Harmonizing the object recognition strategies of deep neural networks with humans

Thomas Fel[*1,2], **Ivan Felipe**[*1], **Drew Linsley**[*1,3], **Thomas Serre**[1,2,3]
{thomas_fel,ivan_felipe_rodriguez,drew_linsley}@brown.edu

Abstract

The many successes of deep neural networks (DNNs) over the past decade have largely been driven by computational scale rather than insights from biological intelligence. Here, we explore if these trends have also carried concomitant improvements in explaining the visual strategies humans rely on for object recognition. We do this by comparing two related but distinct properties of visual strategies in humans and DNNs: *where* they believe important visual features are in images and *how* they use those features to categorize objects. Across 84 different DNNs trained on ImageNet and three independent datasets measuring the *where* and the *how* of human visual strategies for object recognition on those images, we find a systematic trade-off between DNN categorization accuracy and alignment with human visual strategies for object recognition. *State-of-the-art DNNs are progressively becoming less aligned with humans as their accuracy improves*. We rectify this growing issue with our neural harmonizer: a general-purpose training routine that both aligns DNN and human visual strategies and improves categorization accuracy. Our work represents the first demonstration that the scaling laws [1–3] that are guiding the design of DNNs today have also produced worse models of human vision. We release our code and data at `https://serre-lab.github.io/Harmonization` to help the field build more human-like DNNs.

1 Introduction

Rich Sutton stated [4] that the bitter lesson "from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin." Deep learning has been the standard approach to object categorization problems ever since the paradigm shifting success of AlexNet [5] on the ImageNet [6] benchmark a decade ago. As deep neural network (DNN) performance has continued to improve in the intervening years, Sutton's lesson has become more fitting than ever, with recent networks rivaling and likely outperforming humans on the benchmark [7] through brute-force computational scale: increasing the number of network parameters and number of images used for training orders-of-magnitude beyond AlexNet [1–3]. While the successes of so-called "scaling laws" are undeniable, this singular focus on performance in the field has side-stepped an equally important question that will govern the utility of object recognition models for the brain sciences and industry applications alike: *are the visual strategies learned by DNNs aligned with those used by humans?*

The visual strategies that mediate object recognition in humans can be decomposed into two related but distinct processes: identifying *where* the important features for object recognition are in a scene, and determining *how* to integrate the selected features into a categorical decision [8, 9]. It has been

[*]These authors contributed equally.
[1]Department of Cognitive, Linguistic, & Psychological Sciences, Brown University, Providence, RI
[2]Artificial and Natural Intelligence Toulouse Institute (ANITI), Toulouse, France
[3]Carney Institute for Brain Science, Brown University, Providence, RI

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

known for nearly a century [10–13] that different humans attend to similar locations when asked to find and recognize objects. After selecting these important features, human observers are also consistent in how they use those features to categorize objects – the inclusion of a few pixels in an image can be the difference between recognizing an object or not [9, 14].

Has the past decade of DNN development produced any models that are aligned with these human visual strategies for object recognition? Such a model could transform cognitive science by supporting a better mechanistic understanding of how vision works. More human-like models of object recognition would also resolve the problems with predictablity and interpretablity of DNNs [15–18], and control their alarming tendency to rely on "shortcuts" and dataset biases to perform well on tasks [19]. In this work, we perform the first large-scale and systematic comparison of the visual strategies of DNNs and humans for object recognition on ImageNet.

Contributions. In order to compare human and DNN visual strategies, we first turn to the human feature importance maps collected by Linsley et al. [20,21]. Their datasets, *ClickMe* and *Clicktionary*, contain maps of nearly 200,000 unique images in ImageNet that highlight the visual features humans believe are important for recognizing them. These datasets amount to a reverse inference on *where* important visual features are in ImageNet images (Fig. 1). We complement these datasets with new psychophysics experiments that directly test how important visual features are used for object recognition (Fig. 1). **As DNN performance has increased on ImageNet, their alignment with human visual strategies captured in these datasets has worsened.** This trade-off is found over 84 different DNNs representing all popular model classes – from those trained for adversarial robustness to those pushing the scaling laws in network capacity and training data. To summarize our findings:

- The trade-off between DNN object recognition accuracy and alignment with human visual strategies replicates across three unique datasets: *ClickMe* [20], *Clicktionary* [21], and our psychophysics experiments.
- We shift this trade-off with our neural harmonizer, a novel drop-in module for co-training any DNN to align with human visual strategies while also achieving high task accuracy. Harmonized DNNs learn visual strategies that are significantly more aligned with humans *than any other DNN we tested*.
- We release our data and code at `https://serre-lab.github.io/Harmonization/` to help the field tackle the growing misalignment between DNNs and humans.

2 Related work

Do DNNs explain human visual perception? Despite the continued success of DNNs on computer vision benchmarks, there are conflicting accounts on their ability to explain human vision. On the one hand, there is evidence that DNNs are improving as models of human visual perception on challenging tasks, such as recognizing objects obscured by noise [23]. On the other hand, there is also evidence that DNNs struggle to explain perceptual phenomena in human vision like contextual illusions [24], perceptual grouping [19,25,26], and categorical prototypes [27]. Others have found differences between human attention data and DNN models of visual attention [20,28]. Moreover, DNNs have stopped improving as models of the ventral visual system in humans and primates over recent years. While the original theory was that model explanations of object-evoked neural activity patterns improved alongside model categorization accuracy [29], recent large-scale DNNs are worse at explaining neural data than older ones with lower ImageNet accuracy [30].

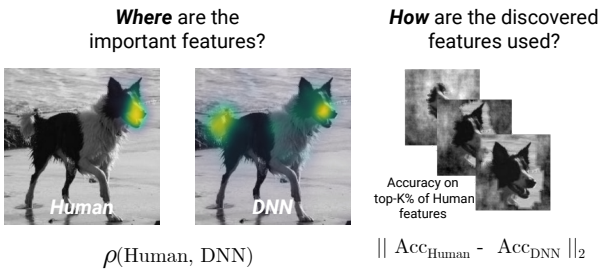

Figure 1: **Visual strategies of object recognition**. We investigate the alignment of human and DNN visual strategies in object categorization. We decompose human visual strategies into descriptions of *where* important features are [20,22], and *how* those features are integrated into visual decisions.

What are the visual strategies underlying human object recognition? Ever since its inception, a goal of vision science has been to characterize the neural processes supporting object recognition in humans. It has been discovered that object recognition can be decomposed into different processing stages that emerge over time [8, 31–36], where the earliest stage is associated with processing through feedforward connections in the visual system, and the later stage is associated with processing through feedback connections. Since the DNNs used today mostly rely on feedforward connections, it is likely that they are better models for that rapid feedforward phase of processing than the subsequent feedback phase [33, 37]. To maximize the likelihood that the visual strategies learned by DNNs align with those used by humans, our experiments focus on the visual strategies of rapid feedforward object recognition in humans.

Most closely related to our work, are studies of "top-down" image saliency and *where* category diagnostic visual features are in images. These studies typically involve asking participants to search for an object in an image, or find visual features that are diagnostic for an object's category or identity [10–13, 20, 22, 38]. In our work, we complement these descriptions of *where* important features are in images with psychophysics testing *how* those features are used to categorize objects.

Comparing visual strategies of humans and machines. As methods in explainable artificial intelligence have developed over the past decade, they have opened up opportunities for comparing the visual regions selected by humans and DNNs when solving tasks. Many of these comparisons have focused on human image saliency measurements captured by eye tracking or mouse clicks during passive or active viewing [20, 22, 39–42]. Others have compared categorical representation distances [40, 43] or combined those distances with measures of human attention [28]. The most direct comparisons between human and DNN visual strategies involved analyzing the minimal image patches needed to recognize objects [9, 44, 45]. However, these studies were limited and compared humans with older DNNs on tens of images. To the best of our knowledge, the largest-scale evaluation of human and DNN visual strategies relied on the *ClickMe* dataset to compare visual regions preferred by humans and attention models trained for object recognition [20]. What is noticeably missing from each of these studies is an large-scale analysis spanning many images and models of how human and DNN alignment has changed as a function of model performance.

Improving the correspondence between humans and machines. Inconsistencies between human and DNN representations can be resolved by directly training models to act more like humans. DNNs have been trained to have more human-like attention, or human-like representational distances in their output layers [20, 40, 43, 46, 47]. Here, we add to these successes with the neural harmonizer, a training routine that automatically aligns the visual strategies (Fig. 1) of any two observers by minimizing the dissimilarity of their decision explanations.

3 Methods

Human feature importance datasets. We focused on the ImageNet dataset to compare the visual strategies of humans and DNNs for object recognition at scale. We relied on the two significant efforts for gathering feature importance data from humans on ImageNet: the *Clicktionary* [22] and *ClickMe* [20] games, which use slightly different methods to collect their data. Both games begin with the same basic setup: two players work together to locate features in an object image that they believe are important for categorizing it. As one of the players selects important image regions, those regions are filled into a blank canvas for the other observer to see and categorize the image as quickly as possible. In *Clicktionary* [22], both players are humans, whereas in *ClickMe* [20], the player selecting features is a human and the player recognizing images is a DNN (VGG16 [48]). For both games, feature importance maps depicting the average object category diagnosticity of every pixel was computed as the probability of it being clicked by a participant. In total, *Clicktionary* [22] contained feature importance maps for 200 images from the ImageNet validation set, whereas *ClickMe* [20] contained feature importance maps for a non-overlapping set of 196,499 images from ImageNet training and validation sets. Thus, *ClickMe* has far more data than *Clicktionary*, but *Clicktionary* data has more reliable human feature importance data than *ClickMe*. Our experiments measure the alignment between human and DNN visual strategies using *ClickMe* and *Clicktionary* feature importance maps captured on the ImageNet validation set. As we describe in §4, *ClickMe* feature importance maps from the ImageNet training set are used to implement our neural harmonizer.

Psychophysics participants and dataset. We complemented the feature importance maps from *Clicktionary* and *ClickMe* with psychophysics experiments on rapid visual categorization. We recruited 199 participants from Amazon Mechanical Turk (`mturk.com`) to complete the experiments. Participants viewed a psychophysics dataset consisting of the 100 animal and 100 non-animal images in the Clicktionary game taken from the ImageNet validation set [22]. We used the feature importance maps for each image as masks for the object images, allowing us to control the proportion of important features observers were shown when asked to recognize objects (Fig. 5a). We generated versions of each image that reveal anywhere between 1% to 100% (at log-scale spaced intervals) of the important object pixels against a phase scrambled noise background (see Appendix §1 for details on mask generation). The total number of revealed pixels was equal for every image at a given level of image masking, and the revealed pixels were centered against the noise background. Each participant saw only one masked version of each object image.

Psychophysics experiment. Participants were instructed to categorize images in the psychophysics dataset as animals or non-animals as quickly and accurately as possible. Each experimental trial consisted of the following sequence of events overlaid onto a white background (SI Fig. 1): (*i*) a fixation cross displayed for a variable time (1,100–1,600ms); (*ii*) an image for 400ms; (*iii*) an additional 150ms of response time. In other words, the experiment forced participants to perform rapid object categorization. They were given a total of 550ms to view an image and press a button to indicate its category (feedback was provided on trials in which responses were not provided within this time limit). Images were sized at 256 x 256 pixel resolution, which is equivalent to a stimulus size approximately between 5 – 11 degrees of visual angle across a likely range of possible display and seating setups we expect participants used for the experiment. Similar paradigms and timing parameters have been shown to capture pre-attentive visual system processing [31, 49–51]. Participants provided informed consent electronically and were compensated $3.00 for their time ($\sim$ 10–15 min; approximately $15.00/hr).

Models. We compared humans with 84 different DNNs representing the variety of approaches used in the field today: 50 CNNs trained on ImageNet [1, 48, 52–63, 63–72], 6 CNNs trained on other datasets in addition to ImageNet (which we refer to as "CNN extra data") [1, 65, 73], 10 vision transformers [74–78], 6 CNNs trained with self-supervision [79, 80], and 13 models trained for robustness to noise or adversarial examples [81, 82]. We used pretrained weights for each of these models supplied by their authors, with a variety of licenses (detailed in SI §2), implemented in Tensorflow 2.0, Keras, or PyTorch.

4 Results

4.1 *Where* are diagnostic object features for humans and DNNs?

To systematically compare the visual strategies of object recognition for humans and DNNs on ImageNet, we first turned to the *ClickMe* dataset of feature importance maps [20]. In order to derive comparable feature importance maps for DNNs, we needed a method that could be efficiently and consistently applied to each of the 84 DNNs we tested without any idiosyncratic hyperparameters. This led us to choose a classic method for explainable artificial intelligence, image feature saliency [83]. We prepared human feature importance maps from *ClickMe* by taking the average importance map produced by humans for every image that also appeared in ImageNet validation. We then used Spearman's rank-correlation to measure the similarity between human feature maps and DNN feature maps for each image [49]. We also computed the inter-rater alignment of human feature importance maps as the mean split-half correlation across 1000 random splits of the participant pool ($\rho = 0.66$). We then normalized each human-DNN correlation by this score [20].

There were dramatic qualitative differences between the features selected by humans and DNNs on ImageNet. In general, humans selected less context and focused more on object parts: for animals, parts of their faces; for non-animals, parts that enable their usage, like the spade of a shovel (see Fig. 2 and SI Fig. 5. The DNN that was most aligned with humans, the DenseNet121, was still only 38% aligned with humans (Fig. 3).

Plotting the relationship between DNNs' top-1 accuracy on ImageNet with their human alignment revealed a striking trade-off: as the accuracy of DNNs has improved beyond DenseNet121, their

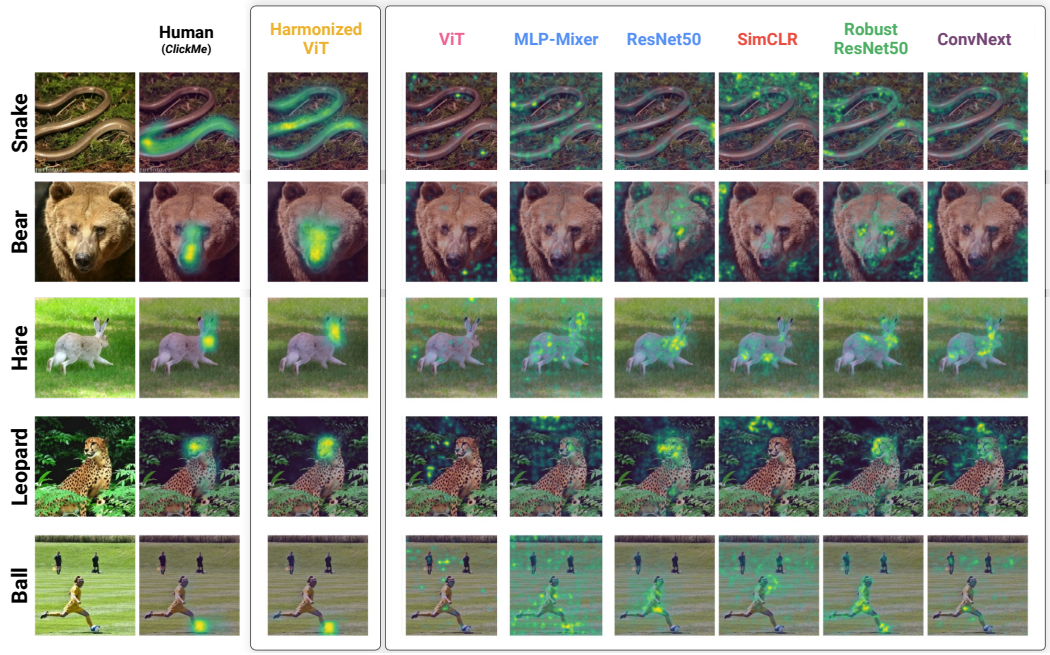

Figure 2: **Human and DNNs rely on different features to recognize objects.** In contrast, our neural harmonizer aligns DNN feature importance with humans. We smooth feature importance maps from humans (*ClickMe*) and DNNs with a Gaussian kernel for visualization.

alignment with humans has worsened (Fig. 3). For example, consider the ConvNext [1], which achieved the best top-1 accuracy in our experiments (85.8%), was only 22% aligned with humans – equivalent to the alignment of the BagNet33 [68] (63% top-1 accuracy). As an additional control, we computed the similarity between the average *ClickMe* map, which exhibits a center bias [84, 85] (SI Fig. 5), and each individual *ClickMe* map. This center-bias control was only outperformed by 42/84 CNNs we tested († in Fig. 3). Overall, we observe that human and DNN alignment has considerably worsened since the introduction of these two models.

The neural harmonizer. While scaling DNNs has immensely helped performance on popular benchmark tasks, there are still fundamental differences in the architectures of DNNs and the human visual system [37] which could part of the reason to blame for poor alignment. While introducing biological constraints into DNNs could help this problem, there is plenty of evidence that doing so would hurt benchmark performance and require bespoke development for every different architecture [86–88]. *Is it possible to align a DNN's visual strategies with humans without hurting its performance?*

Such a general-purpose method for aligning human and DNN visual strategies should satisfy the following criteria: (***i***) The method should work with any fully-differentiable network architecture. (***ii***) It should not present optimization issues that interfere with learning to solve a task, and the task-accuracy of a model trained with the method should not be worse than a model trained without the method. We created the neural harmonizer to satisfy these criteria.

Let us consider a supervised categorization problem with an input space, $\mathcal{X}$ an output space $\mathcal{Y} \subseteq \mathbb{R}^c$ and a predictor function $\boldsymbol{f}_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ, which maps an input vector $\boldsymbol{x} \in \mathcal{X}$ to an output $\boldsymbol{f}_\theta(\boldsymbol{x})$. We denote $\boldsymbol{g} : \mathcal{F} \times \mathcal{X} \to \mathcal{X}$ an explanation functional that, given a predictor $\boldsymbol{f}_\theta \in \mathcal{F}$ and an input, returns a feature importance map $\boldsymbol{\phi} = \boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x})$. Here, we focus on DNN saliency $\boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x}) \triangleq \nabla_{\boldsymbol{x}} \boldsymbol{f}_\theta(\boldsymbol{x})$ as our method for computing feature importance in DNNs, but the method can in principle work with any differentiable network explanation method.

To satisfy criterion (***i***), the neural harmonizer introduces a differentiable loss that will enforce alignment across feature importance map scales from any neural network. Let $\mathcal{P}_i(.)$ be the function mapping a feature importance map $\boldsymbol{\phi}$ to it is representation in the N levels of a Gaussian pyramid, with $i \in \{1, ..., N\}$. The function $\mathcal{P}_i(\boldsymbol{\phi})$ is computed by downsampling $\mathcal{P}_{i-1}(\boldsymbol{\phi})$ using a Gaussian

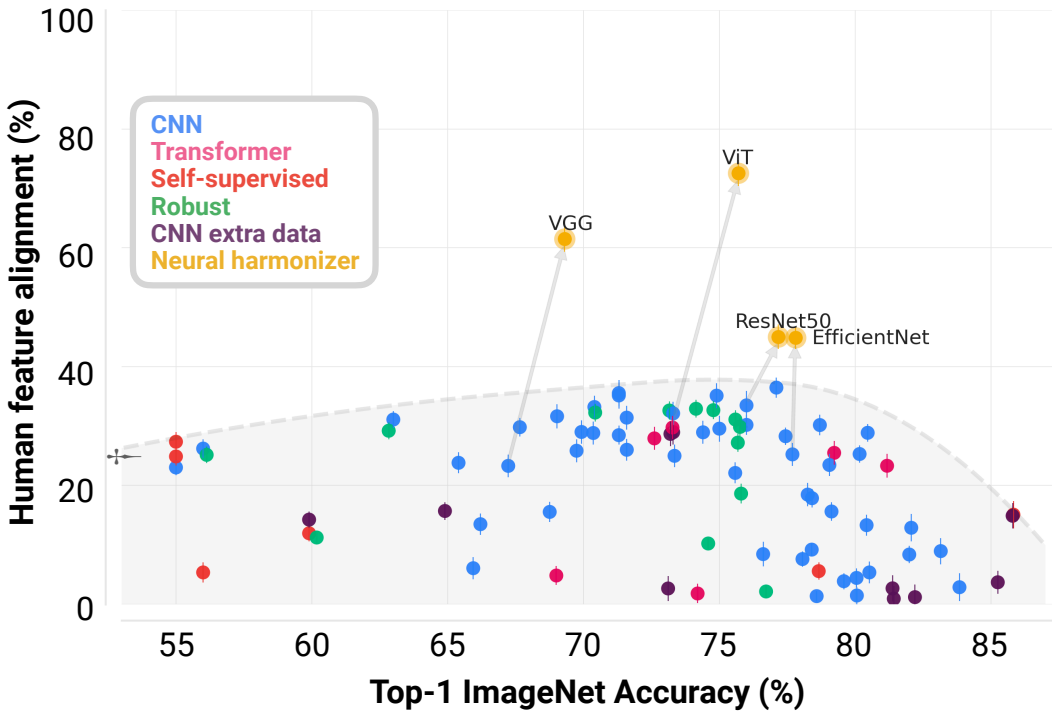

Figure 3: **The trade-off between DNN performance and alignment with human feature importance from *ClickMe* [20].** Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The shaded region denotes the pareto frontier of the trade-offs between ImageNet accuracy and human feature alignment for unharmonized models. Harmonized models (VGG16, ResNet50, ViT, and EfficientNetB0) are more accurate and aligned than versions of those models trained only for categorization. Error bars are bootstrapped standard deviations over feature alignment. Arrows show a shift in performance after training with the neural harmonizer. The feature alignment of an average of *ClickMe* maps with held-out maps is denoted by †.

kernel, with $\mathcal{P}_1(\phi) = \phi$. We then seek to minimize $\sum_i^N ||\mathcal{P}_i(\boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x})) - \mathcal{P}_i(\phi)||^2$, which will align feature importance maps between humans and DNNs at every scale of the pyramid.

To satisfy criterion (***ii***), the neural harmonizer should work well with training routines designed for large-scale computer vision challenges like ImageNet. This means that the neural harmonizer loss must avoid optimization issues at scale. To do this, we need a way of comparing feature importance maps between humans and DNNs that is invariant to the norm of either map. We therefore standardize feature importance maps from humans and DNNs before comparing them, and only measure alignment on the most important areas of the image for each observer. Formally, let $\boldsymbol{z}(.)$ be a standardization function over feature importance maps that takes the mean and standard deviation computed for each map ϕ such that $\boldsymbol{z}(\phi)$ has 0 activation on average and unit standard deviation. To focus alignment on important regions, let $\boldsymbol{z}(\phi)^+$ denote the positive part of the standardized explanation $\boldsymbol{z}(\phi)$. Finally, we include a task loss, the familiar cross entropy objective, to yield the complete neural harmonization loss and train models that are at least as accurate as those trained without harmonization:

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

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

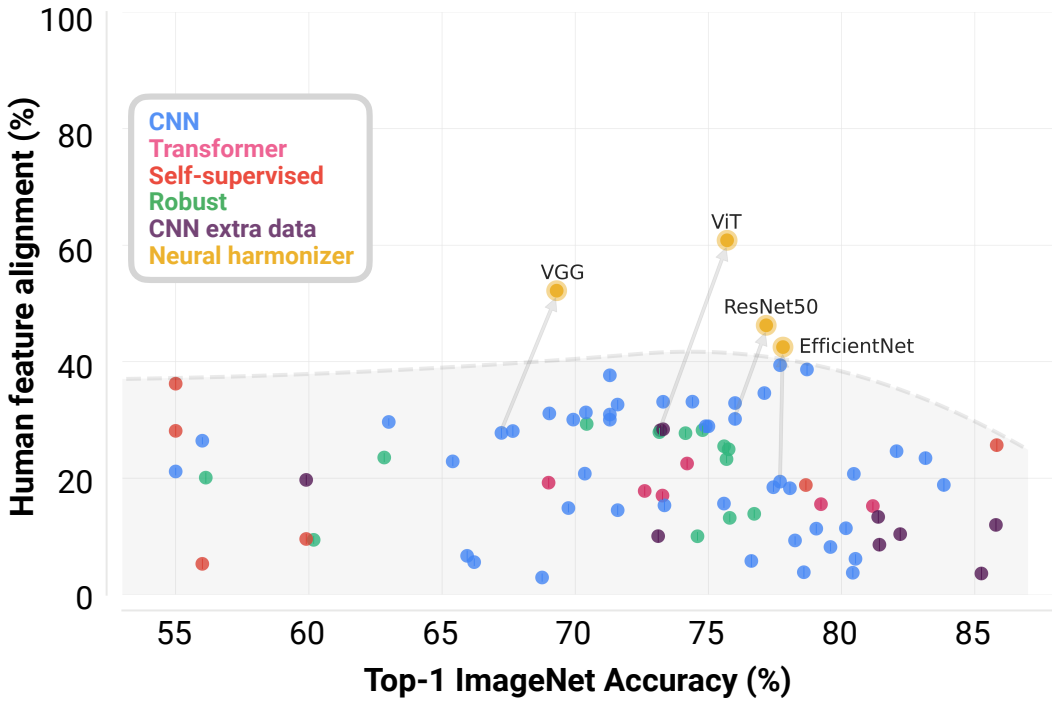

Figure 4: **The trade-off between DNN performance and alignment with human feature importance from *Clicktionary* [22].** Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The shaded region denotes the pareto frontier of the trade-offs between ImageNet accuracy and human feature alignment for unharmonized models. Harmonized models (VGG16, ResNet50, MobileNetV1, and EfficientNetB0) are more accurate and aligned than versions of those models trained only for categorization. Error bars are bootstrapped standard deviations over feature alignment. Arrows denote a shift in performance after training with the neural harmonizer.

Training. We trained four different DNNs with the neural harmonizer: VGG16, ViT, ResNet50, and EfficientNetB0. These models were selected because they are popular convolutional and transformer networks with open-source architectures that are straightforward to train and also sit near the boundary of the trade-off between DNN performance and alignment with humans. Models were trained using the neural harmonizer to optimize categorization performance on ImageNet and feature importance map alignment with human data from *ClickMe*. We trained models on all images in the ImageNet training set, but because *ClickMe* only contains human feature importance maps for a portion of those images, we computed the categorization loss but not the neural harmonizer loss for images without importance maps. Models were trained using 8 cores V4 TPUs on the Google Cloud Platform, and training lasted approximately one day. Models were trained with an augmented ResNet training recipe (built from `https://github.com/tensorflow/tpu/`). Models were optimized with SGD and momentum over batches of 512 images, a learning rate of 0.3, and label smoothing [89]. Images were augmented with random left-right flips and mixup [90]. The learning rate was adjusted over the course of training with a schedule that began with an initial warm-up period of 5 epochs and then decaying according to a cosine function over 90 epochs, with decay at step 30, 50 and 80. We validated that a ResNet50 and VGG16 trained with these hyperparameters and schedule using standard cross-entropy (but not the neural harmonizer) matched published performance.

The neural harmonizer aligns human and DNN visual strategies. We found that harmonized models broke the trade-off between ImageNet accuracy and model alignment with *ClickMe* human feature importance maps (Fig. 3). Harmonized models were significantly more aligned with feature importance maps and also performed better on ImageNet. The changes in *where* harmonized models find important features in images were dramatic: a harmonized ViT had feature importance maps that are far less reliant on context (Fig. 2) and approximately 150% more aligned with humans (Fig. 3;

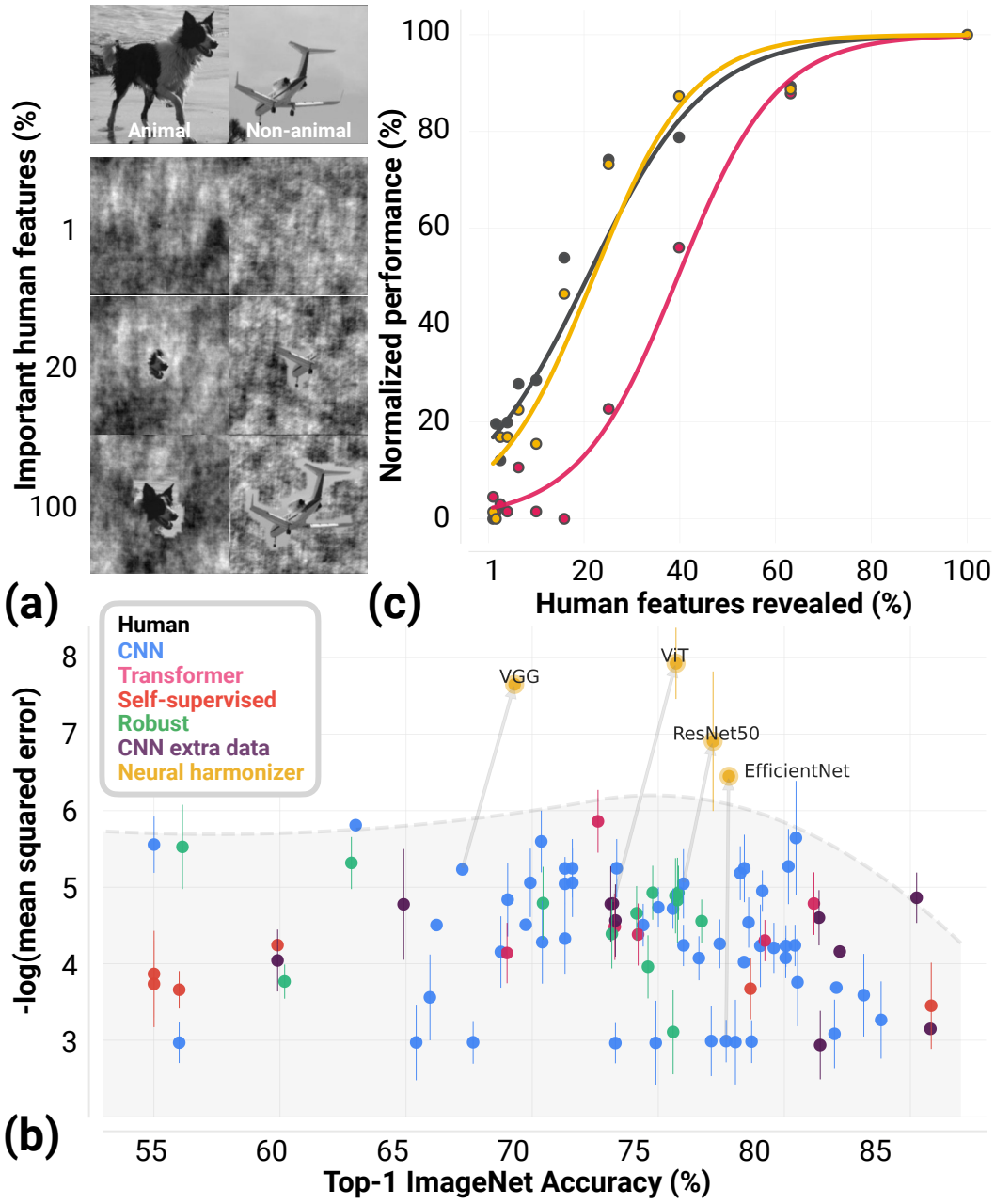

Figure 5: **Comparing *how* humans and DNNs use visual features during object recognition**. **(a)** Humans and DNNs categorized ImageNet validation images as animals or non-animals. The images revealed only a portion of the most important visual features according to the *Clicktionary* game [91]. **(b)** There was a trade-off between DNN top-1 accuracy on ImageNet and alignment with human visual decision making. The shaded region denotes the pareto frontier of the trade-off between ImageNet accuracy and human feature alignment for unharmonized models. Arrows denote a shift in performance after training with the neural harmonizer. Error bars are bootstrapped standard deviations over decision-making alignment. **(c)** A state-of-the-art DNN like the ViT learned a different strategy for integrating visual features into decisions than humans or a harmonized ViT.

ViT goes from 28.7% to 72.6% alignment after harmonization). The same model also performed 4% better in top-1 accuracy without any changes to its architecture. Similar improvements were found for

the harmonized VGG16 and ResNet50. While the EfficientNetB0 had only a minimal improvement in accuracy, it too exhibited a large boost in human feature alignment.

Clicktionary. To test if the trade-off between DNN ImageNet accuracy and alignment with humans is a general phenomenon we next turned to *Clicktionary* [22]. Indeed, we observed a similar trade-off on this dataset as we found for *ClickMe*: alignment with human feature importance from *Clicktionary* has worsened as DNN accuracy has improved on ImageNet (Fig. 4). As with *ClickMe*, harmonized DNNs shift the accuracy-alignment trade-off on this dataset.

4.2 *How* do humans and DNNs integrate diagnostic object features into decisions?

The trade-off we discovered between DNN accuracy on ImageNet and alignment with human visual feature importance suggests that the two use different visual strategies for object classification. However, there is potential for an even deeper problem. Even if two observers deem the same regions of an image as important for recognizing it, there is no guarantee that they use the selected features in the same way to render their decisions. We posit that if two observers have aligned visual strategies, the will agree on both *where* important features are in an image and *how* they use those features for decisions.

We developed a psychophysics experiment to measure how different humans use features in ImageNet images to recognize objects. Participants viewed versions of these images where only a proportion of the features that were deemed most important in the *Clicktionary* game were visible (Fig. 5a). Participants had to accurately detect whether or not the image contained an animal within 550ms, which forced them to rely on feedforward processing as much as possible [33]. Each of the 200 images we used were shown to a single participant only once. We accumulated responses from all participants to construct decision curves that showed how accurately the average human converted any given proportion of image features into an object decision. We performed the same experiment on DNNs as we did on humans, recording animal vs. non-animal decisions according to whether or not the most probable category in the model's 1000-category output was an animal. Because the experiment was speeded, humans did not achieve perfect accuracy. Thus, we normalized performance for humans and DNNs to compare the rate at which each integrated features into accurate decisions.

We discovered a similar trade-off between ImageNet accuracy and alignment with human visual decision making in this experiment as we did in *ClickMe* and *Clicktionary* (Fig. 5b). Indeed, the model that was most aligned with human decision-making – the BagNet33 [68] – only achieved 63.0% accuracy on ImageNet. Surprisingly, harmonized models broke this trend, particularly the harmonized ViT (Fig. 5b, top-right), despite no explicit constraints in that procedure which forced consistent decision-making with humans. In contrast, an unharmonized ViT integrates visual information into accurate decisions less efficiently than humans or harmonized models (Fig. 5c).

5 Conclusion

Models that reliably categorize objects like humans do would shift the paradigms of the cognitive sciences and artificial intelligence. But despite continuous progress over the past decade on the ImageNet benchmark, DNNs are becoming *worse* models of human vision. Our solution to this problem, the neural harmonizer, can be applied to any DNN to align their visual strategies with humans and even improve performance.

We observed the greatest benefit of harmonization on the visual transformer, the ViT. This finding is particularly surprising given that transformers eschew the locality bias of convolutional neural networks that has helped them become the new standard for modeling human vision and cognition [37]. Thus, we suspect that the neural harmonizer is especially well-suited for large-scale training on low-inductive bias models, like transformers. We also hypothesize that the improvements in human alignment provided by the neural harmonizer will yield a variety of downstream benefits for a model like the ViT, including better predictions of perceptual similarity, stimulus-evoked neural responses, and even performance on visual reasoning tasks. We leave these analyses for future work.

The field of computer vision today is following Sutton's prescient lesson: benchmark tasks can be scaling architectural capacity and the size of training data. However, as we have demonstrated here, these scaling laws are exchanging performance for alignment with human perception. We encourage

the field to re-analyze the costs and benefits of this exchange, particularly in light of the growing concerns about DNNs leveraging shortcuts and dataset biases to achieve high performance [19]. Alignment with human vision need not be exchanged with performance if DNNs are harmonized. Our codebase (`https://serre-lab.github.io/Harmonization/`) can be used to incorporate the neural harmonizer into any DNN created and measure its alignment with humans on the datasets we describe in this paper.

Limitations. One possible explanation for the misalignment between DNNs and humans that we observe is that recent DNNs have achieved superhuman accuracy on ImageNet. Superhuman DNNs have been described in biomedical applications [92, 93] where there is definitive biological ground-truth labels, but ImageNet labels are noisy, making it unclear if such an achievement is laudable (`https://labelerrors.com/`). Thus, an equally likely explanation is that the continued improvements of DNNs at least partially reflect their exploitation of shortcuts in ImageNet [19].

The scope of our work is also limited in that it focuses on object recognition in ImageNet. It is possible that models trained on other tasks, such as segmentation, may be more aligned with humans.

Finally, our modeling efforts were hamstrung for the largest-scale models in existence. Our work does not answer how much harmonization would help a model like CLIP because of the massive investment needed to train it. The neural harmonizer can be applied to CLIP but it is possible that more *ClickMe* human feature importance maps are needed for successful harmonization.

Broader impacts. A persistent issue in the field of artificial intelligence is the tendency of models to exploit dataset biases. A central theme of our work is that there are facets of human perception that are not captured by DNNs, particularly those which follow the scaling laws which have been so embraced by industry leaders. Forcing DNNs to rely on similar visual strategies as humans could represent a scalable path forward to correcting the insidious biases which have assailed under-constrained models of artificial intelligence.

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

[2] Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. (June 2021)

[3] 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)

[4] Sutton, R.: The bitter lesson. (May 2019)

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems. Volume 25., Curran Associates, Inc. (2012)

[6] 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

[7] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. (September 2014)

[8] DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition? Neuron **73**(3) (February 2012) 415–434

[9] 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

[10] Buswell, G.T.: How people look at pictures: a study of the psychology and perception in art. **198** (1935)

[11] Yarbus, A.L.: Eye Movements and Vision. Springer US

[12] Posner, M.I.: Orienting of attention. Q. J. Exp. Psychol. **32**(1) (February 1980) 3–25

[13] Mannan, S.K., Kennard, C., Husain, M.: The role of visual salience in directing eye movements in visual object agnosia. Curr. Biol. **19**(6) (March 2009) R247–8

[14] Gruber, L.Z., Ullman, S., Ahissar, E.: Oculo-retinal dynamics can explain the perception of minimal recognizable configurations. Proc. Natl. Acad. Sci. U. S. A. **118**(34) (August 2021)

[15] Fel, T., Colin, J., Cadene, R., Serre, T.: What I cannot predict, I do not understand: A Human-Centered evaluation framework for explainability methods. Advances in Neural Information Processing Systems (NeurIPS) (December 2021)

[16] Fel, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., Serre, T.: Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. Advances in Neural Information Processing Systems (NeurIPS) (November 2021)

[17] Fel, T., Ducoffe, M., Vigouroux, D., Cadene, R., Capelle, M., Nicodeme, C., Serre, T.: Don't lie to me! robust and efficient explainability with verified perturbation analysis. Workshop, Proceedings of the International Conference on Machine Learning (ICML) (February 2022)

[18] Fel, T., Vigouroux, D., Cadène, R., Serre, T.: How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (September 2020)

[19] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11) (November 2020) 665–673

[20] 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)

[21] Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017) 936–944

[22] 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

[23] 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)

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

[25] Kim*, J., Linsley*, D., Thakkar, K., Serre, T.: Disentangling neural mechanisms for perceptual grouping. International Conference on Representation Learning (2020)

[26] Linsley, D., Malik, G., Kim, J., Govindarajan, L.N., Mingolla, E., Serre, T.: Tracking without re-recognition in humans and machines. (May 2021)

[27] Golan, T., Raju, P.C., Kriegeskorte, N.: Controversial stimuli: Pitting neural networks against each other as models of human cognition. Proc. Natl. Acad. Sci. U. S. A. **117**(47) (November 2020) 29330–29337

[28] 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

[29] Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A. **111**(23) (June 2014) 8619–8624

[30] Schrimpf, M., Kubilius, J., Lee, M.J., Ratan Murty, N.A., Ajemian, R., DiCarlo, J.J.: Integrative benchmarking to advance neurally mechanistic models of human intelligence. Neuron **108**(3) (November 2020) 413–423

[31] Fabre-Thorpe, M.: The characteristics and limits of rapid visual categorization. Front. Psychol. **2** (October 2011) 243

[32] Roelfsema, P.R., Lamme, V.A., Spekreijse, H.: The implementation of visual routines. Vision Res. **40**(10-12) (2000) 1385–1411

[33] Serre, T., Oliva, A., Poggio, T.: A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. U. S. A. **104**(15) (April 2007) 6424–6429

[34] Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K.A., Cichy, R.M., Hauk, O., Kriegeskorte, N.: Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. U. S. A. **116**(43) (October 2019) 21854–21863

[35] Jagadeesh, A.V., Gardner, J.L.: Texture-like representation of objects in human visual cortex. Proceedings of the National Academy of Sciences **119**(17) (2022) e2115302119

[36] Berrios, W., Deza, A.: Joint rotational invariance and adversarial training of a dual-stream transformer yields state of the art Brain-Score for area V4. (March 2022)

[37] Serre, T.: Deep learning: The good, the bad, and the ugly. Annu Rev Vis Sci **5** (September 2019) 399–426

[38] Koehler, K., Guo, F., Zhang, S., Eckstein, M.P.: What do saliency models predict? J. Vis. **14**(3) (March 2014) 14

[39] 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

[40] 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

[41] 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)

[42] 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

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

[44] 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

[45] Srivastava, S., Ben-Yosef, G., Boix, X.: Minimal images in deep neural networks: Fragile object recognition in natural images. (February 2019)

[46] Boyd, A., Tinsley, P., Bowyer, K., Czajka, A.: CYBORG: Blending human saliency into the loss improves deep learning. (December 2021)

[47] Bomatter, P., Zhang, M., Karev, D., others: When pigs fly: Contextual reasoning in synthetic and natural scenes. Proceedings of the (2021)

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

[49] Eberhardt, S., Cader, J.G., Serre, T.: How deep is the feature analysis underlying rapid visual categorization? In Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., eds.: Advances in Neural Information Processing Systems 29. Curran Associates, Inc. (2016) 1100–1108

[50] Kirchner, H., Thorpe, S.J.: Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision Res. **46**(11) (May 2006) 1762–1776

[51] Muriel, B., Simon, T., Holle, K.: Rapid object categorization without conscious recognition: aneuropsychological study. J. Vis. **7**(9) (June 2007) 1033–1033

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

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

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

[55] 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)

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

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

[58] 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)

[59] 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

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

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

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

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

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

[65] 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)

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

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

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

[69] 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

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

[71] 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)

[72] 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)

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

[74] d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., Sagun, L.: ConViT: Improving vision transformers with soft convolutional inductive biases. (March 2021)

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

[76] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: MLP-Mixer: An all-MLP architecture for vision. (May 2021)

[77] 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. (October 2020)

[78] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your ViT? data, augmentation, and regularization in vision transformers. (June 2021)

[79] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. (February 2020)

[80] Zeki Yalniz, I., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. (May 2019)

[81] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. (November 2018)

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

[83] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. (December 2013)

[84] Deza, A., Konkle, T.: Emergent properties of foveated perceptual systems. (June 2020)

[85] Wang, P., Cottrell, G.W.: Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. J. Vis. **17**(4) (April 2017) 9

[86] Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., Kreiman, G.: Recurrent computations for visual pattern completion. Proc. Natl. Acad. Sci. U. S. A. **115**(35) (August 2018) 8835–8840

[87] Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D.L., DiCarlo, J.J.: Brain-Like object recognition with High-Performing shallow recurrent ANNs. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R., eds.: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 12805–12816

[88] Schrimpf, M., Kubilius, J., Lee, M.J., Ratan Murty, N.A., Ajemian, R., DiCarlo, J.J.: Integrative benchmarking to advance neurally mechanistic models of human intelligence. Neuron **108**(3) (November 2020) 413–423

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

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

[91] Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., Serre, T.: What are the visual features underlying human versus machine vision? (January 2017)

[92] Linsley, J.W., Linsley, D.A., Lamstein, J., Ryan, G., Shah, K., Castello, N.A., Oza, V., Kalra, J., Wang, S., Tokuno, Z., Javaherian, A., Serre, T., Finkbeiner, S.: Superhuman cell death detection with biomarker-optimized neural networks. Sci Adv **7**(50) (December 2021) eabf8142

[93] Lee, K., Zung, J., Li, P., Jain, V., Sebastian Seung, H.: Superhuman accuracy on the SNEMI3D connectomics challenge. (May 2017)

[94] Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Halpern, D., Hamrick, J.B., Chan, P.: psiturk: An open-source framework for conducting replicable behavioral experiments online. Behav. Res. Methods **48**(3) (September 2016) 829–842

[95] Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. Proc. IEEE **69**(5) (May 1981) 529–541

[96] Thomson, M.G.: Visual coding and the phase structure of natural scenes. Network **10**(2) (May 1999) 123–132

[97] Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. (May 2020)

A Psychophyics

The psychophysics experiments of §4.2 were implemented with the psiTurk framework [94] and custom javascript functions. Each trial sequence was converted to a HTML5-compatible video for the fastest reliable presentation time possible in a web browser. Videos were cached before each trial to optimize reliability of experiment timing within the web browser. A photo-diode verified the reliability of stimulus timing in our experiment was consistently accurate within $\sim 10\text{ms}$ across different operating system, web browser, and display type configurations.

Participants: We recruited 199 participants from Amazon Mechanical Turk (`mturk.com`) for the experiments. Participants were based in the United States, used either the Firefox or Chrome browser on a non-mobile device, and had a minimal average approval rating of 95% on past Mechanical Turk tasks.

Stimuli: Experiment images were taken from the *Clicktionary* dataset [22]. Images were sampled from 5 target and 5 distractor categories: border collie, sorrel (horse), great white shark, bald eagle, and panther; trailer truck, sports car, speedboat, airliner, and school bus. Images were presented to human participants (and DNNs) either intact or with a perceptual phase scrambled mask that exposed a proportion of their most important visual features, as described in the main text. Images were cast to greyscale to control for trivial color-based cues for classification and blend the scrambled mask background into the foreground. Responses to intact images were used to normalize the performance of each observer on masked images relative to their maximum performance on these images.

Image masks were created for each image to reveal only a proportion of the most important visual features. For each image, we created masks that revealed between 1% and 100% (at log-scale spaced intervals) of the object pixels in the corresponding image's *Clicktionary* feature importance map. We generated these masks in two steps. First, we computed a phase-scrambled version of the image [95, 96]. Next, we used a novel "stochastic flood-fill" algorithm to reveal a contiguous region of the most important visual features in the image according to humans. Our flood-fill algorithm was seeded on the pixel deemed most important by humans in the image, then grew outwards anisotropically and biased towards pixels with higher feature importance scores (Figure S1). The revealed region was always centered on the image. Each participant saw every category exemplar only once, with its amount of image revelation randomly selected from all possible configurations.

After providing online consent, participants were instructed to complete a rapid visual categorization task in which they had to classify stimuli revealing a portion of the most diagnostic object features (Fig. S3). Each experimental trial began with a cross for participants to fixate for a variable time (1,100–1,600ms), then a stimulus for 400ms, then another cross and additional time for participants to render a decision. Participants were instructed to provide a decision after the first fixation cross, but that they only had 650ms to answer. If they were too slow to respond they were told to respond faster and the trial was discarded.

B Harmonization loss

The neural harmonizer loss Fig. S2 uses several components crucial to its performance: a pyramidal representation of decision explanation maps and normalizing those maps.

When computing the difference between model explanations for an image and the human feature importance map for that image, we rely on a pyramid representation of each to compute these differences Fig. S2). This pyramid allows for a model to align its feature representations with humans at multiple scales and corrects for an important problem in datasets like *ClickMe*: the human data is an approximation and not precise at the pixel level. This lack of precision can present optimization issues, and computing a pyramid representation alleviates those issues because it allows a model to learn to focus on regions that are important for humans without pixel-level precision.

*These authors contributed equally.
[1]Department of Cognitive, Linguistic, & Psychological Sciences, Brown University, Providence, RI
[2]Artificial and Natural Intelligence Toulouse Institute (ANITI), Toulouse, France
[3]Carney Institute for Brain Science, Brown University, Providence, RI

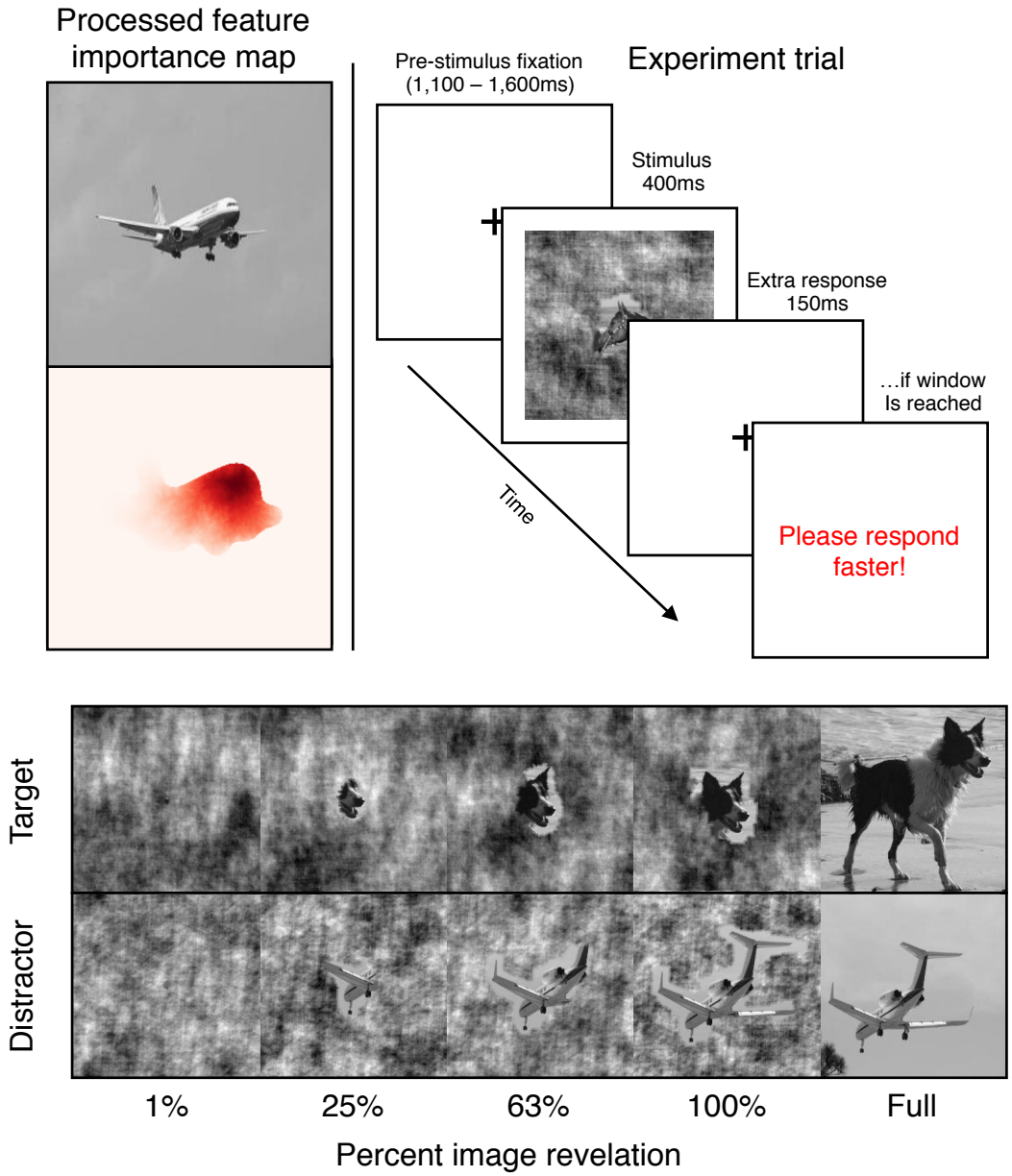

Figure S1: **Overview of the psychophysics paradigm.** Participants performed a rapid animals vs. vehicles categorization paradigm (top). Stimuli were created using feature importance maps derived from humans or DNNs via a "stochastic flood-fill" algorithm that revealed image regions of different sizes centered on important features. Sample stimuli are shown (bottom) for different percentages of image revelation. Note that 100% revelation corresponds to all non-zero pixels in a feature importance map.

Standardization tackles a similar problem: because of the imprecision of human data, we choose to focus harmonization on only the most important areas selected by humans in *ClickMe*. By standardizing then rectifying before comparing human and model explanations, we reduce noise in the harmonization procedure.

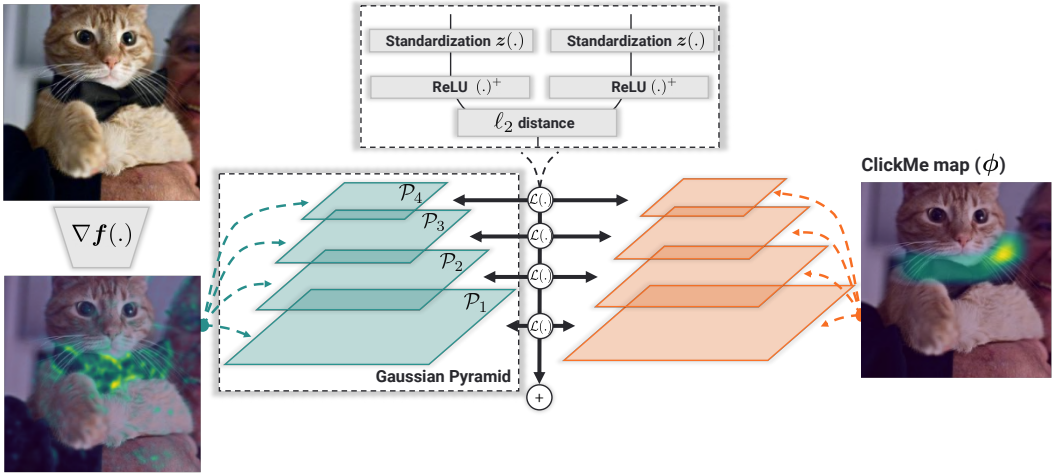

Figure S2: **Computing the neural harmonizer loss.**.

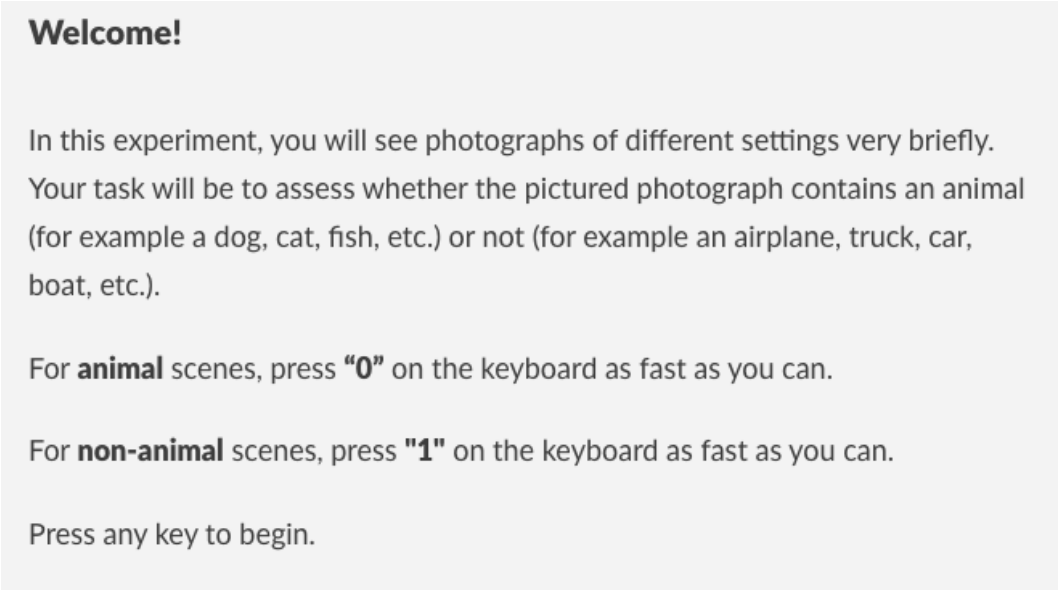

Figure S3: **Psychophysics experiment instructions.**

C Additional Results

C.1 *ClickMe*

The *ClickMe* game by [20] was used to identify category diagnostic features in ImageNet images. These feature importance maps largely focus on object regions rather than context, and in contrast to segmentation maps select features on the "front" or "face" of objects (Fig. S4).

As discussed in the main text, we found a trade-off between DNN top-1 ImageNet accuracy and the alignment of their feature importance maps with humans importance maps from *ClickMe*. This trade-off persists across multiple scales of feature importance maps, including $4\times$ (Fig. S6) and $16\times$ (Fig. S7) sub-sampled maps, meaning that simple smoothing is not sufficient to fix the trade-off.

Figure S4: **Example *ClickMe* feature importance maps on ImageNet images.**

C.2 *ViT attention*

While in the main text we investigate alignment between humans and models using gradient feature importance visualizations, the attention maps in transformer models like the ViT provide another avenue for investigation. To understand whether or not attention maps from ViT are more aligned with humans than their gradient-based decision explanation maps, we computed attention rollouts for harmonized and unharmonized ViTs [97]. We found that both versions of the ViT had similar correlations between their attention rollouts and human *ClickMe* maps: 0.38 for the harmonized ViT and 0.393 for the unharmonized model. This surprising result suggests that the harmonizer affects the process by which ViTs integrate visual information into their decisions rather than how they allocate attention. Through manipulating ViT decision making processes, the harmonizer can induce the large changes in gradient-based visualizations and psychophysics that we describe in the main text.

C.3 *Correlations between measurements of human visual strategies*

Our results rely on three independent datasets measuring different features of human visual strategies: *ClickMe*, *Clicktionary*, and the psychophysics experiments we introduce in this manuscript. The fact that all three evoke similar trade-offs between top-1 accuracy and human alignment is a surprising result that deserves further attention. We investigated these trade-offs by measuring the correlation between human alignment on each dataset, with and without models trained with the neural harmonizer. We found that correlations between datasets were lower across the board when neural harmonizer models were not included. The association between model alignments with *Clicktionary* versus psychophysics results were not significant ($\rho = 0.21$, n.s.; Fig. S9), but the associations between model alignments with *ClickMe* versus psychophysics ($\rho = 0.51, p < 0.001$; Fig. S8) and *ClickMe* versus *Clicktionary* ($\rho = 0.77, p < 0.001$; Fig. S10) were both significant. Each correlation improved when the neural harmonizer models were included in the calculation. This finding indicates that the neural harmonizer successfully aligned visual strategies between humans and DNNs, and was not merely benefiting from either *where* humans versus DNNs considered important visual features to be or *how* humans versus DNNs incorporated those features into their decisions.

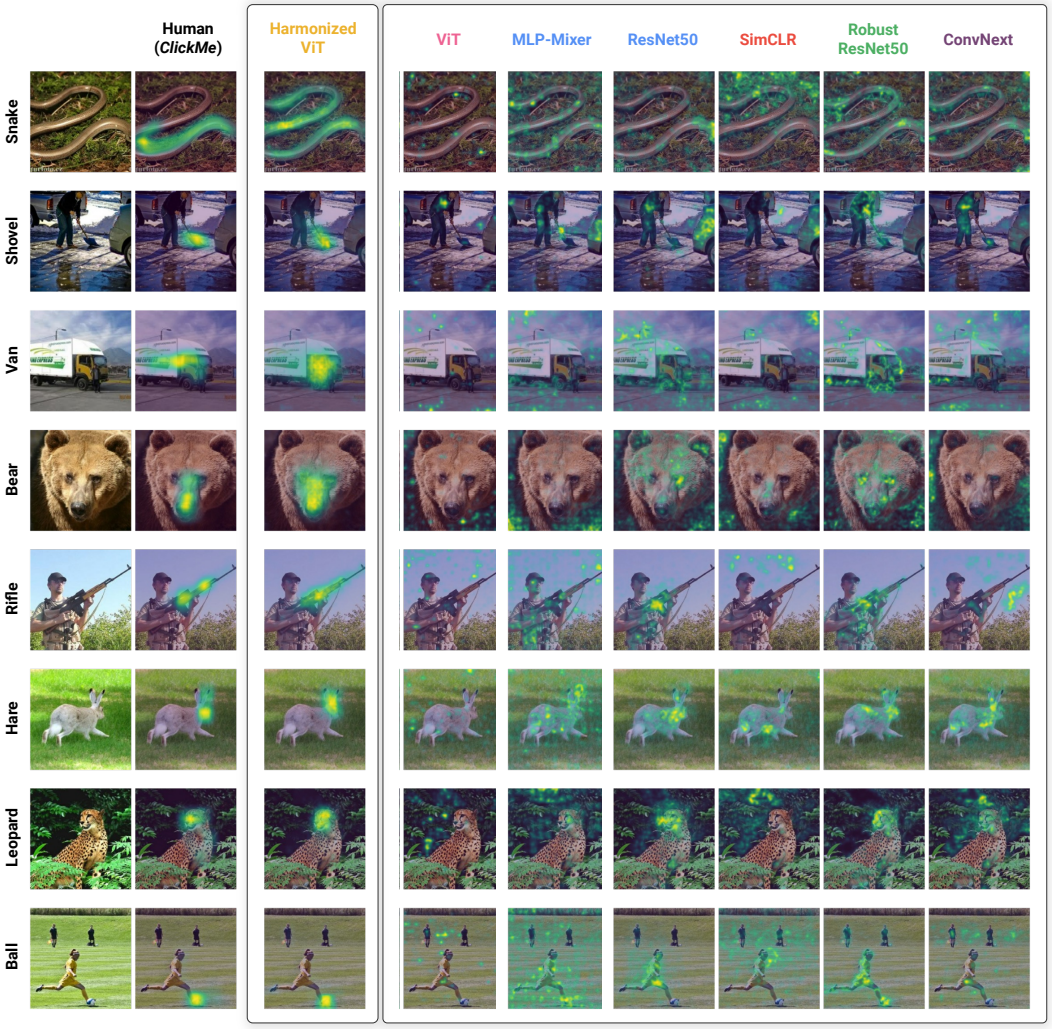

Figure S5: **Feature importance maps of humans, harmonized, and unharmonized models on ImageNet**.

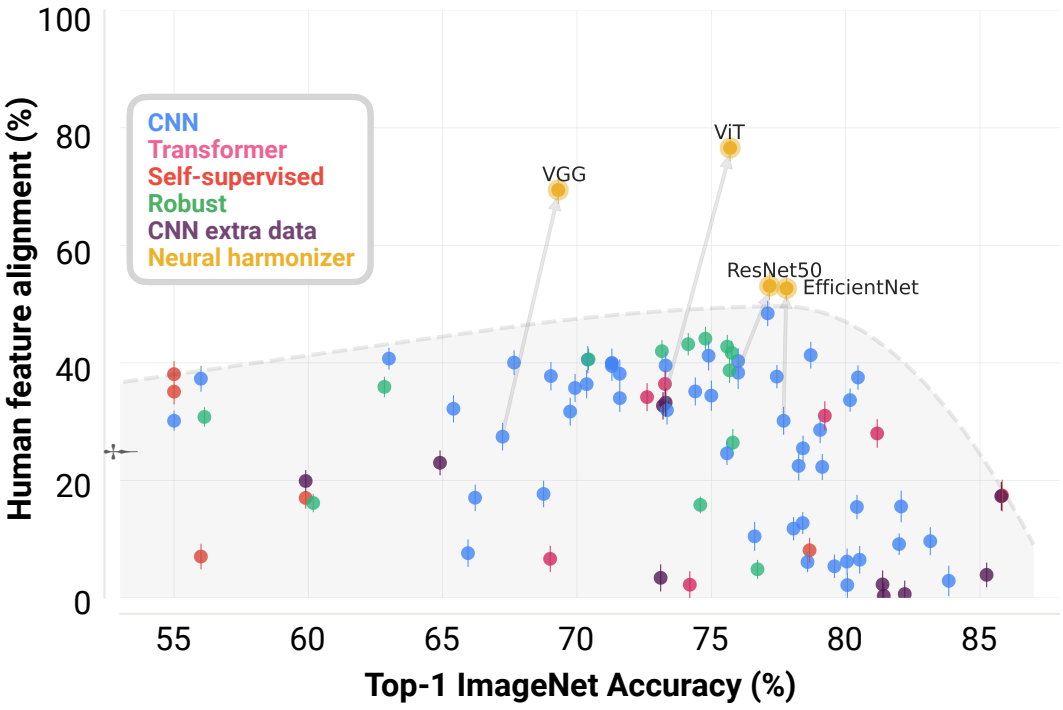

Figure S6: **The neural harmonizer's effect is robust across image scales.** Here, we show that the trade-off between ImageNet accuracy and alignment with humans holds across downsizing by a factor of *4*. The Neural harmonizer once again yields the model with the best alignment with humans. Grey-shaded area captures the trade-off between accuracy and alignment in standard DNNs. Error bars are bootstrapped standard deviations over feature alignment.

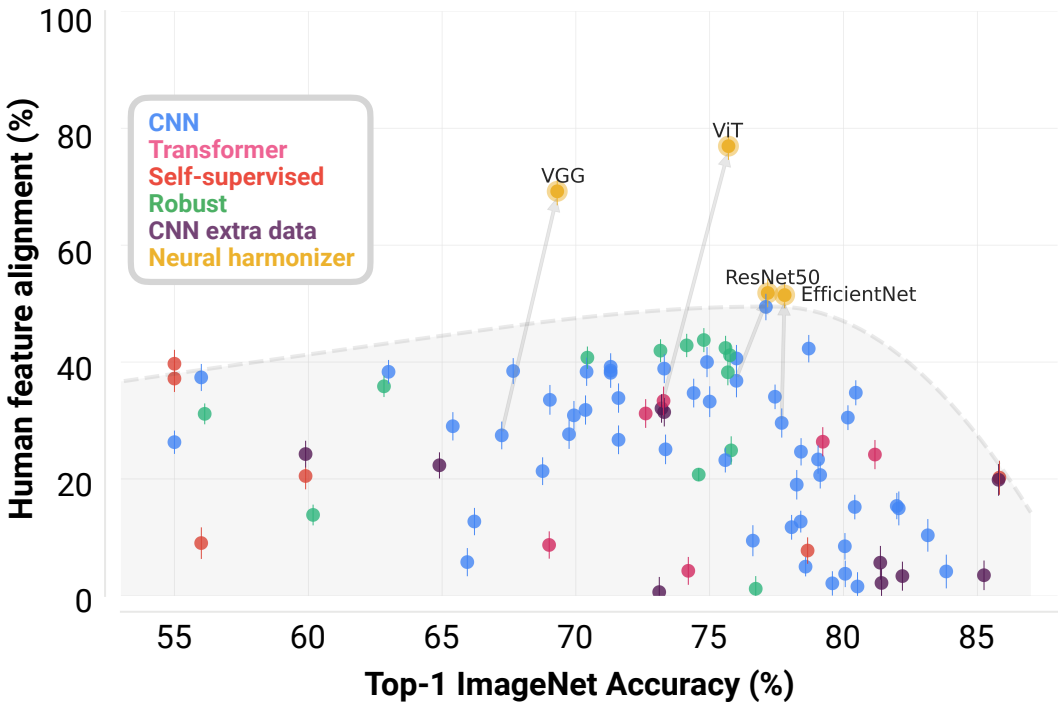

Figure S7: **The neural harmonizer's effect is robust across image scales.** Here, we show that the trade-off between ImageNet accuracy and alignment with humans holds across downsizing by a factor of *16*. The Neural harmonizer once again yields the model with the best alignment with humans. Grey-shaded area captures the trade-off between accuracy and alignment in standard DNNs. Error bars are bootstrapped standard deviations over feature alignment.

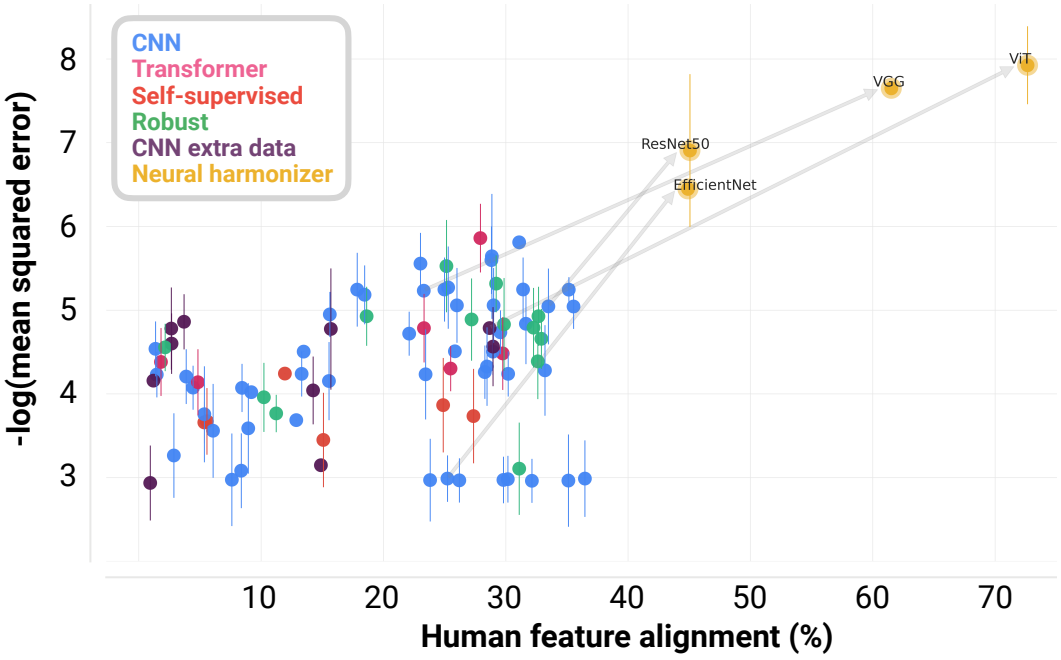

Figure S8: **The association between *ClickMe* alignment versus psychophysics alignment.** These scores are significantly correlated, $\rho = 0.68, p < 0.001$. Error bars are bootstrapped standard deviations over feature alignment.

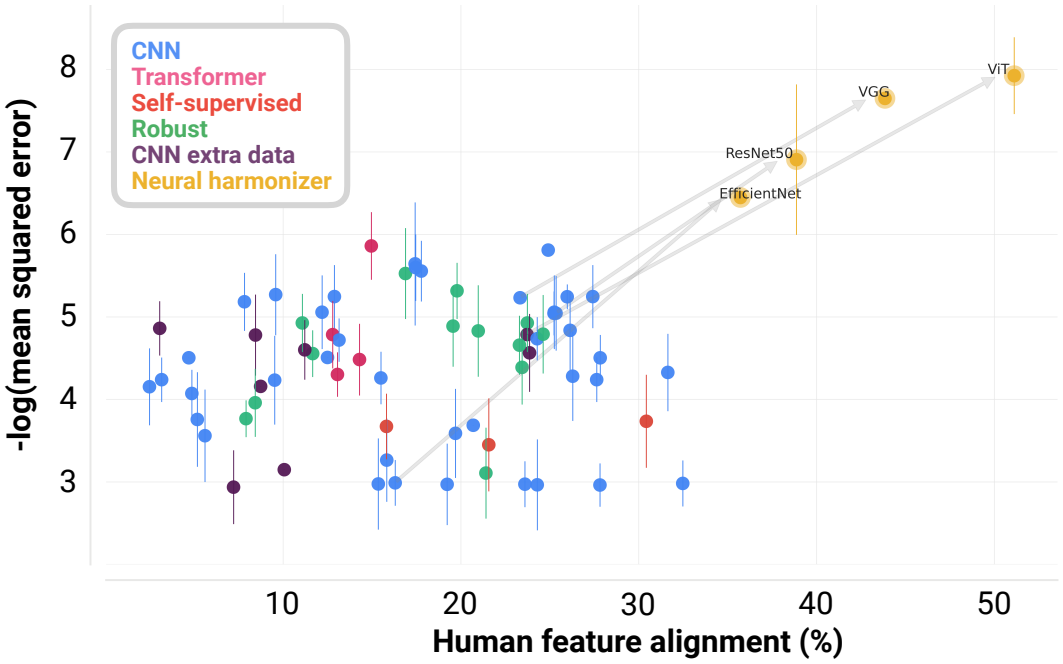

Figure S9: **The association between *Clicktionary* alignment versus psychophysics alignment.** These scores are significantly correlated, $\rho = 0.53, p < 0.001$.

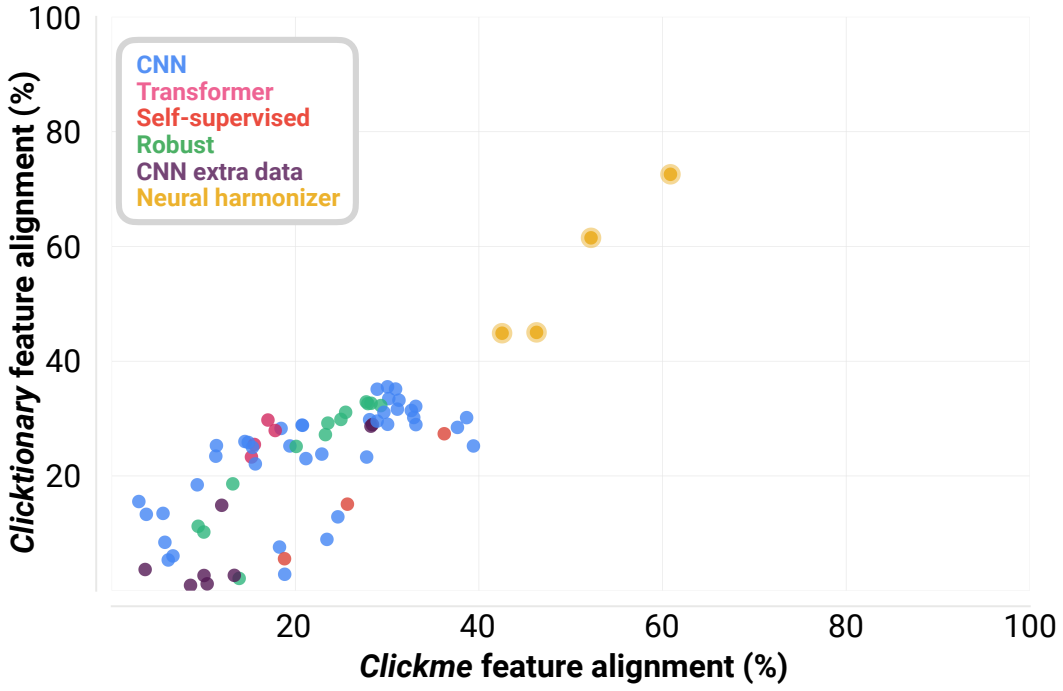

Figure S10: **The association between *ClickMe* alignment versus *Clicktionary* alignment.** These scores are significantly correlated, $\rho = 0.85, p < 0.001$. Error bars are bootstrapped standard deviations over feature alignment.

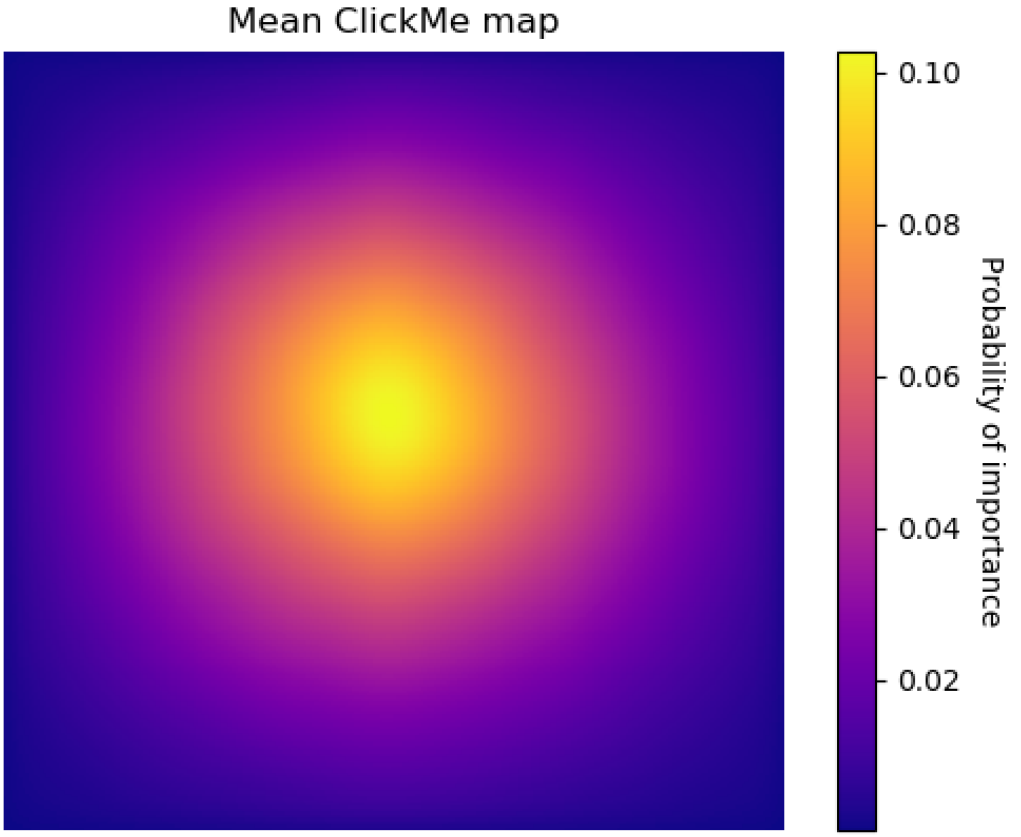

Figure S11: **The mean of *ClickMe* feature importance maps exhibits a center bias, likely due to the positioning of objects in ImageNet images rather than a purely spatial bias of human participants (compare to individual maps shown in S4)**.