BEYOND ADVERSARIAL ROBUSTNESS: BREAKING THE ROBUSTNESS-ALIGNMENT TRADE-OFF IN OBJECT RECOGNITION

Pinyuan Feng[1]* **Drew Linsley**[2,3]* **Thibaut Boissin**[4] **Alekh Karkada Ashok**[2]

Thomas Fel[5] **Stephanie Olaiya**[2] **Thomas Serre**[2,3]

[1]Department of Psychology, Columbia University, USA
[2]Department of Cognitive and Psychological Sciences, Brown University, USA
[3]Carney Institute for Brain Science, Brown University, USA
[4]Institut de Recherche Technologique Saint-Exupéry, France
[5]Kempner Institute, Harvard University, USA

ABSTRACT

A well-known limitation of deep neural networks (DNNs) is their sensitivity to adversarial attacks (Szegedy et al., 2014). That DNNs can easily be fooled by minute image perturbations imperceptible to humans has long been considered a significant vulnerability of deep learning, which may eventually force a shift towards modeling paradigms that are faithful to biology. Nevertheless, the ever-evolving capabilities of DNNs have largely eclipsed these early concerns. Do adversarial perturbations continue to pose a threat to DNNs?

Here, we investigate whether DNN improvements in image categorization have led to concurrent improvements in robustness to adversarial perturbations. We evaluated DNN adversarial robustness in two ways. First, we measured the *tolerance* of DNNs to adversarial perturbations by recording the norm of the smallest image perturbation needed to change a model's decision using a standard "minimum-norm" robustness metric. Second, we measured *alignment* of perturbations and the degree to which they target pixels that are diagnostic for human observers. We uncover a surprising trade-off: as DNNs have improved on ImageNet, they have grown more tolerant to adversarial perturbations. However, these perturbations are also progressively less aligned with features critical to humans for object recognition.

To better understand the source of this trade-off, we turn to DNN training methods that have previously been reported to align DNNs with human vision, namely *adversarial training* (Goodfellow et al., 2014) and *harmonization* (Fel et al., 2022). Our results show that both methods improve this trade-off, significantly increasing the tolerance to adversarial perturbations and alignment of DNN perturbations with human visual features. Harmonized models, unlike adversarially trained ones, are also able to maintain their ImageNet accuracy in the process. Our findings suggest that, the vulnerability of DNNs to adversarial perturbations can be at least partially mitigated by augmenting the trends in model scaling that are driving development today with training routines that align models with biological intelligence. We release our code and data to support continued progress toward studying the adversarial behavior of DNNs.

*These authors contributed equally to this work. Email: pf2477@columbia.edu
†https://github.com/TonyFPY/Adversarial_Alignment

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 (Szegedy et al., 2014; Biggio & Roli, 2017). Given the myriad of ways in which DNNs are increasingly being deployed in our everyday lives, this vulnerability may pose a significant security threat. In recent years, the dangers of adversarial perturbations have been gradually overshadowed as they have continuously been scaled-up. Billion-parameter DNNs that have been trained on internet-scale datasets rival or surpass humans performance across vision, language, and reasoning tasks. However, it is not yet known how this scaling of DNNs has affected their sensitivity to adversarial perturbations.

There are several ways to make DNNs more "robust" to adversarial attacks, meaning that it will take a larger image perturbation (in terms of pixel change) than usual to change a model's decision (Kurakin et al., 2018). For example, there are algorithmic defenses that can be incorporated into DNN inference (Cohen et al., 2019) and training routines that increase the adversarial robustness of DNNs (Madry et al., 2018; Zhang et al., 2019; Cisse et al., 2017). These approaches carry two key drawbacks. First, there is a well-established trade-off between a model's adversarial robustness and task accuracy (Yang et al., 2020; Stutz et al., 2019). Second, while improving a DNN's robustness means that a stronger perturbation is needed to attack it, there is no constraint on which image pixels are attacked. Humans rely on some visual features more than others to recognize objects (Schyns & Oliva, 1994; Ullman et al., 2016; Linsley et al., 2019a; Fel et al., 2022). If a perturbation targets features that are less important to humans for recognition, it may be challenging to notice regardless of the size of the perturbation (Malhotra et al., 2020) regardless of the perturbation strength (Fig. 1). We propose that for DNNs to be genuinely robust to adversarial attacks, perturbations should be as noticeable to humans as possible. That is, the minimal perturbation needed to change a model's decision should result in large changes to object features that are diagnostic to humans for recognition (Fig. 1).

There are reasons to believe that the scaling laws that have helped DNNs reach their many recent successes in vision and language may at least partially improve their adversarial robustness (Bubeck & Sellke, 2023). Large-scale vision transformers are as robust to random image perturbations as humans are (Dehghani et al., 2023; Geirhos et al., 2021), and it is possible that this means larger adversarial perturbations are needed to attack these models. However, state-of-the-art DNNs also learn to recognize objects using different visual features than humans (Fel et al., 2022). It is unclear how these attributes 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 perturbed images, which we refer to as perturbation *tolerance*, and (*ii*) the alignment of these attacks with object features that humans find diagnostic for recognition, which we refer to as human *alignment*. We discover the following:

- DNNs have grown signficantly more tolerant to the size of adversarial perturbations as they have improved on ImageNet. This means that the scaling of DNNs over the past several years has brought some defense against adversarial attacks.
- In contrast, successful attacks on DNNs are becoming significantly less aligned with humans as they target image pixels that are less important for human vision.
- Most importantly, there is a pareto-front governing the trade-off between *Perturbation Tolerance* and *Adversarial Human Alignment*, suggesting that new approaches are needed for human-like adversarial robustness.
- We find that this pareto front can be broken by training routines that have previously been found effective at aligning DNNs with human perception: the *harmonization* (Fel et al., 2022) and *adversarial training* (Madry et al., 2018). Both approaches lead to DNNs with significantly improved *Perturbation Tolerance* and *Adversarial Human Alignment*; however, only harmonization leads to models that maintain (or slightly improve) their accuracy.

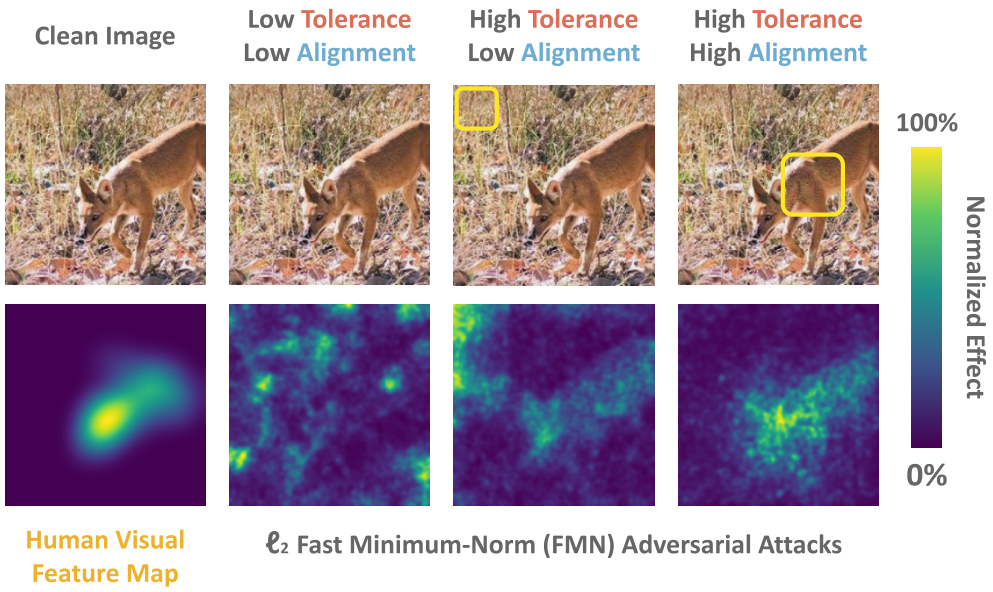

Figure 1: **We propose a new goal for adversarial robustness: Robust models should be tolerant to strong adversarial image perturbations, and successful attacks should target object features that are diagnostic to humans for classification.** Adversarial perturbations that are large in size and aligned with human perception are more likely to be noticed by users, which will help neutralize them. Depicted here are ImageNet images, the corresponding human feature importance map from *ClickMe* (Linsley et al., 2019a), and adversarial perturbations from ℓ_2 Fast Minimum-Norm (FMN) attacks on three different DNNs that are designed to change the models' decisions from animal→non-animal. 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 (Low Tolerance/Low Alignment). Another DNN is more tolerant to perturbations, but successful perturbations affect features that are different than those that humans rely on for object recognition (High Tolerance/Low Alignment). A third DNN approaches our ideal outcome: successful attacks are only possible with strong perturbations, and those perturbations affect features humans find diagnostic (High Tolerance/High Alignment). Please zoom in for more details.

2 Related work

2.1 Adversarial perturbations and human perception

The vulnerability of DNNs to adversarial perturbations represents a major threat to the safety and security of real-world DNN-based applications. That adversarial perturbations are imperceptible to humans has also made them a popular source for study in the vision sciences. There is evidence that adversarial perturbations on convolutional neural networks (CNNs) can transfer to humans in rapid psychophysics experiments (Elsayed et al., 2018; Gaziv et al., 2023). Others have found that neurons in primate inferotemporal cortex share a similar tolerance to adversarial perturbations as adversarially trained DNNs (Guo et al., 2022). On the other hand, others have claimed that the similarities between the adversarial robustness of DNN and human perception can be arbitrarily controlled by experimental design and stimulus choices (Malhotra et al., 2020; Dujmović et al., 2020; Malhotra et al., 2022). Moreover, researchers have shown that scaling up can help adversarial training achieve better performance on adversarial defenses, but the generated adversarial examples become biased relative to human perception (Bartoldson et al., 2024). Unlike prior work, our study enriches and reconciles 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 Human Alignment* will bring us one step closer toward artificial vision systems that see as humans do.

2.2 Aligning the visual strategies of humans and machines

Taken to its extreme, it is possible for a DNN to have an arbitrarily high tolerance to adversarial perturbations. However, we argue that tolerance alone is an incomplete description of DNN robustness, since a strong perturbation could be achieved through large changes to unnoticeable pixels (for instance, on an image boundary) (Malhotra et al., 2020). This is one of the many reasons why there is a growing urgency in computer vision to ensure that DNNs that rival human performance on image benchmarks can succeed with visual strategies that are interpretable and approximately 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 (Linsley et al., 2017; 2019b; Jiang et al., 2015; Lai et al., 2019; Ebrahimpour et al., 2019). Others have found success by comparing the behaviors of DNNs with humans, either by optimizing for distances between patterns of behavior (Peterson et al., 2018; Roads & Love, 2020; Muttenthaler et al., 2022; Sucholutsky & Griffiths, 2023), or by combining behavioral data with human eye tracking (Langlois et al., 2021). Another direct comparison of human and DNN alignment involved identifying the minimal image patch needed by each for object recognition (Ullman et al., 2016; Funke et al., 2018). In this work, we turn to *harmonization*, an approach that forces DNNs to solve tasks by relying on similar features as humans (Fel et al., 2022).

3 Methods

3.1 Experimental stimuli

We evaluated models on a standard ImageNet subset used for evaluating adversarial attacks (Engstrom et al., 2019) (see A.1 for details), consisting of 960 images from 240 categories. To simplify the adversarial perturbation space, and control for potential confounds related to the perceptual distance between ImageNet's 1000 categories, we reduced the task to animal vs. non-animal classification. Next, we paired each image with a visual feature importance map from human participants that highlights parts of objects relevant for recognition taken from *ClickMe* (Linsley et al., 2019a). These maps highlight parts of the faces of animals, the wheels and front grilles of cars, and the wings and cockpits of airplanes.

3.2 DNN model zoo

We evaluated 309 DNNs that are representative of the variety of approaches used in computer vision today. Each model was implemented in PyTorch with the TIMM toolbox (Wightman, 2019), using pre-trained weights downloaded from TIMM. There were 125 **convolutional neural networks** (CNNs) trained on ImageNet, 78 **vision transformers** (ViT), and 52 **hybrid architectures** that used a combination of CNN and ViT. Also, we incorporated 54 models pre-trained using representative **self-supervised learning** methods, such as CLIP (Radford et al., 2021) and DINO (Caron et al., 2021; Oquab et al., 2023), which have emerged as effective training paradigms for learning generalizable visual representations on large-scale unlabeled data. Additional details on these DNNs, including individual licenses, can be found in A.2.

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 (Fel et al., 2022; Kumar et al., 2022; Bowers et al., 2022). 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 (Fel et al., 2022). As this approach has significantly improved the alignment of DNNs with humans, we hypothesized that it would also improve the *Adversarial Human Alignment* of DNNs without inhibiting their ability to accurately recognize objects. In our experiment, we evaluated the impact of harmonization (**harmonized**) on 9 models (see A.3 for details on training).

Robustified DNNs. We also tested 33 adversarially-trained (**Robust**) DNNs. We trained *ResNet18, ResNet50, Wide ResNet-50-2* (He et al., 2016; Zagoruyko & Komodakis, 2016) to be tolerant to ℓ_∞-bounded and ℓ_2-bounded PGD perturbations (Madry et al., 2018) using the *robustness* package (Engstrom et al., 2019). A DNN's robustness to these attacks is controlled by a hyperparameter ϵ, which is the maximum allowable perturbation. For ℓ_2-bounded perturbationa, we

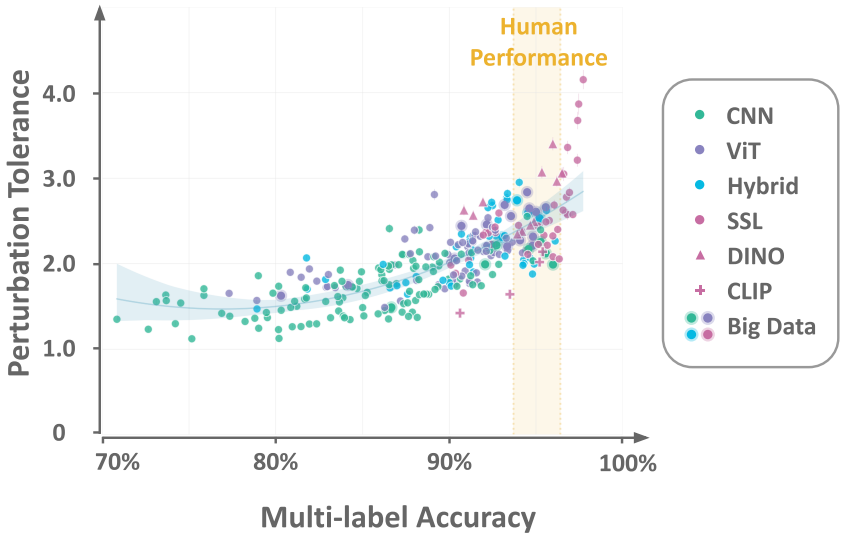

Figure 2: **The *Perturbation Tolerance* of DNNs has significantly increased as they have improved on ImageNet.** Each dot denotes a DNN's performance on multi-label ImageNet (Shankar et al., 2020) vs. its average ℓ_2 robustness radius to ℓ_2 minimum-norm attacks, which we call *Perturbation Tolerance*. Error bars denote standard error, and the variance is so small for some models that it is invisible. The yellow region represents human performance, with mean=95.14% and std=2.02%.

trained models with $\epsilon \in \{0.01, 0.03, 0.05, 0.1, 0.25, 0.5\}$ and for ℓ_∞-bounded attacks, we used $\epsilon \in \{0.5/255, 1.0/255, 2.0/255, 4.0/255, 8.0/255\}$. In evaluation, we used a different perturbation (see Section 3.3) to ensure fair comparisons across models, as training on a specific attack would inherently bias the models toward better performance against that same attack during evaluation (see A.4 for more details).

3.3 Adversarial attacks and Evaluation

Ever since the introduction of adversarial attacks (Szegedy et al., 2014), 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 prediction, and (*iii*) yielded continuous-valued perturbations that could be compared to *ClickMe* feature importance maps to measure their alignment with human visual perception.

One group of candidates we considered is bounded-norm attacks, such as FGSM (Goodfellow et al., 2014), BIM (Kurakin et al., 2018) , and PGD (Madry et al., 2018). These are popular methods that are widely used for evaluating adversarial robustness due to their simplicity and efficiency. However, they generate adversarial examples by perturbing inputs within a predefined norm constraint. This means we need to additionally search for the minimum perturbation needed for an attack, which can sometimes result in less accurate or suboptimal solutions. To address this limitation, we turned to the Fast Minimum-Norm (FMN) adversarial attack (Pintor et al., 2021), which outperforms other minimum-norm attacks in terms of speed, reliability, and effectiveness.

Following the FMN approach of Pintor et al. (2021), we ran ℓ_2 FMN attack for 1000 iterations, using an annealing step size that starts at 1.0 and decreases to 10^{-5}. The initial step size for epsilon update γ was set to 5.0. The algorithm performs normalized gradient descent and projects into an adaptive epsilon to find the minimum norm. All attacks were successful for every image and model. We used 1 NVIDIA A6000 GPU for the attacks, which took between 30 and 240 minutes per model.

Perturbation Tolerance. We compute the ℓ_2 distance between a clean image and the minimum ϵ attacked counterpart, and we report the metric as the average distance from the evaluation over the entire dataset to show how resistant models are to adversarial attacks. Higher *Perturbation Tolerance* indicates larger changes to the input are required to change the model's decision, and vice versa.

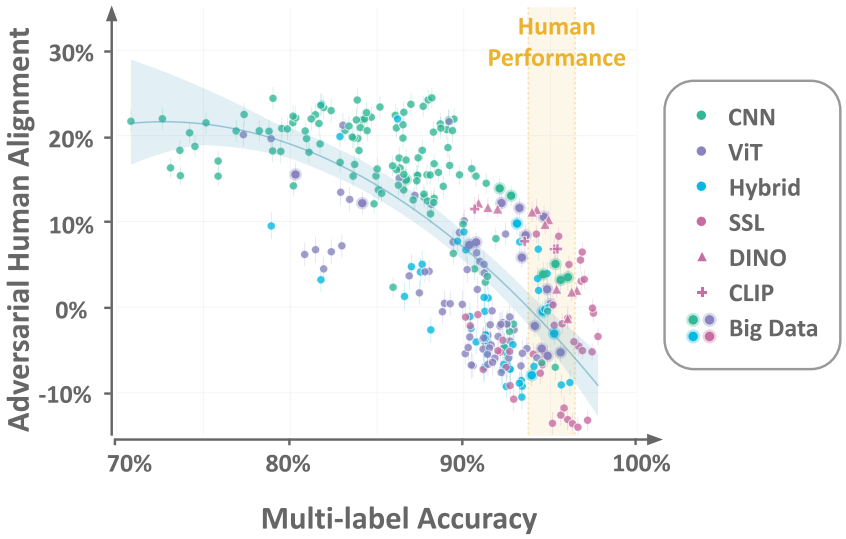

Figure 3: **Successful adversarial attacks on DNNs are becoming less aligned with human perception as they have improved on ImageNet.** Each dot denotes a DNN's ImageNet performance on multi-label accuracy (Shankar et al., 2020) vs. the average Spearman correlation between successful attacks an images' human feature importance maps from *ClickMe*. We call this correlation a DNN's *Adversarial Human Alignment*. Error bars denote standard error, and variance may be so small for some models that they are not visible. We fit a regression line to show the decreasing trend.

Adversarial Human Alignment. We measure the average Spearman Rank Correlation between the perturbation pattern from an attacked image and the human visual feature map of the corresponding clean image. *Adversarial Human Alignment* is quantified by averaging the correlation across all pairs. While humans and machines exhibit different levels of sensitivity to perturbation intensity, our primary focus is on *where* the perturbations occur. We hypothesize that the spatial alignment reveals whether models "fail in human-like ways." In other words, we probe whether the perturbed features overlap with the features humans find diagnostic for recognition, in order to determine whether they share similar perceptual biases or not.

Multi-label Accuracy. We quantify the model's performance on ImageNet using *Multi-label Accuracy* (Shankar et al., 2020; Vasudevan et al., 2022). Unlike traditional top-1 and top-5 accuracy, which evaluate models based on a single predicted label or a narrow set of candidates, multi-label accuracy accounts for all semantically correct labels for an image. This metric also incorporates human performance, allowing for a more direct comparison between humans and models.

4 Results

DNNs are becoming more tolerant to adversarial attacks as they improve on ImageNet. We used ℓ_2 FMN to attack DNNs in our model zoo to change their object recognition decisions on each image from our stimulus set. We computed *Perturbation Tolerance* scores for each DNN as the average ℓ_2 distance between clean images and the minimum-attacked versions found by FMN. We found that, as DNNs have improved on ImageNet, their *Perturbation Tolerance* has also improved, significantly (Fig. 2, $\rho_s = 0.81$, $p < 0.001$). We use a second-degree polynomial regression line to fit the trends, but it fails to capture the continued growth beyond human-level ImageNet performance. The most accurate DNN[†] we tested, rivaled the *Perturbation Tolerance* of several adversarially-trained models despite being approximately $11.24\% \sim 46.62\%$ more accurate on ImageNet. We also found a shift in *Perturbation Tolerance* based on model architecture. ViTs are significantly more tolerant to attacks than CNNs (Fig. 2, ViT vs. CNN, $T(163) = 9.11$, $p < 0.001$). For the learning paradigm, models pre-trained through self-supervised learning, which achieve higher accuracy on ImageNet, are significantly more robust than those trained through supervised learning (Fig. 2, SSL vs. Other,

[†]eva02_large_patch14_448.mim_in22k_ft_in22k_in1k

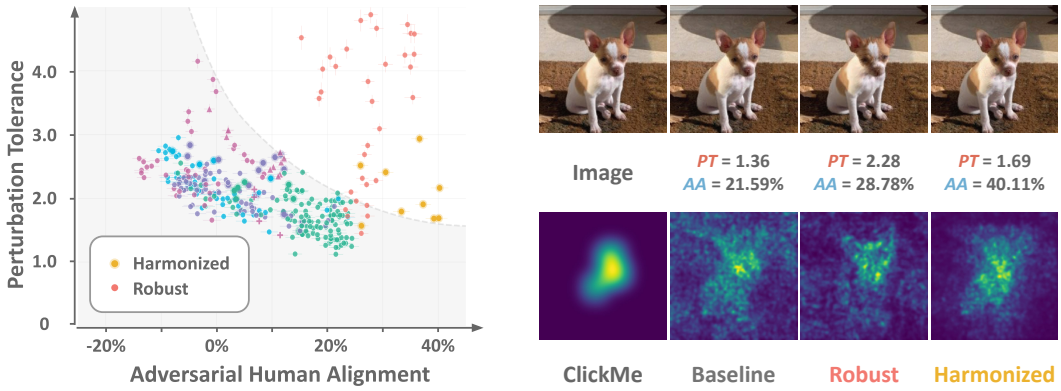

Figure 4: **DNNs trade-off between *Adversarial Human Alignment* and *Perturbation Tolerance*. Left** Each dot denotes a DNN's average Spearman correlation between successful attacks and human feature importance maps from *ClickMe* vs. the ℓ_2 distance between successfully attacked and clean images. There is a pareto-front governing the trade-off between these ways of measuring DNNs' behavior under adversarial attacks. DNNs that are harmonized or adversarially trained are able to break the trade-off. **Right** Images of a Chihuahua with perturbations from different models, along with the human feature importance map from *ClickMe* and maps depicting the adversarial perturbations for a baseline, adversarially robust, and harmonized *ResNet18*. The *Adversarial Human Alignment* (AA) and *Perturbation Tolerance* (PT) scores are also depicted.

$T(65) = 7.85$, $p < 0.001$). Additionally, data diet plays a key role because we observed that models pre-trained on larger datasets and then finetuned on ImageNet are more tolerant to perturbations compared to models solely trained on ImageNet (Fig. 2, Big Data vs. Other, $T(32) = 5.65$, $p < 0.001$). Implying from the above results, 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.

As DNNs improve on ImageNet, their adversarial perturbations are becoming less aligned with visual features that are diagnostic to humans. To measure the alignment between a DNN's adversarial perturbations and human perception, we turned to *ClickMe*, a large-scale dataset of human feature importance maps for ImageNet. We then measured a DNN's *Adversarial Human Alignment* as the average Spearman correlation between *ClickMe* maps and successful adversarial perturbations 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, $\rho_s = -0.74, p < 0.001$). Again, we see that the regression line, fitted with a second-degree polynomial, follows a downward trend, but fails to account for the sharp drop after reaching human-level ImageNet performance. The most accurate model has a $\rho_s = -0.03$ *Adversarial Human Alignment*, whereas the least accurate model[†] has a higher alignment of $\rho_s = 0.22$. In contrast to our findings with *Perturbation Tolerance*, CNNs are on average significantly more aligned with human visual features than ViTs (Fig. 3, ViT vs. CNN, $T(153) = -12.5$, $p < 0.001$). Similarly, self-supervised models, which are widely believed to learn more robust and generalizable representations, shows significantly lower *Adversarial Human Alignment* than supervised models (Fig. 3, SSL vs. Other, $T(98) = -7.49$, $p < 0.001$). However, increasing the amount of training data in the pre-training stage does not significantly affect the spatial feature alignment (Fig. 3, Big Data vs. Other, $T(40) = -2.54$, $p = 0.015$, not significant at the $p < 0.001$ threshold).

DNNs trade-off between *Perturbation Tolerance* and *Adversarial Human Alignment*. After plotting the *Perturbation Tolerance* of each DNN in our zoo against its *Adversarial Human 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. There is a pareto-front governing the trade-off between these two adversarial metrics in our study. The existence of this pareto-front indicates a fundamental constraint with the development of DNNs: improving robustness or alignment comes at the cost of the other.

[†]lcnet_050.ra2_in1k

Human-aligned models and robustified models show surprising adversarial behavior. We reasoned that a potential approach for breaking the robustness-alignment trade-off is to turn to models that are explicitly trained to be aligned with human perception. One successful strategy is *harmonization*, which can improve the representational alignment of DNNs with human perception while also maintaining or slightly improving model accuracy on ImageNet (Fel et al., 2022). Indeed, we found that all *harmonized* DNNs show significantly higher *Perturbation Tolerance* and *Adversarial Human Alignment* than their baselines (Appendix B).

Another strategy that has been used to align models with human perception is adversarial training Goodfellow et al. (2014). These models are also significantly more tolerant than other standard models (Fig. 4, Robust vs. Other, $T(41) = 6.41$, $p < 0.001$), which can be explained by the fact that they are trained on adversarial examples to withstand a certain level of perturbation. Besides, we observed that they achieve significantly higher *Adversarial Human Alignment* than other models (Fig. 4, Robust vs. Other, $T(106) = 20.85$, $p < 0.001$), which is consistent with the findings in recent work Gaziv et al. (2023); Bartoldson et al. (2024). However, unlike *harmonized* models, the adversarially robust models trade-off accuracy for their improvements in *Tolerance* and *Adversarial Human Alignment* (Robust vs. Other, $T(45) = -8$, $p < 0.001$).

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 (Krizhevsky et al., 2012) 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 (Kaplan et al., 2020; Dehghani et al., 2023). 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, adversarial attacks on accurate models have begun to consistently affect parts of object images that humans find less important or completely irrelevant for recognition. Scaling up DNNs may improve robustness, but it does not necessarily enhance alignment with human perception, which means there is a fundamental misalignment of the training routines used for large-scale DNNs today. It is important for the field of vision research to explore new approaches to alignment to ensure that adversarial attacks target features humans rely on for perception and action. As this issue has broad implications for interpretability, vision modeling, and the use of DNNs as a computational tool to study human vision.

Human-aligned models and robustified models as partial solutions to break the trade-off. Robustified DNNs achieve the best of both worlds of adversarial vulnerability. This is probably because adversarial training enforces models to ignore perturbations on spurious features while focusing on semantically meaningful features that more closely align with human object recognition. The observation can be further supported by recent findings in the literature (Etmann et al., 2019; Geirhos et al., 2021; Gaziv et al., 2023). However, their performance on ImageNet is considerably compromised. Although harmonized DNNs demonstrate more effective adversarial behavior than their baselines, they still fall behind models through adversarial training with larger perturbation strength. 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. Another promising approach is to train visual models with auxiliary losses that integrate both harmonization and adversarial training, potentially allowing models to enhance robustness and alignment while maintaining the performance.

Limitations. We relied on a ℓ_2-norm attack for our experiments because it is fast, easy-to-optimize, and widely used in the adversarial robustness community (Akhtar & Mian, 2018). Whether our results translate to other ℓ_p-norm attacks remains for future research. While we have explored the alignment between humans and machines in adversarial attacks, we have not conducted psychophysical experiments to see if adversarial features can transfer across different entities (e.g., humans and models). Our approach, instead, focuses on the behavioral patterns under spatial constraints, in-

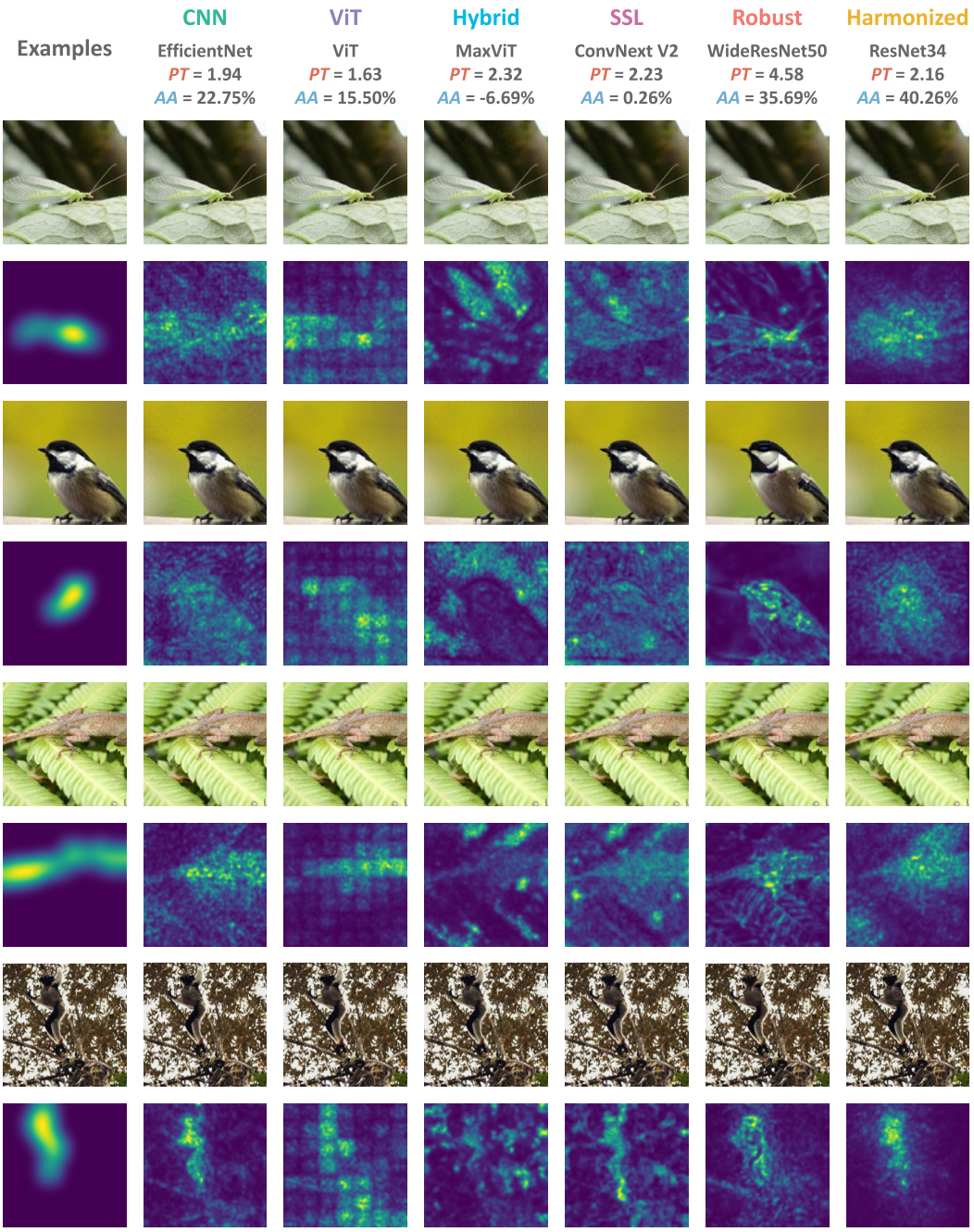

Figure 5: ℓ_2 **FMN adversarial attacks for DNNs.** Plotted here are ImageNet images and adversarial attacks for a variety of DNNs. The images and corresponding *ClickMe* maps are presented at the leftmost column (snow leopard, bagel, American chameleon, cradle. zoom in to see attack details). Perturbations are best viewed by zooming in.

vestigating whether models and humans share similar perceptual biases in terms of which image regions are most susceptible to perturbations. Prior work (Elsayed et al., 2018; Veerabadran et al., 2023) suggests that human perception can still be influenced by subtle changes in images, although humans exhibit greater robustness to such perturbations compared to neural networks. We leave further exploration of these aspects for future work.

Broader impacts. Adversarial attacks reveal the fundamental limitations in modeling biological vision with DNNs. Our findings highlight that the scaling trends that are driving progress in computer vision today exhibit a misalignment in adversarial behavior between humans and machines. To address this, new approaches for inducing representational alignment between DNNs and humans are needed to close the gap, ultimately improving both interpretability and real-world reliability of data-driven vision modeling.

ACKNOWLEDGMENTS

Our work is supported by ONR (N00014-24-1-2026), NSF (IIS-2402875), and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004). Additional support was provided by the Carney Center for Computational 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

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. ISSN 2169-3536. doi: 10.1109/access.2018.2807385. URL http://dx.doi.org/10.1109/ACCESS.2018.2807385.

Brian R. Bartoldson, James Diffenderfer, Konstantinos Parasyris, and Bhavya Kailkhura. Adversarial robustness limits via scaling-law and human-alignment studies. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. December 2017.

Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, Benjamin D Evans, Jeffrey Mitchell, and Ryan Blything. Deep problems with neural network models of human vision. *Behav. Brain Sci.*, pp. 1–74, December 2022.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Journal of the ACM*, 70(2):1–18, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. URL https://api.semanticscholar.org/CorpusID:233444273.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pp. 854–863. PMLR, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. February 2023.

Marin Dujmović, Gaurav Malhotra, and Jeffrey S Bowers. What do adversarial images tell us about human vision? *Elife*, 9, September 2020.

Mohammad K Ebrahimpour, J Ben Falandays, Samuel Spevack, and David C Noelle. Do humans look where deep convolutional neural networks "attend"? In *Advances in Visual Computing*, pp. 53–65. Springer International Publishing, 2019.

Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and Time-Limited humans. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

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

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1823–1832. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/etmann19a.html`.

Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Adv. Neural Inf. Process. Syst.*, 2022.

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

Guy Gaziv, Michael J. Lee, and James J. DiCarlo. Strong and precise modulation of human percepts via robustified anns. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. June 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Chong Guo, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James Dicarlo. Adversarially trained neural representations are already as robust as biological neural representations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8072–8081. PMLR, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 2016.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. *Adversarial examples are not bugs, they are features*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1072–1080, June 2015.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. January 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Manoj Kumar, Neil Houlsby, Nal Kalchbrenner, and Ekin Dogus Cubuk. Do better ImageNet classifiers assess perceptual similarity better? September 2022.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pp. 195–231. Springer, 2018.

Qiuxia Lai, Salman Khan, Yongwei Nie, Jianbing Shen, Hanqiu Sun, and Ling Shao. Understanding more about human and machine attention in deep neural networks. June 2019.

Thomas Langlois, Haicheng Zhao, Erin Grant, Ishita Dasgupta, Tom Griffiths, and Nori Jacoby. Passive attention in artificial neural networks predicts human visual selectivity. In M Ranzato, A Beygelzimer, Y Dauphin, P S Liang, and J Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27094–27106. Curran Associates, Inc., 2021.

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

Drew Linsley, Dan Shiebler, S Eberhardt, and Thomas Serre. Learning what and where to attend. *International Conference on Learning Representations (ICLR)*, 2019a.

Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend with humans in the loop. In *International Conference on Learning Representations*, 2019b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Gaurav Malhotra, Benjamin D Evans, and Jeffrey S Bowers. Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Res.*, 174:57–68, September 2020.

Gaurav Malhotra, Marin Dujmović, and Jeffrey S Bowers. Feature blindness: A challenge for understanding and modelling visual object recognition. *PLoS Comput. Biol.*, 18(5):e1009572, May 2022.

George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, November 1995. ISSN 1557-7317. doi: 10.1145/219717.219748. URL `http://dx.doi.org/10.1145/219717.219748`.

L Muttenthaler, J Dippel, L Linhardt, and others. Human alignment of neural network representations. *arXiv preprint arXiv*, 2022.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

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

Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints. *Advances in Neural Information Processing Systems*, 34, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Brett D Roads and Bradley C Love. Enriching ImageNet with human similarity judgments and psychological embeddings. November 2020.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. *Image synthesis with a single (robust) classifier*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. {BREEDS}: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=mQPBmvyAuk.

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

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on ImageNet. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8634–8644. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/shankar20c.html.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6976–6987, 2019.

Ilia Sucholutsky and Thomas L Griffiths. Alignment with human representations supports robust few-shot learning. January 2023.

Christian Szegedy, Google Inc, Wojciech Zaremba, Ilya Sutskever, Google Inc, Joan Bruna, Dumitru Erhan, Google Inc, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *In ICLR*, 2014.

Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proc. Natl. Acad. Sci. U. S. A.*, 113(10):2744–2749, March 2016.

Vijay Vasudevan, Benjamin Caine, Raphael Gontijo-Lopes, Sara Fridovich-Keil, and Rebecca Roelofs. When does dough become a bagel? analyzing the remaining mistakes on imagenet. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Vijay Veerabadran, Josh Goldman, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, Jonathon Shlens, Jascha Sohl-Dickstein, Michael C. Mozer, and Gamaleldin F. Elsayed. Subtle adversarial image manipulations influence both human and machine perception. *Nature Communications*, 14(1), August 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-40499-0. URL http://dx.doi.org/10.1038/s41467-023-40499-0.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith (eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

A APPENDIX

A.1 EXPERIMENTAL STIMULI

We used the *robustness* package (Engstrom et al., 2019) to build a customized ImageNet validation dataset, grouping semantically similar classes into 12 representative superclasses based on the WordNet hierarchy (Miller, 1995). Since ImageNet categories are imbalanced (e.g., more dog than bird classes), this grouping ensured a more balanced distribution. We then evenly sampled 960 images to prevent specific subcategories (e.g., certain dog breeds) from biasing evaluation. Customizing ImageNet dataset is widely used in the adversarial attack context (Ilyas et al., 2019; Santurkar et al., 2019; 2021; Gaziv et al., 2023). It provides a balanced trade-off, preserving the diversity and complexity of natural images while significantly reducing the label space. Also, using a constrained subset helps ensure more reliable and interpretable adversarial analysis by mitigating label ambiguities that arise in fine-grained classification.

Images were preprocessed and normalized using the standard TIMM routine. Additionally, to adapt ImageNet models for use with our customized dataset, we introduced a class-mapping layer that aggregates predictions from fine-grained ImageNet classes into animate/inanimate classes (Ilyas et al., 2019; Santurkar et al., 2019; 2021; Gaziv et al., 2023). This is achieved by constructing a mapping between the new labels and the corresponding ImageNet labels, which ensures that the model outputs are compatible with the customized dataset while maintaining the pre-trained ImageNet model's capabilities.

A.2 DNN MODEL ZOO

We comprehensively evaluated the adversarial robustness of DNNs on a large sample of models from the TIMM toolbox (Wightman, 2019). 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 viewed at `https://anonymous.4open.science/r/Adversarial_Alignment-CF28`.

A.3 NEURAL HARMONIZERS

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 $\boldsymbol{p}$ 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 9 DNNs for object recognition on ImageNet while relying on category-diagnostic features captured by *ClickMe* (Fel et al., 2022): `resnet18`, `resnet34`, `resnet50`, `resnetv2_50`, `resnet101`, `resnet152`, `vit_tiny_patch16_224`, `convnext_tiny`, `mobilenetv3_small_050`. Each was trained with different settings of λ_1 and λ_2, which controlled the relative strength of losses for object recognition and alignment, respectively.

Neural Harmonizer	Perturbation Tolerance (PT)	Adversarial Human Alignment (AA)
resnetv2_50	2.9369 ± 0.0785	$36.62\% \pm 1.44\%$
vit_tiny_patch16_224	2.4101 ± 0.0563	$30.51\% \pm 1.40\%$
convnext_tiny	2.5164 ± 0.0605	$25.97\% \pm 1.38\%$
mobilenetv3_small_050	1.5694 ± 0.0328	$26.14\% \pm 1.13\%$
resnet18	1.6884 ± 0.0156	$40.11\% \pm 1.56\%$
resnet34	2.1633 ± 0.0430	$40.26\% \pm 1.56\%$
resnet50	1.6827 ± 0.0313	$39.22\% \pm 1.48\%$
resnet101	1.7925 ± 0.0413	$33.35\% \pm 1.49\%$
resnet152	1.9059 ± 0.0365	$37.29\% \pm 1.40\%$

Table 1: Perturbation Tolerance (PT) and Adversarial Human Alignment (AA) for various harmonized models. The values represent the mean PT and AA for each model, with standard error of the mean (SEM) shown as ±.

A.4 Adversarially-trained Models

Adversarial training is a robust optimization technique aimed at improving the resilience of deep neural networks (DNNs) against adversarial attacks. These attacks introduce small, often imperceptible perturbations to input images, causing significant misclassification. Given a neural network parameterized by θ, trained on a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, standard training optimizes the empirical risk:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(x), y)], \tag{3}$$

where $\mathcal{L}$ is the loss function, typically cross-entropy for classification tasks.

An adversarial example x' is generated by applying a perturbation δ to the input x such that the classifier is fooled while maintaining perceptual similarity:

$$x' = x + \delta, \quad \text{s.t.} \quad \|\delta\|_p \leq \epsilon, \tag{4}$$

where $\|\cdot\|_p$ denotes the p-norm constraint, and ϵ is a small perturbation budget.

Adversarial training improves model robustness by explicitly incorporating adversarial examples into the training process. Instead of minimizing the loss on clean samples, the model is trained to minimize the worst-case loss within a perturbation budget:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(f_\theta(x+\delta), y) \right]. \tag{5}$$

This min-max formulation seeks to find adversarial perturbations δ that maximize the loss while simultaneously updating θ to minimize the worst-case loss.

Below are results from a list of adversarially-trained models:

Model	**Perturbation Tolerance (PT)**	**Adversarial Human Alignment (AA)**
resnet18_l2_eps0.01	1.4511 ± 0.026775	$26.12\% \pm 1.26\%$
resnet18_l2_eps0.03	1.7228 ± 0.035167	$27.25\% \pm 1.26\%$
resnet18_l2_eps0.05	1.8900 ± 0.041815	$27.13\% \pm 1.30\%$
resnet18_l2_eps0.1	2.2793 ± 0.054803	$28.78\% \pm 1.29\%$
resnet18_l2_eps0.25	3.0940 ± 0.085464	$29.41\% \pm 1.27\%$
resnet18_l2_eps0.5	4.1068 ± 0.126492	$30.48\% \pm 1.32\%$
resnet50_l2_eps0.01	1.7186 ± 0.031991	$23.63\% \pm 1.23\%$
resnet50_l2_eps0.03	1.9581 ± 0.040712	$25.20\% \pm 1.15\%$
resnet50_l2_eps0.05	2.2171 ± 0.052840	$27.07\% \pm 1.21\%$
resnet50_l2_eps0.1	2.6884 ± 0.065770	$26.70\% \pm 1.29\%$
resnet50_l2_eps0.25	3.5240 ± 0.098367	$28.06\% \pm 1.27\%$
resnet50_l2_eps0.5	4.6717 ± 0.139479	$28.87\% \pm 1.28\%$
wide_resnet50_2_l2_eps0.01	1.8233 ± 0.034353	$23.27\% \pm 1.21\%$
wide_resnet50_2_l2_eps0.03	2.0990 ± 0.044210	$24.25\% \pm 1.18\%$
wide_resnet50_2_l2_eps0.05	2.3147 ± 0.050032	$26.29\% \pm 1.11\%$
wide_resnet50_2_l2_eps0.1	2.8313 ± 0.067706	$27.27\% \pm 1.15\%$
wide_resnet50_2_l2_eps0.25	3.8358 ± 0.100168	$27.33\% \pm 1.15\%$
wide_resnet50_2_l2_eps0.5	4.8839 ± 0.132979	$27.82\% \pm 1.24\%$
resnet18_linf_eps0.5_255	3.5849 ± 0.106339	$35.41\% \pm 1.50\%$
resnet18_linf_eps1.0_255	4.5922 ± 0.151595	$35.08\% \pm 1.50\%$
resnet18_linf_eps2.0_255	4.3452 ± 0.148342	$23.48\% \pm 1.11\%$
resnet18_linf_eps4.0_255	3.5693 ± 0.126666	$18.49\% \pm 0.85\%$
resnet18_linf_eps8.0_255	3.6703 ± 0.146257	$18.86\% \pm 0.89\%$
resnet50_linf_eps0.5_255	4.0640 ± 0.122031	$35.00\% \pm 1.39\%$
resnet50_linf_eps1.0_255	4.2507 ± 0.132856	$33.98\% \pm 1.40\%$
resnet50_linf_eps2.0_255	4.0732 ± 0.140176	$21.49\% \pm 1.01\%$
resnet50_linf_eps4.0_255	4.0304 ± 0.143759	$19.13\% \pm 0.92\%$
resnet50_linf_eps8.0_255	4.5225 ± 0.184825	$15.29\% \pm 0.79\%$
wide_resnet50_2_linf_eps0.5_255	4.2642 ± 0.122851	$35.65\% \pm 1.42\%$
wide_resnet50_2_linf_eps1.0_255	4.5811 ± 0.142719	$35.69\% \pm 1.43\%$
wide_resnet50_2_linf_eps2.0_255	4.7295 ± 0.163614	$34.46\% \pm 1.39\%$
wide_resnet50_2_linf_eps4.0_255	4.7939 ± 0.169146	$26.01\% \pm 1.20\%$
wide_resnet50_2_linf_eps8.0_255	4.2206 ± 0.149968	$20.46\% \pm 0.86\%$

Table 2: Perturbation Tolerance (PT) and Adversarial Human Alignment (AA) for Robustified Models.

B APPENDIX

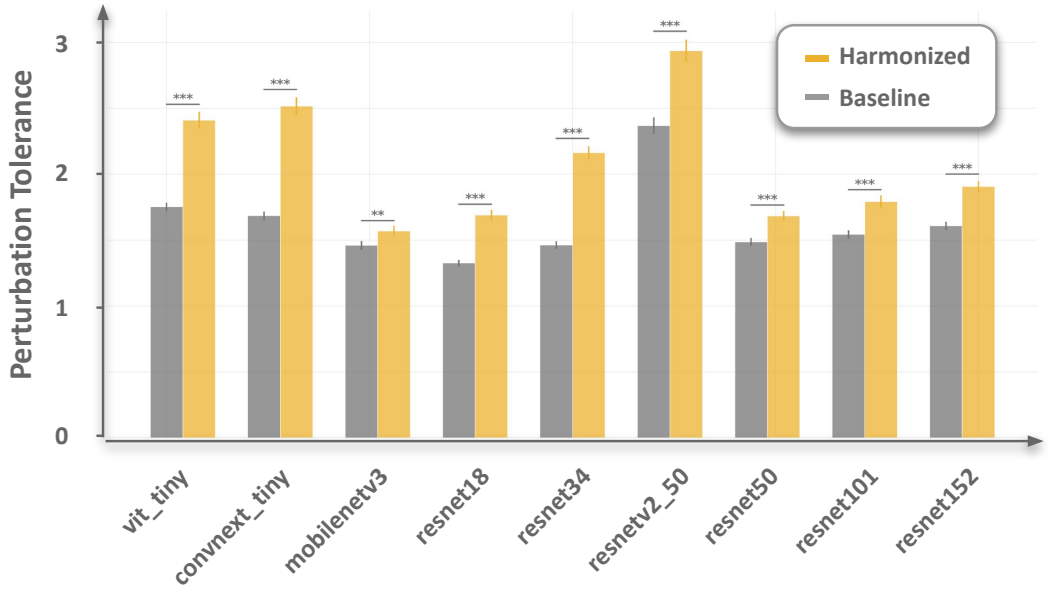

Figure 6: *Perturbation Tolerance* of neural harmonizers and their baseline models. Among them, one harmonized model `mobilenetv3_small_050.lamb_in1k_harmonized` shows a *Perturbation Tolerance* increase that does not reach significance at the strict $p < 0.001$ threshold but is still statistically significant at $p < 0.01$.

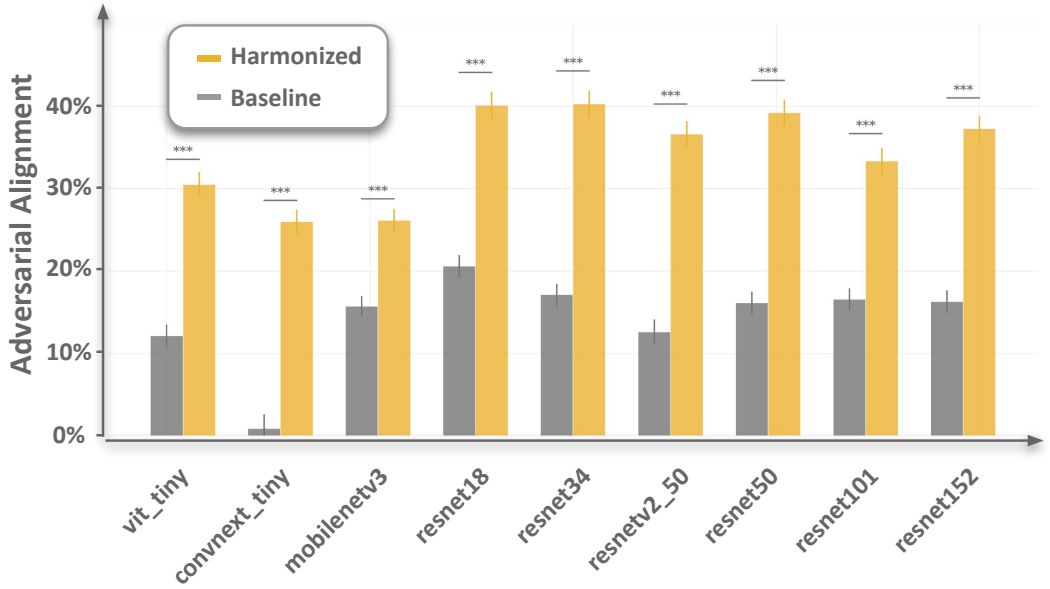

Figure 7: *Adversarial Human Alignment* of harmonized models and their baseline models. All harmonized models perform significantly better than their baselines.

C APPENDIX

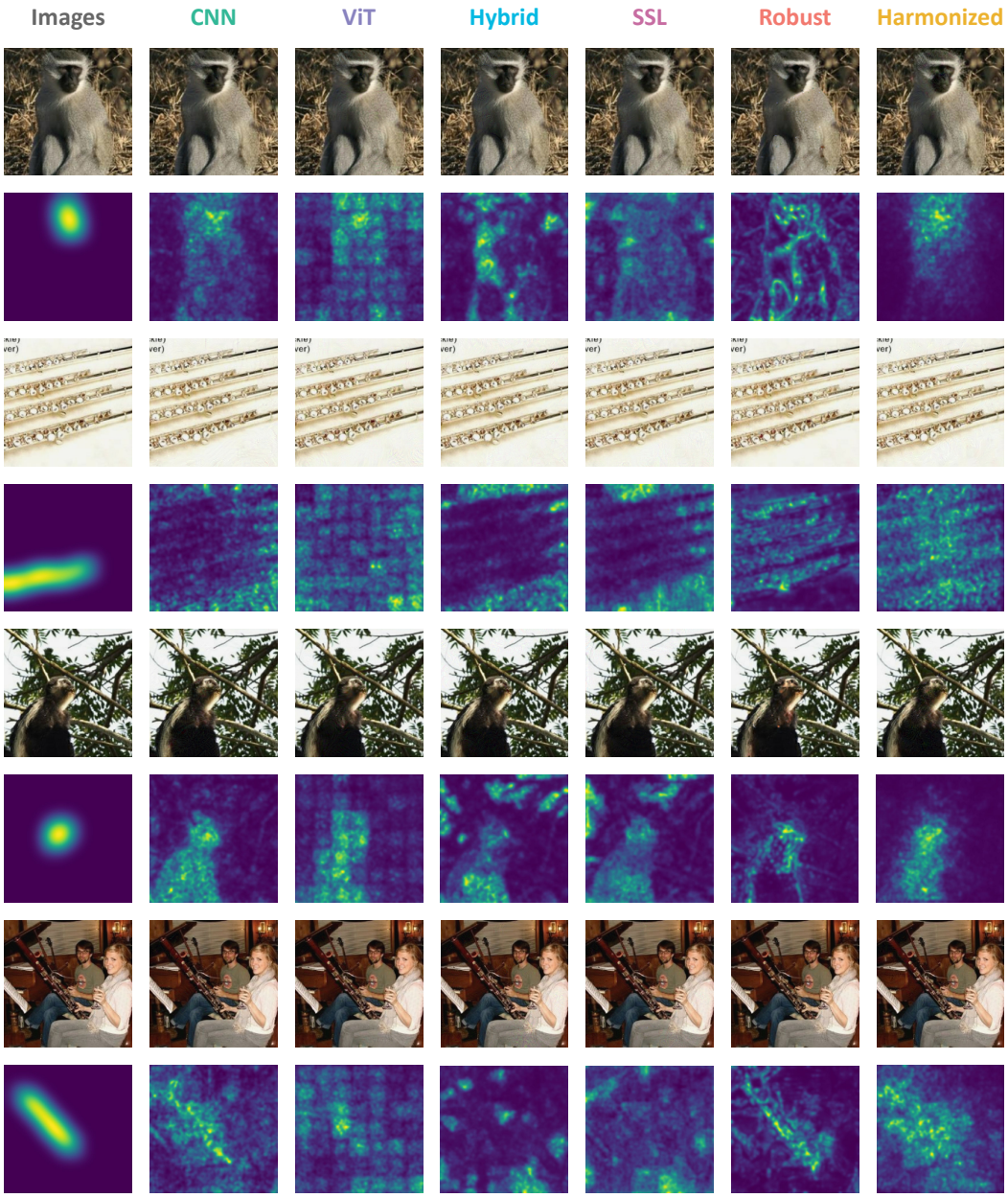

Figure 8: Visualization of stimuli overlaid by visual feature importance maps and adversarial perturbations generated from *efficientnet_b0*, *vit_tiny*, *maxvit_tiny*, *convnextv2_base*, *wide_resnet50_2_linf_eps1.0*, *resnet34_harmonized* (from left to right).

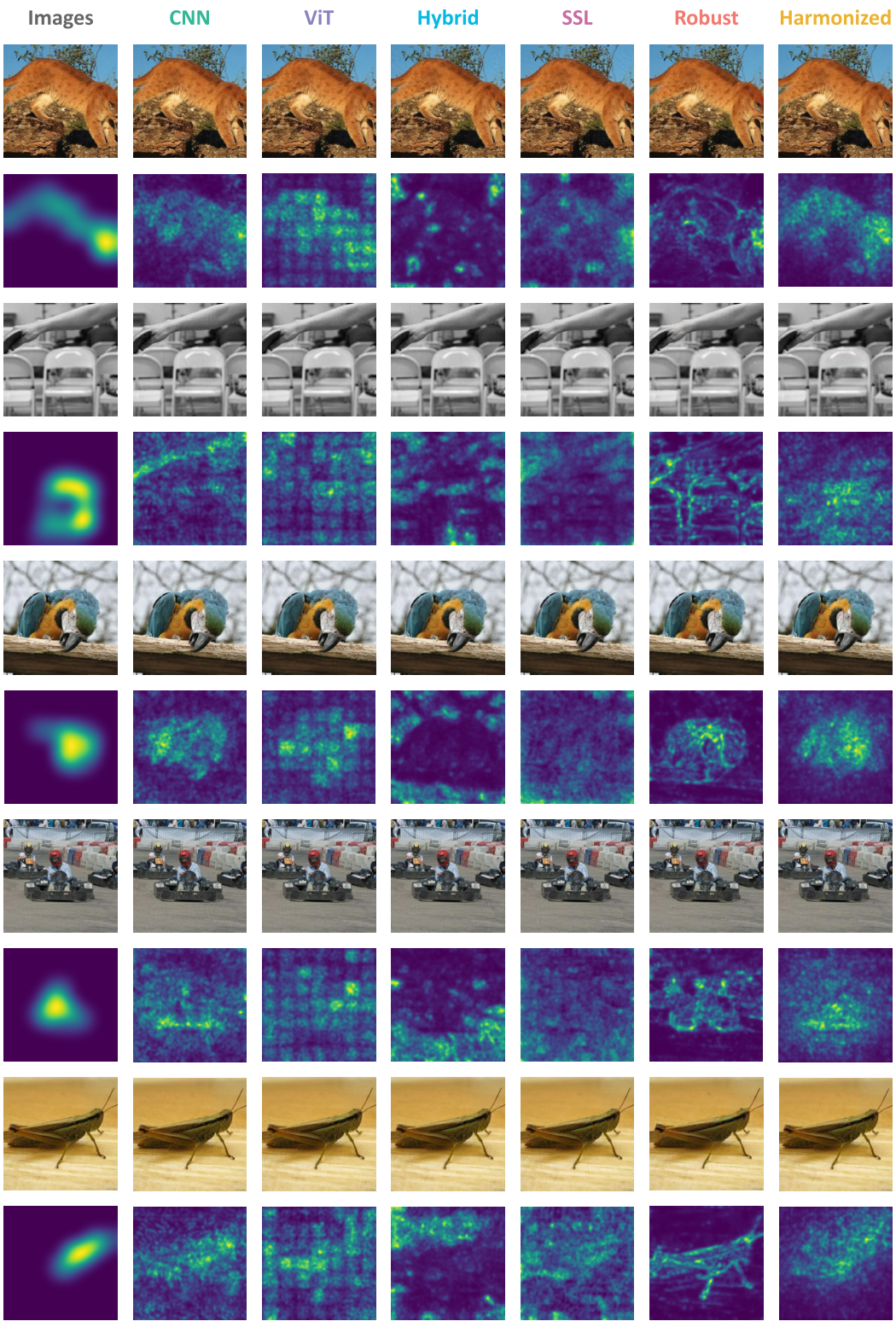

Figure 9: Visualization of stimuli overlaid by visual feature importance maps and adversarial perturbations generated from *efficientnet_b0*, *vit_tiny*, *maxvit_tiny*, *convnextv2_base*, *wide_resnet50_2_linf_eps1.0*, *resnet34_harmonized* (from left to right).