Adversarial alignment: Breaking the trade-off between the strength of an attack and its relevance to human perception

Drew Linsley, Pinyuan Feng, Thibaut Boissin, Alekh Karkada Ashok, Thomas Fel, Stephanie Olaiya, Thomas Serre

▶ **To cite this version:**

Drew Linsley, Pinyuan Feng, Thibaut Boissin, Alekh Karkada Ashok, Thomas Fel, et al.. Adversarial alignment: Breaking the trade-off between the strength of an attack and its relevance to human perception. 2025. hal-04919766

HAL Id: hal-04919766
https://hal.science/hal-04919766v1

Preprint submitted on 29 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:2306.03229v1 [cs.CV] 5 Jun 2023

Adversarial Alignment: breaking the trade-off between the strength of an attack and its relevance to human perception

Drew Linsley[*1,2], **Pinyuan Feng**[*3], **Thibaut Boissin**[4], **Alekh Karkada Ashok**[1],
Thomas Fel[5], **Stephanie Olaiya**[1], **Thomas Serre**[1,2,3,5]
drew_linsley@brown.edu

Abstract

Deep neural networks (DNNs) are known to have a fundamental sensitivity to adversarial attacks, perturbations of the input that are imperceptible to humans yet powerful enough to change the visual decision of a model [1]. Adversarial attacks have long been considered the "Achilles' heel" of deep learning, which may eventually force a shift in modeling paradigms. Nevertheless, the formidable capabilities of modern large-scale DNNs have somewhat eclipsed these early concerns. Do adversarial attacks continue to pose a threat to DNNs?

In this study, we investigate how the robustness of DNNs to adversarial attacks has evolved as their accuracy on ImageNet has continued to improve. We measure adversarial robustness in two different ways: First, we measure the smallest adversarial attack needed to cause a model to change its object categorization decision. Second, we measure how aligned successful attacks are with the features that humans find diagnostic for object recognition. We find that adversarial attacks are inducing bigger and more easily detectable changes to image pixels as DNNs grow better on ImageNet, but these attacks are also becoming less aligned with the features that humans find diagnostic for object recognition. To better understand the source of this trade-off and if it is a byproduct of DNN architectures or the routines used to train them, we turn to the *neural harmonizer*, a DNN training routine that encourages models to leverage the same features humans do to solve tasks [2]. Harmonized DNNs achieve the best of both worlds and experience attacks that are both detectable and affect object features that humans find diagnostic for recognition, meaning that attacks on these models are more likely to be rendered ineffective by inducing similar effects on human perception. Our findings suggest that the sensitivity of DNNs to adversarial attacks can be mitigated by DNN scale, data scale, and training routines that align models with biological intelligence. We release our code and data to support this goal.

1 Introduction

For at least a decade, it has been known that the behavior of deep neural networks (DNNs) can be controlled by small "adversarial" perturbations of the input that are imperceptible to humans [1, 3].

[*]These authors contributed equally.
[1]Department of Cognitive, Linguistic, & Psychological Sciences, Brown University, Providence, RI
[2]Carney Institute for Brain Science, Brown University, Providence, RI
[3]Department of Computer Science, Brown University, Providence, RI
[4]Institut de Recherche Technologique Saint-Exupéry, Toulouse, France
[5]Artificial and Natural Intelligence Toulouse Institute, Toulouse, France

Preprint. Under review.

As DNNs are increasingly being incorporated into software and tools we use in our everyday lives, their vulnerability to adversarial attacks is potentially an unsolved existential threat to the security and safety of these architectures. However, over recent years, the danger of adversarial attacks has been overshadowed by the ever-increasing scale-up of DNNs, and their resulting remarkable achievements across vision, language, and robotics. Billion-parameter DNNs are being trained on internet-scale datasets to perform tasks at levels that rival or surpass humans, bringing us tantalizingly closer to intelligent systems that can transform our lives for the better. It is not known how the scale of DNNs has affected their sensitivity to adversarial attacks.

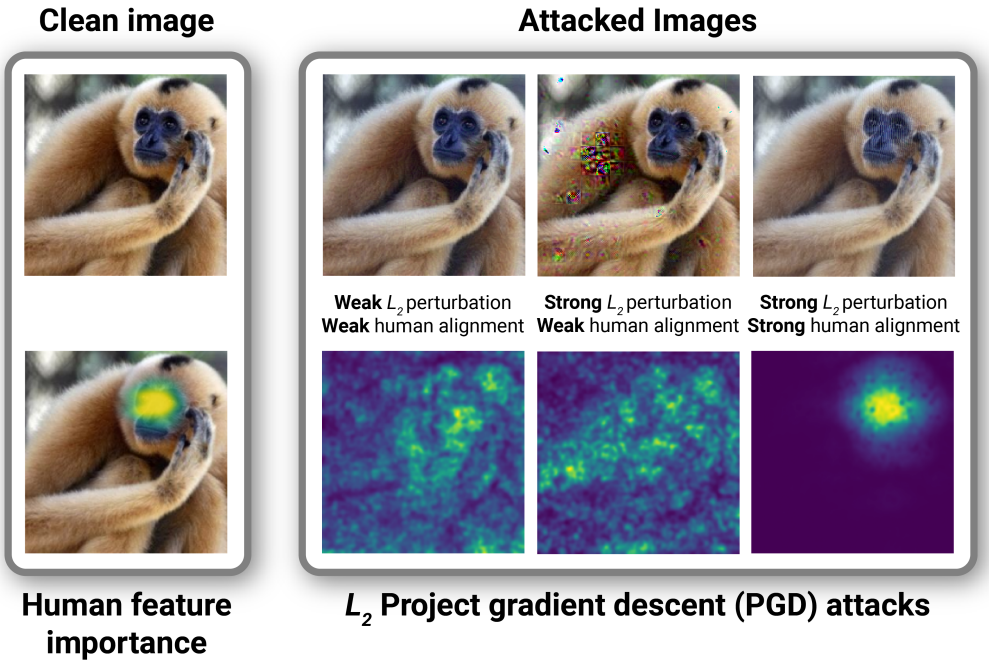

Figure 1: **We propose a new goal for adversarial robustness: Robust models are not only tolerant to strong adversarial image perturbations, successful attacks also target object features that humans find diagnostic for classification.** Adversarial attacks that are large in size and aligned with human perception are more likely to affect humans like they do models, which will neutralize their effectiveness. Shown here are an image of a snow monkey, its corresponding human feature importance map from *ClickMe* [4], and "untargeted" adversarial attacks from ℓ_2 projected gradient descent (PGD) on three different DNNs. One DNN can be attacked with a weak perturbation (as measured by ℓ_2 distance between clean and attacked images), and the successful attack is misaligned with the *ClickMe* feature importance map according to the Spearman correlation between the two. Another DNN is more tolerant to perturbations, but successful perturbations are still misaligned with human perception (strong perturbation/weak alignment). A third DNN approaches our ideal result: strong perturbations are needed for successful attacks, and those perturbations affect features humans use for recognizing the object (the face), which renders these attacks more easily detectable and less effective. Zoom in to see details of each attack.

There are a number of known ways to make DNNs more "robust" to adversarial attacks, meaning that it will take a larger change in pixels between an attacked and clean image to trick a model [5]. For example, there are algorithmic defenses that can be incorporated into DNN inference [6] and training routines that increase the adversarial robustness of DNNs [7–9]. These approaches carry two key drawbacks. First, there is a well-established trade-off between a model's adversarial robustness and its task accuracy [10, 11]. Second, while improving a DNN's robustness means that a stronger perturbation is needed to attack it, there is no constraint on what parts of images are attacked. Humans rely on certain features more than others to recognize objects [2,4,12,13], and if a DNN attack affects features that are less important to humans for recognition, it may still prove be ignored or difficult to notice [14] regardless of the perturbation strength (Fig. 1). We propose that for DNNs to be truly

robust to adversarial attacks, then perturbations should induce large and detectable changes to the object features that humans find diagnostic for recognition (Fig. 1).

There is reason to believe that the scaling laws which have helped DNNs reach their many recent successes in vision and language may at least partially improve their adversarial robustness [15]. Large-scale vision transformers are as robust to non-adversarial image perturbations as humans are, and it is possible that this means larger adversarial perturbations are needed to attack these models [16, 17]. However, DNNs with high accuracy on ImageNet are also learning to recognize objects with features that are misaligned with those used by humans [2]. It is not clear how these features of large-scale DNN vision interact and whether or not they affect the adversarial robustness of models.

Contributions. In this work, we evaluate a large and representative sample of DNNs from the past decade to understand how their adversarial robustness has changed as they have evolved and improved on ImageNet. We measure adversarial robustness in two ways: (*i*) the average ℓ_2 distance between clean and attacked images, which we refer to as "perturbation tolerance", and (*ii*) the alignment of attacks with object features that humans find diagnostic for recognition, which we refer to as "adversarial alignment". We discover the following:

- DNNs have experienced a significant increase in perturbation tolerance as they have improved on ImageNet. In other words, the scale-up of DNNs that has happened over the past several years has partially helped defend them from adversarial attacks.
- In contrast, successful attacks on DNNs are becoming significantly less aligned with the object features that humans rely on for recognition [2, 18] as models grow more accurate at ImageNet.
- Vision transformers [16, 19] (ViTs) and convolutional neural networks (CNNs) are robust in significantly different ways: ViTs have greater perturbation tolerance but CNNs have better adversarial alignment. Most importantly, there is a pareto-front governing the trade-off between these ways of measuring adversarial robustness, indicating that new approaches are needed for human-like adversarial robustness.
- We achieve a partial solution to this goal with the *neural harmonizer* [2], a routine for aligning DNN representations with humans that significantly improves perturbation tolerance and adversarial alignment.

2 Methods

DNN model zoo. We measured the adversarial robustness of 283 DNNs, which are representative of the variety of approaches used in computer vision today. There were 127 convolutional neural networks (CNNs) trained on ImageNet [20–31, 31–33, 33–35, 35–43], 123 vision transformers [16, 19, 44–52] (ViT), and 15 CNN/ViT hybrid architectures that used a combination of both types of circuits [53, 54]. Each model was implemented in PyTorch with the TIMM toolbox (`https://github.com/huggingface/pytorch-image-models`), using pre-trained weights downloaded from TIMM. Additional details on these DNNs, including the licenses of each, can be found in Appendix §A.

Neural Harmonizer. There is a growing body of work indicating that the representations and perceptual behaviors of DNNs are becoming less aligned with humans as they improve on ImageNet [2, 55, 56]. It has also been found that this misalignment can be partially addressed by the *neural harmonizer*, a training routine that forces DNNs to learn object recognition using features that are diagnostic for humans. As this approach has significantly improved the alignment of DNNs with humans [2], we hypothesized that it would also improve the adversarial alignment of DNNs without inhibiting their ability to accurately recognize objects.

Training DNNs for ImageNet with the *neural harmonizer* involves adding an another loss to cross-entropy for object recognition optimization. The additional loss forces a model's gradients to be as similar as possible to feature importance maps collected from humans. Distances between DNN and human feature imporance maps are computed at multiple scales by a function $\mathcal{P}_i(.)$, which downsamples each map $\boldsymbol{\phi}$ to N levels of a pyramid using a Gaussian kernel, with $i \in \{1, ..., N\}$. To train a DNN with the *neural harmonizer* we seek to minimize $\sum_i^N ||\mathcal{P}_i(\boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x})) - \mathcal{P}_i(\boldsymbol{\phi})||^2$ and align DNN feature importance maps with humans at every level of the pyramid. To facilitate learning,

feature importance maps from DNNs and humans are normalized and rectified before distances are computed using $\boldsymbol{z}(.)$, a preprocessing function that takes a feature importance map $\boldsymbol{\phi}$ and transforms it to have 0 mean and unit standard deviation. Putting these pieces together, the completed *neural harmonizer* loss involves computing the following:

$$\mathcal{L}_{\text{Harmonization}} = \lambda_1 \sum_{i}^{N} ||(\boldsymbol{z} \circ \mathcal{P}_i \circ \boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x}))^+ - (\boldsymbol{z} \circ \mathcal{P}_i(\boldsymbol{\phi}))^+||_2 \tag{1}$$

$$+ \mathcal{L}_{CCE}(\boldsymbol{f}_\theta, \boldsymbol{x}, \boldsymbol{y}) + \lambda_2 \sum_{i} \theta_i^2 \tag{2}$$

We follow the original *neural harmonizer* training recipe to optimize 14 DNNs for object recognition on ImageNet while relying on category-diagnostic features captured by *ClickMe* [2]: one `VGG16`, one `ResNet50_v2`, one `ViT_b16`, one `EfficientNet_b0`, six versions of `ConvNext Tiny`, and four versions of `MaxViT Tiny`. Each version was trained with different settings of λ_1 and λ_2, which controlled the relative strength of losses for object recognition and alignment, respectively (see Appendix §B for details).

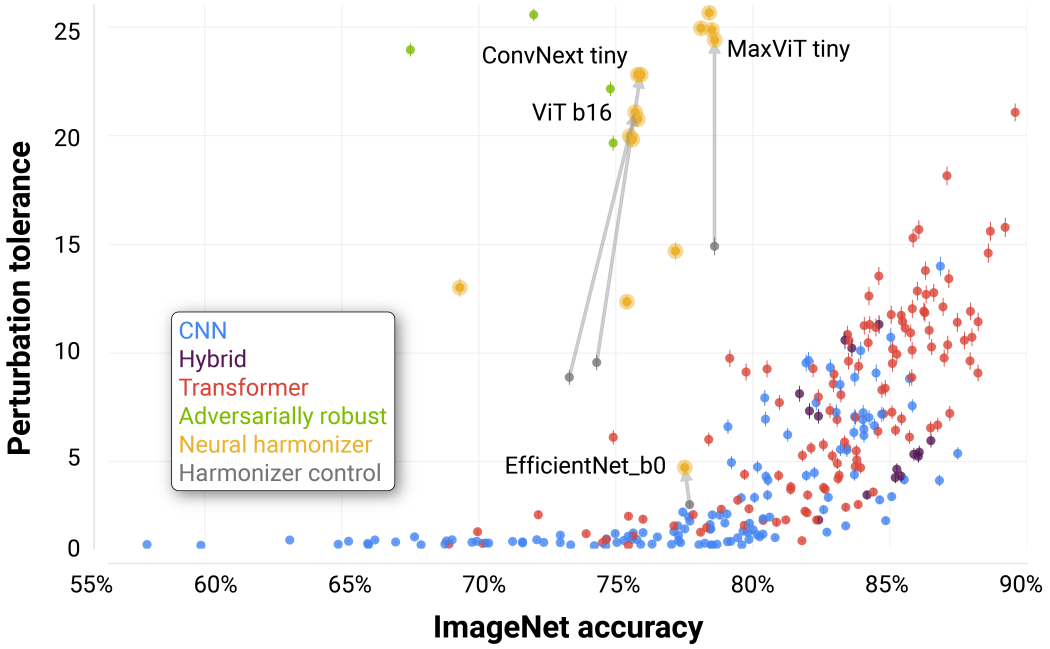

Figure 2: **The perturbation tolerance of DNNs has significantly increased as they have improved on ImageNet.** Each dot denotes a DNN's ImageNet accuracy vs. its average ℓ_2 robustness radius to ℓ_2 PGD attacks, which we call "perturbation tolerance". Arrows show the change of a DNN in both dimensions after it has been trained with the neural harmonizer. There is a significant positive correlation between ImageNet accuracy and perturbation tolerance ($\rho_s = 0.70$, $p < 0.001$). Error bars denote standard error, and variance may be so small for some models that they are not visible.

ClickMe is a large-scale effort for capturing feature importance maps from human participants that highlight parts of objects that are relevant and irrelevant for recognition. For example, these maps focus on the faces of animals, the wheels and fronts of cars, and the wings and cockpits of airplanes [57]. Models were trained for object recognition using the ImageNet training set and *ClickMe* human feature importance maps for the nearly 200,000 images that had annotations. The training was done on Tensorflow 2.0 with 8 V4 TPU cores per model. An object recognition loss was computed for every image, and the full harmonization loss was only computed for those images that had human feature importance maps. Batches of 512 images and feature importance maps were augmented with random left-right flips and mixup [58] during training. Model weights were optimized using SGD with momentum, label smoothing [59], a learning rate of 0.3, and a cosine

learning rate schedule consisting of a five epoch warm-up followed by learning rate decays at steps 30, 50, and 80. We also trained versions `ViT_b16`, `EfficientNet_b0`, `ConvNext Tiny`, and `MaxViT Tiny` with crossentropy but not the complete *neural harmonizer* as controls, and refer to these as harmonizer control models.

Adversarially robust DNNs We also tested the perturbation tolerance and adversarial alignment of robust DNNs. We trained four `Robust ResNetv2-50s` to be tolerant to ℓ_∞-bounded attacks using a standard procedure [7, 60] (code from `https://github.com/microsoft/robust-models-transfer`). A DNN's robustness to these attacks is controlled by a hyperparameter ϵ, and we trained versions with $\epsilon \in 0.001, 0.01, 0.05, 0.1$ and the same training setup used for models trained with the *neural harmonizer*.

Experimental stimuli. We selected 1000 images at random from the ImageNet validation set which also had *ClickMe* feature importance maps. Each image was from a different ImageNet category, and images were preprocessed with each model's specific procedure before computing adversarial attacks.

Adversarial attacks. Ever since the introduction of adversarial attacks [1], the field has exploded with variations that trade-off speed for effectiveness. In our study, we were interested in using attacks that (*i*) could be applied to our model zoo and stimulus set in a reasonable amount of time, (*ii*) would approach the smallest perturbation needed to change a model's behavior, and (*iii*) yielded continuous-valued perturbations that could be compared to *ClickMe* feature importance maps to measure their alignment with human perception. One candidate for these criteria is the popular Fast Gradient Sign Method (FGSM [61]), which is renowned for its time efficiency. However, its attacks are suboptimal [7], and it belongs to the L_∞ attack category which only captures the sign of an attack at every pixel and is poorly suited for computing correlations with human feature importance maps. We instead turned to ℓ_2 Projected Gradient Descent (PGD [7]), which fits each of our criteria. ℓ_2 PGD iteratively searches for the smallest possible image perturbation within a fixed ϵ ball that changes model behavior.

For each model in our zoo, we ran a single ℓ_2 PGD attack for 3 iterations and used binary search to find the minimum perturbation tolerance with $\epsilon \in 0.001, 10$ on every ImageNet image in our stimulus dataset. We also constrained attacks to fall within the pixel range of natural images (i.e. $[0, 255]$). We report perturbation tolerance as the ℓ_2 distance between a clean version of an image and the minimum ϵ attacked-version. All attacks were successful for every image and model. We used NVIDIA TITAN 8 GPUs for generating adversarial attacks, which took between 30 and 240 minutes per model for the complete 1000-image stimulus dataset.

3 Results

DNNs are becoming more tolerant to adversarial attacks as they improve on ImageNet. We used ℓ_2 PGD to attack the object recognition decisions of every DNN in our model zoo for the 1000 images in our stimulus set. We computed perturbation tolerance scores for each DNN as the average ℓ_2 distance between clean images and the attacked versions found by PGD that changed its recognition decision. Surprisingly, as DNNs have improved on ImageNet, their perturbation tolerance has also improved, significantly (Fig. 2, $\rho_s = 0.70$, $p < 0.001$). As a point of comparison, the most accurate DNN we tested, the `eva_giant_patch14_336.m30m_ft_in22k_in1k`, rivaled the perturbation tolerance of `Robust ResNetv2-50s` (i.e., trained for perturbation tolerance) despite being approximately 22% more accurate on ImageNet. We also found a shift in perturbation tolerance based on model architecture. ViTs were significantly more tolerant to perturbations than CNNs (Fig. 2, red vs. blue, $T(122) = 9.12$, $p < 0.001$). We found that this pattern of results replicated when using ℓ_∞ PGD instead of ℓ_2 PGD (Appendix §C, $\rho_s = 0.72$, $p < 0.001$). In other words, the continued optimization of DNNs for performance on ImageNet holds promise for building models that are as robust to image perturbations as any approach designed specifically to build such tolerance.

Successful adversarial attacks are becoming less aligned with human perception. We propose that an adversarially robust DNN should not only be tolerant to strong image perturbations, successful attacks should also target features that humans find diagnostic for object recognition. In this way, even if an attack is successful, it will affect humans like it does DNNs, making the image more difficult to

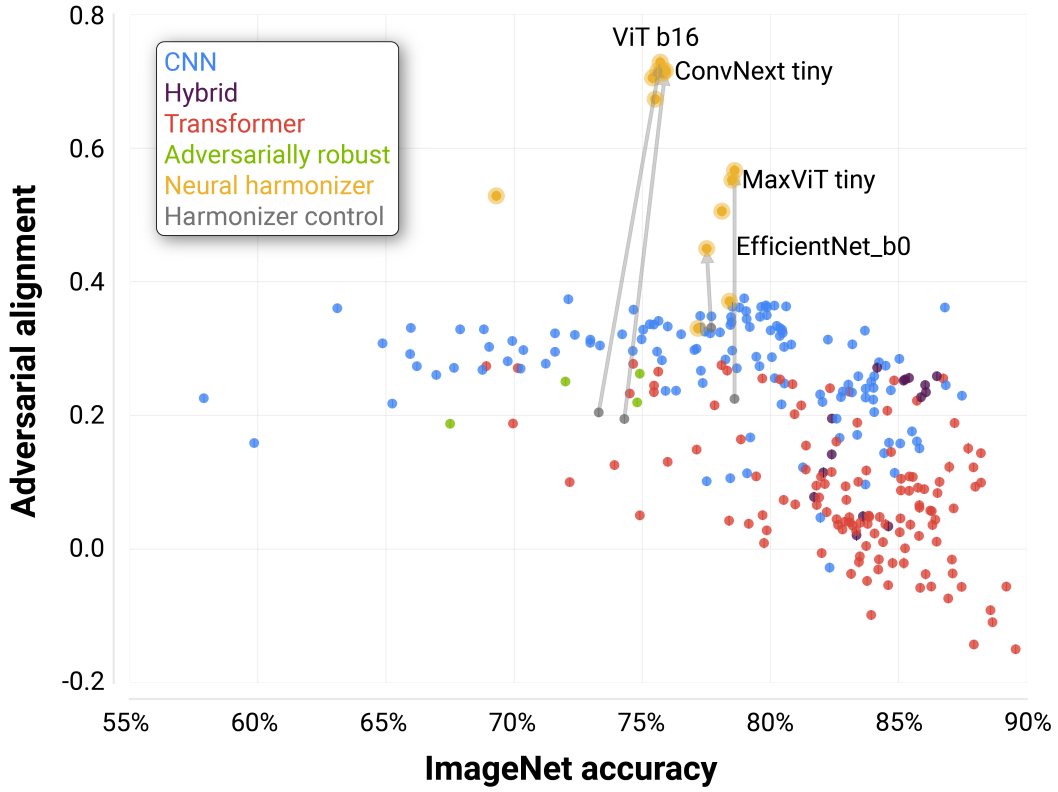

Figure 3: **Successful adversarial attacks on DNNs are becoming less aligned with human perception as they have improved on ImageNet.** Each dot denotes a DNN's ImageNet accuracy vs. the average Spearman correlation between successful attacks an images' human feature importance maps from *ClickMe*. We call this correlation a DNN's adversarial alignment. Arrows show the change of a DNN in both dimensions after it has been trained with the neural harmonizer. Error bars denote standard error, and variance may be so small for some models that they are not visible.

recognize and reducing the potency of the attack. To measure the alignment of a model's adversarial attacks with humans, we turned to *ClickMe*, a large-scale dataset of human feature importance maps for ImageNet [57]. We then measured a DNN's adversarial alignment with humans as the average Spearman correlation between *ClickMe* maps and successful adversarial attacks for every image in our stimulus set.

As DNNs have improved on ImageNet, the alignment of their attacks with human perception has dropped significantly (Fig. 3, $\rho_s = -0.53, p < 0.001$). The 89.57% accurate `eva_giant_patch14_336.m30m_ft_in22k_in1k` has a $\rho_s = -0.15$ adversarial alignment with humans, whereas the 78.98% accurate `MixNet-L` has a $\rho_s = 0.38$ adversarial alignment with humans. In contrast to our findings with perturbation tolerance, CNNs were on average significantly more adversarially aligned with humans than ViTs (Fig. 3, red vs. blue, $T(122) = -18.73$, $p < 0.001$).

DNNs trade-off between perturbation tolerance and adversarial alignment. After plotting the perturbation tolerance of each DNN in our zoo against its adversarial alignment, we found a striking pattern: DNNs either have a strong tolerance to perturbations and misaligned attacks *or* successful attacks are weak in strength but moderately aligned with human perception. The partial outlier to this pattern is DNNs trained for adversarial robustness, which are tolerant to strong perturbations and have moderate adversarial alignment, but are also relatively inaccurate on ImageNet.

We reasoned that another approach for breaking the perturbation strength and adversarial alignment trade-off we observed is to train models for alignment with human perception. One solution to this problem is the *neural harmonizer*, which can significantly improve the representational alignment of DNNs with human perception while also maintaining or slightly improving model accuracy on

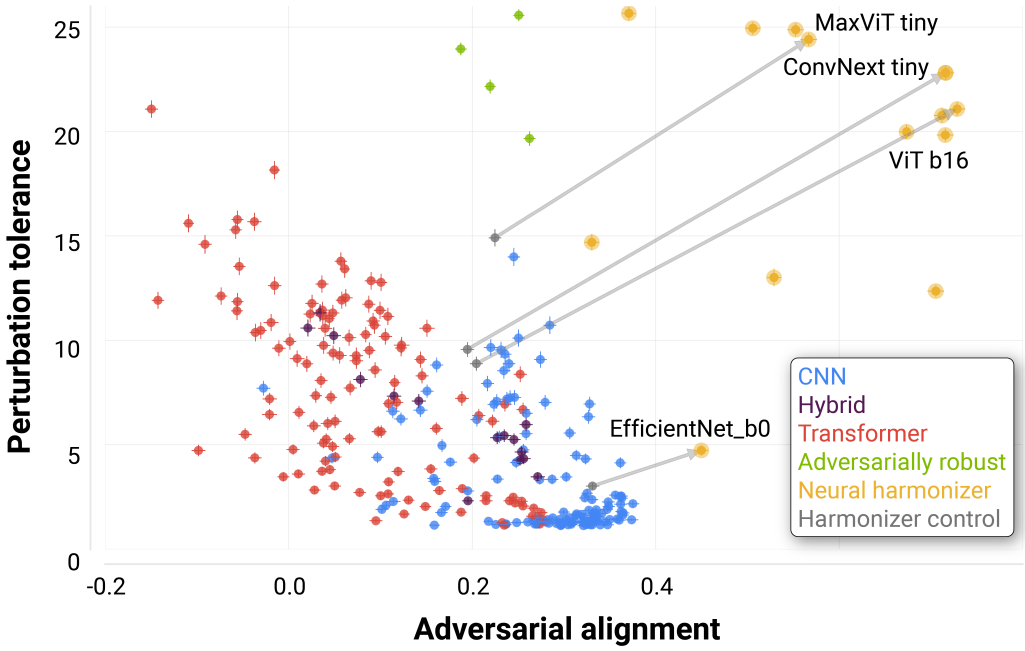

Figure 4: **DNNs trade-off between adversarial alignment perturbation tolerance.** Each dot denotes a DNN's average Spearman correlation between successful attacks and images' human feature importance maps from *ClickMe* vs. the ℓ_2 distance between successfully attacked and clean images. We call these scores adversarial alignment and perturbation tolerance, respectively. Arrows show the change of a DNN in both dimensions after it has been trained with the neural harmonizer. Error bars denote standard error, and variance may be so small for some models that they are not visible.

ImageNet [2], unlike adversarial robustness training [7]. Indeed, we found that a `harmonized ResNet50` approached an `adversarially-trained ResNet50` in average perturbation tolerance (14.69 vs. 23.95) while being 14% more accurate on ImageNet (Fig. 2). We also observed that successful attacks on harmonized DNNs were significantly more aligned with human perception than any other DNN, and they break the perturbation tolerance and adversarial alignment trade-off faced by nearly all other DNNs (average alignment of harmonized DNNs vs. the most aligned unharmonized DNN, $T(999) = 15.63$, $p < 0.001$, Fig. 4).

Successful adversarial attacks on harmonized models target features that humans rely on for recognition: for example, distorting the face of a monkey but leaving the rest of an image untouched (Fig. 5). All other DNNs, including ones trained for adversarial robustness, have attacks that affect image context as much or more than they do the foreground object. While large-scale and highly accurate DNNs like the ViT have high perturbation tolerance, meaning that successful attacks can be visible and detectable by eye (Appendix Fig. S1-2), these patterns of noise may be ignored as inconsequential image distortions [62] since they rarely affect features that are diagnostic for humans.

4 Related work

Adversarial attacks and human perception. Adversarial attacks represent a major threat to safety and security because they are hard or impossible to detect by eye. This feature of adversarial attacks – their perceptibility or lack thereof – has also made them a popular source for study in the vision sciences. It has been suggested that even though adversarial attacks look nonsensical, humans can nevertheless decipher their meaning [63]. Similarly, there is evidence that adversarial attacks on CNNs can transfer to humans in rapid psychophysics experiments [64] and that DNNs trained for adversarial robustness and neurons in primate inferotemporal cortex share a similar tolerance to adversarial perturbations [65]. On the other hand, others have claimed that the similarities between the adversarial robustness of DNN and human vision can be arbitrarily controlled by experimental

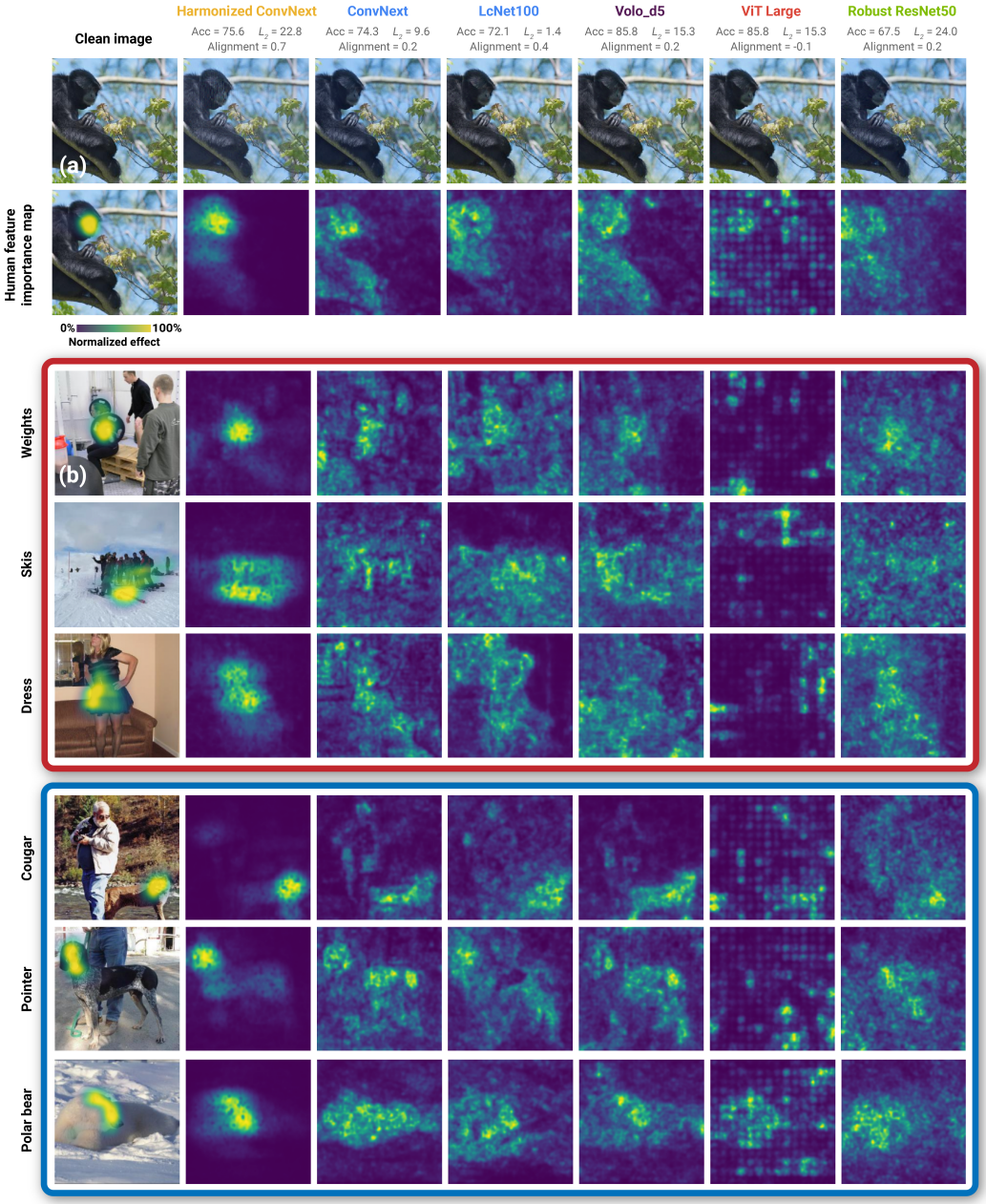

Figure 5: ℓ_2 **PGD adversarial attacks for DNNs.** Plotted here are ImageNet images, human feature importance maps from *ClickMe*, and adversarial attacks for a variety of DNNs. Attacked images are included for the image of a monkey at the top (zoom in to see attack details). The red box shows inanimate categories, and the blue box shows animate categories.

design and stimulus choices [14, 62, 66]. Our findings enrich and reconcile these disparate claims by demonstrating that adversarial robustness, as it is commonly used to describe perturbation tolerance, need not entail alignment with humans. DNNs that achieve perturbation tolerance and adversarial alignment will bring us one step closer towards artificial vision systems that see like humans do.

Aligning the visual strategies of humans and machines. Taken to its extreme, it is possible that a DNN can have arbitrarily high tolerance to adversarial perturbations, but those perturbations could occur in a single, unnoticeable, pixel on the boundary of an image [14]. This is one of the many

reasons why there is a growing urgency in the field of computer vision to ensure that DNNs that rival human performance on image benchmarks can achieve their successes with visual strategies that are interpretable and at least partially consistent with those of humans. There has been progress made towards this goal by evaluating or co-training DNNs with data on human attention and saliency, gathered from eye tracking or mouse clicks during passive or active viewing [57, 67–70]. Others have achieved similar success by comparing the behaviors of models to humans, either by computing and optimizing for distances between patterns of behavior [71–74], or by combining behavioral data with human eye tracking [75]. Another direct comparison of human and DNN alignment involved identifying the minimal image patch needed by each for object recognition [13, 76]. While the *ClickMe* data we used here for harmonizing DNNs is significantly larger than any of these other efforts, we suspect that they hold similar promise in helping DNNs improve the human perceptual alignment of adversarial attacks.

5 Discussion

DNN scale provides valuable protection against the strength of adversarial attacks. Perhaps the biggest breakthrough in artificial intelligence since the release of AlexNet is the finding that scaling the number of parameters in DNNs and the size of their datasets for training can help them rival and outperform humans on challenging tasks [16, 77]. Here, we show that scale also provides concomitant benefits to the perturbation tolerance of DNNs: the size of an adversarial attack needed to affect today's most largest-scale and most-accurate DNNs is significantly greater than ever before. This trend also appears to be accelerating, with ViTs growing tolerant at a faster rate than ever before. DNN scale may be sufficient for "defanging" adversarial attacks by making them detectable to humans.

DNN scale worsens their adversarial alignment with human perception. As the perturbation tolerance of DNNs has improved with ImageNet accuracy, successful attacks on accurate models have begun to consistently affect parts of object images that humans find less important or completely irrelevant for recognition. In other words, DNN scale is at best only a partial solution to adversarial robustness, and it is important for the field of computer vision to explore new approaches to alignment to ensure that adversarial attacks target features humans rely on for behavior. Thus, even if adversarial attacks are successful, they will be ineffective because they induce the same behaviors in humans as they do in DNNs.

The *neural harmonizer* is a short-term solution to adversarial robustness. Harmonized DNNs achieve the best of both worlds of adversarial robustness: they have high perturbation tolerance, and successful attacks target features humans rely on for object recognition. We suspect that scaling the *neural harmonizer* to larger and more accurate DNNs, and expanding the size of *ClickMe* (potentially with pseudo-labels on internet-scale datasets), will bring the field closer to models that are sufficiently robust to adversarial attacks. The success of the *neural harmonizer* also suggests that there is a fundamental misalignment of the training routines used for large-scale DNNs today, and it is possible that advances could also be made without *ClickMe* feature importance maps by inducing more human-like developmental principles onto models. We release our code and data to support continued progress towards adversarial robustness (`https://serre-lab.github.io/Adversarial-Alignment/`).

Limitations. We relied on ℓ_2 PGD for our experiments because it is relatively fast and a "universal first-order adversary" [7], meaning that it is the strongest possible adversarial attack on a DNN that relies on first-order information. While this might suggest that our results are specific to ℓ_2 PGD, we found they translate to ℓ_∞ PGD (Appendix §C). Moreover, in a very small-scale experiment we found a similar pattern of results with the highly-effective but extremely slow-to-compute Carlini-Wagner (CW) attack [78] (Appendix §C). Thus, our findings are likely a general feature of adversarial attacks on DNNs.

Broader impacts. Adversarial attacks have posed an immense problem for the security and safety of DNNs since their discovery. If a DNN's behavior can be controlled by an imperceptible pattern of noise added by a bad actor, then how can they ever be trusted in our everyday lives? We show that the scaling trends that are driving progress in computer vision today offer a partial solution to these

attacks, and new approaches for inducing representational alignment between DNNs and humans can potentially close the remaining gap.

Acknowledgments and Disclosure of Funding

This work was supported by ONR (N00014-19-1-2029), NSF (IIS-1912280 and EAR-1925481), DARPA (D19AC00015), NIH/NINDS (R21 NS 112743), and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). Additional support provided by the Carney Institute for Brain Science and the Center for Computation and Visualization (CCV). We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud (TFRC) program as well as computing hardware supported by NIH Office of the Director grant S10OD025181.

References

[1] Szegedy, C., Google Inc, Zaremba, W., Sutskever, I., Google Inc, Bruna, J., Erhan, D., Google Inc, Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: In ICLR. (2014)

[2] Fel*, T., Felipe*, I., Linsley*, D., Serre, T.: Harmonizing the object recognition strategies of deep neural networks with humans. Adv. Neural Inf. Process. Syst. (2022)

[3] Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. (December 2017)

[4] Linsley, D., Shiebler, D., Eberhardt, S., Serre, T.: Learning what and where to attend. International Conference on Learning Representations (ICLR) (2019)

[5] Kurakin, A., Goodfellow, I., Bengio, S., Dong, Y., Liao, F., Liang, M., Pang, T., Zhu, J., Hu, X., Xie, C., et al.: Adversarial attacks and defences competition. In: The NIPS'17 Competition: Building Intelligent Systems, Springer (2018) 195–231

[6] Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: international conference on machine learning, PMLR (2019) 1310–1320

[7] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations

[8] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning, PMLR (2019) 7472–7482

[9] Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., Usunier, N.: Parseval networks: Improving robustness to adversarial examples. In: International Conference on Machine Learning, PMLR (2017) 854–863

[10] Yang, Y.Y., Rashtchian, C., Zhang, H., Salakhutdinov, R.R., Chaudhuri, K.: A closer look at accuracy vs. robustness. Advances in neural information processing systems **33** (2020) 8588–8601

[11] Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6976–6987

[12] Schyns, P.G., Oliva, A.: From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. Psychol. Sci. (1994)

[13] Ullman, S., Assif, L., Fetaya, E., Harari, D.: Atoms of recognition in human and computer vision. Proc. Natl. Acad. Sci. U. S. A. **113**(10) (March 2016) 2744–2749

[14] Malhotra, G., Evans, B.D., Bowers, J.S.: Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. Vision Res. **174** (September 2020) 57–68

[15] Bubeck, S., Sellke, M.: A universal law of robustness via isoperimetry. Journal of the ACM **70**(2) (2023) 1–18

[16] Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme, C., Minderer, M., Puigcerver, J., Evci, U., Kumar, M., van Steenkiste, S., Elsayed, G.F., Mahendran, A., Yu, F., Oliver, A., Huot, F., Bastings, J., Collier, M.P., Gritsenko, A., Birodkar, V., Vasconcelos, C., Tay, Y., Mensink, T., Kolesnikov, A., Pavetić, F., Tran, D., Kipf, T., Lučić, M., Zhai, X., Keysers, D., Harmsen, J., Houlsby, N.: Scaling vision transformers to 22 billion parameters. (February 2023)

[17] Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., Brendel, W.: Partial success in closing the gap between human and machine vision. (June 2021)

[18] Linsley, D., Kim, J., Ashok, A., Serre, T.: Recurrent neural circuits for contour detection. International Conference on Learning Representations (2020)

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

[20] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. (March 2021)

[21] Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. (May 2019)

[22] Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. (March 2020)

[23] Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. (May 2019)

[24] Simonyan, K., Zisserman, A.: Very deep convolutional networks for Large-Scale image recognition. (September 2014)

[25] Huang, L., Zhao, X., Huang, K.: GOT-10k: A large High-Diversity benchmark for generic object tracking in the wild. (October 2018)

[26] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. (December 2015)

[27] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A.: ResNeSt: Split-Attention networks. (April 2020)

[28] Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2Net: A new Multi-Scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2) (February 2021) 652–662

[29] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (BiT): General visual representation learning. (December 2019)

[30] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. (January 2018)

[31] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. (January 2022)

[32] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. (February 2016)

[33] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. (December 2015)

[34] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. (October 2016)

[35] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. (February 2021)

[36] Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A., Le, Q.V.: Adversarial examples improve image recognition. (November 2019)

[37] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. (November 2016)

[38] Brendel, W., Bethge, M.: Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. (March 2019)

[39] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: XNect: real-time multi-person 3D motion capture with a single RGB camera. ACM Trans. Graph. **39**(4) (July 2020) 82:1–82:17

[40] Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. (July 2017)

[41] Wang, C.Y., Liao, H.Y.M., Yeh, I.H., Wu, Y.H., Chen, P.Y., Hsieh, J.W.: CSPNet: A new backbone that can enhance learning capability of CNN. (November 2019)

[42] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: MnasNet: Platform-Aware neural architecture search for mobile. (July 2018)

[43] Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. (November 2019)

[44] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. (December 2020)

[45] El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., Jegou, H.: XCiT: Cross-Covariance image transformers. (June 2021)

[46] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. (November 2021)

[47] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. (March 2021)

[48] Mehta, S., Rastegari, M.: MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. (October 2021)

[49] Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: MaxViT: Multi-Axis vision transformer. (April 2022)

[50] Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA-02: A visual representation for neon genesis. (March 2023)

[51] Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA: Exploring the limits of masked visual representation learning at scale. (November 2022)

[52] Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. (March 2021)

[53] Xu, W., Xu, Y., Chang, T., Tu, Z.: Co-Scale Conv-Attentional image transformers. (April 2021)

[54] Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S.: VOLO: Vision outlooker for visual recognition. (June 2021)

[55] Kumar, M., Houlsby, N., Kalchbrenner, N., Cubuk, E.D.: Do better ImageNet classifiers assess perceptual similarity better? (September 2022)

[56] Bowers, J.S., Malhotra, G., Dujmović, M., Montero, M.L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J.E., Heaton, R.F., Evans, B.D., Mitchell, J., Blything, R.: Deep problems with neural network models of human vision. Behav. Brain Sci. (December 2022) 1–74

[57] Linsley, D., Shiebler, D., Eberhardt, S., Serre, T.: Learning what and where to attend with humans in the loop. In: International Conference on Learning Representations. (2019)

[58] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. (October 2017)

[59] Müller, R., Kornblith, S., Hinton, G.: When does label smoothing help? (June 2019)

[60] Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., Madry, A.: Do adversarially robust ImageNet models transfer better? (July 2020)

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

[62] Dujmović, M., Malhotra, G., Bowers, J.S.: What do adversarial images tell us about human vision? Elife **9** (September 2020)

[63] Zhou, Z., Firestone, C.: Humans can decipher adversarial images. Nat. Commun. **10**(1) (March 2019) 1334

[64] Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., Sohl-Dickstein, J.: Adversarial examples that fool both computer vision and Time-Limited humans. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., eds.: Advances in Neural Information Processing Systems. Volume 31., Curran Associates, Inc. (2018)

[65] Guo, C., Lee, M., Leclerc, G., Dapello, J., Rao, Y., Madry, A., Dicarlo, J.: Adversarially trained neural representations are already as robust as biological neural representations. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., eds.: Proceedings of the 39th International Conference on Machine Learning. Volume 162 of Proceedings of Machine Learning Research., PMLR (2022) 8072–8081

[66] Malhotra, G., Dujmović, M., Bowers, J.S.: Feature blindness: A challenge for understanding and modelling visual object recognition. PLoS Comput. Biol. **18**(5) (May 2022) e1009572

[67] Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., Serre, T.: What are the visual features underlying human versus machine vision? In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). (October 2017) 2706–2714

[68] Jiang, M., Huang, S., Duan, J., Zhao, Q.: SALICON: Saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015) 1072–1080

[69] Lai, Q., Khan, S., Nie, Y., Shen, J., Sun, H., Shao, L.: Understanding more about human and machine attention in deep neural networks. (June 2019)

[70] Ebrahimpour, M.K., Falandays, J.B., Spevack, S., Noelle, D.C.: Do humans look where deep convolutional neural networks "attend"? In: Advances in Visual Computing, Springer International Publishing (2019) 53–65

[71] Peterson, J.C., Abbott, J.T., Griffiths, T.L.: Evaluating (and improving) the correspondence between deep neural networks and human representations. Cogn. Sci. **42**(8) (November 2018) 2648–2669

[72] Roads, B.D., Love, B.C.: Enriching ImageNet with human similarity judgments and psychological embeddings. (November 2020)

[73] Sucholutsky, I., Griffiths, T.L.: Alignment with human representations supports robust few-shot learning. (January 2023)

[74] Muttenthaler, L., Dippel, J., Linhardt, L., others: Human alignment of neural network representations. arXiv preprint arXiv (2022)

[75] Langlois, T., Zhao, H., Grant, E., Dasgupta, I., Griffiths, T., Jacoby, N.: Passive attention in artificial neural networks predicts human visual selectivity. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W., eds.: Advances in Neural Information Processing Systems. Volume 34., Curran Associates, Inc. (2021) 27094–27106

[76] Funke, J., Tschopp, F.D., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., Turaga, S.C.: Large scale image segmentation with structured loss based deep learning for connectome reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. (2018) 1–1

[77] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. (January 2020)

[78] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp), Ieee (2017) 39–57

[79] Wightman, R.: Pytorch image models. `https://github.com/rwightman/pytorch-image-models` (2019)

[80] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. (June 2009) 248–255

A DNN Model Zoo

We comprehensively evaluated the adversarial robustness of DNNs on a large sample of models from the TIMM toolbox [79]. These DNNs, available under the Apache 2.0 license, are intended for non-commercial research purposes. The complete list of DNNs we evaluated on can be in Table $S1$ below.

Architecture	Model	Versions
CNN	VGG	8
	ResNet	8
	EfficientNet	7
	ConvNext	6
	MobileNet	10
	Inception	3
	DenseNet	4
	RegNet	22
	Xception	4
	MixNet	4
	DPN	6
	DarkNet	1
	NFNet	11
	TinyNet	5
	LCNet	3
	DLA	12
	MnasNet	4
	CSPNet	3
ViT	General ViT	8
	MobileViT	10
	Swin	22
	MaxViT	14
	DeiT	24
	CaiT	10
	XCiT	28
	EVA	5
Hybrid	VOLO	8
	CoAtNet	13

Table S1: **A list of models selected from TIMM library.**

B *Neural Harmonizer* Training

In our work, we followed the original neural harmonizer training recipe to train and harmonize 14 DNNs for object recognition on the ImageNet [80] dataset. By adjusting the regularization terms λ_1 and λ_2, we controlled the relative importance of losses for object recognition and human feature alignment during training. We sampled as many λ_1 and λ_2 settings as possible given our resources, and included all versions in our experiments. In total, we trained one `VGG16`, one `ResNet50_v2`,

Model	Accuracy (%)	Human Alignment (%)	Note
VGG	69.3	61.5	$\lambda = 2$
ResNet 50	77.17	45.0	$\lambda = 2$
EfficientNet B0	77.51	52.3	$\lambda = 20$
ViT B16	75.7	72.6	$\lambda = 5$
ConvNext Tiny v1	75.9	73.2	$\lambda = 1$
ConvNext Tiny v2	75.8	73.3	$\lambda = 2$
ConvNext Tiny v3	75.8	74.5	$\lambda = 3$
ConvNext Tiny v4	75.5	72.1	$\lambda = 5$
ConvNext Tiny v5	75.6	71.1	$\lambda = 8$
ConvNext Tiny v6	75.4	73.2	$\lambda = 10$
MaxViT Tiny v1	78.6	45.3	$\lambda = 1$
MaxViT Tiny v2	78.4	46.8	$\lambda = 2$
MaxViT Tiny v3	78.5	57.6	$\lambda = 5$
MaxViT Tiny v4	78.1	59.0	$\lambda = 10$

Table S2: **DNN architectures trained with the *neural harmonizer.***

one `ViT_b16`, one `EfficientNet_b0`, six variations of `ConvNext Tiny`, and four variations of `MaxViT Tiny` (Table S2). Note that we did not attempt to harmonize models pre-trained on datasets other than ImageNet because the *ClickMe* feature importance dataset we used contained annotations on a subset of images in ImageNet.

C Adversarial attacks

ℓ_2 PGD. The core idea of the Projected Gradient Descent (PGD) attack is to cast adversarial attacks as a constrained optimization problem. PGD leverages the (first-order) gradient information of the model to optimize adversarial attacks while keeping the perturbation size δ within certain constraints. Each step of the PGD attack that we used in our experiments can be presented as follows:

$$\delta := \mathcal{P}\left(\delta + \alpha \nabla_\delta \operatorname{loss}\left(f_\theta(x+\delta), y\right)\right) \tag{3}$$

, where f_θ refers to DNN, α means step size, and $\mathcal{P}$ denotes the projection onto the ball of interest. In our experiments, we used the ℓ_2 PGD to attack the object recognition decisions of DNNs. This attack aims to generate adversarial images by perturbing the input data within a bounded ℓ_2 norm or Euclidean ball, iteratively moving an attacked image representation to the closest point on the circle of a particular radius ϵ centered at the image representation origin (i.e $\|\delta\|_2 = \epsilon$).

To obtain the perturbation tolerance of DNNs, the objective is to find the minimum ϵ that causes the failure of model identification for each image. In our approach, we iteratively refined the value of ϵ using binary search, to efficiently find the minimum ϵ. We started by setting a lower-bound epsilon value ϵ_l and an upper-bound epsilon value ϵ_u based on empirical knowledge. The lower-bound value represents the minimum perturbation that we assumed could result in misclassification, while the upper-bound value was initially set to a large value that we expected would always result in model failure. In each iteration of the binary search, we perturbed the clean image with a midpoint value between ϵ_l and ϵ_u. Then, we evaluated the generated adversarial example by feeding it into the DNN model. If the model correctly identified the adversarial examples, we adjusted the lower-bound epsilon value ϵ_l to the midpoint. However, if the model made the wrong prediction, we updated the upper-bound epsilon value ϵ_u to the midpoint, narrowing down the search range accordingly. We repeated this process until the difference between the upper-bound and lower-bound epsilon values was less than a predefined threshold, indicating that we had converged to a minimum perturbation

Algorithm 1 Find the minimum perturbation tolerance of an DNN model.

Input: a DNN $\mathcal{F}_\theta$, image-label pairs $(\mathcal{X}, \mathcal{Y})$, ℓ_2 PGD function $\mathcal{P}$, ℓ_2 norm function $\mathcal{N}$, lower-bound epsilon ϵ_l, upper-bound epsilon ϵ_u, threshold k.
Output: minimum perturbation tolerance t.

$\mathcal{T} \leftarrow [\,]$
for x_i, y_i in $\mathcal{X}, \mathcal{Y}$ **do**
  $l, r \leftarrow \epsilon_l, \epsilon_u$
  while $r - l \geq k$ **do**
    $m \leftarrow l + (r - l)/2$
    $\hat{x}_i, \hat{y}_i \leftarrow \mathcal{P}(x_i, y_i, m, \mathcal{F}_\theta)$
    if $y_i \neq \hat{y}_i$ **then**
      $r \leftarrow m$
    else
      $l \leftarrow m$
  $\epsilon \leftarrow r$
  $\hat{x}_i, \hat{y}_i \leftarrow \mathcal{P}(x_i, y_i, m, \mathcal{F}_\theta)$
  $\mathcal{T}.\text{append}(\mathcal{N}(x_i, \hat{x}_i))$
$t \leftarrow \text{mean}(\mathcal{T})$
return t

tolerance of that single image. At this point, we averaged the ℓ_2 distortions between the clean image and their corresponding adversarial example. The pseudocode is shown in Alg. 1.

ℓ_2 PGD Attack on Large-scale DNNs. Large-scale and highly accurate DNNs, especially ViT-based models, exhibit a high perturbation tolerance; successful attacks can often result in visible perturbations. However, these visually noticeable patterns of noise do not always affect image features that humans rely on for recognizing objects.

Our study highlights the behavior of six large-scale ViT-based models that demonstrate exceptional performance on the ImageNet dataset, achieving Top-1 accuracy of over 86% (Fig. $S1$ and Fig. $S2$). These DNNs exhibit high perturbation tolerance, but also have low alignment with human perception, complicating their detection and interpretation by human observers. For example, `eva_giant_patch14_336.m30m_ft_in22k_in1k` has the highest object recognition performance among these six ViT-based models, along with the highest perturbation tolerance. However, its adversarial alignment score is negative and anticorrelated with features humans rely on for recognition: the attack primarily affects the background regions instead of the foreground object. The observation suggests that there is a trade-off between perturbation tolerance and adversarial alignment, especially for large-scale and high-accuracy models.

Results of ℓ_∞ PGD Attack. Our earlier findings highlight the valuable protection that the DNN scale offers against adversarial attacks. Notably, we also found this finding holds true when considering the ℓ_∞ PGD attack, as can been seen from Fig. S3. This translation of results to ℓ_∞ PGD attacks (correlation between ℓ_2 and ℓ_∞ perturbation tolerance: $\rho_s = 0.72, p < 0.001$, Fig. S3) provides additional evidence supporting the promising impact of optimizing DNNs for ImageNet performance in building robust models that can withstand image perturbations. This further strengthens the idea that DNNs with higher accuracy on ImageNet tend to display increased resilience against adversarial perturbations.

ℓ_2 Carlini-Wagner & ℓ_2 PGD Attacks. Despite the power of the Carlini-Wagner (C&W) attack, it is known for being extremely slow to compute, which not only requires more gradient steps than PGD but also requires the tuning of an extra parameter denoted c. To understand if they are correlated with the PGD attacks we relied on throughout this work, we ran a small-scale survey of DNN tolerance to CW attacks, involving 50 CNNs and 50 ViTs from our model zoo, and 100 images from our 1000 image stimulus set. We found a similar pattern of results with CW as we did with the ℓ_2 PGD attack (perturbation tolerance: $\rho_s = 0.71, p < 0.001$, Fig. S4; adversarial alignment: $\rho_s = 0.87, p < 0.001$, Fig. S5).

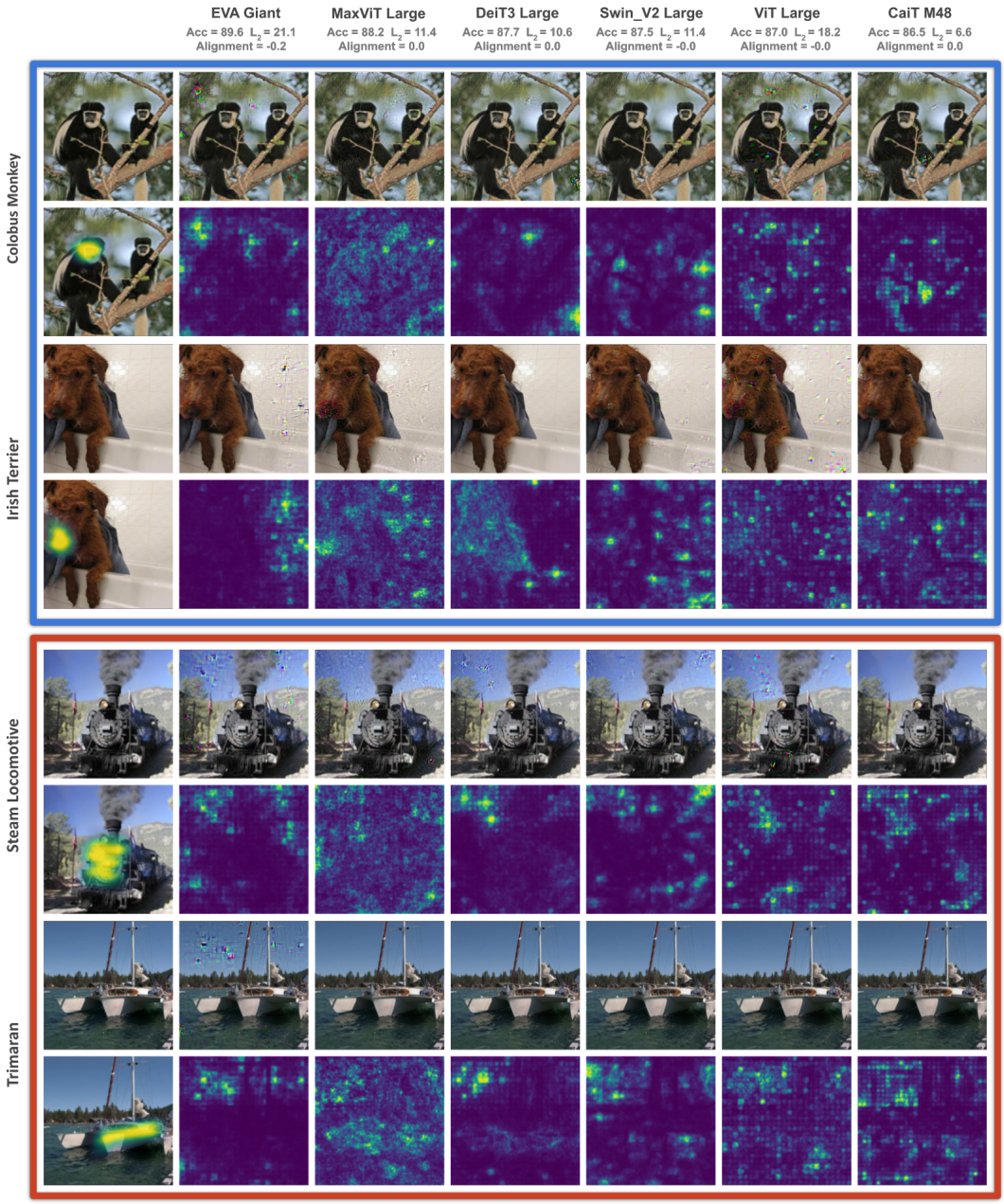

Figure S1: ℓ_2 **PGD attack on large-scale models.** Plotted here are ImageNet images, human feature importance maps from ClickMe, and adversarial attacks for 6 large-scale and high-accuracy DNNs. The red box shows inanimate categories, and the blue box shows animate categories.

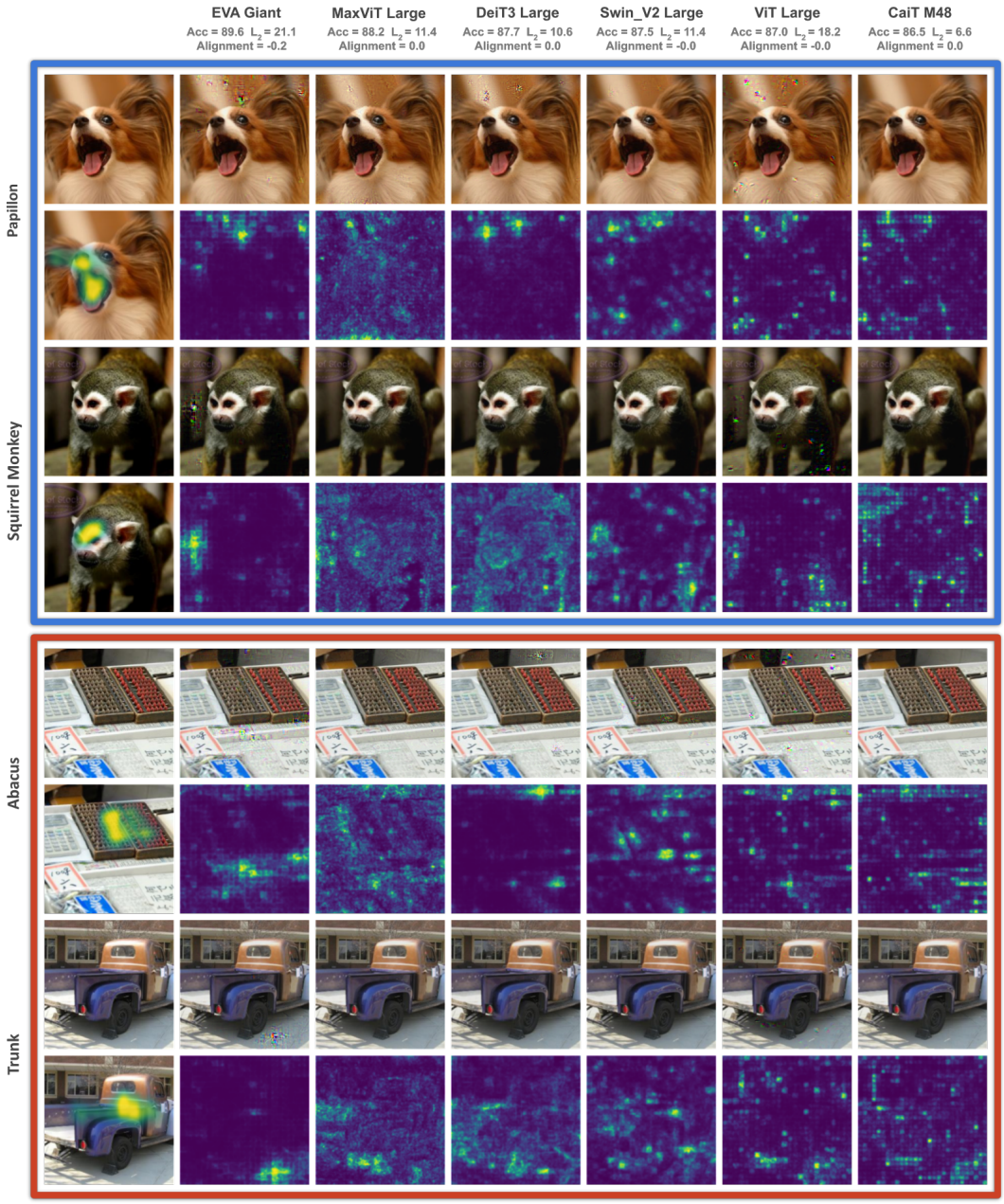

Figure S2: ℓ_2 **PGD attack on large-scale models.** Plotted here are ImageNet images, human feature importance maps from ClickMe, and adversarial attacks for 6 large-scale and high-accuracy DNNs. The red box shows inanimate categories, and the blue box shows animate categories.

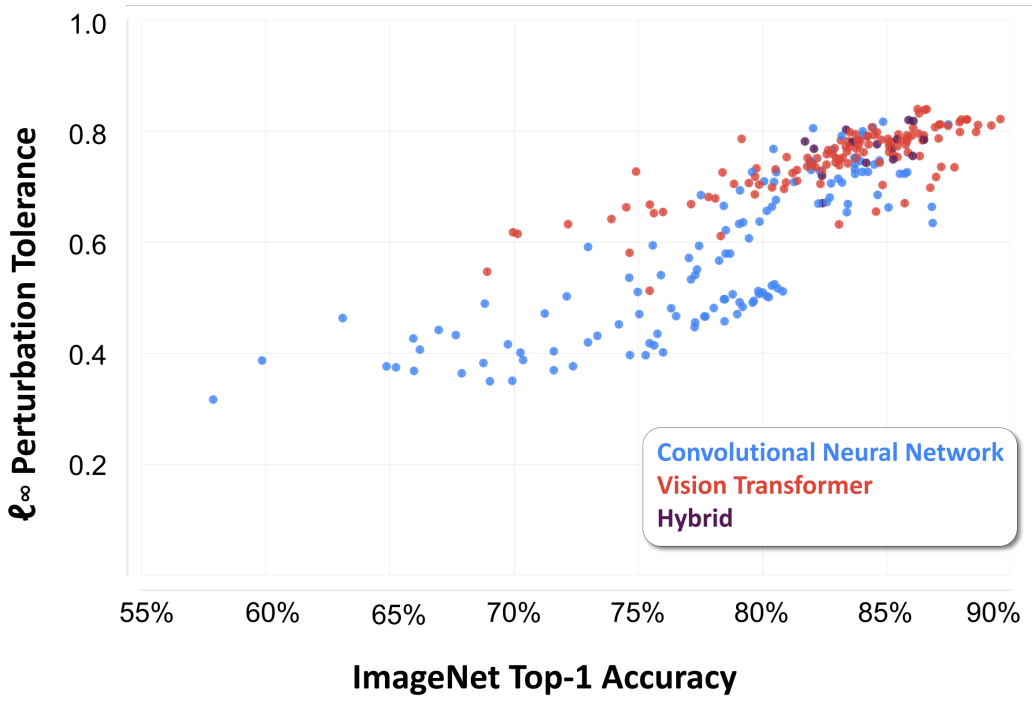

Figure S3: **The perturbation tolerance of DNNs based on ℓ_∞ PGD attack increases as they have improved on ImageNet.**

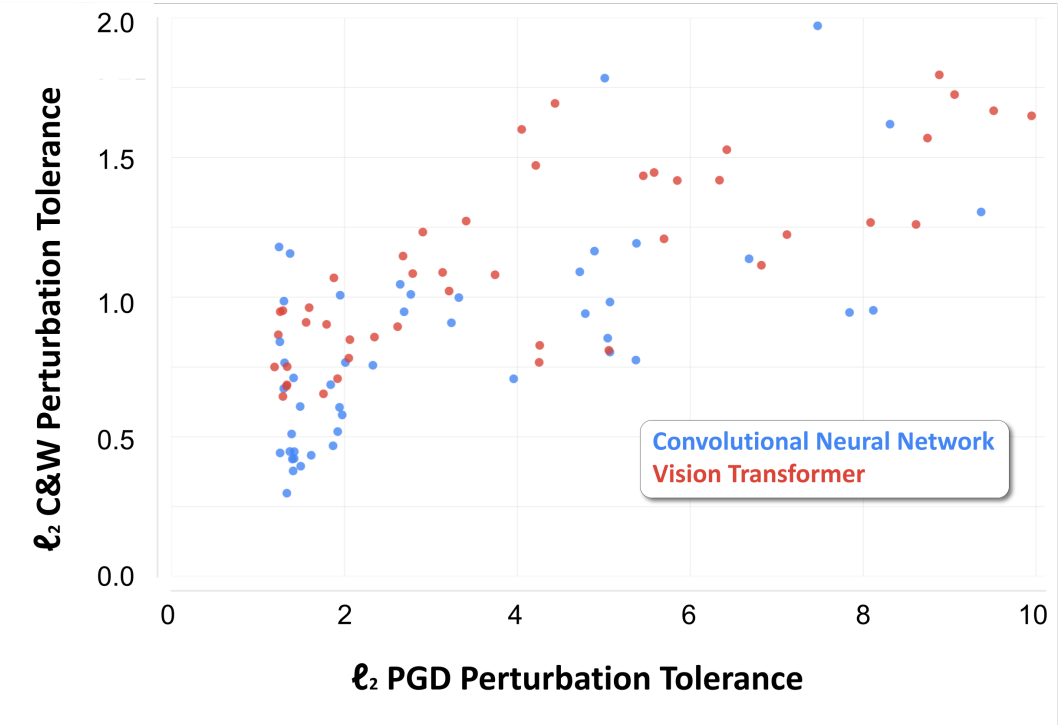

Figure S4: **A comparison between ℓ_2 PGD attack and ℓ_2 C&W attack on perturbation tolerance.**

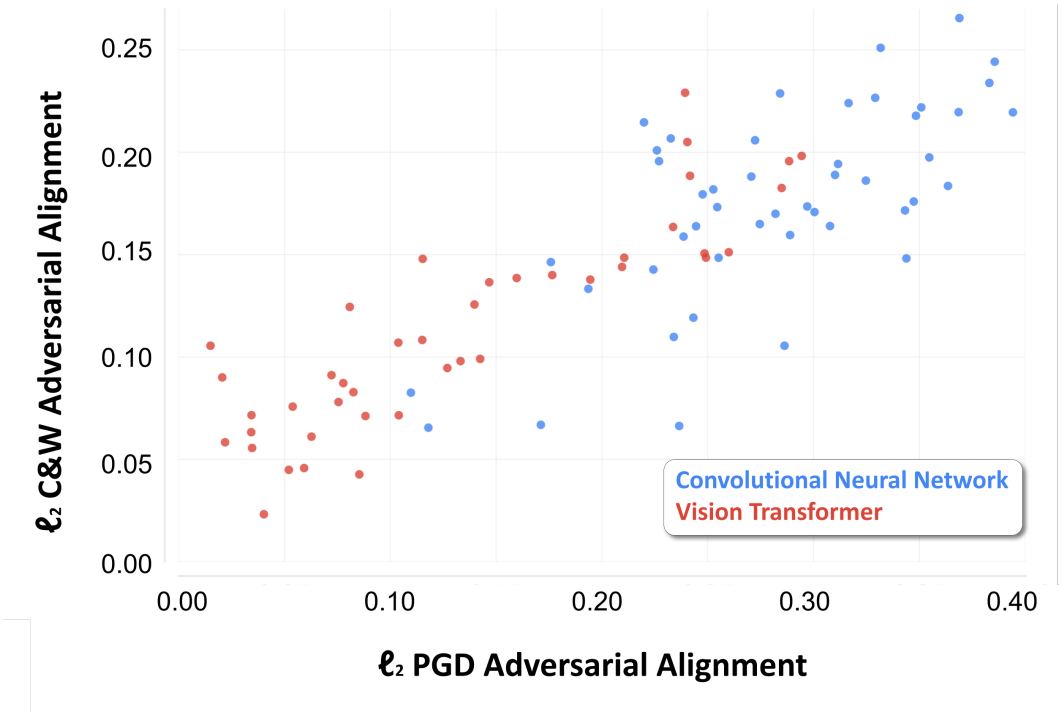

Figure S5: **A comparison between ℓ_2 PGD attack and ℓ_2 C&W attack on adversarial alignment.**