Small-scale adversarial perturbations expose differences between predictive encoding models of human fMRI responses

Nikolas McNeal[1,2,*], **Mainak Deb**[3,*], and **N. Apurva Ratan Murty**[4,5]

[1]Machine Learning, Georgia Tech
[2]School of Mathematics, Georgia Tech
[3]Independent contributor
[4]CoE in Computational Cognition, Georgia Tech
[5]Cognition and Brain Science, Georgia Tech

Editors: Marco Fumero, Clementine Domine, Zorah Lähner, Donato Crisostomi, Luca Moschella, Kimberly Stachenfeld

Abstract

Artificial neural network-based vision encoding models have made significant strides in predicting neural responses and providing insights into visual cognition. However, progress appears to be slowing, with many encoding models achieving similar levels of accuracy in predicting brain activity. In this study, we show that encoding models of human fMRI responses are highly vulnerable to small-scale adversarial attacks, revealing differences not captured using predictive accuracy alone. We then test adversarial sensitivity as a complementary evaluation measure and show that it offers a more effective way to distinguish between highly predictive encoding models. While explicit adversarial training can increase robustness of encoding models, we find that it comes at the cost of brain prediction accuracy. Our preliminary findings also indicate that the choice of model features-to-brain mapping might play a role in optimizing both robustness and accuracy, with sparse mappings typically resulting in more robust encoding models of neural activity. These findings reveal key vulnerabilities of current models, introduce a novel evaluation procedure, and offer a path toward improving the balance between robustness and predictive accuracy for future encoding models [1].

1 Introduction

Artificial neural networks (ANNs), loosely inspired by the architecture of the visual cortex, have become the leading models for understanding human vision [1–3]. These models excel not only at complex tasks like object recognition (e.g., ImageNet classification) but also provide a valuable framework for studying visual cognition more broadly [4–6]. ANN-based encoding models, which map neural network features to brain activity, have unlocked a key ability to predict responses at the level of single neurons [7], voxels [8], entire brain regions [9, 10], and even human and non-human primate behavior [11–15]. Early work established a link between a model's performance on complex tasks (like ImageNet) and the ability to predict brain responses: better task performance

[*]These authors contributed equally to this work.
[1]Code is available at `https://github.com/murtylab/adversarial-attacks-brainmodels/`

Proceedings of the II edition of the Workshop on Unifying Representations in Neural Models (UniReps 2024).

typically translated to better brain/behavioral predictions [16, 1, 17]. However, this relationship has plateaued; despite continual improvements in task performance, gains in brain prediction accuracy (henceforth predictivity) have largely stalled. This observation raises critical questions: Are models with similar predictivity learning the same features, or are key differences going unnoticed? Is there a more effective metric that can reveal these differences and help us identify the better models, even when their predictivity appears to be equally high? In this work, we show that small, imperceptible (to humans) adversarial attacks on predictive encoding models can reveal meaningful differences, providing a sharper lens to evaluate their fidelity as models of the brain.

The concern that encoding model predictivity has plateaued is not new [16, 9, 10, 17, 18]. This stagnation has sparked two major responses within the field. On one front, this challenge has driven the development of *entirely new models* incorporating aspects of brain-like operations (like recurrence [19, 20, 2, 21, 22]) or by directly aligning with behavioral or neural data [23–25]. The second front challenges predictivity as the primary metric altogether, advocating for *alternative evaluation methods* like centered kernel alignment [26–29] or single-neuron selectivity [30, 31] to capture more nuanced aspects of brain-model alignment. In this work, we are advocating a slightly different strategy. Predictivity must remain a vital benchmark measure of our models: predictive models have enabled new understanding of brain function including the ability to modulate responses in the visual cortex [32–35]. However, predictivity alone is insufficient, especially when we are limited by data. We propose complementing predictivity measures with additional evaluation metrics. Here we introduce adversarial sensitivity as a potential tool for stronger model evaluations.

Adversarial perturbations have long plagued AI systems. Previous work has shown that tiny, imperceptible changes to an image can drastically alter model predictions [36–42]. This issue has driven extensive research into making AI models more robust, particularly for mission-critical applications. Yet the impact of adversarial perturbations has received surprisingly little attention in vision neuroscience. Some work has explored "robustified" encoding models, either through training directly on neural data [43] to estimate neural robustness or by employing explicit robust pre-training to modify percepts [44–46]. To our knowledge, no study has directly examined the vulnerability of encoding models to targeted adversarial perturbations, the relationship between adversarial sensitivity and predictivity, or the impact of model mapping choices on the model's adversarial robustness. Understanding the bounds of our encoding models is crucial for progress. If imperceptible changes can distort model predictions, it raises concerns about their reliability in capturing true neural processes and ability to generalize to unseen data.

The central contribution of our work is threefold: (A) we demonstrate that encoding models are susceptible to small-scale adversarial attacks (Figures 1, 3), (B) we show that adversarial sensitivity is a potentially more effective way to differentiate between encoding models than predictive accuracy alone (Figure 3), and (C) we find that the choice of feature-to-brain mapping in encoding model can impact adversarial sensitivity, with sparse mappings producing relatively more robust models of neural activity (Figure 5).

2 Methods

2.1 fMRI Dataset

We used publicly available 7T fMRI data from the Natural Scenes Dataset (NSD) [47] for all analyses in this study. Specifically, we focused on the responses to 515 shared stimuli obtained from fMRI scans of eight subjects in category-selective brain regions. Each subject viewed these images three times over multiple experimental sessions. All analyses were conducted using version 3 of the dataset (betas_fithrf_GLMdenoise_RR), obtained directly from the NSD website. In this work we focused on the category-selective areas: fusiform face area (FFA) [48], extrastriate body area (EBA) [49], parahippocampal place area (PPA) [50], and the visual word form area (VWFA) [51]. To ensure the inclusion of only the most category-selective voxels, we applied a stringent threshold of tval > 7 for all analyses. Here, "tval" represents a conventional t-statistic derived from a two-sample t-test, quantifying the contrast between responses to each category (Word, Number, Body, Limb, Adult, Child, Corridor, House, Car, Instrument) against all other stimuli. Models were trained to predict the voxel and trial-averaged responses across subjects, as in previous work [9].

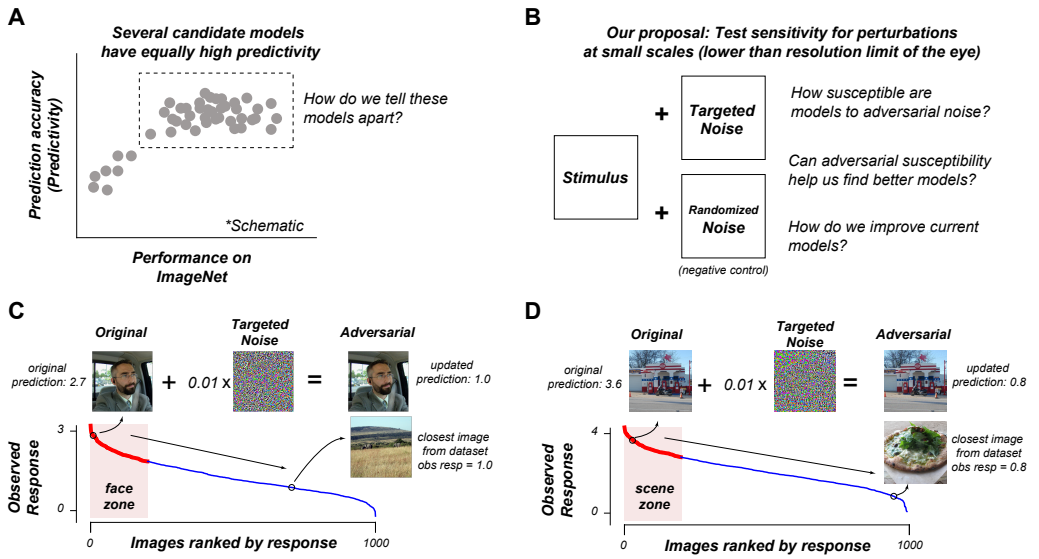

Figure 1: **Motivation, central questions and example adversarial perturbations on encoding models. A.** Schematic illustrating the trend observed in previous studies: many encoding models show similarly high predictions on brain data. Performance on ImageNet is shown on the x-axis, and prediction accuracy on brain data is shown on the y-axis. A similar figure with actual data can be found in previous work [1]. **B.** Strategy and central questions. For a given stimulus, we generate a targeted noise pattern (and a randomized control) to assess how sensitive the encoding model is to adversarial noise. **C.** Example of an adversarial perturbation applied to the FFA encoding model (VGG16). The model's prediction for the original face stimulus (top left) is significantly altered when a targeted imperceptible noise pattern is added (top middle). The modified image produces a much lower response (top right), similar to that for a non-preferred stimulus (bottom). The x-axis shows images sorted by response, and the y-axis represents the observed FFA response (percent signal change). The red region highlights the response for the preferred category (faces). **D.** Same as C, but for an example scene stimulus and the PPA.

2.2 Encoding Model

Typical ANN-based encoding models consist of two components: embeddings from a specific layer of the artificial neural network (serving as the representational basis) and a trainable mapping function. This mapping is typically done through regularized linear regression, which projects the features into the response subspace of neural activity.

Formally, we input each of our training images (see cross-validation schema next) into a representational encoder f and extract the latent feature vector $z_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. These features are then passed through our mapping function $g : \mathbb{R}^{C_l \times H_l \times W_l} \to \mathbb{R}^n$, where n is the dimensionality of the predicted neural data. To build the encoder model, we freeze f (the weights of the representational encoder) and train the mapping g.

Model architectures: We considered eight pre-trained artificial neural network architectures that have been previously validated against brain data. These include ResNet-50 [52], VGG16 [53], Inception v3 [54], SqueezeNet v1 [55] , AlexNet [56], CORnet-RT [57], DenseNet [58], and MobileNet-v2 [59].

To investigate whether increasing robustness improves the prediction accuracy of the encoding models, we also used publicly available models that were robustified through adversarial training [60]. These models share the same architecture (ResNet-50) and learning rule but differ in the degree to which they are trained adversarially. More details on the robust models and their training can be found in [61].

Encoding model mapping procedures: In this study, we experiment with five different mapping functions: ordinary least squares regression (OLS), lasso regression, ridge regression, a two-layer multi-layer perceptron (MLP), and a convolutional neural network (CNN). The first three mapping functions generate direct brain predictions, while the latter two involve learning at least one additional layer of features. These new features may enhance the model's ability to predict brain responses and could provide more representational robustness. However, the regression methods are computationally faster and do not require extensive hyperparameter tuning for convergence. Our two-layer MLP and CNN both include one hidden layer with 128 units. Note that we used OLS mapping for the first half of the paper because it is the most computationally efficient and does not rely on any assumptions.

Encoding model cross-validation procedure: We used the 515 shared images across all 8 subjects from the NSD dataset. We trained the model on a randomly chosen set of 400 images and all results in the study are based on predicted responses based on the held-out 115 images.

In Section 3.5, we investigate the effect of L_1 readout regularization on the adversarial robustness of the encoding model. We fit each model to the data using only one randomly chosen subject (subj2), testing six different values of the regularization coefficient α $(0.0001, 0.001, 0.005, 0.01, 0.05, 0.1)$. The α value that maximized predictive accuracy for this subject was selected for further analysis. Importantly, all model evaluations were conducted using an independent metric (adversarial sensitivity) and across all subjects.

2.3 Evaluating encoding model robustness

We evaluated adversarial robustness against the Fast Gradient Sign Method (FGSM) [37]. FGSM attacks are bounded by the L_∞ norm. That is, we find the maximum change δ (bounded by a "*perturbation budget*" ϵ) predicted to change the response of a given voxel. A successful attack would drastically (and unrealistically) change the predicted response of the encoding model. We quantified the adversarial sensitivity s_i for the i-th brain region/voxel using the method described in [43]. Specifically, we use a sensitivity metric s_i defined as:

$$s_i = \max_{||\delta|| \leq \epsilon} (g(f((x))) - g(f((x + \delta))))$$

Due to the computational cost of our adversarial attacks, we consider the mean sensitivity over subjects/regions (i.e., "sensitivity" is computed for the mean of the vector returned by $g()$). There are two things to note about this metric. First, since s_i is a measure of model *sensitivity*, high values on this metric would indicate lower adversarial robustness. We indicate this in several of our plots. The second is that since the metric does not have an upper bound, the results must not be interpreted across regions. While other forms of adversarial attacks exist in the literature, we focus on FGSM for simplicity and consistency.

2.4 Encoding Model Discriminability

We evaluate the ability of both metrics – adversarial robustness and model predictivity – to discrimininate encoding models of the brain. For each of the eight models evaluated, we compute the average sensitivity across all subjects and brain regions. We explore whether the spread of the adversarial robustness distribution of the encoding models will be greater than the spread of the model predictivity distribution (i.e., "adversarial robustness" serves as a better discriminative tool). To evaluate this, we test the variance and sparseness of both adversarial sensitivity and predictivity.

Normalized Variance: Since the scale of "sensitivity" (unbounded) and "predictivity" (bounded -1 to 1) are different, we cannot directly compare the variances. Instead, we first divide all accuracy and sensitivity values by their respective maximum value before reporting the variances (hence normalized variance).

Sparseness: We use the sparseness metric defined in [62, 63]. Specifically, for a distribution of values $P(r)$, sparseness (S) is computed with the following:

$$S = 1 - \frac{E[r]^2}{E[r^2]},$$

where $E[\cdot]$ denotes the expectation operator.

3 Results

Our investigation focuses on category-selective regions—specifically face, body, scene, and word-selective areas (FFA, EBA, PPA, and VWFA, respectively) from the Natural Scenes Dataset (NSD). These regions were chosen because of the extensive work on developing encoding models for them and because they provide the necessary foundational intuition for interpreting changes due to adversarial perturbations (Figure 1C, 1D). The extensive literature on category-selective regions offers strong priors on the expected response magnitudes. For example, the fusiform face area (FFA) is known to exhibit strong responses to faces and much weaker responses to non-preferred categories, such as scenes. This prior knowledge about its selectivity can be used to illustrate how adversarial noise can disrupt these predictable response patterns. Focusing on category-selective regions makes it easier to demonstrate the vulnerability of encoding models, which is why we prioritized these regions in out study.

We specifically focus on *very small image perturbations* ($\epsilon \leq 3/255$) lower than the resolution limit of the human eye and hence imperceptible to humans. This is because the response of brain voxels to these targeted noise patterns remains currently unknown. Restricting our analysis to small magnitudes ensures that the adversarial sensitivities we detect are real and meaningful.

3.1 Several ANN-based encoding models predict voxel responses equally well

We first set out to replicate the previous finding that encoding models exhibit similar accuracy in predicting brain responses. To do this, we examined eight pre-trained neural network architectures that have been reported extensively in prior work [16, 10, 9]. For each model, we focused on features from an intermediate layer, selecting the layer that had previously been shown to achieve the highest cross-validated accuracy in predicting responses from category-selective regions based on an independent fMRI dataset [9]. This choice removed experimenter degrees of freedom. Next, we constructed encoding models by mapping the features from a subset of images to brain responses using a linear mapping function (see Methods for details on cross-validation procedures). This entire process is depicted schematically in Figure 2A.

Overall, we found that these ANN-based encoding models were highly effective at predicting brain responses to held-out images (replicating previous findings [10, 9]). This is illustrated for an example brain region (EBA) in Figure 2B (R = 0.76, P<0.00001). Across all regions we considered, the models were able to predict nearly all of the explainable variance in the observed data. The prediction accuracy for each model architecture (Figure 2C, bars) was very close to the estimated noise ceiling (Figure 2C, sideways triangle, derived from corrected split-half correlations). Importantly, all models appeared to perform similarly well at predicting responses to unseen images. These results replicate the earlier observation that a wide range of encoding models are approximately equal in their ability to predict responses in the brain.

3.2 All ANN-based encoding models are susceptible to small scale adversarial attacks

How susceptible are encoding models to adversarial attacks? To address this, we engineered an imperceptible noise pattern specifically designed to alter the predicted response for a given brain region, along with a randomized noise pattern of the same magnitude and statistical properties as a control. We discovered that even the slightest targeted noise, unseen by the human eye, could completely derail the encoding model's predicted response. This is shown for an example encoding model (VGG16) for the FFA and PPA in Figures 1C and 1D. Initially, the model's prediction for the unaltered image from the preferred category (faces for FFA, scenes for PPA) was high. This agrees with our expectation about images from the preferred category. However, adding a small amount of targeted noise was enough to push the predicted response well outside the preferred category range to the extreme end of the observed response spectrum. As a negative control, we used a shuffled version of the same targeted noise. Importantly, this shuffled noise pattern, despite having the same summary statistical properties of the noise, did not alter the predicted response to the same extent (delta = 0.01).

We quantified the adversarial sensitivity for each model by measuring the change in predicted response to the adversarially perturbed image. For a given subject and brain region, we obtain responses to both clean and perturbed images and calculate the average difference (s_i). We average

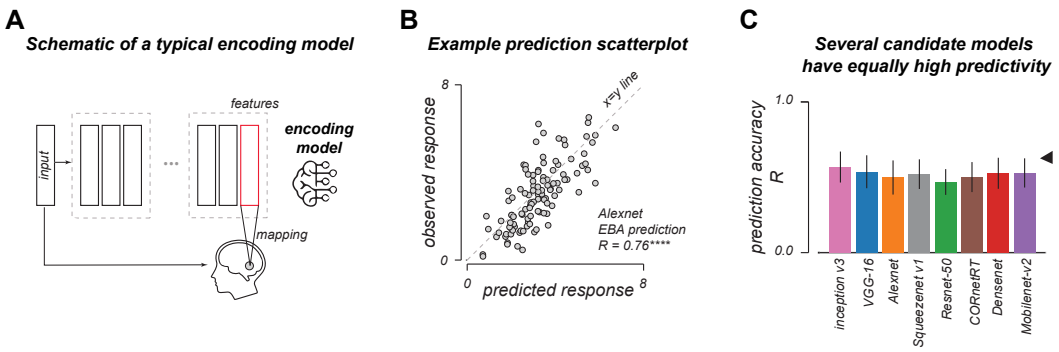

Figure 2: **Several encoding models have equally high predicitivity on fMRI data A.** Schematic outlining the construction of a typical encoding model. Features from an intermediate model layer (shown in red) are used to build a linear mapping function (indicated as "mapping") to predict responses in specific brain regions. **B.** Example scatterplot showing predicted (x-axis) versus observed responses for AlexNet in the EBA. The dotted line represents the x = y line, and each dot corresponds to a stimulus that was not used in model training (cross-validated). **C.** Bar plot showing various candidate encoding model architectures (x-axis) and their ability to predict responses to unseen images (y-axis). The black sideways triangle indicates the ceiling performance (median Spearman-Brown corrected split-half correlation across subjects and brain regions). Bars represent the mean response, with error bars showing the SEM across models and brain regions.

over i brain regions and subjects, yielding our estimate for the model change in response. Figure 3B shows these results for all encoding models. As the strength of the perturbation (ϵ) increased (x-axis), the adversarial sensitivity also increased (as expected). Note that in this context, higher sensitivity indicates lower adversarial robustness for the model. These findings demonstrate that all tested ANN models were vulnerable to targeted adversarial attacks. In fact, for most models, even a small perturbation with $\epsilon = 3/255$ was enough to significantly alter the predicted response.

3.3 Adversarial sensitivity better discriminates between ANN-models than predictivity

Next, we evaluated whether adversarial sensitivity could serve as a more effective tool for distinguishing between candidate encoding models of the brain. We present these analyses for $\epsilon = 3/255$, although all subsequent inferences hold across other values as well. The results for adversarial sensitivity across all encoding models at $\epsilon = 3/255$ are displayed in Figure 3C. To facilitate comparison, the models are arranged in the same order as shown in Figure 2C.

To assess the effectiveness of adversarial sensitivity compared to predictivity, we employed two different measures. First, we measured the sparseness [63] of the adversarial sensitivity and predictivity metrics across models. Sparseness was chosen since it is a scale invariant measure and can be used to directly compare between predictivity and adversarial sensitivity (see Methods for details). Figure 3D (top) shows that model sparseness was significantly higher for adversarial sensitivity than for predictivity, indicating better discriminability across models. A problem with sparseness however is that it is highly sensitive to outliers. To allay this concern, we adopted a second, more intuitive variance measure (normalized to match the scale between sensitivity and predictivity). As shown in Figure 3D (bottom), the normalized variance was also higher for adversarial sensitivity compared to predictivity. Together, these measures present a consistent picture: adversarial sensitivity distinguishes between encoding models more effectively than predictivity alone.

3.4 Increasing model robustness via adversarial training does not improve model predictivity

So far, we have demonstrated two key findings: 1) commonly used encoding models are sensitive to imperceptible adversarial noise, and 2) adversarial sensitivity can serve as a tool to distinguish between predictive models. How can we build better, more robust encoding models? The natural thing to try is to simply replace the current model architecture with a more robust one. In this section, we explored what happens when we use robustified models. To test this question, we fixed the model

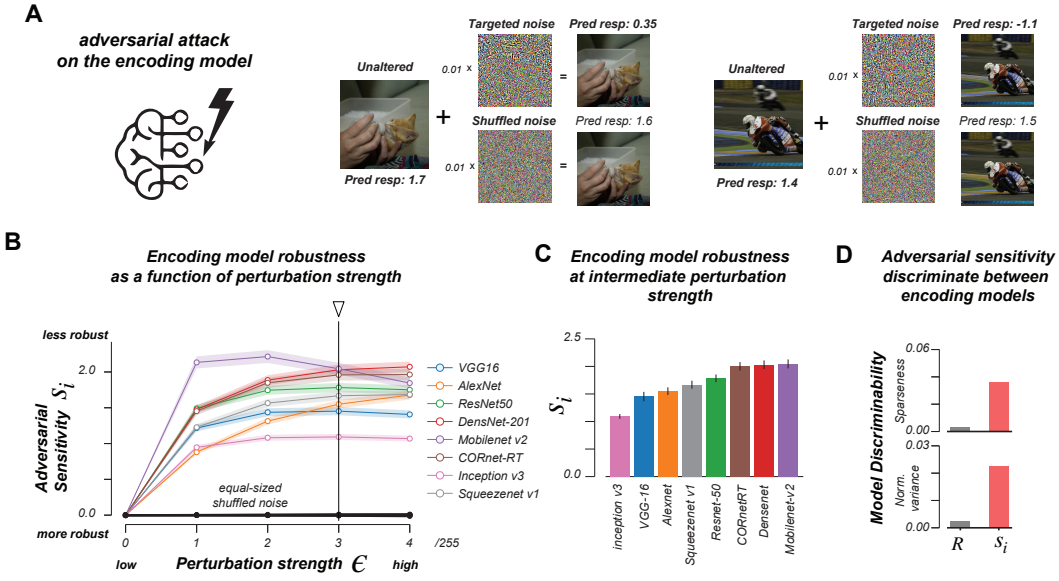

Figure 3: **Adversarial sensitivity effectively discriminates between encoding models. A.** Example of adversarial attacks applied directly to the encoding models. Each attack shows the unaltered image (left), the targeted noise and shuffled noise (middle panels top and bottom respectively), and model predictions for these images **B.** The effect of perturbation strength (x-axis) on the model's adversarial sensitivity (y-axis). Each colored line represents a different candidate encoding model architecture. The dots show the mean sensitivity, and the shaded areas represent the standard error across subjects and brain regions. The black line indicates the negative control using randomized noise. The triangle above marks the perturbation strength used for the subsequent analyses. **C.** Bar plots showing the adversarial sensitivity (y-axis) for all encoding models at a perturbation strength of $3/255$. The models are arranged in the same order as in Figure 2C for direct comparison. **D.** Barplots showing the discriminability between models using adversarial sensitivity and predictivity. Top: Bar plots illustrating model discriminability using predictivity (R) and adversarial sensitivity (s_i). Top: Discriminability based on the sparseness measure (y-axis). Bottom: Discriminability based on a normalized variance measure (y-axis).

architecture (ResNet50) and parametrically varied the strength of adversarial training using publicly available robustified models [60]. This strategy is illustrated schematically in Figure 4A.

As expected, we found that robust models were indeed less vulnerable to added adversarial noise. Figure 4B shows how adversarial sensitivity decreases as the strength of adversarial training increases. Are robustified models effective at predicting fMRI responses? Here, we observed a trade-off: as the models became more robust, their ability to predict fMRI responses declined. This reduction was quite significant and is shown across all models and regions. These results suggest that while adversarial training does improves robustness, it may do so at the cost of reduced predictivity for brain data. This trade-off underscores a deeper issue – achieving both high predictivity and robustness requires more than simple adversarial training.

3.5 Sparse mappings tend to improve adversarial robustness of encoding models without sacrificing model predictivity

A less well-understood aspect of encoding models is the effect of the specific choice of mapping between model features and neural responses. We wondered if certain mapping functions could improve an encoding model's sensitivity to targeted noise. There are many potential linear and non-linear mapping functions to explore. To constrain our choices, we first evaluated five different mapping methods: ordinary least squares (no regularization), Lasso (L_1) regression (sparse), Ridge (L_2) regression, a two-layer multi-layer perceptron (MLP), and a convolutional neural network. We chose two candidate encoding models (VGG16 and ResNet50) for this initial exploration of mapping methods. An issue with these is that many of these methods involve choosing appropriate

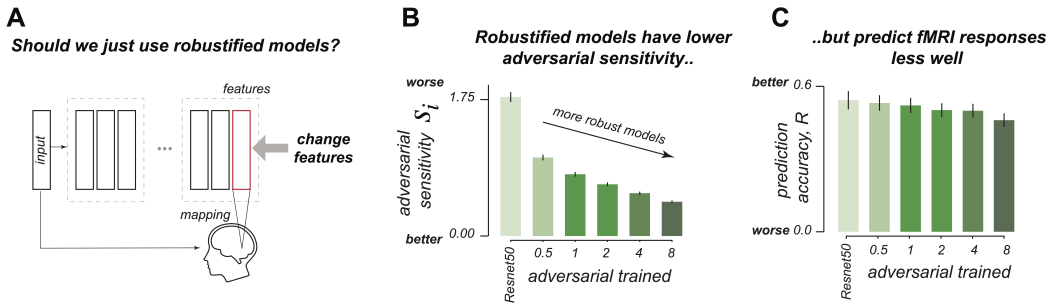

Figure 4: **Robust training reduces the predictive accuracy of fMRI encoding models. A.** Schematic illustration of the analysis. In this step, we replace the original features (shown in red) with features that have been robustified through adversarial training. **B.** Bar plots showing that increasing the level of adversarial training (x-axis) improves the adversarial sensitivity of the encoding models (y-axis). **C.** Bar plots showing that increasing the level of adversarial training (x-axis) reduces the predictive accuracy of encoding models on fMRI data (y-axis).

hyperparameters. Hyperparameters were selected based on prediction accuracy (see Methods), but we focus our attention on an independent metric: adversarial sensitivity. The results are presented in Table 1. Across both models, we found evidence of a significant boost in adversarial sensitivity when using a sparse mapping.

Adversarial sensitivity for different model-to-brain mapping functions

Model	OLS	**L1 (Lasso)**	L2 (Ridge)	2-layer MLP	CNN
VGG16	1.453	**.891**	1.453	1.358	1.734
ResNet50	1.782	**.821**	1.782	1.286	1.051

Table 1: Effect of readout functions on adversarial sensitivity. L_1 regularization on the readout performed best. The weight of the regularization term, α, was chosen as the value which maximized predictive accuracy from a set of candidate values; see Methods. Note that the "Lasso" column summarizes the results for two of the models in Figure 5B.

Would this observation generalize to other models? To explore this, we compared the sensitivity of all eight models using L_1 (sparse) and OLS (no regularization) mapping-based encoding models across all architectures. Note that the hyperparameters here were determined based on prediction accuracy from one subject, and the results were independently evaluated on adversarial sensitivity from all subjects (see Methods). This preliminary analysis revealed an interesting trend: sparse mappings produced significantly more robust models in 5 out of 8 model architectures. While this suggests that sparse mappings may enhance adversarial robustness, it is important to emphasize that these results are still preliminary and additional testing is needed to confirm whether this pattern holds across a larger sample, different model types, and independent analysis methods. Nonetheless, these early findings hint at the potential of sparse mappings to provide a meaningful boost in robustness.

4 Discussion

In this study, we investigated how susceptible commonly used ANN-based vision encoding models were to small-scale adversarial perturbations. We found that all high-performing models were vulnerable to imperceptible, small-scale adversarial noise (Figure 3). We also demonstrated that adversarial sensitivity, more effectively than prediction accuracy, could be used to differentiate between models (Figure 3). However, increasing model robustness through adversarial training came at the expense of reducing their ability to predict fMRI responses (Figure 4). Finally, we found early evidence that a simple sparse mapping approach on the mapping function could significantly improve adversarial robustness (Figure 5). These findings reveal key limitations of current encoding models and suggest new strategies for enhancing their performance.

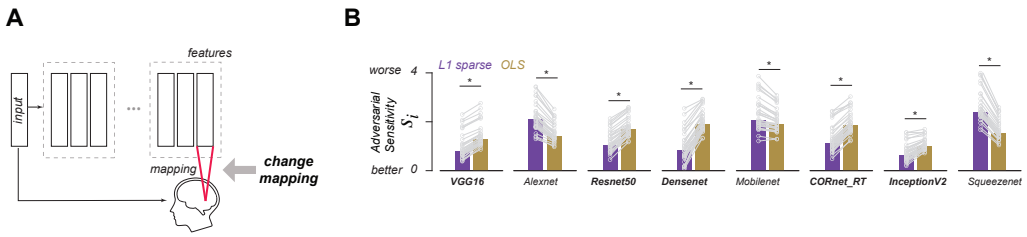

Figure 5: **Sparse model-to-brain mappings tend to lower adversarial sensitivity of encoding models . A.** Schematic illustration of the analysis. Here, we evaluate the model-to-brain mappings (highlighted in red) **B.** Bar plots showing adversarial sensitivity (y-axis) for all models tested. The dots and connected lines represent an encoding model for a specific subject and brain region. * indicates statistical significance (paired t-test, $P < 0.001$) between OLS and sparse mappings. Models in bold indicate improved adversarial sensitivity for sparse (L_1) mappings compared to OLS-based mappings.

Our adversarial attacks had two key features. First, the perturbations were deliberately kept small to focus on imperceptible changes. Our pilot analyses, based on an 8-degree viewing angle, suggest the detection threshold for adversarial images to be around $\epsilon = 8/255$. While a formal estimate on a larger sample is underway, we assumed that small perturbations, as those used in this study, would not alter brain voxel responses (though see [43]). This allowed us to test the model's vulnerability in a regime where the visual system should remain stable, highlighting its susceptibility to subtle adversarial noise. However, these assumptions require formal testing in future work. The second key feature is that our method targeted the encoding models directly (instead of the model features). This approach enabled us to assess vulnerabilities in the model's representational mappings to brain activity, not just the image embeddings. While previous studies have examined the relationship between model robustness and neural predictions in monkeys [46], or the link between spatial features and neural representations [64, 65], our work extends these findings by exploring how adversarial perturbations *directly* affect model representations most predictive of human fMRI brain responses.

One interpretation of our results is that current high-performing, predictive encoding models are fundamentally flawed given how drastically they fail when exposed to targeted adversarial noise. While this is true, our aim is not merely to highlight these vulnerabilities. It is not entirely unexpected that these models are susceptible to adversarial perturbations, given what we know about neural networks in general. However, we propose leveraging adversarial sensitivity as a tool to guide the development of more accurate and resilient models. In fact, we find that adversarial sensitivity provides an additional layer of insight into model performance, helping to distinguish between highly predictive encoding models.

By analyzing how different models respond to adversarial perturbations, we start to uncover their limitations and use new insights into the development of more robust brain models. To this end, we tested two strategies. While adversarial training is widely used in the AI community to enhance model resilience, we found that it came at a significant cost to model predictivity (see also [46]). As models became more robust, their ability to accurately predict brain responses declined substantially. This trade-off highlights a compromise that must be carefully considered when developing models for neuroscience applications. In contrast, we found that a relatively simple sparse mapping between model features and brain representations was enough to significantly reduce the adversarial sensitivity of most encoding models, usually outperforming more complex non-linear mapping methods. We hope to explore these differences further in future work.

Taken together, our results expose the critical vulnerabilities of ANN-based predictive encoding models to adversarial perturbations, highlight adversarial sensitivity as a powerful tool for differentiating between models, and suggest a promising path for enhancing model robustness. As we continue our search for brain-like models, striking the right balance between robustness and predictivity will be crucial. Our work provides a foundation for tracking this balance, offers new model evaluations, and offers prescriptions to guide the development of more accurate and resilient models that can be applied to study human cognition even beyond vision.

References

[1] Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020.

[2] Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064, 2021.

[3] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

[4] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1(1):417–446, 2015.

[5] Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, 2023.

[6] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[7] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[8] Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[9] N. Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J. DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, 12, Sep 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25409-6.

[10] Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, 2023. doi: 10.1101/2022.03.28.485868.

[11] Charles Y. Zheng, Francisco Pereira, Chris I. Baker, and Martin N. Hebart. Revealing interpretable object representations from human behavior. In *International Conference on Learning Representations*, 2019.

[12] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

[13] Katharina Dobs, Joanne Yuan, Julio Martinez, and Nancy Kanwisher. Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences*, 120(32):e2220642120, 2023.

[14] Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H McDermott. Model metamers illuminate divergences between biological and artificial neural networks. *Nature Neuroscience*, 2023.

[15] Anne Harrington and Arturo Deza. Finding biological plausibility for adversarially robust features via metameric tasks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=yeP_zx9vqNm`.

[16] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018.

[17] Drew Linsley, Ivan F. Rodriguez Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret S. Livingstone, and Thomas Serre. Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[18] Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46:e385, 2023.

[19] Jonas Kubilius, Martin Schrimpf, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, and R. Garnett, editors, *Neural Information Processing Systems (NeurIPS)*, 2019.

[20] Tim C. Kietzmann, Courtney J. Spoerer, Lynn K. A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019. doi: 10.1073/pnas.1905544116.

[21] Ruben S van Bergen and Nikolaus Kriegeskorte. Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193, 2020.

[22] Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel L Yamins. Task-driven convolutional recurrent models of the visual system. *Advances in neural information processing systems*, 31, 2018.

[23] Thomas Fel, Ivan F. Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[24] Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Geary, Michael Ferguson, David D Cox, and James J DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. *bioRxiv*, pages 2022–07, 2022.

[25] Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, pages 2022–03, 2022.

[26] Yena Han, Tomaso A. Poggio, and Brian Cheung. System identification of neural systems: If we got it right, would we know? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 12430–12444. PMLR, 2023.

[27] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 2019.

[28] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *CoRR*, abs/2310.13018, 2023.

[29] Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Baldwin Geary, Michael Ferguson, David Daniel Cox, and James J. DiCarlo. Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[30] Meenakshi Khosla and Alex H. Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In Marco Fumero, Emanuele Rodolà, Clémentine Dominé, Francesco Locatello, Karolina Dziugaite, and Mathilde Caron, editors, *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models, 15 December 2023, Ernest N. Morial Convention Center, New Orleans, USA*, volume 243 of *Proceedings of Machine Learning Research*, pages 326–341. PMLR, 2023.

[31] Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, 2022. doi: 10.1101/2022.03.16.484578.

[32] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.*, 38(33):7255–7269, August 2018.

[33] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

[34] Subha Nawer Pushpita, Elizabeth Mieczkowski, Bradley Duchaine, and N. Apurva Ratan Murty. Intensive fmri scanning and computational models can provide insight into the neural basis of developmental prosopagnosia. *Journal of Vision*, 23(9):5838, 2023. doi: https://doi.org/10.1167/jov.23.9.5838.

[35] Carlos R. Ponce, Will Xiao, Peter F. Schade, Till S. Hartmann, Gabriel Kreiman, and Margaret S. Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009.e10, 2019. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2019.04.005.

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[37] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[38] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[39] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

[40] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.*, 23(5):828–841, 2019.

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[42] Yatie Xiao, Chi-Man Pun, and Kongyang Chen. Towards evaluating the robustness of deep neural semantic segmentation networks with feature-guided method. *Knowl. Based Syst.*, 281:111063, 2023.

[43] Chong Guo, Michael J. Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James J. DiCarlo. Adversarially trained neural representations may already be as robust as corresponding biological neural representations. In *39th International Conference on Machine Learning, ICML 2022, Baltimore, MD, USA, 2015, Conference Track Proceedings*, 2022.

[44] Guy Gaziv, Michael J. Lee, and James J. DiCarlo. Strong and precise modulation of human percepts via robustified anns. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[45] Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks. *Nature Communications*, 15(1):1989, 2024.

[46] Yifei Ren and Pouya Bashivan. How well do models of visual cortex generalize to out of distribution samples? *PLOS Computational Biology*, 20(5):e1011145, 2024.

[47] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Logan T. Dowdle, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *Nature Neuroscience*, 2022. doi: 10.1038/s41593-021-00962-x.

[48] Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, 17(11):4302–4311, June 1997.

[49] P E Downing, Yuhong Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science (New York, N.Y.)*, 293(5539):2470–3, September 2001. ISSN 0036-8075.

[50] Russell A. Epstein and Nancy G. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392:598–601, 1998.

[51] Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2): 291–307, 02 2000. ISSN 0006-8950.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016.

[55] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.

[56] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[57] Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, 2018.

[58] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[59] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4510–4520. Computer Vision Foundation / IEEE Computer Society, 2018.

[60] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL `https://github.com/MadryLab/robustness`.

[61] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[62] Ben D B Willmore, James A Mazer, and Jack L Gallant. Sparse coding in striate and extrastriate visual cortex. *J. Neurophysiol.*, 105(6):2907–2919, June 2011.

[63] William E. Vinje and Jack L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000. doi: 10.1126/science.287.5456.1273.

[64] Zhe Li, Josue Ortega Caro, Evgenia Rusak, Wieland Brendel, Matthias Bethge, Fabio Anselmi, Ankit B Patel, Andreas S Tolias, and Xaq Pitkow. Robust deep learning object recognition models rely on low frequency information in natural images. *PLOS Computational Biology*, 19(3):e1010932, 2023.

[65] Ajay Subramanian, Elena Sizikova, Najib J. Majaj, and Denis G. Pelli. Spatial-frequency channels, shape bias, and adversarial robustness. In Alice Oh andt Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.