (UMAP)[18] (Figure 2A,D,F). Neuronal representations of clean human and monkey faces were clearly separable in all monkeys, while adversarial images were shifted away from human faces towards monkey faces. To quantify whether the adversarial images were represented as the target category, we trained support vector machines (SVMs) to classify clean images as human or monkey based on neuronal responses. Five hundred different subsets of 80% of the data were used to train 500 SVMs. The SVMs were then used to classify the held-out 20% of clean images along with the adversarial images.

We first present results in monkey P (Figure 2A, B), whose neuronal responses were modeled by the substitute model. The SVMs achieved high test accuracy ($97.4\pm0.8\%$) on clean human faces, indicating that the recorded neurons distinctly represented human and monkey faces, consistent with qualitative UMAP visualization. In contrast, many human$\rightarrow$monkey adversarial images were classified as monkey faces. We defined *success rate* for the adversarial attack as the fraction of images classified as the targeted category. For comparison, clean images achieved a null 'success rate' of $2.6\pm0.8\%$. The success rate of adversarial images increased with noise level, reaching $32\pm7\%$ at level 10 noise (Figure 2B,E,G). We also calculated the distance between adversarial images and both the original and target classes in neural state space (Figures 2A, S5A). Each image was localized along the norm of the SVM decision hyperplane.

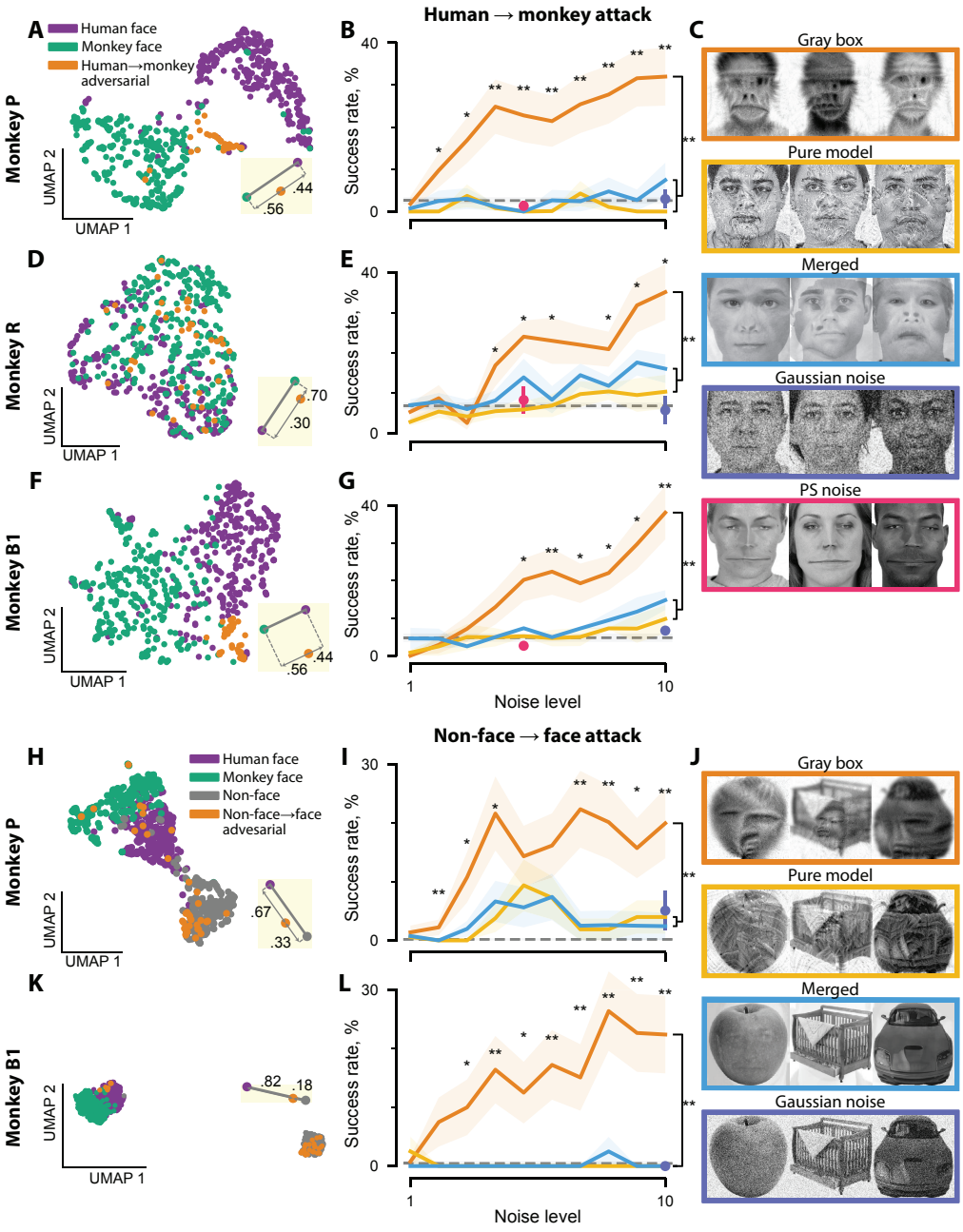

Figure 2: Neuron-level results of adversarial attack. (A–G), Human→monkey attack. (A), UMAP visualization of neuronal representation of images in monkey P. 'Gray box attack human face' corresponds to noise level 10. Inset shows average distances from adversarial images to clean human faces and clean monkey faces, along the direction that best separates the latter two and normalized to the distance between them. Points in inset show centers of mass of UMAP points for illustration only, and do not correspond to the distance quantification. (B) Success rates of attack and control images (pure model, merged, Gaussian noise, and PS noise images) at different noise levels. Legend and example images are in (C). Shading and error bars show s.e.m. over bootstrap samples. *, $p<0.05$ and **, $p<0.01$. (D, E), Same as (A, B) for monkey R. (F, G), Same as (A, B) for monkey B1. (H–L), Same as (A–G) for non-face→face attack in monkeys P and B1.

Compared to an average distance (normalized to 1) between monkey and human faces, the average distance was reduced by nearly half to 0.56 between human→monkey adversarial images and monkey faces.

Were adversarial images being misclassified simply because noise degraded image quality? Could similar success rates be obtained by creating adversarial images without fitting the substitute model to neuronal data? To test these possibilities, we generated attack images using 4 other methods (Figure 2C): (1) adding random *Gaussian noise*; (2) interpolating between human and monkey face images on a pixel level (*merged*); (3) generating adversarial images for our full fine-tuned ResNet-101 without mapping to neuronal responses (*pure model*); (4) manually editing images in Photoshop (*PS noise*) to introduce common monkey features we subjectively assessed in the adversarial images (namely, wider mouth and narrower nose). Pure model and merged images were generated at the same 10 noise levels; Gaussian noise images were only tested at level 10; PS noise images were only tested at level 5.

All types of control images achieved generally lower success rates than our adversarial images (Figure 2B). At level 10 noise, Gaussian noise, pure model, and merged images achieved 3$\pm$2%, 0$\pm$0%, and 7$\pm$4% success rate respectively, in comparison to 32$\pm$7% achieved by gray box adversarial images. At level 5 noise, PS noise images achieved 1$\pm$1% success rate, compared to 23$\pm$7% for level 5 gray box adversarial images. Starting at level 2, success rates for gray box images were significantly higher than those of merged and pure model images ($p = 0.047, 0.022$ at levels 2, 3, $p \leq 0.005$ at higher levels; one-tailed permutation test, FDR-corrected across 30 tests in 3 monkeys). We also compared the effectiveness of attack methods as quantified by the linear regression slope between noise level and success rate. Gray box adversarial images were significantly more effective than pure model and merged images ($p < 10^{-5}$, FDR-corrected across 3 tests in 3 monkeys). Finally, for Gaussian noise, pure model, and merged images respectively, the normalized SVM distances to monkey faces were 0.94,

0.98 and 0.98, which were little reduced from the distance (1) for clean human faces and larger than the distance (0.56) for the gray box adversarial images. Thus, results together from UMAP visualization, SVM performance, and cluster distance indicate that human→monkey adversarial images were represented as more similar to monkey faces. This performance was not matched by simple image manipulations or pure ANN-based adversarial attack.

Neuronal responses in monkey R were less selective between monkey and human faces than those in monkey P (Figures 2D, S3). Nevertheless, an SVM classifier achieved 93.1±0.8% validation accuracy in categorizing clean human faces. The same gray box adversarial images (generated based on a model of responses in monkey P) achieved 35±7% success rate at level 10 noise, with increasing success rates at higher noise levels (Figure 2E). The success rates of adversarial images were significantly higher than those of control images starting at noise level 4 ($p \leq 0.047$), except at noise level 7 ($p = 0.059$); success rates also improved with noise level significantly more for gray box attack than control images ($p < 10^{-5}$). The normalized distance between level 10 adversarial images and monkey faces was reduced to 0.7 along the SVM direction (Figures S5, 2D inset). Similarly, in monkey B1, gray box adversarial images were more effective than control images starting from noise level 5 ($p \leq 0.016$; Figure 2G), and improved more steeply as noise level increased ($p < 10^{-5}$). Gray box images achieved 38±7% success rate at level 10 and reduced the distance to monkey faces to 0.56 (Figure 2F inset). Thus, the same adversarial images, created by a substitute model fitted to monkey P neuronal responses, were transferable across monkeys.

Misclassification of Adversarial Images Was Not Due to Confusion with Non-face Images

So far, we examined only the relationship between adversarial and clean face images. However, if the adversarial images became dissimilar to faces, the images might also become difficult to

categorize by an SVM trained on face images. To test this possibility, we recorded neuronal responses in monkeys P and R to 300 non-face images (30 object categories, Figure S4) along with human faces, monkey faces, and adversarial images (noise level 10). UMAP visualization of neuronal responses showed that non-face images were clearly separated from human and monkey faces, while adversarial images clustered with face images (Figure S6). SVMs trained to classify clean face from non-face images achieved 97±2% and 94±3% validation accuracy in monkeys P and R, respectively. The same SVMs classified 98±2% and 98±3% of adversarial images as face. Thus, the adversarial images remained represented within the broader category of faces, despite a change in their sub-categorization from humans to monkeys.

The Same Model Can Perform Attack to Different Target Categories

So far, we described human faces adversarially attacked to look like monkey faces (human→monkey). We attempted a more challenging attack objective: modifying non-face images to elicit face-like responses. The images (examples in Figure 2J) were tested with neurons recorded from monkeys P and B1. Neuronal responses in both monkeys were highly face-selective (Figure 2H, K, I, L): Clean non-face images were misclassified only 0.2±0.2% of the time in monkey P and 0.5±0.4% in monkey B1. In comparison, gray box attack images at level 10 noise achieved 20±6% and 22±7% success in the two monkeys. Pure model images achieved 4±3% and 0±0% success, while merged images achieved 2±2% and 0±0% success. Gray box attack images were significantly more successful than both alternative methods at most noise levels in both monkeys. The success rate also improved more rapidly with noise for gray box attack than for alternative methods ($p = 4\times10^{-5}$ for monkey P and $p < 10^{-5}$ for monkey B1).

We also attempted monkey→human attack, generating 40 images for each noise level using the same substitute model. Monkey→human images achieved generally lower success rates than human→monkey images (Figure S7B, C, D), and comparisons on most individual noise

levels did not reach statistical significance, possibly due to the small number of images used and large variance among images. However, gray box attack still improved significantly more rapidly with noise than alternative methods ($p = 4\times10^{-4}$, 5×10^{-5}, 5×10^{-5} for monkeys P, R, and B1 respectively).

Adversarial Images Specifically Changed Monkey Behavioral Classification

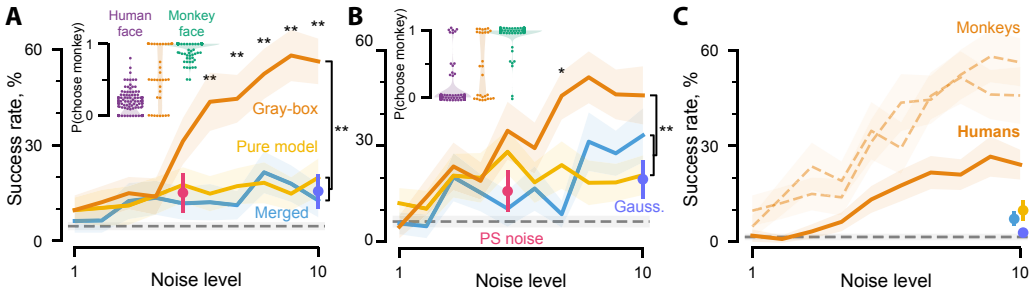

Figure 3: Behavioral results of human→monkey adversarial attack. (A) Behavioral attack success rates in monkey B for attack and control images (pure model, merged, Gaussian noise, and PS noise images) at different noise levels. Inset shows the probability (over repeat trials) of choosing monkey for individual images of clean human faces, clean monkey faces, and level 10 attack images respectively; each point represents one image, and shading represents kernel density estimates. (B) Same as (A) for monkey O. (C) Behavioral attack success rates on human subjects performing the categorization task on Amazon Mechanical Turk. Color scheme is the same as in (A, B). Dashed lines shows success rates in monkeys for comparison.

Can the adversarial images change visual categorization by monkeys on a behavioral level? To answer this question, we trained monkeys B5 and O to categorize human and monkey faces (two-way categorization) on a touchscreen (Figure 1B). Both monkeys achieved around 95% accuracy after training. Next, the monkeys were tested on control images and on the adversarial images generated by the model based on the neuronal recordings from monkey P. To minimize alternative strategies during testing, adversarial images were shown randomly and infrequently (5% of images) among a large number of clean faces. Correct reward was given for clean images, while random reward (50% probability regardless of choice) was given for adversarial

and control images. Each image was tested in 1–8 repeat trials. To calculate success rate, we randomly sampled one response per image, repeating 100,000 times.

The monkeys seldom made mistakes on clean images (5±1% and 6±2% in monkeys B5 and O respectively). In comparison, gray box human→monkey adversarial images achieved high success rates, fooling both monkeys 56±7% and 46±9% of the time, respectively (level 10 noise; Figure 3A, B). Gray box adversarial attack was significantly more efficient over noise levels than pure model and merged images ($p < 10^{-5}$ in B5, p = 0.001 in O). Although the average success rate was about 50%, individual adversarial images were widely distributed in their categorization performance (Figure 3A, B inset).

We also tested the reverse monkey→human attack images in monkey O. Gray box adversarial images achieved 22±7% success rate, compared to 8±5% and 11±5% for pure model and merged images (level 10 noise; Figure S7B). Over noise levels, gray box adversarial images were significantly more effective than both alternative approaches (p = 0.001).

In summary, neuronal and behavioral results show that monkey visual recognition can be selectively biased by adversarial images. The effectiveness of attack is greatly aided by modeling responses of neurons related to the task. Even though the model was based on just 22 neurons recorded from one monkey, the generated images successfully transferred to neurons and behavior across monkeys.

Adversarial Images Specifically Changed Human Categorization

Could the same adversarial images also mislead human judgment? We recruited human subjects through Amazon Mechanical Turk to categorize the same adversarial images evaluated in the previous two sections, using a similar two-way categorization task as in the monkey experiments (Figure 4B, right). The adversarial images were shown infrequently among clean images, and subjects had ample time to examine the images (1 s).

From 73 subjects who completed the task, we included in the analysis 40 subjects who had >95% accuracy on clean images (high performance was not incentivized; most of the remaining 33 subjects had accuracy between 90–95%). Each subject was tested on only one noise level of images, and 4 subjects saw each noise level. Attack success rates of the adversarial images again increased with higher noise, reaching 24±4% at noise level 10. This success rate was higher than those of control images (2%–6%), which were in turn close to the error rates on clean images (3±1%) (Figure 3C). Overall, attack success rates for humans were lower than for monkeys. These results demonstrate that gray box adversarial images, created based on monkey face-selective neuron responses, can transfer to non-time-limited human visual categorization.

Adversarial Training Immunized Monkeys against Future Attack

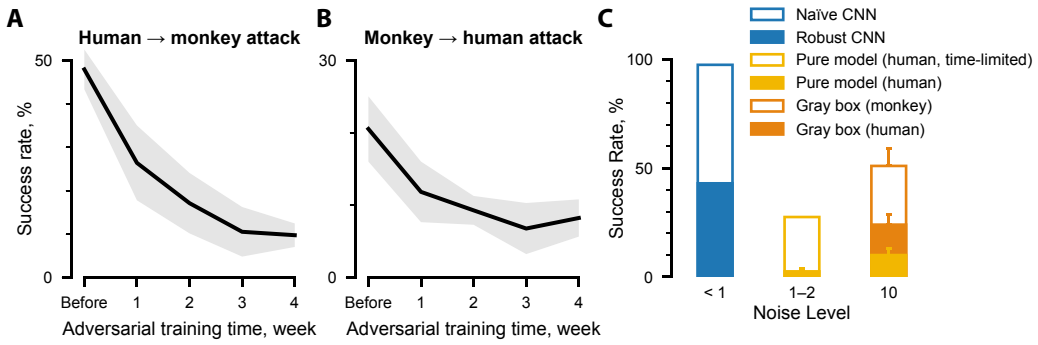

Figure 4: (A, B) Results of adversarial training in monkey O. Over training time, behavioral attack success rate decreased for held-out level 10 adversarial images. Results are shown separately for human→monkey and monkey→human images. (C) Comparison of adversarial attack robustness of different visual systems: naive CNN, robust CNN[19], time-limited[7] and time-unlimited human with Pure Model Images, and time-unlimited human and monkey with gray box adversarial images.

When ANNs are trained with correctly labelled adversarial images, a process termed *adversarial training*, future adversarial attack is much less successful. In other words, adversarial training improves the robustness of ANNs to adversarial attack[20]. Can monkey performance on adversarial images also be improved by adversarial training? To test this hypothesis, we

trained monkey O on adversarial images rewarded according to the original labels. Eighty each of adversarial human images and monkey images were generated with level 10 noise using the same substitute model above, then split evenly for training and testing. Monkey O was trained for 1 month, during which time we measured test performance weekly.

Figure panels 4A, B summarize the results of adversarial training. The success rate of adversarial human→monkey images decreased from 48±4% before training to 27±7% after 1 week of training, eventually reaching 10±5% after 1 month of training (Figure 4A) and approaching the error rate on clean human images (5±2%). Similarly, the success rate of adversarial monkey→human images decreased from 23±6% to 7±5% over 1 month, close to the error rate on clean monkey images (4±4%). These results demonstrate that adversarial training can improve the robustness of a monkey's behavioral categorization against adversarial attack.

Discussion

Perception does not merely mirror reality, as evinced by centuries of visual illusions and an even longer history of art. We designed a specific kind of visual illusion, namely minimally changed images that were originally unambiguous, yet come to be categorized differently. This illusion is distinct from phenomena such as pareidolia, which pertain to images intrinsically ambiguous or misleading. Our adversarial images fooled both humans and monkeys in practically time-unlimited visual categorization. Further, the altered categorization is reflected in monkey category-selective neuronal responses.

The dramatic sensitivity of ANNs to adversarial images raises important challenges for the safety of deep learning applications. Meanwhile, this susceptibility to small noise has also been cited as evidence that ANNs make inferences in a fundamentally different way than primate brains[21,22]. We established an upper bound on the minimum amount of distortion needed to alter primate visual perception. The amount of change (mse = 100) was modest, much smaller

than the average difference between the 600 clean human and monkey faces used (mse = 7000). On the other hand, the noise level was much higher than that needed for attacking ANNs (Figure 4C), although a direct comparison is unfair because we lack complete knowledge of the biological circuits. Incrementally bridging the gap in adversarial vulnerability can lead to better artificial intelligence systems and also to better computational models of visual cortex. On the one hand, more robust algorithms may be designed that leverage neural mechanisms of robustness. On the other hand, better models of the neural circuits may offer ever finer abilities to alter perception.

Emerging evidence suggests that adversarial vulnerability in ANNs may not be a defect, but rather reflects the use of non-robust yet highly informative features for categorization[23,24]. In this light, our adversarial images may also reveal the use of subtle features by primate visual recognition. This hypothesis is consistent with observations that category-selective neurons respond to diagnostic visual features even when they are dissociated from the preferred categories[25,26,27].

Our method depends on a computational model of visual neurons, outperforming a method based purely on ANNs. Previous work has shown that computational models of neurons can be used to predictively control their activity[28,29]. Other studies have outlined a link between visual neuronal activity and behavior[30,31,32]. We bridge the gap, showing that primate behavior can be controlled specifically using a computational model of neuronal activity. These results suggest a conceptual possibility for designing visual stimulation to precisely affect mental states and potentially treat neurological disorders[33].

Methods

Data and code availability The data underlying all figures and code for gray box attack will be made available upon publication at `https://github.com/yuanli2333/Adversarial-images-`

Subjects Four adult male macaca mulatta (9–12 kg; 5–13 years old) and one adult male macaca nemestrina (13 kg, 11 years old) were socially housed in standard quad cages on 12/12 hr light/dark cycles. All procedures on non-human primates were approved by the Harvard Medical School Institutional Animal Care and Use Committee, and conformed to NIH guidelines provided in the Guide for the Care and Use of Laboratory Animals.

Human psychophysics experiments were conducted online on Amazon Mechanical Turk. All participants provided written informed consent and received monetary compensation for participation in the experiments. All experiments were conducted according to protocols approved by the Institutional Review Board at Children's Hospital.

Neuronal recording Monkeys P, R, and B1 were implanted with chronic microelectrode arrays (MicroProbes, Gaithersburg, MD). Recording sites were targeted to the medial lateral (ML) face patch for monkeys P and R and the anterior medial (AM) face patch for monkey B1. Array targets were localized using fMRI aligned to CT scans, and during surgery, using landmarks from the CT scans. Localization was confirmed after surgery by CT scans. Extracellular electrical signals were amplified and recorded using the Omniplex data acquisition system (Plexon, Dallas, TX).

Animals were implanted with custom-made titanium or plastic headposts before fixation training. After several weeks of fixation training, the animals underwent a second surgery for array or chamber implantation. All surgeries were performed under general anesthesia using sterile technique.

Substitute model To generate adversarial images, a substitute model was trained to predict IT neuronal responses[3]. The model was fixed before testing any of the adversarial images at the neuronal and behavioral levels. The substitute model comprised a pre-trained CNN (ResNet-101[11] trained on ImageNet) and a linear mapping model (Figure 1A). Because the 1,000 pre-

trained categories do not include our categories of interest (human face and monkey face), we collected thousands of images for these two categories and fine-tuned the ResNet-101 on them. Next, a linear model was trained to map extracted image features (layer conv5_3, the highest convolutional layer) to neuronal responses. The linear model was factorized in the spatial and feature dimensions[10]. The spatial module was a convolutional kernel W_s. The feature module was a mixing pointwise convolution W_t, i.e., a weighted sum over the feature dimension. Both W_s and W_t were parameterized separately for each IT neuron. In sum, the response for neuron $n = 1, \ldots, 22$ to image x is modeled as

$$\hat{y}_n = \sum (W_s^n * \text{ResNet}_{\text{conv5\_3}}(x)) \cdot W_t^n + W_d^n, \tag{1}$$

where $*$ is the convolution operation and W_d^n is a bias parameter. The parameters were jointly optimized to minimize a loss function $\mathcal{L}_e$ composed of the prediction error $\mathcal{L}_p$, an L2-regularization loss $\mathcal{L}_2$, and a spatial smoothness loss $\mathcal{L}_{\text{laplace}}$:

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

$$\mathcal{L}_2 = \lambda_s \sum \|W_s\|_2 + \lambda_t \sum \|W_t\|_2 \tag{3}$$

$$\mathcal{L}_{\text{laplace}} = \lambda_l \sum W_s * \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{4}$$

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

The hyper-parameters $\{\lambda_s, \lambda_t, \lambda_l\} = \{1, 0.1, 0.7\}$ were obtained from a grid-search to produce the highest prediction accuracy. The substitute model achieved an average correlation of 0.7

between predicted and actual IT neuron responses on held-out test images The preferred features of neurons modeled by the substitute model (Figure S2) were visualized using activation maximization[34].

Generating adversarial and control images The same adversarial and control images were used in neuron-level and behavioral experiments. Adversarial images were generated using the substitute model and multi-step fast gradient attack[6] (Figure 1B). Specifically, each adversarial image x_{adv} was iteratively generated from an unmodified (*clean*) image x as

$$x_{\text{adv}}^{t+1} = x_{\text{adv}}^{t} + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(\theta, x_{\text{adv}}^{t}, P)), \tag{6}$$

where x_{adv}^{t} is the adversarial image at iteration t; $\text{sign}(\nabla_x \mathcal{L}(\cdot))$ is the gradient of the cost function $\mathcal{L}$; ϵ is similar to a learning rate and controls the amount of change in each step; θ represents the substitute model parameters; and P is the neuronal response pattern of the target category. For example, in human→monkey attack, $x_{\text{adv}}^{0} = x$ is a clean human face, and P is the mean neuronal response to 300 monkey faces. The cost function at step t is

$$\mathcal{L} = \left\| M_\theta(x_{\text{adv}}^{t}) - P \right\|_2^2 + \mathcal{L}_v, \quad \mathcal{L}_v = \sum_{i,j} \left(|x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}| \right) \tag{7}$$

where M_θ is the substitute model, $\|\cdot\|_2^2$ is the L2 loss between model-predicted and target response patterns, $\mathcal{L}_v$ is the total variation loss, and (i, j) indexes pixels of the image. The total variation loss prevents high frequency noise from dominating the generated image features.

The amount of noise was measured as the mean squared error (mse) on pixel values (range 0–255) between the original and modified images. We defined 10 noise levels. The mse was 55 at level 1 and increased linearly to 100 at level 10. The adversarial noise is small relative to the difference between clean faces, which averaged about 7000 among the 400 clean human and monkey faces used in our work.

The four types of control images were generated from the same 40 clean human faces underlying gray box adversarial images. Pure Model Images were iteratively generated as

$$x_{\text{pure}}^{t+1} = x_{\text{pure}}^{t} + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(\theta_p, x_{\text{pure}}^{t}, y)) \tag{8}$$

where θ_p represents the parameters of the fine-tuned ResNet-101 (without mapping to IT neuronal responses), y ={human, monkey} is the label of the target category, and $\mathcal{L}$ is the cross-entropy loss.

For human→monkey Merged Images, we randomly sampled one monkey face x_m and merged it with each clean human image x_h: $x_{\text{merged}} = x_h + \lambda x_m$, where λ effectively controls the noise level. Similarly, monkey→human Merged Images were generated as $x_{\text{merged}} = x_m + \lambda x_h$.

Gaussian Noise Images were generated as $x_{\text{gauss}} = x + \epsilon g_x$, where g_x was sampled from the standard normal distribution.

For PS noise images, we used the Liquify tool in Photoshop to introduce features that we noticed in activation maximization visualization (Figure S2), such as wider mouth and narrower nose. The mse of the modification was calibrated to noise level 5.

Neuron-level attack experiments We recorded neuronal responses to clean human faces, clean monkey faces, and adversarial images in monkeys P, B1 and R. Images were presented in a randomized order in the receptive fields of the neurons (Figure 1B). Each image was presented for 100 ms, with a 150 ms interval between images. Monkeys were required to fixate within a window of 1.5–3 degrees in exchange for juice reward. Eye position was monitored using an infrared eye tracker (ISCAN, Woburn, MA).

Neuronal responses to an image were calculated as firing rates averaged within a response window (100–200 ms after stimulus onset for monkeys P and B1; 120–290 ms for monkey R) and over repeated trials. Visually selective neurons were selected using a split-half self-correlation criterion (raw correlation >0.6 for Monkey P, >0.3 for B1, and >0.2 for Monkey R;

thresholds were chosen to ensure enough neurons were selected, and differed due to different numbers of repeated trials completed by the monkeys).

Attack success was quantified using linear Support Vector Machines (SVMs). Each SVM was trained to categorize the images (two-way categorization) based on the corresponding neuronal responses. For example, in monkey→human attack, the SVM was trained to classify each image as either a monkey face or a human face. Each SVM was trained on 80% of clean images. Then, the trained SVM was used to classify the remaining 20% of clean images, attack images, and control images. Attack success rate was calculated as the percentage of images that were classified as the target category.

Behavior-level attack experiments in monkeys Monkeys B and O were trained on a touchscreen to categorize human and monkey faces (Figure 1B, right column). Each image was presented for 1000 ms, after which two choice buttons were presented (left button: human face; right button: monkey face). The monkey had to make a choice within 4500 ms (actual reaction times were much shorter, on average 700 ms for both monkeys) and was rewarded for making the correct choice. Both monkeys reached high performance after training (>95%). During testing, adversarial images were randomly intermixed with clean images and occurred with a low probability (5%) to minimize alternative strategies; reward was given for a correct choice on clean images and randomly (50% of the time) regardless of choice for adversarial images.

Behavior-level attack experiments in humans Subjects on Amazon Mechanical Turk were invited to perform a two-way image categorization task. Subjects were instructed to determine whether each image is a human face or a monkey face. To answer, subjects pressed the left arrow key for human face and the right array key for monkey face. Each image was presented for 1000 ms. No feedback was provided. Subjects performed with generally high accuracy (on clean images). We selected subjects with >95% accuracy to include in further analyses. Adversarial

images were randomly intermixed with clean images and occurred with a low probability (5%) to minimize alternative strategies.

Statistics In neuron-level attack, 500 random splits of clean images were used to fit 500 SVMs. Mean success rate was the average over 500 SVMs and images in one category. Standard error of the mean was calculated across 100,000 bootstrap sample means, where each bootstrap was sampled with replacement over SVMs and over images within a category.

We compared the per-noise level success rates of gray box attack images to merged and pure model images. Tests were separate for each alternative method, noise level, and monkey. P-values were obtained from a one-tailed permutation test (100,000 permutations). To test the null hypothesis that gray box attack success rate was indistinguishable from both alternative methods, the two p-values were corrected for multiple comparisons using the Holm-Šídák method, and the larger p-value was used. Combined p-values were corrected for false discovery rate over monkeys and noise levels using the two-stage step-up method of Benjamini, Krieger and Yekutieli (30 comparisons in Figure 2B, E, G; 20 in Figure 2I, L).

We also compared gray box adversarial images to merged and pure model images in terms of the linear regression slope between noise level and success rate. The distribution of regression slopes for each alternative hypothesis were approximated by the distribution of slopes fit to each bootstrap sample. The p-value was thus the fraction of slopes in the alternative distribution equal to or larger than the observed slope for gray box images (a one-tailed test). The combined p-value was obtained as above, namely by taking the larger p-value of the two alternative hypotheses after Holm-Šídák correction. Combined p-values were corrected for false discovery rate over monkeys as above (3 comparisons in Figure 2B, E, G; 2 in Figure 2I, L; 2).

In monkey behavior-level attack, the success rate of each method at each noise level was defined as the percentage of images that were categorized by the monkey as the target category.

Each monkey responded 1–8 times to each image across repeated trials. We randomly sampled one choice per image and calculated the success rate over images, repeating 100,000 times. We reported the bootstrap mean and standard error of the mean. Statistical tests of per-noise level success rate and success rate improvement with noise were conducted as in neuron-level attack experiments (100,000 permutations; FDR correction over 20 comparisons for per-level tests and 2 comparisons for slope tests in Figure 3A, B).

Adversarial training Monkey O was trained to classify clean human and monkey face images as well as human→monkey and monkey→human gray box adversarial images. During training, clean and attack images were evenly mixed. The correct label for attack images is the label of the original images. 1000–2000 trials were completed per day (3–8 hours). Once per week, we tested the monkey's accuracy on held-out adversarial images. During testing, adversarial images occurred infrequently among clean images (1 in 20).

Acknowledgments

This work was supported by NIH grants R01EY16187, P30EY012196, R01EY026025, and NSF STC award CCF-1231216 to the Center for Brains, Minds and Machines at MIT. Jiashi Feng was supported by grants AI.SG R-263-000-D97-490 and MOE Tier-II R-263-000-D17-112. We thank Hu Zhang, David Castro, Ariana Sherdil and Peter F. Schade for discussion and assistance during the work.

Author contributions

LY, WX, and JF conceived of the study. LY, WX, GK, and MSL designed the experiments. LY developed the software for creating adversarial images. LY, WX, and MSL acquired the data. LY and WX analyzed the data and wrote the draft manuscript. All authors interpreted the data

and revised the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034 (2015).

[2] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).

[3] Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619–8624 (2014).

[4] Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology* **15**, e1006897 (2019).

[5] Szegedy, C. *et al.* Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[6] Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[7] Elsayed, G. *et al.* Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, 3910–3920 (2018).

[8] Zhou, Z. & Firestone, C. Humans can decipher adversarial images. *Nature communications* **10**, 1–9 (2019).

[9] Berardino, A., Laparra, V., Ballé, J. & Simoncelli, E. Eigen-distortions of hierarchical representations. In *Advances in neural information processing systems*, 3530–3539 (2017).

[10] Klindt, D., Ecker, A. S., Euler, T. & Bethge, M. Neural system identification for large populations separating "what" and "where". In *Advances in Neural Information Processing Systems*, 3506–3516 (2017).

[11] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

[12] Schrimpf, M. *et al.* Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv* 407007 (2018).

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

[14] Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

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

[16] Ma, D. S., Correll, J. & Wittenbrink, B. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* **47**, 1122–1135 (2015).

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

[18] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[19] Zhang, H. *et al.* Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573* (2019).

[20] Tramèr, F. *et al.* Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* (2017).

[21] Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* **19**, 356 (2016).

[22] Serre, T. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science* **5**, 399–426 (2019).

[23] Ilyas, A. *et al.* Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 125–136 (2019).

[24] Xie, C. *et al.* Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 819–828 (2020).

[25] Bracci, S., Ritchie, J. B., Kalfas, I. & de Beeck, H. P. O. The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience* **39**, 6513–6525 (2019).

[26] Wardle, S. G., Taubert, J., Teichmann, L. & Baker, C. I. Rapid and dynamic processing of face pareidolia in the human brain. *Nature Communications* **11**, 1–14 (2020).

[27] Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* 1–6 (2020).

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

[29] Walker, E. Y. *et al.* Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience* **22**, 2060–2065 (2019).

[30] Sheinberg, D. L. & Logothetis, N. K. The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences* **94**, 3408–3413 (1997).

[31] Afraz, S.-R., Kiani, R. & Esteky, H. Microstimulation of inferotemporal cortex influences face categorization. *Nature* **442**, 692–695 (2006).

[32] Rajalingham, R. & DiCarlo, J. J. Reversible inactivation of different millimeter-scale regions of primate it results in different patterns of core object recognition deficits. *Neuron* **102**, 493–505 (2019).

[33] Iaccarino, H. F. *et al.* Gamma frequency entrainment attenuates amyloid load and modifies microglia. *Nature* **540**, 230–235 (2016).

[34] Mahendran, A. & Vedaldi, A. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision* **120**, 233–255 (2016).

[35] Papernot, N. *et al.* Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 506–519 (2017).

[36] Liu, Y., Chen, X., Liu, C. & Song, D. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2017).

[37] Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J. & Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 15–26 (2017).

[38] Ilyas, A., Engstrom, L., Athalye, A. & Lin, J. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598* (2018).

[39] Narodytska, N. & Kasiviswanathan, S. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1310–1318 (2017).

Extended data figures

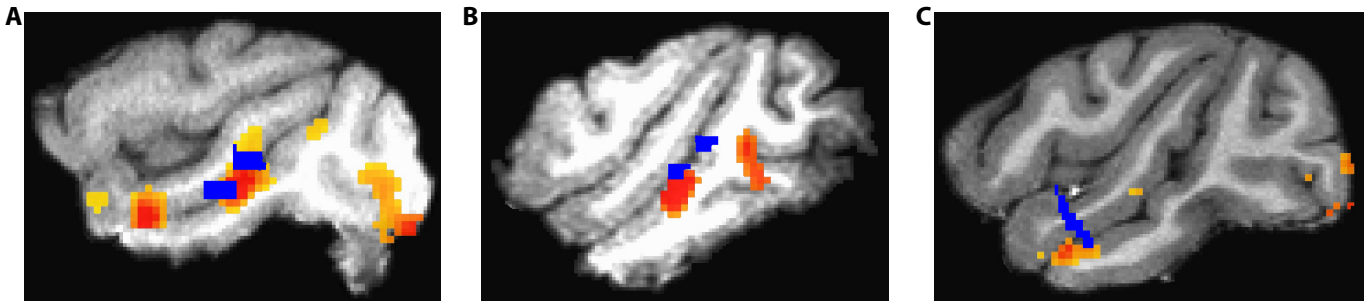

Figure S1: Location of recording arrays in monkeys P (A), R (B), and B1 (C) in relation to fMRI-defined patches selective for faces over objects. Blue indicates arrays localized by CT; red indicates face selectivity (q = 0.01).

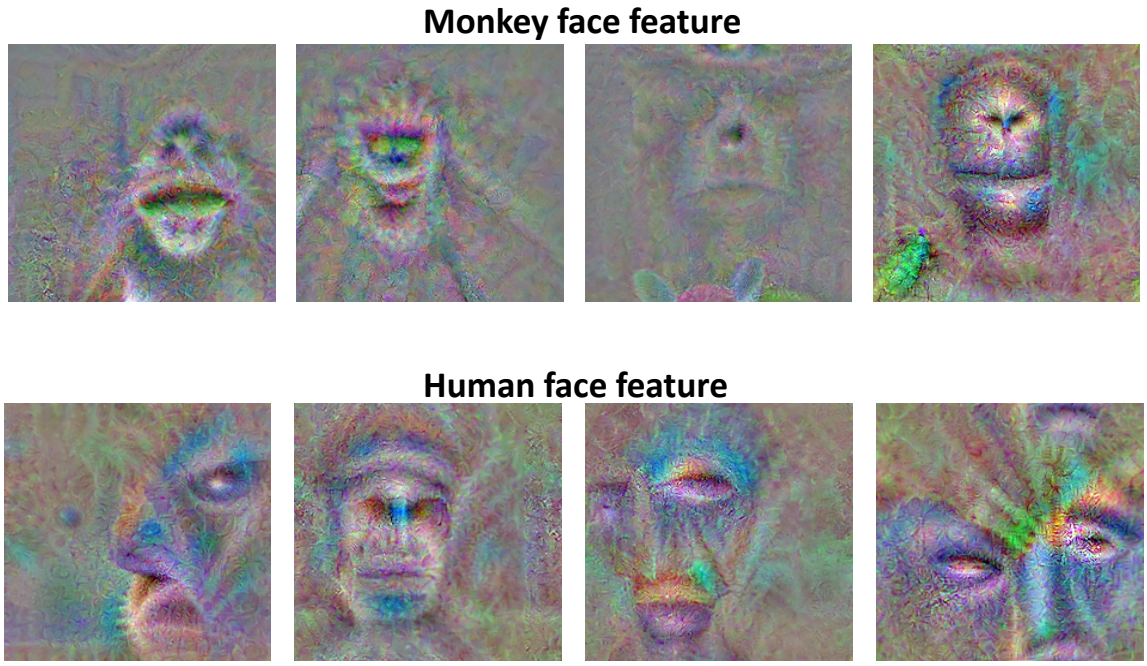

Figure S2: Visualization of the activation patterns for the substitute model fitted to 8 example neurons. Some of the substitute model features resemble monkey face features (first row) and others resemble human face features (second row).

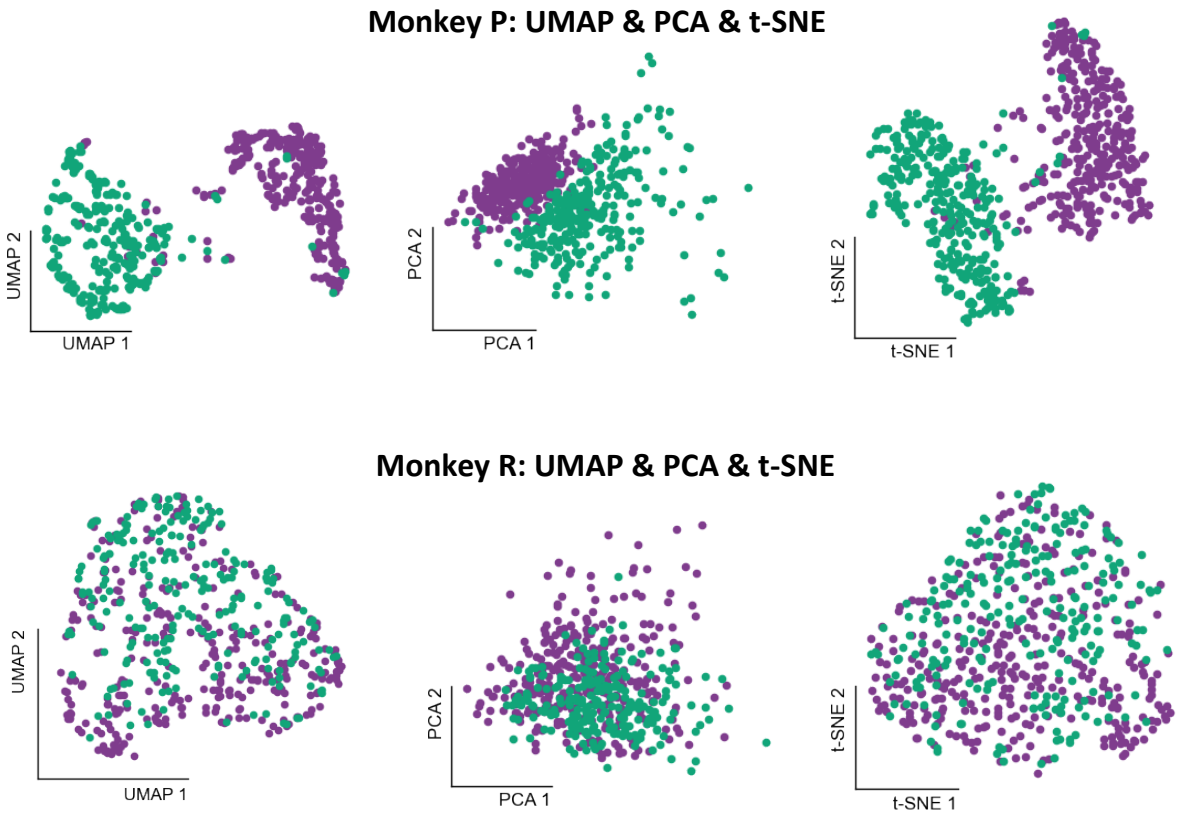

Figure S3: Neuronal representations of monkey and human face images visualized with different dimensionality-reduction techniques. Neuronal responses in monkeys P and R were projected in 2-D by UMAP, PCA, or t-SNE. Green and purple points represent monkey and human faces respectively.

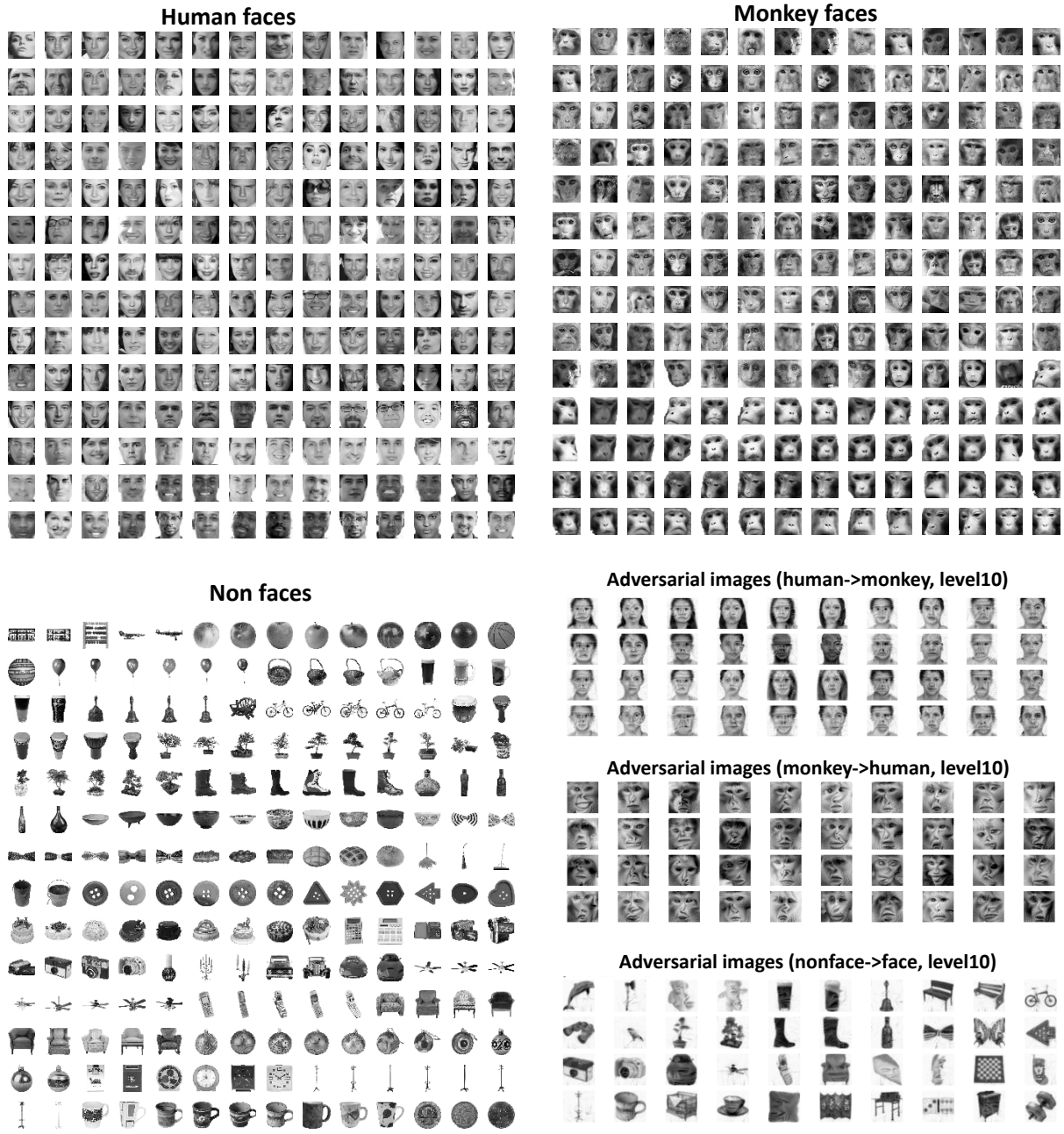

Figure S4: Images used in this study: human faces, monkey faces, non-face images, and adversarial images with level 10 noise.

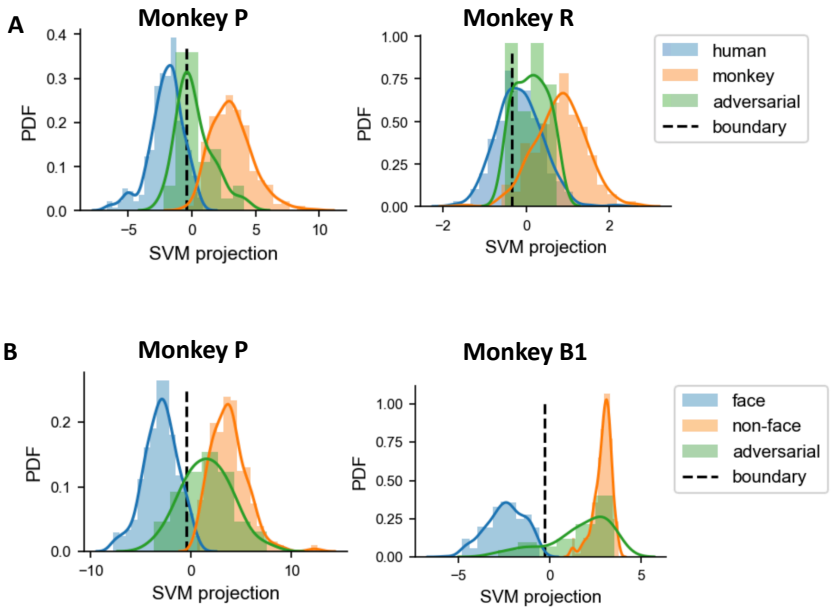

Figure S5: Distribution of neuronal response distance to the SVM boundary (black dotted line), grouped by image category. Panels A shows monkeys P and R neuronal responses in human→monkey adversarial attack. Panels B shows monkeys P and B1 neuronal responses in non-face→face adversarial attack.

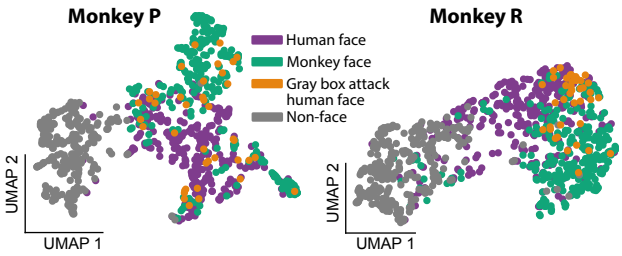

Figure S6: UMAP visualization of neuronal responses in monkeys P and R in human→monkey attack, showing that attack images remain more similar to other face images. Note that global shape may not be preserved when visualizing different data subsets by UMAP, which is a stochastic nonlinear method based on nearest-neighbor relations (cf. Figure 2A,D,H).

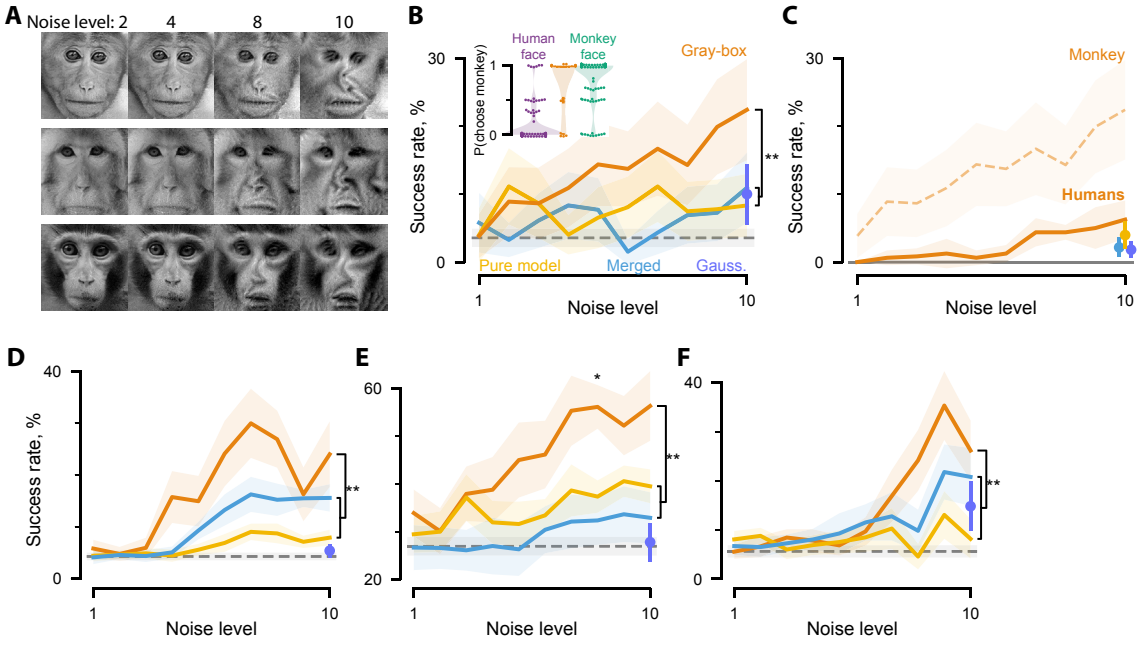

Figure S7: Results of monkey→human adversarial attack. (A) Example adversarial monkey faces at different noise levels. (B) Behavioral results in monkey O. (C) Behavioral attack success rates on human subjects performing the categorization task on Amazon Mechanical Turk. (D, E, F) Neuron-level results in monkeys P, R and B1, respectively.

Supplementary information

Attack success pattern of individual images across neuron-level and behavioral experiments. The adversarial images were generated by modeling only the neuronal responses in monkey P, but these images successfully extrapolated to neuron- and behavior-levels in other monkeys. Was the pattern of success over individual adversarial images also consistent across monkeys and experiments? We calculated the correlation, across individual adversarial images, between behavioral success rates in monkeys B and O. The correlation coefficient was weakly positive, although not statistically significant (Pearson's r = 0.28, Spearman's r = 0.31, Kendall's τ_b = 0.26; p >0.1 for all cases). There were also weakly positive but not statistically significant correlations between the distances to the SVM hyperplane in neuron-level experiments (monkey P) and the behavioral success rates (monkeys B and O) ($r_{Pearson}$ = 0.24, 0.12; $r_{Spearman}$ = 0.26, 0.13; $\tau_{Kendall}$ = 0.2, 0.1; p >0.1). Therefore, the consistent, weakly positive correlations suggest that attack success patterns over individual images could be partly general and partly idiosyncratic across monkeys and readout modes. Since we tested only 40 adversarial images, the statistical tests may not have enough power to detect a weak effect. Since there are multiple patches in the primate face-processing system (typically 6 in each hemisphere)[13], our recording from a limited number of neurons from only one face patch (ML) in monkey P may not strongly correlate with behavioral choices of other monkeys.

Related methods for adversarial attack. Several methods for black-box adversarial attack have been proposed. One approach, akin to ours, is to train a substitute model to mimick the target model (specifically, to mimic the output of the target model for each specific input)[6,35]. This method requires >2,000 target model queries to attack MNIST classifiers, which have simple inputs (28×28 pixel pictures of handwritten digits). On larger datasets of natural images such as ImageNet, adversarial images based on substitute networks do not transfer well to the

original model, for both targeted and non-targeted attack[36]. Another approach is to directly estimate attack gradients around each clean image, without using a substitute model[37,38,39]. This method requires on the order of 2×10^6 queries *per image* for a targeted attack objective comparable to ours[38], impractical for attacking primate recognition. Our gray box method lacks full knowledge about the decision-relevant neural circuit, but does leverage recording from a (small) subset of neurons likely involved in the task. Neuronal data were essential, as pure model images did not transfer well (Figures 2, 3, S7), consistent with prior work[36]. Additionally, our method benefits from existing knowledge of the ventral visual cortex. The backbone of our substitute model, a CNN pre-trained on object recognition tasks, has been shown to be successful in not only explaining but also predictively driving neuronal responses[3,28,29].