Alignment and Adversarial Robustness: Are More Human-Like Models More Secure?

Blaine Hoak[1,*,†], Kunyang Li[1,†] and Patrick McDaniel[1]

[1]University of Wisconsin-Madison, USA

Abstract

A small but growing body of work has shown that machine learning models which better align with human vision have also exhibited higher robustness to adversarial examples, raising the question: *can human-like perception make models more secure?* If true generally, such mechanisms would offer new avenues toward robustness. In this work, we conduct a large-scale empirical analysis to systematically investigate the relationship between representational alignment and adversarial robustness. We evaluate 144 models spanning diverse architectures and training paradigms, measuring their neural and behavioral alignment and engineering task performance across 105 benchmarks as well as their adversarial robustness via AutoAttack. Our findings reveal that while average alignment and robustness exhibit a weak overall correlation, *specific* alignment benchmarks serve as strong predictors of adversarial robustness, particularly those that measure selectivity toward texture or shape. These results suggest that different forms of alignment play distinct roles in model robustness, motivating further investigation into how alignment-driven approaches can be leveraged to build more secure and perceptually-grounded vision models.

1. Introduction

Representational alignment—how closely a model resembles biological vision—has been studied extensively with the goal of measuring, bridging, or increasing alignment in machine learning models [1]. Recent observations [2, 3] suggest that alignment may have implications beyond neuroscience: models that are more aligned with human perception have also exhibited increased robustness to adversarial examples—inputs with near-imperceptible perturbations that induce model misclassification— hinting at a deeper connection between alignment and security.

However, the relationship between representational alignment and adversarial robustness remains poorly understood. While the former seeks to align models with human cognition, adversarial examples in security highlight a fundamental misalignment: imperceptible perturbations can drastically degrade model accuracy while leaving human perception unaffected. Prior robustness techniques, such as adversarial training [4], are computationally expensive and potentially vulnerable to new attack strategies. Although recent work has suggested links between perceptual alignment and robustness, especially in models incorporating biological priors [2, 3], a broad and systematic evaluation of this relationship across diverse models and alignment benchmarks is still lacking. A fundamental question remains: do these objectives complement each other, leading to better-aligned and more robust models, or do they introduce conflicting trade-offs?

In this work, we investigate the relationship between human alignment and robustness to adversarial examples in vision models through a diverse, large-scale empirical analysis. In our analysis, we study 144 models across different architectures and training schemes, and measure their alignment across 105 different benchmarks on neural, behavioral, and engineering tasks via the BrainScore library [5]. We then evaluate the robustness of these models using AutoAttack [6], a state-of-the-art ensemble attack.

Our analysis reveals that while robustness is only weakly correlated with *average* vision alignment, *specific* alignment benchmarks are strongly predictive of adversarial robustness. We also find that models

SPAIML'25: International Workshop on Security and Privacy-Preserving AI/ML, October 26, 2025, Bologna, Italy

[*]Corresponding author.

[†]These authors contributed equally.

bhoak@cs.wisc.edu (B. Hoak)

0000-0003-2960-0686 (B. Hoak); 0009-0007-4264-855X (K. Li); 0000-0003-2091-7484 (P. McDaniel)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4154/paper1.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-01-28

with similar alignment profiles exhibit similar robustness. Together, these findings suggest that different forms of alignment contribute differently to robustness, highlighting the value of alignment-driven approaches for improving security in vision systems.

- *Average* alignment, particularly for behavioral benchmarks, poorly predicts robustness (explaining 6% of variance), demonstrating that not all methods of alignment increase robustness.
- Individual benchmarks are much better indicators of robust accuracy, but the *specific* benchmark matters; we found instances of benchmarks that contributed positively *and* ones that contributed negatively to robustness in every alignment category.
- The top benchmarks that are most indicative of robust accuracy show clear trends that inform the factors leading to higher robustness: (1) robust models process *texture information* specifically more similarly to humans than non-robust models to and (2) models' ability to recognize objects without global structures intact actually hurts their robustness.
- t-SNE visualizations reveal that robustness is structured in alignment space, with similar models forming robustness-consistent clusters and demonstrating that alignment on certain benchmarks is a good indicator of robustness.

In summary, we uncover specific features of alignment that closely tie with model robustness and show that increasing alignment on these benchmarks offers a new avenue for building more robust and human-aligned models.

2. Background

2.1. Representational Alignment

Representational alignment studies the extent to which internal representations of machine learning models correspond to human cognitive processes. Early studies found that deep neural networks (DNNs) trained on large-scale image datasets develop hierarchical feature representations similar to those observed in the primate ventral stream, particularly in high-level visual areas like the inferior temporal (IT) cortex [7, 5]. This led to efforts to quantify the alignment between artificial and biological vision, using techniques such as Representational Similarity Analysis (RSA) [8] and Centered Kernel Alignment (CKA) [9]. To improve alignment, researchers have proposed strategies that incorporate cognitive constraints or psychological priors into model architectures [2]. Supervised fine-tuning with human-annotated datasets [10] ensures that learned representations align more closely with human-understandable features. Furthermore, novel techniques [11, 3, 12] have been developed to encourage similarity between model activations and human neural responses as recorded through fMRI and EEG experiments. In this study, we use a comprehensive set of neural, behavioral, and engineering alignment metrics to quantify representational alignment.

2.2. Adversarial Examples

Although machine learning models have shown strong capabilities to achieve high accuracy in various tasks [13, 10, 14, 15], they remain vulnerable to adversarial examples [6, 4, 16, 17, 18]. Adversarial examples are specially crafted inputs that contain perturbations which are imperceptible to humans, yet can significantly decrease model accuracy. In computer vision systems, there have been many studies on developing attack algorithms, such as FGSM [17], PGD [4], and AutoAttack [6]. These methods aim to maximize model's loss subject to constraints of perturbations defined by certain ℓ_p-norms:

$$x_{adv} = \arg\max_{\|\delta\|_p \leq \epsilon} L(x + \delta, y)$$

where x and y represent the original image and its predicted label, respectively, δ is the perturbation to solve for, and L is the model's loss function. The perturbation constraint ϵ is measured through an

ℓ_p-norm—most commonly ℓ_∞. While many works have historically evaluated the robustness of their model through PGD [4], it has been shown that "robust" models can often suffer from gradient masking, causing gradient-based attacks like PGD to fail [19], and leading to a sense of overestimated robustness. To overcome this, multiple attacks, including both white- and black-box attacks should be used [20]. Thus, the AutoAttack ensemble [6] has become the de-facto standard for evaluating robustness.

3. Methods

This section details how we measure the alignment and robustness of machine learning models with the goal of assessing if more human-like vision models are also more resilient to security vulnerabilities.

3.1. Alignment.

To measure alignment and download candidate models, we leverage the BrainScore [5] library. BrainScore provides a standardized framework for evaluating model similarity to biological vision through a set of neural, behavioral, and engineering benchmarks, supplying 106 benchmarks in total. These benchmarks quantify how closely a model's internal representations and outputs correspond to neurophysiological recordings, human psychophysical behavior, and performance on engineered vision tasks. Neural alignment is measured by comparing activations from DNNs to neural recordings from primate visual cortex regions (e.g., V1, V2, V4, and IT), using similarity metrics like Representational Similarity Analysis (RSA) [8]. Behavioral alignment assesses whether models replicate human psycho-physical responses in object recognition and perturbation tests, while engineering alignment evaluates model robustness to controlled distortions, such as contrast reductions, or performance on out-of-distribution data. We use 105 benchmarks by discarding one of them, which has NaN value for all the selected models.

In total, the BrainScore library has documented benchmark scores for 434 models. Out of those, there are 240 models available in their registry (the remaining models were either submitted privately or have been deprecated). From the 240 models in the registry, we further removed an additional 89 models that were either incompatible with ImageNet (e.g., do not output 1000 classes or expect video streams) or could not be run on RobustBench due to gradient masking. After this, we had to discard an additional 7 models, which represented all the VOne class models [2] because they were not able to run on AutoAttack due to gradient alteration or masking, suggesting that previous results finding that VOne models are more robust to adversarial examples could have been due to overestimated robustness and highlighting the importance of evaluating robustness under comprehensive attack strategies. After this filtering process, we were left with 144 models for our evaluation.

Model diversity is critical for evaluating the generalization of alignment-robustness relationships. Thus, the 144 models selected for our evaluation are representative by covering a broad spectrum of architectures and training recipes. The majority of architectures are convolutional neural networks (CNNs) [21] such as various ResNet [15] and VGG variants [22]. We also include more recent architectural designs such as Vision Transformers (ViTs) [10], which uses self-attention mechanisms to capture global dependencies in images, its variants [23], and hybrid models [13, 24] (i.e., a combination of CNNs and ViTs). This architectural diversity is complemented by models trained with different strategies such as standard supervised learning and self-supervised learning (e.g., contrastive learning [25]). Notably, the evaluated models also include those specifically designed or trained with properties relevant to the study's central hypothesis: some models have undergone adversarial training [4], which improves robustness against adversarial examples, while others use mechanisms to emphasize shape bias [26, 27, 28], shifting model reliance from texture to shape information. This design has been shown to be more aligned with human perception. The comprehensive model set allows for a thorough exploration of how architectural design and training methods impact the relationship between representational alignment and adversarial robustness.

3.2. Robustness.

To evaluate the robustness of our models, we use AutoAttack [6, 29], which serves as the standard for evaluating the robustness of neural networks due to its strong attack performance and fully automated parameter-free design. AutoAttack contains 4 attacks: APGD-CE, APGD-DLR, FAB, and Square Attack. By evaluating on AutoAttack, we are not only evaluating on the most performant attacks, but also integrating in both white-box attacks and black-box attacks which has been recommended in previous works to combat reporting overestimated robustness due to gradient masking or obfuscation [16].

To better understand how the relationship between adversarial robustness and alignment changes as attacks change, we evaluate the ℓ_∞ robustness of our models at three different epsilon levels: $\epsilon = \{\frac{0.25}{255}, \frac{0.5}{255}, \frac{1}{255}\}$ to represent adversaries at different capability levels and small, medium, and large image distortion levels. While these values are typically lower than what would be benchmarked on platforms such as RobustBench [29], we choose these values with the goal of having a wide distribution of robust accuracies to identify separability between models, rather than the goal of bringing the model down to 0% accuracy as is typically done.

4. Results

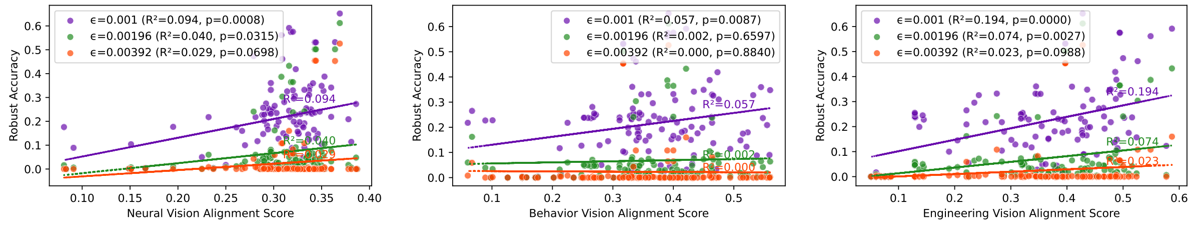
In this work, we hypothesize that there is a relationship between model robustness and alignment, due to the inherent similarity of the goals in each of these spaces. Here, we focus on answering the question *are more aligned machine learning models more robust to adversarial examples?*

To address this, we conduct a series of experiments. We use the BrainScore library v2.2.4 to measure alignment [5] and load models. Details on models evaluated can be found in section 3. Once these models have been loaded and their alignment has been measured across the 105 alignment benchmarks, we evaluate their robustness using AutoAttack [6] from the TorchAttacks [30] library v3.5.1. The ImageNet [31] validation set is used for clean inputs to the model and serves as the starting point to generate adversarial examples. All experiments are run across 12 A100 GPUs with 40 GB of VRAM and CUDA version 11.1 or greater. Our code is available for download at https://github.com/kyangl/alignment-security.

4.1. Average Alignment

We first investigate how well different classes of alignment—neural alignment, behavioral alignment, and engineering task performance—predict model robustness. For each class, we take the average score across all the benchmarks, giving us a single score for each model in the class. While many works have typically studied average vision alignment overall (i.e., the average of all the benchmarks across all classes), it has been shown that this can overemphasize behavioral alignment at the cost of neural alignment [32].

We then assessed each model's robust accuracy against AutoAttack at three different values of epsilon $\epsilon = \{0.001, 0.00196, 0.00392\}$, which corresponds to $\{\frac{0.25}{255}, \frac{0.5}{255}, \frac{1}{255}\}$, respectively. In Figure 1, we analyze the average score for neural alignment, behavioral alignment, and engineering task performance



(a) Neural Alignment (b) Behavioral Alignment (c) Engineering Task Performance

Figure 1: Average vision alignment score vs robust accuracy on neural, behavioral, and engineering benchmarks.

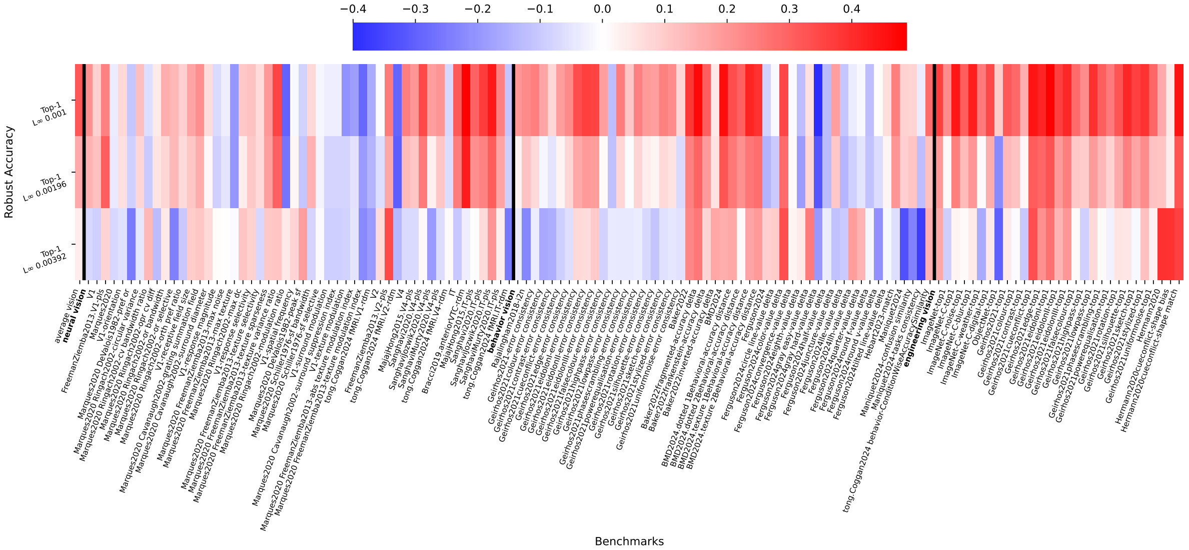


Figure 2: Heatmap of each of the BrainScore benchmarks, ordered and separated (black bars) by area of alignment (neural, behavioral, engineering) vs the robust accuracy. Each cell represents the correlation between a benchmark and the robust accuracy across models.

on the x-axis and the robust accuracy on the y-axis. Each dot represents a model, and the 3 colors correspond to the model's robust accuracy at the three different epsilon values. We perform linear fits for each epsilon value and report the statistical significance.

We find statistically significant ($p \leq 0.05$) correlations for: the two lowest ϵ values for neural alignment (explaining up to 9% of variance), $\epsilon = 0.001$ for behavioral alignment (6% of variance), and at the two lowest ϵ values for engineering task performance (up to 19% of variance). Overall, the relatively low R^2 values, coupled with the difficulty of getting statistically significant correlations at higher epsilon values, suggest that average alignment scores are, at best, a weak indicator of robust accuracy. This counter-intuitive finding leads us to study further on individual benchmarks.

4.2. Individual Benchmarks

Motivated by the previous experiment where we find that average alignment is weakly correlated with robust accuracy, we hypothesize that averaging scores across different benchmarks may obscure that some individual benchmarks are stronger predictors of robust accuracy than others. Here, certain alignment benchmarks could be more indicative of robust accuracy than others.

To examine this hypothesis, we collect all models' scores on individual benchmarks for the three classes (neural alignment, behavioral alignment, and engineering task performance) and compute the correlation between each of these scores and robust accuracy at our three different ϵ values. Figure 2 shows a heatmap of the 105 different benchmarks on the x-axis and robust accuracy at three different ϵ values on the y-axis. In each cell, we report the Spearman correlation coefficient between the selected benchmark score and robust accuracy across models.

From this figure, we find multiple interesting trends. First, we see a wide range of correlations between different benchmarks, confirming our hypothesis that not every current alignment metric is a good indicator of robust accuracy. Additionally, we sometimes see significant changes to the correlation of robust accuracy and a benchmark as the ϵ value increases (and thus becomes a stronger attack). These changes appear to cluster by class of alignment. Roughly speaking, the neural alignment benchmarks (shown from the first to the second black bar) seem highly dependent on the task, with benchmarks in this category having correlations at both ends of the spectrum. The behavioral benchmarks (shown from the second to third black bars) tend to be, surprisingly, often negatively correlated with robust accuracy at mid and high ϵ values, and the correlation mostly decreases as ϵ increases. Finally, engineering task performance (shown from the third black bar to the end of the figure) tends to have more stable (and more positive) correlations as ϵ increases.

Furthermore, we identified specific benchmarks that are strongly correlated with robust accuracy. Interestingly, many of these benchmarks that exhibit strong positive correlations with robust accuracy even at high epsilon values measure, to some degree, a model's selectivity towards texture information, meaning that *models which are more robust to adversarial examples tend to process texture information more similarly to humans.* Below, we highlight and discuss the benchmarks we found to be most strongly (positive or negative) correlated with robust accuracy.

In the neural category, we found the correlation between robust accuracy and alignment to be highly dependent on the area of visual processing. At small to medium values of ϵ, correlations are strongest in benchmarks that measure alignment to V4 and IT areas of the visual system. At later values of ϵ, which represents a stronger attack, `FreemanZiemba2013.V2-pls` which measures responses to naturalistic texture stimuli in V2 [33], was the metric with the strongest positive correlation. This suggests that processing texture more similarly to biological vision systems in later visual areas is even more important than early visual areas to yield robust models. Interestingly, we found that metrics utilizing fMRI data were typically the most negatively correlated.

In the behavioral category, we see many strong positive correlations at low values of ϵ. The sets of benchmarks that stand out the most are the BMD and Baker [34] groups, which measure the ability to recognize objects by silhouettes when the shape has been distorted. In Baker and Elder [34], it was found that humans are less able to recognize objects when shape information is distorted than neural networks are, demonstrating that machine learning models have a certain degree of insensitivity to changes in shape information and thus don't rely as heavily on it as humans do for object recognition. This finding further supports the notion that neural networks may rely more on texture information rather than shape information [26]. The strong positive relationship between these benchmarks and performance on adversarial examples suggests that *models' ability to recognize objects without spatial relationships actually hurts their overall robustness.*

Finally, on engineering tasks, we see strong positive correlations across nearly all benchmarks with low ϵ. However, these correlations decrease as ϵ increases for most benchmarks. The benchmarks that remain most strongly positively correlated with high ϵ are both related to bias towards shape over texture information. First is the `Geirhos2021cueconflict-top1` benchmark [26], which measures the probability of a model classifying an object using shape information rather than texture via texture-shape conflicted images. The other is the set of benchmarks from [28]: `Hermann 2020 cue conflict shape _bias` and `Hermann2020cueconflictshape_match`, which similarly measures the probability of a model classifying an object using shape information and the percentage of the times the model classifies according to the shape class, rather than texture or other classes. In all, these results show that when models are able to classify inputs according to their shape information, as humans do, rather than texture information, they will be more robust to adversarial examples.

Across all categories of visual alignment, we find that models which exhibit a stronger reliance on shape information and robust processing of texture cues (i.e., more aligned to how humans process information) tend to be more resilient to adversarial examples. These results suggest that increasing alignment towards preferring and processing high-level visual features, particularly textures, in the way biological vision systems do, serves as a fruitful direction for creating more robust and more aligned models.

4.3. Robustness in Alignment Space

Building on our finding that certain alignment benchmarks (especially those measuring texture sensitivity) are strongly correlated with adversarial robustness, we next explore whether models with similar alignment profiles also exhibit similar levels of robust accuracy. The goal of this analysis is to see if model similarity on alignment benchmarks is predictive of model performance against adversarial examples, which would support the hypothesis that aligning models under certain metrics results in better robustness.

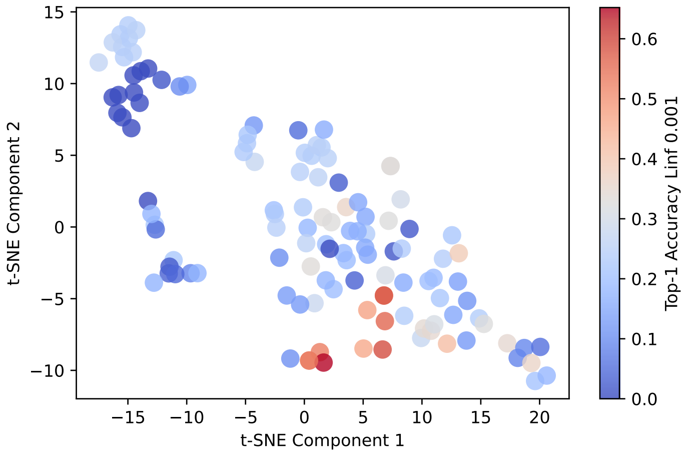


Figure 3: t-SNE plot of principal components from benchmark features for each model colored by the model's robust accuracy.

To compare the similarity of model performance across benchmarks, we utilize t-SNE to reduce the dimensions of the results and project them into a subspace that can be visualized. In Figure 3, each point represents one of our 144 models, colored by its robust accuracy under ℓ_∞ adversarial perturbations with $\epsilon = 0.001$. We observe that the models are distributed across the embedding space with clustering patterns that are unique to their robustness.

Analyzing the non-robust models (blue dots), we find that while some cluster in the upper left, many are scattered broadly throughout the space. This dispersion suggests that poor robustness is not associated with a single alignment profile. Rather, there are *many different ways in which models can fail to defend against adversarial examples.* That is, vulnerability appears to be distributed across a range of misaligned configurations.

In contrast, the robust models (red dots) form a tight, well-separated cluster. These models not only share a similar alignment profile, but their separation from the rest of the models *suggests that robust behavior is tied to a specific region of alignment space.* This strongly supports the view that certain alignment characteristics are predictive of robustness.

These findings offer a new perspective on adversarial robustness; rather than unique quirks of the model resulting in vulnerability to adversarial examples, there are specific, isolated properties that allow models to more robustly process inputs. In other words, *it's not that specific misalignment will lead to model vulnerability, but rather specific alignment will lead to robustness.* Furthermore, these properties can be measured in alignment metrics, and thus optimized for in order to build more robust models.

5. Related Work

There has been substantial progress in bridging the representational differences between humans and machine learning models over the last few years. Geirhos et al. [35] show that many of the high-performance models match or exceed human feedforward performance on OOD datasets. Several works suggest that more human-aligned model architectures may also be more robust. Dapello et al. [2] design the biologically inspired VOne block to simulate V1 processing, improving robustness to adversarial and common corruptions. Li et al. [3] regularize models based on neural recordings from mice to increase robustness and alignment.

Furthermore, Subramanian et al. [36] show that differences in spatial frequency processing explain both shape bias and adversarial vulnerability. Models with higher human alignment show improved performance on datasets like ImageNet-A [37]. Other works highlight that models tend to rely on texture over shape [26, 28], a bias that increases susceptibility to natural adversarial samples [38, 39].

Our work expands on these findings with a large-scale empirical study across diverse architectures and benchmarks. Unlike prior work that focuses on specific models or alignment methods, we systematically analyze how different forms of alignment relate to adversarial robustness, discovering meaningful connections between model perception and security.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, A. Bobu, B. Kim, B. C. Love, C. J. Cueva, E. Grant, I. Groen, J. Achterberg, J. B. Tenenbaum, K. M. Collins, K. L. Hermann, K. Oktar, K. Greff, M. N. Hebart, N. Cloos, N. Kriegeskorte, N. Jacoby, Q. Zhang, R. Marjieh, R. Geirhos, S. Chen, S. Kornblith, S. Rane, T. Konkle, T. P. O'Connell, T. Unterthiner, A. K. Lampinen, K.-R. Müller, M. Toneva, T. L. Griffiths, Getting aligned on representational alignment, 2024. URL: http://arxiv.org/abs/2310.13018. doi:10.48550/arXiv.2310.13018, arXiv:2310.13018 [q-bio].

[2] J. Dapello, T. Marques, M. Schrimpf, F. Geiger, D. Cox, J. J. DiCarlo, Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 13073–13087. URL: https://proceedings.neurips.cc/paper/2020/hash/98b17f068d5d9b7668e19fb8ae470841-Abstract.html.

[3] Z. Li, W. Brendel, E. Walker, E. Cobos, T. Muhammad, J. Reimer, M. Bethge, F. Sinz, Z. Pitkow, A. Tolias, Learning from brains how to regularize machines, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/hash/70117ee3c0b15a2950f1e82a215e812b-Abstract.html.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, 2019. URL: http://arxiv.org/abs/1706.06083.

[5] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, J. J. DiCarlo, Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, 2018. URL: http://biorxiv.org/lookup/doi/10.1101/407007. doi:10.1101/407007.

[6] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, in: Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020. URL: https://proceedings.mlr.press/v119/croce20b.html.

[7] D. L. Yamins, H. Hong, C. Cadieu, J. J. DiCarlo, Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream, in: Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: https://papers.nips.cc/paper_files/paper/2013/hash/9a1756fd0c741126d7bbd4b692ccbd91-Abstract.html.

[8] N. Kriegeskorte, M. Mur, P. A. Bandettini, Representational similarity analysis - connecting the branches of systems neuroscience, Frontiers in Systems Neuroscience 2 (2008). URL: https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full.

[9] S. Kornblith, M. Norouzi, H. Lee, G. Hinton, Similarity of Neural Network Representations Revisited, 2019. URL: http://arxiv.org/abs/1905.00414. doi:10.48550/arXiv.1905.00414.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. URL: http://arxiv.org/abs/2010.11929. doi:10.48550/arXiv.2010.11929.

[11] L. Muttenthaler, L. Linhardt, J. Dippel, R. A. Vandermeulen, K. Hermann, A. K. Lampinen, S. Kornblith, Improving neural network representations using human similarity judgments, 2023. URL: http://arxiv.org/abs/2306.04507.

[12] Y.-A. Cheng, I. F. Rodriguez, S. Chen, K. Kar, T. Watanabe, T. Serre, RTify: Aligning Deep Neural Networks with Human Behavioral Decisions, 2024. URL: http://arxiv.org/abs/2411.03630. doi:10.48550/arXiv.2411.03630.

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

[14] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60 (2017). URL: https://dl.acm.org/doi/10.1145/3065386.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: CVPR 2016, 2015. URL: http://arxiv.org/abs/1512.03385.

[16] N. Carlini, D. Wagner, Towards Evaluating the Robustness of Neural Networks, 2017. URL: http://arxiv.org/abs/1608.04644. doi:10.48550/arXiv.1608.04644.

[17] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, 2015. URL: http://arxiv.org/abs/1412.6572. doi:10.48550/arXiv.1412.6572.

[18] R. Sheatsley, B. Hoak, E. Pauley, P. McDaniel, The Space of Adversarial Strategies, in: 32nd USENIX Security Symposium, 2023. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/sheatsley.

[19] A. Athalye, N. Carlini, D. Wagner, Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, in: Proceedings of the 35th ICML, PMLR, 2018. URL: https://proceedings.mlr.press/v80/athalye18a.html.

[20] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, A. Kurakin, On Evaluating Adversarial Robustness, 2019. URL: http://arxiv.org/abs/1902.06705. doi:10.48550/arXiv.1902.06705.

[21] K. O'Shea, R. Nash, An Introduction to Convolutional Neural Networks, 2015. URL: http://arxiv.org/abs/1511.08458.

[22] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2014. URL: http://arxiv.org/abs/1312.6034. doi:10.48550/arXiv.1312.6034.

[23] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, A. Kolesnikov, Knowledge Distillation: A Good Teacher Is Patient and Consistent, 2022, pp. 10925–10934. URL: https://openaccess.thecvf.com/content/CVPR2022/html/Beyer_Knowledge_Distillation_A_Good_Teacher_Is_Patient_and_Consistent_CVPR_2022_paper.html.

[24] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021. URL: https://proceedings.mlr.press/v139/touvron21a.html.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th ICML, PMLR, 2021. URL: https://proceedings.mlr.press/v139/radford21a.html.

[26] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, in: ICLR, 2019. URL: http://arxiv.org/abs/1811.12231.

[27] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, F. A. Wichmann, Generalisation in humans and deep neural networks, in: NeurIPS 2018, 2020. URL: http://arxiv.org/abs/1808.08750.

[28] K. L. Hermann, T. Chen, S. Kornblith, The Origins and Prevalence of Texture Bias in Convolutional Neural Networks, in: NeurIPS 2020, 2020. URL: http://arxiv.org/abs/1911.09071.

[29] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein, RobustBench: a standardized adversarial robustness benchmark, in: NeurIPS, arXiv, 2021. URL: http://arxiv.org/abs/2010.09670.

[30] H. Kim, Torchattacks: A PyTorch Repository for Adversarial Attacks, 2021. URL: http://arxiv.org/abs/2010.01950.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, in: IJCV 2015, arXiv, 2015. URL: http://arxiv.org/abs/1409.0575.

[32] J. Ahlert, T. Klein, F. Wichmann, R. Geirhos, How Aligned are Different Alignment Metrics?, 2024. URL: http://arxiv.org/abs/2407.07530. doi:10.48550/arXiv.2407.07530.

[33] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, J. A. Movshon, A functional and perceptual signature of the second visual area in primates, Nature Neuroscience 16 (2013) 974–981. URL: https://www.nature.com/articles/nn.3402.

[34] N. Baker, J. H. Elder, Deep learning models fail to capture the configural nature of human shape perception, iScience 25 (2022). URL: https://www.sciencedirect.com/science/article/pii/S2589004222011853.

[35] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, W. Brendel, Partial success in closing the gap between human and machine vision, in: 35th Conference on NeurIPS, NeurIPS, 2021. URL: http://arxiv.org/abs/2106.07411.

[36] A. Subramanian, E. Sizikova, N. J. Majaj, D. G. Pelli, Spatial-frequency channels, shape bias, and adversarial robustness, in: Conference on Neural Information Processing Systems, NeurIPS, 2023.

[37] I. Sucholutsky, T. L. Griffiths, Alignment with human representations supports robust few-shot learning, in: Advances in Neural Information Processing Systems, volume 36, 2023, pp. 73464–73479. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/e8ddc03b001d4c4b44b29bc1167e7fdd-Abstract-Conference.html.

[38] B. Hoak, R. Sheatsley, P. McDaniel, Err on the Side of Texture: Texture Bias on Real Data, in: 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), IEEE Computer Society, 2025, pp. 661–680. URL: https://www.computer.org/csdl/proceedings-article/satml/2025/171100a661/26Vnph1kKxW.

[39] B. Hoak, P. McDaniel, Explorations in Texture Learning, in: ICLR 2024, Tiny Papers Track, 2024. URL: http://arxiv.org/abs/2403.09543.