Partial success in closing the gap between human and machine vision

Anonymous Author(s)
Affiliation
Address
email

Abstract

A few years ago, the first CNN surpassed human performance on ImageNet. However, it soon became clear that machines lack robustness on more challenging test cases, a major obstacle towards deploying machines "in the wild" and towards obtaining better computational models of human visual perception. Here we ask: Are we making progress in closing the gap between human and machine vision? To answer this question, we tested human observers on a broad range of out-of-distribution (OOD) datasets, adding the "missing human baseline" by recording 85,120 psychophysical trials across 90 participants. We then investigated a range of promising machine learning developments that crucially deviate from standard supervised CNNs along three axes: objective function (self-supervised, adversarially trained, CLIP language-image training), architecture (e.g. vision transformers), and dataset size (ranging from 1M to 1B).

Our findings are threefold. (1.) The longstanding *robustness gap* between humans and CNNs is closing, with the best models now matching or exceeding human performance on most OOD datasets. (2.) There is still a substantial image-level *consistency gap*, meaning that humans make different errors than models. In contrast, most models systematically agree in their categorisation errors, even substantially different ones like contrastive self-supervised vs. standard supervised models. (3.) In many cases, human-to-model consistency improves when training dataset size is increased by one to three orders of magnitude. Our results give reason for cautious optimism: While there is still much room for improvement, the behavioural difference between human and machine vision is narrowing. In order to measure future progress, 17 OOD datasets with image-level human behavioural data are provided as a benchmark.

1 Introduction

Looking back at the last decade, deep learning has made tremendous leaps of progress by any standard. What started in 2012 with AlexNet [1] as the surprise winner of the ImageNet Large-Scale Visual Recognition Challenge quickly became the birth of a new AI "summer", a summer lasting much longer than just a season. With it, just like with any summer, came great expectations: the hope that the deep learning revolution will see widespread applications in industry, that it will propel breakthroughs in the sciences, and that it will ultimately close the gap between human and machine perception.

We have now reached the point where deep learning has indeed become a significant driver of progress in industry [e.g. 2, 3], and where many disciplines are employing deep learning for scientific discoveries [4–9]—*but are we making progress in closing the gap between human and machine vision?*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

For a long time, the gap between human and machine vision was mainly approximated by comparing benchmark accuracies on IID (independent and identically distributed) test data: as long as models are far from reaching human-level performance on challenging datasets like ImageNet, this approach is adequate [10]. Currently, models are routinely matching and in many cases even outperforming humans on IID data. At the same time, it is becoming increasingly clear that models systematically exploit shortcuts shared between training and test data [11–14]. Therefore we are witnessing a major shift towards measuring model performance on out-of-distribution (OOD) data rather than IID data alone, which aims at testing models on more challenging test cases where there is still a ground truth category, but certain image statistics differ from the training distribution. Many OOD generalisation datasets have been proposed, such as ImageNet-C [15] for corruption robustness, ImageNet-Sketch [16] for robustness towards sketches, and Stylized-ImageNet [17] for robustness towards image style changes. What most of these datasets have in common, however, is that they lack human comparison data. This is unfortunate, since we can no longer assume that humans reach near-ceiling accuracies on these challenging test cases as they do on standard noise-free IID object recognition datasets. In order to address this issue, we carefully tested human observers in a psychophysics laboratory on a broad range of OOD datasets, adding the "missing human baseline" by providing some 85K psychophysical trials across 90 participants. Crucially, we showed exactly the same images to multiple observers, which means that we are able to compare human and machine vision on the fine-grained level of individual images [18–20]).

The resulting 17 OOD datasets with large-scale human comparion data enable us to investigate recent exciting machine learning developments that crucially deviate from "vanilla" CNNs along three axes: objective function (supervised vs. self-supervised, adversarially trained, and CLIP's joint language-image training), architecture (convolutional vs. vision transformer) and training dataset size (ranging from 1M to 1B images). Taken together, these are some of the most promising directions our field has developed to date—but this field would not be machine learning if new breakthroughs weren't within reach in the next few weeks, months and years. Therefore, we will open-source `modelvshuman`, a Python project that enables testing both PyTorch and TensorFlow models on our comprehensive benchmark suite of OOD generalisation data in order to measure future progress (a github link will be added after the anonymous review period; code is included in the supplementary material). Even today, our results give cause for (cautious) optimism. After a method overview (Section 2), we are able to report that the human-machine *robustness gap* is closing: the best models now match or exceed human performance on most OOD datasets (Section 3). While there is still a substantial image-level *consistency gap* between humans and machines, this gap is narrowing on some—but not all—datasets when the size of the training dataset is increased (Section 4).

2 Methods: datasets, psychophysical experiments, models, metrics, toolbox

OOD datasets with consistency-grade human data. We collected human data for 17 generalisation datasets (visualized in Figures 7 and 8 in the Appendix) on a carefully calibrated screen in a dedicated psychophysical laboratory (a total of 85,120 trials across 90 observers). Five datasets each correspond to to a single manipulation (sketches, edge-filtered images, silhouettes, images with a texture-shape cue conflict, and stylized images where the original image texture is replaced by the style of a painting); the remaining twelve datasets correspond to parametric image degradations (e.g. different levels of noise or blur). Those OOD datasets have in common that they are designed to test ImageNet-trained models. OOD images were obtained from different sources: sketches from ImageNet-Sketch [16], stylized images from Stylized-ImageNet [17], edge-filtered images, silhouettes and cue conflict images from [17][1], and the remaining twelve parametric datasets were adapted from [21]. For these parametric datasets, [21] collected human accuracies but unfortunately, they showed different images to different observers implying that we cannot use their human data to assess image-level consistency between humans and machines. Thus we collected psychophysical data for those images ourselves by showing exactly the same images to multiple observers for each of those twelve datasets. Additionally, we cropped the images from [21] to 224×224 pixels to allow for a fair comparison to ImageNet models (all models included in our comparison receive 224×224 input images; [21] showed 256×256 images to human observers in many cases).

[1]For those three datasets consisting of 160, 160 and 1280 images respectively, consistency-grade psychophysical data was already collected by the authors and included in our benchmark with permission from the authors.

Psychophysical experiments. 90 observers were tested in a darkened chamber. Stimuli were presented at the center of a 22" monitor with 1920×1200 pixels resolution (refresh rate: 120 Hz). Viewing distance was 107 cm and target images subtended 3×3 degrees of visual angle. Human observers were presented with an image and asked to select the correct category out of 16 basic categories (such as chair, dog, airplane, etc.). Stimuli were balanced w.r.t. classes and presented in random order. For ImageNet-trained models, in order to obtain a choice from the same 16 categories, the 1,000 class decision vector was mapped to those 16 classes using the WordNet hierarchy [22].[2] We closely followed the experimental protocol defined by [21], who presented images for 200 ms followed by a $1/f$ backward mask to limit the influence of recurrent processing (otherwise comparing to feedforward models would be difficult). Further experimental details are provided in Appendix C.

Models. In order to disentangle the influence of objective function, architecture and training dataset size, we tested a total of 52 models: 24 standard ImageNet-trained CNNs [23], 8 self-supervised models [24–29],[3] 6 Big Transfer models [30], 5 adversarially trained models [31], 5 vision transformers [32, 33], two semi-weakly supervised models [34] as well as Noisy Student [35] and CLIP [36]. Technical details for all models are provided in the Appendix.

Metrics. In addition to *OOD accuracy* (averaged across conditions and datasets), the following three metrics quantify how closely machines are aligned with the decision behaviour of humans.

Accuracy difference $A(m)$ is a simple aggregate measure that compares the accuracy of a machine m to the accuracy of human observers in different out-of-distribution tests,

$$A(m) : \mathbb{R} \to [0,1], m \mapsto \frac{1}{|D|} \sum_{d \in D} \frac{1}{|H_d|} \sum_{h \in H_d} \frac{1}{|C_d|} \sum_{c \in C_d} (\mathrm{acc}_{d,c}(h) - \mathrm{acc}_{d,c}(m))^2 \tag{1}$$

where $\mathrm{acc}_{d,c}(\cdot)$ is the accuracy of the model or the human on dataset $d \in D$ and condition $c \in C_d$ (e.g. a particular noise level), and $h \in H_D$ denotes a human observer tested on dataset d. Analogously, one can compute the average accuracy difference between a human observer h_1 and all other human observers by substituting h_1 for m and $h \in H_D \setminus \{h_1\}$ for $h \in H_D$ (which can also be applied for the two metrics defined below).

Aggregated metrics like $A(m)$ ignore individual image-level decisions. Two models with vastly different image-level decision behaviour might still end up with the same accuracies on each dataset and condition. Hence, we include two additional metrics in our benchmark that are sensitive to decisions on individual images.

Observed consistency $O(m)$ [20] measures the fraction of samples for which humans and a model m get the same sample either both right *or* both wrong. More precisely, let $b_{h,m}(s)$ be one if both a human observer h and m decide either correctly or incorrectly on a given sample s, and zero otherwise. We calculate the average observed consistency as

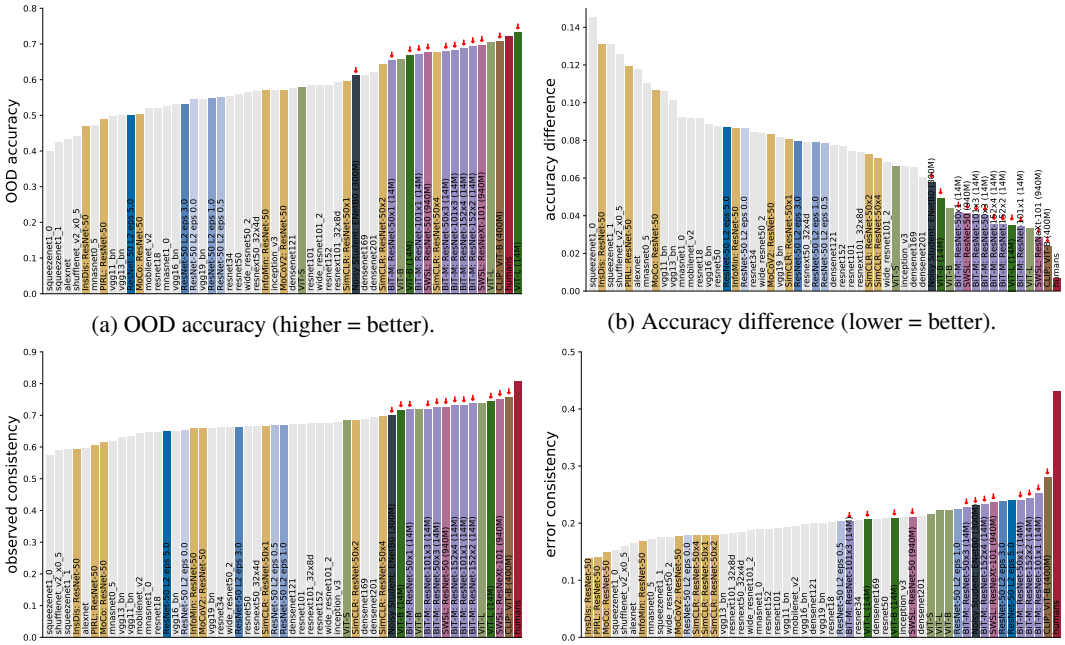
$$O(m) : \mathbb{R} \to [0,1], m \mapsto \frac{1}{|D|} \sum_{d \in D} \frac{1}{|H_d|} \sum_{h \in H_d} \frac{1}{|C_d|} \sum_{c \in C_d} \frac{1}{|S_{d,c}|} \sum_{s \in S_{d,c}} b_{h,m}(s) \tag{2}$$

where $s \in S_{d,c}$ denotes a sample s (in our case, an image) of condition c from dataset d. Note that this measure can only be zero if the accuracy of h and m are exactly the same in each dataset and condition.

Error consistency $E(m)$ [20] tracks whether there is above-chance consistency. This is an important distinction, since e.g. two decision makers with 95% accuracy each will have at least 90% observed consistency, even if their 5% errors occur on non-overlapping subsets of the test data (intuitively, they both get most images correct and thus observed overlap is high). To this end, error consistency (a.k.a. Cohen's kappa, cf. [37]) indicates whether the observed consistency is larger than what could have been expected given two independent binomial decision makers with matched accuracy, which we denote as $\hat{o}_{h,m}$. This can easily be computed analytically [e.g. 20, equation 1]. Then, the average

[2] The appendix of [21] describes in more detail how this mapping is optimal using the 'average' probability aggregation method ("Aggregating probabilities from coarse classes").

[3] We presented a preliminary and much less comprehensive version of this work at a NeurIPS workshop. A reference will be added after the anonymous review period.



(a) OOD accuracy (higher = better).

(b) Accuracy difference (lower = better).

(c) Observed consistency (higher = better).

(d) Error consistency (higher = better).

Figure 1: Core results, aggregated over 17 out-of-distribution (OOD) datasets: The OOD robustness gap between human and machine vision is closing (top), but an image-level consistency gap remains (bottom). Results compare humans, standard supervised CNNs, self-supervised models, adversarially trained models, vision transformers, noisy student, BiT, SWSL and CLIP. For convenience, ↓ marks models that are trained on large-scale datasets. Metrics defined in Section 2. Best viewed on screen.

error consistency is given by

$$E(m) : \mathbb{R} \to [-1, 1], m \mapsto \frac{1}{|D|} \sum_{d \in D} \frac{1}{|H_d|} \sum_{h \in H_d} \frac{1}{|C_d|} \sum_{c \in C_d} \frac{\left(\frac{1}{|S_{d,c}|} \sum_{s \in S_{d,c}} b_{h,m}(s)\right) - \hat{o}_{h,m}(S_{d,c})}{1 - \hat{o}_{h,m}(S_{d,c})} \tag{3}$$

Benchmark & toolbox. $A(m)$, $O(m)$ and $E(m)$ each quantify a certain aspect of the human-machine gap. We use the mean rank order across these metrics to determine an overall model ranking (Table 2 in the Appendix). However, we would like to emphasise that the primary purpose of this benchmark is to generate insights, not winners. Since insights are best gained from detailed plots and analyses, we open-source `modelvshuman`, a Python project to benchmark models against human data.[4] The current model zoo already includes 50+ models, and an option to add new ones (both PyTorch and TensorFlow). Evaluating a model will produce a 15+ page report on model behaviour. All plots in this paper can be generated for future models—to track whether they narrow the gap towards human vision, or to determine whether an algorithmic modification to a baseline model (e.g., an architectural improvement) changes model behaviour.

3 Robustness across models: the OOD robustness gap between human and machine vision is closing

We are interested in measuring whether we are making progress in closing the gap between human and machine vision. For a long time, CNNs were unable to match human robustness in terms of generalisation beyond the training distribution—a large OOD *robustness gap* [14, 21, 38, 39]. Having tested human observers on 17 OOD datasets, we are now able to compare the latest developments in machine vision to human perception. Our core results are shown in Figure 1: the OOD robustness gap

[4] Of course, comparing human and machine vision is not limited to object recognition behaviour: other comparisons may be just as valid and interesting.

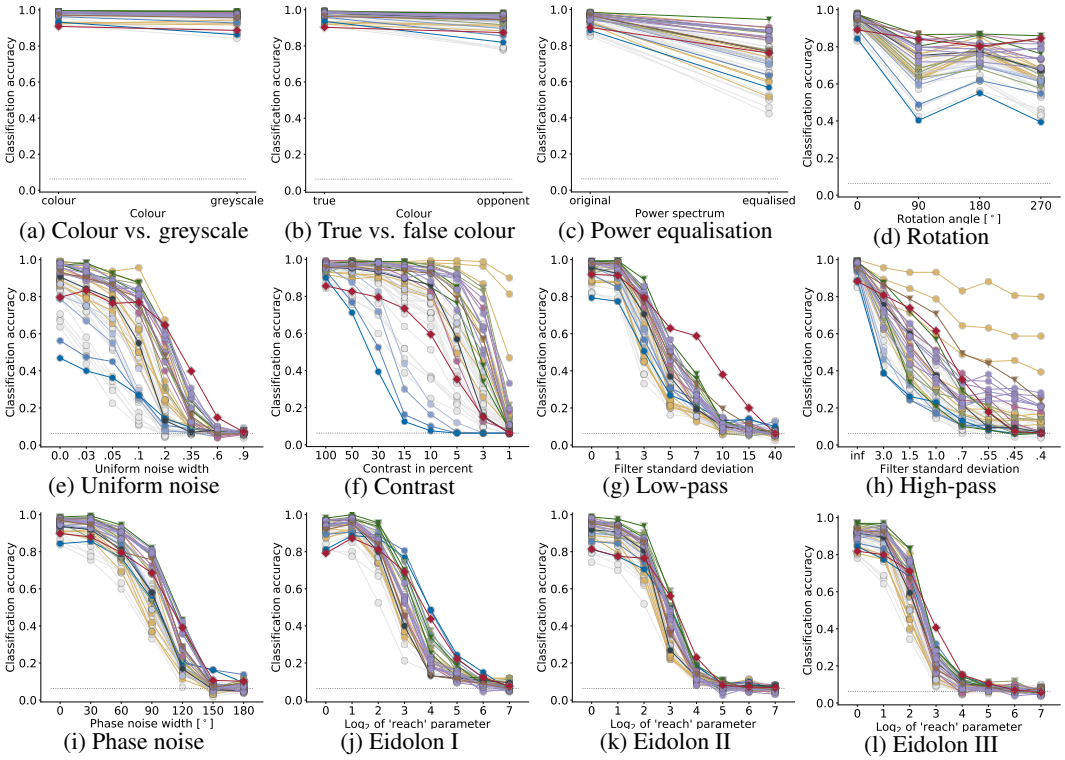


Figure 2: The OOD robustness gap between human and machine vision is closing. Robustness towards parametric distortions for humans, standard supervised CNNs, self-supervised models, adversarially trained models, vision transformers, noisy student, BiT, SWSL, CLIP. Symbols indicate architecture type (○ convolutional, ▽ vision transformer, ◊ human); best viewed on screen.

between human and machine vision is closing (1a, 1b), especially for models trained on large-scale datasets. On the individual image level, a human-machine consistency gap remains (especially 1d), which will be discussed later.

Self-supervised models *"If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning and the cherry on the cake is reinforcement learning"*, Yann LeCun said in 2016 [40]. A few years later, the entire cake is finally on the table—the representations learned via self-supervised learning[5] now compete with supervised methods on ImageNet [29] and outperform supervised pre-training for object detection [27]. But how do recent self-supervised models differ from their supervised counterparts in terms of their behaviour? Do they bring machine vision closer to human vision? Humans, too, rapidly learn to recognise new objects without requiring hundreds of labels per instance; additionally a number of studies reported increased similarities between self-supervised models and human perception [41–45]. Figure 2 compares the generalisation behaviour of eight self-supervised models in orange (PIRL, MoCo, MoCoV2, InfoMin, InsDis, SimCLR-x1, SimCLR-x2, SimCLR-x4)—with 24 standard supervised models (grey). We find only marginal differences between self-supervised and supervised models: Across distortion types, self-supervised networks are well within the range of their poorly generalising supervised counterparts. However, there is one exception: the three SimCLR variants show strong generalisation improvements on uniform noise, low contrast, and high-pass images, where they are the three top-performing self-supervised networks—quite remarkable given that SimCLR models were trained on a different set of augmentations (random crop with flip and resize, colour distortion, and Gaussian blur). Curious by the outstanding performance of SimCLR, we asked whether the self-supervised objective function or the choice of training data augmentations was the defining factor. When comparing self-supervised SimCLR models with augmentation-matched baseline models trained in the standard

[5] "Unsupervised learning" and "self-supervised learning" are sometimes used interchangeably. We use the term "self-supervised learning"' since those methods use (label-free) supervision.

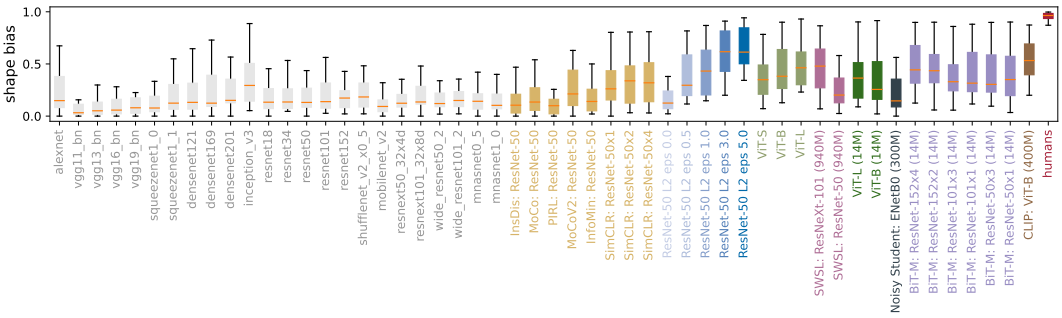


Figure 3: Shape (top) vs. texture (bottom) biases of different models. While human shape bias is not yet matched, several approaches improve over vanilla CNNs. Box plots show category-dependent distribution of shape/texture biases.

supervised fashion (Figure 14 in the Appendix), we find that the augmentation scheme (rather than the self-supervised objective) indeed made the crucial difference: supervised baselines show just the same generalisation behaviour, a finding that fits well with [46], who observed that the influence of training data augmentations is stronger than the role of architecture or training objective. In conclusion, our analyses indicate that the "cake" of contrastive self-supervised learning currently (and disappointingly) tastes much like the "icing".

Adversarially trained models The vulnerability of CNNs to adversarial input perturbations is, arguably, one of the most striking shortcomings of this model class compared to robust human perception. A successful method to increase adversarial robustness is *adversarial training* [e.g. 47, 48]. The resulting models were found to transfer better, have meaningful gradients [49], and enable interpolating between two input images [50]: "robust optimization can actually be viewed as inducing a *human prior* over the features that models are able to learn" [51, p. 10]. Therefore, we include five models with a ResNet-50 architecture and different accuracy-robustness tradeoffs, adversarially trained on ImageNet with Microsoft-scale resources by [31] to test whether models with "perceptually-aligned representations" also show human-aligned OOD generalisation behaviour—as we would hope. This is not the case: the stronger the model is trained adversarially (darker shades of blue in Figure 2), the more susceptible it becomes to (random) image degradations. Most strikingly, a simple rotation by 90 degrees leads to a 50% drop in classification accuracy. Adversarial robustness seems to come at the cost of increased vulnerability to large-scale perturbations.[6] On the other hand, there is a silver lining: when testing whether models are biased towards texture or shape by testing them on cue conflict images (Figure 3), in accordance with [53, 54] we observe a perfect relationship between shape bias and the degree of adversarial training, a big step in the direction of human shape bias (and a stronger shape bias than any of the other models).

Vision transformers In computer vision, convolutional networks have become by far the dominating model class over the last decade. Vision transformers [32] break with the long tradition of using convolutions and are rapidly gaining traction [55]. We find that the best vision transformer (ViT-L trained on 14M images) even *exceeds* human OOD accuracy (Figure 1a shows the average across 17 datasets). There appears to be an additive effect of architecture and data: vision transformers trained on 1M images (light green) are already better than standard convolutional models; training on 14M images (dark green) gives another performance boost.

Standard models trained on more data: BiT-M, SWSL, Noisy Student Interestingly, the biggest effect on OOD robustness we find simply comes from training on larger datasets, not from advanced architectures. When standard models are combined with large-scale training (14M images for BiT-M, 300M for Noisy Student and a remarkable 940M for SWSL), OOD accuracies reach levels not known from standard ImageNet-trained models; these models even outperform a more powerful architecture (vision transformer ViT-S) trained on less data (1M) as shown in Figure 1a. Simply training on (substantially) more data substantially narrows the gap to human OOD accuracies (1b).

[6] This might be related to [52], who studied a potentially related tradeoff between selectivity and invariance.

CLIP CLIP is special: trained on 400M images (more data) with joint language-image supervision (novel objective) and a vision transformer backbone (non-standard architecture), it is the most human-like model across all of our metrics presented in Figure 1; most strikingly in terms of error consistency to humans (which will be discussed in the next section). We tested a number of hypotheses to disentangle why CLIP is "special". *H1: because CLIP is trained on a lot of data?* No: Noisy Student—a model trained on a comparably large dataset of 300M images—does not perform as well. *H2: because CLIP receives higher-quality labels?* About 6% of ImageNet labels are plainly wrong [56]. Could it be the case that CLIP simply performs better since it doesn't suffer from this issue? In order to test this, we used CLIP to generate new labels for all 1.3M ImageNet images: (a) hard labels, i.e. the top-1 class predicted by CLIP; and (b) soft labels, i.e. using CLIP's full posterior distribution as a target. We then trained ResNet-50 from scratch on CLIP hard and soft labels. However, this does not show any robustness improvements over a vanilla ImageNet-trained ResNet-50, thus different/better labels are not a likely root cause. *H3: because CLIP has a special image+text loss?* Yes and no: CLIP training on ResNet-50 leads to astonishingly poor OOD results, so training a standard model with CLIP loss alone is insufficient. However, while neither architecture nor loss alone sufficiently explain why CLIP is special, we find a clear interaction between architecture and loss (described in more detail in the Appendix along with the other "CLIP ablation" experiments mentioned above).

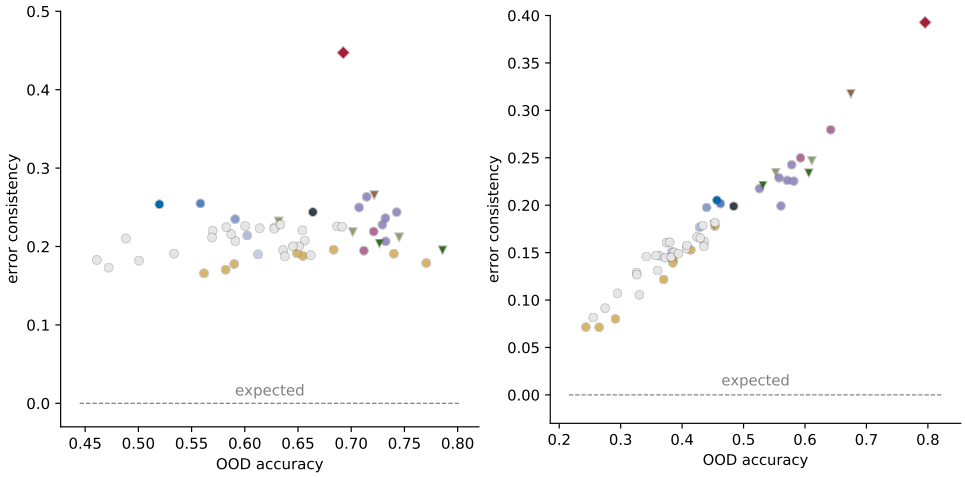
Figure 4: Data-rich models narrow the substantial image-level consistency gap between humans and machines. Error consistency analysis on a single dataset (sketch images; for other datasets see Appendix, Figures 9, 10, 11, 12, 13) shows that most models cluster (dark red = highly consistent errors) irrespective of their architecture and objective function; humans cluster differently (high human-to-human consistency, low human-to-model consistency); but some data-rich models including CLIP and SWSL blur the boundary, making more human-like errors than standard models.

4 Consistency between models: data-rich models narrow the substantial image-level consistency gap between human and machine vision

In the previous section we have seen that while self-supervised and adversarially trained models lack OOD robustness, models based on vision transformers and/or trained on large-scale datasets now match or exceed human performance on most datasets. Behaviourally, a natural follow-up question is to ask not just how many, but *which* errors models make—i.e., do they make errors on the same individual images as humans on OOD data (an important characteristic of a "human-like" model)? This is quantified via *error consistency* (defined in Section 2); which additionally allows us to compare models with each other, asking e.g. which model classes make similar errors. In Figure 4, we compare all models with each other and with humans, asking whether they make errors on the same images. On this particular dataset (sketch images), we can see one big model cluster. Irrespective of whether one takes a standard supervised model, a self-supervised model, an adversarially trained model or a vision transformer, all those models make highly systematic errors (which extends the results of [20, 57] who found similarities between standard vanilla CNNs). Humans, on the other hand, show a very different pattern of errors. Interestingly, the boundary between humans and some data-rich models at the bottom of the figure—especially CLIP (400M images) and SWSL (940M)—is blurry: some (but not all) data-rich models much more closely mirror the patterns of errors that humans make, and we identified the first models to achieve higher error consistency with humans than with other (standard) models. Are these promising results shared across datasets, beyond the sketch images? In Figures 1c and 1d, aggregated results over 17 datasets are presented. Here, we can see that data-rich models approach human-to-human observed consistency, but not error consistency. Taken in isolation, *observed* consistency is not a good measure of image-level consistency since it does not take consistency by chance into account; *error* consistency tracks whether there is consistency beyond chance; here we see that there is still a substantial image-level *consistency gap* between human and machine vision. However, several models improve over vanilla CNNs, especially BiT-M (trained on 14M images) and CLIP (400M images). This progress is non-trivial; at the same time, there is ample room for future improvement.

How do the findings from Figure 4 (showing nearly human-level error consistency for sketch images) and from Figure 1d (showing a substantial consistency gap when aggregating over 17 datasets) fit together? Upon closer inspection, we discovered that there are two distinct cases. On 12 datasets (stylized, colour/greyscale, contrast, high-pass, phase-scrambling, power-equalisation, false colour, rotation, eidolonI, -II and -III as well as uniform noise), the human-machine gap is large; here, more robust models do not show improved error consistency (as can be see in Figure 5a). On the other hand, for five datasets (sketch, silhouette, edge, cue conflict, low-pass filtering), there is a completely different result pattern: Here, OOD accuracy is a near-perfect predictor of error consistency, which means that improved generalisation robustness leads to more human-like errors (Figure 5b). Furthermore, training on large-scale datasets leads to considerable improvements along both axes for standard CNNs. Within those models trained on larger datasets, CLIP scores best, but models with a standard architecture (SWSL: based on ResNet-50 and ResNeXt-101) closely follow suit.

It remains an open question why the training dataset appears to have the most important impact on a model's decision boundary as measured by error consistency (as opposed to other aspects of a model's inductive bias). Datasets contain various shortcut opportunities [14], and if two different models are trained on similar data, they might converge to a similar solution simply by exploiting the same shortcuts—which would also fit well to the finding that adversarial examples typically transfer very well between different models [58, 59]. Making models more flexible (such as transformers, a generalisation of CNNs) wouldn't change much in this regard, since flexible models can still exploit the same shortcuts. Two predictions immediately follow from this hypothesis: (1.) error consistency between two identical models trained on very different datasets, such as ImageNet vs. Stylized-ImageNet, is much lower than error consistency between very different models (ResNet-50 vs. VGG-16) trained on the same dataset. (2.) error consistency between ResNet-50 and a highly flexible model (e.g., a vision transformer) is much higher than error consistency between ResNet-50 and a highly constrained model like BagNet-9 [60]. We provide evidence for both predictions in Appendix B, which makes the shortcut hypothesis of model similarity a potential starting point for future analyses.



(a) 12 "disappointing" datasets

(b) 5 "promising" datasets

Figure 5: Partial failure, partial success: Error consistency with humans aggregated over multiple datasets. Left: 12 datasets where model accuracies exceed human accuracies; here, there is still a substantial image-level consistency gap to humans. Right: 5 datasets (sketch, silhouette, edge, cue conflict, low-pass) where humans are more robust. Here, OOD accuracy is a near-perfect predictor of image-level consistency; especially data-rich models (e.g. CLIP, SWSL, BiT) narrow the consistency gap to humans. Symbols indicate architecture type (○ convolutional, ▽ vision transformer, ◊ human).

5 Discussion

Summary We set out to answer the question: *Are we making progress in closing the gap between human and machine vision?* In order to quantify progress, we performed large-scale psychophysical experiments on 17 out-of-distribution datasets (which will be open-sourced along with Python evaluation code as a benchmark to track future progress). We then investigated models that push the boundaries of traditional deep learning (different objective functions, architectures, and dataset sizes ranging from 1M to 1B), asking how they perform relative to human visual perception. We found that the OOD robustness gap between human and machine vision is closing, as the best models now match or exceed human accuracies. At the same time, an image-level consistency gap remains, however, this gap that is at least in some cases narrowing for models trained on large-scale datasets.

Limitations Model robustness is studied from many different viewpoints, including adversarial robustness [58], theoretical robustness guarantees [e.g. 61], or label noise robustness [e.g. 62]. The focus of our study is robustness towards non-adversarial out-of-distribution data, which is particularly well-suited for comparisons with humans. Furthermore, human and machine vision can be compared in many different ways. This includes comparing against neural data [63, 64], contrasting Gestalt effects [e.g. 65], object similarity judgments [66], or mid-level properties [45] and is of course not limited to studying object recognition. By no means do we mean to imply that our behavioural comparison is the only feasible option—on the contrary, we believe it will be all the more exciting to investigate whether our behavioural findings have implications for other means of comparison!

Discussion We have to admit that we view our results concerning the benefits of increasing dataset size by one-to-three orders of magnitude with mixed feelings. On the one hand, "simply" training standard models on (a lot) more data certainly has an intellectually disappointing element—particularly given many rich ideas in the cognitive science and neuroscience literature on which architectural changes might be required to bring machine vision closer to human vision [67–71]. Additionally, large-scale training comes with infrastructure demands that are hard to meet for many academic researchers. On the other hand, we find it truly exciting to see that machine models are narrowing not just the OOD robustness gap to humans, but that also, at least for some datasets, those models are actually making more human-like decisions on an individual image level; image-level response consistency is a much stricter behavioural requirement than just e.g. matching overall accuracies. Taken together, our results give reason to celebrate partial success in closing the gap between human and machine vision. In those cases where there is still ample room for improvement, our psychophysical benchmark datasets may prove useful in quantifying future progress.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[2] Jinjiang Wang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, 48:144–156, 2018.

[3] Javier Villalba-Diez, Daniel Schmidt, Roman Gevers, Joaquín Ordieres-Meré, Martin Buchwitz, and Wanja Wellbrock. Deep learning for industrial computer vision quality control in the printing industry 4.0. *Sensors*, 19(18):3987, 2019.

[4] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[5] Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, 10:94, 2016.

[6] Garrett B Goh, Nathan O Hodas, and Abhinav Vishnu. Deep learning for computational chemistry. *Journal of computational chemistry*, 38(16):1291–1307, 2017.

[7] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141): 20170387, 2018.

[8] Dan Guest, Kyle Cranmer, and Daniel Whiteson. Deep learning and its application to LHC physics. *Annual Review of Nuclear and Particle Science*, 68:161–181, 2018.

[9] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[11] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

[12] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pages 456–473, 2018.

[13] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, 2019.

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

[15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[16] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. *arXiv preprint arXiv:1905.13549*, 2019.

[17] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[18] David M. Green. Consistency of auditory detection judgments. *Psychological Review*, 71(5):392–407, 1964.

[19] Kristof Meding, Dominik Janzing, Bernhard Schölkopf, and Felix A. Wichmann. Perceiving the arrow of time in autoregressive motion. *Advances in Neural Information Processing Systems (NeurIPS)*, 32: 2303–2314, 2019.

[20] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[22] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[23] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1485–1488, 2010.

[24] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[26] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[27] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[28] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[30] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6 (2):8, 2019.

[31] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better? *arXiv preprint arXiv:2007.08489*, 2020.

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[33] Ross Wightman. PyTorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[34] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[35] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[37] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[38] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *26th International Conference on Computer Communication and Networks*, pages 1–7. IEEE, 2017.

[39] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399–426, 2019.

[40] Yann LeCun. Predictive learning, 2016. URL `https://www.youtube.com/watch?v=0unt2Y4qxQo`.

[41] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2(4):210–219, 2020.

[42] A Emin Orhan, Vaibhav V Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *arXiv preprint arXiv:2007.16189*, 2020.

[43] Talia Konkle and George A Alvarez. Instance-level contrastive learning yields human brain-like representation without category-supervision. *bioRxiv*, 2020.

[44] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael Frank, James DiCarlo, and Daniel Yamins. Unsupervised neural network models of the ventral visual stream. *bioRxiv*, 2020.

[45] Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, pages 1–16, 2021.

[46] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[48] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. *arXiv preprint arXiv:1511.03034*, 2015.

[49] Simran Kaur, Jeremy Cohen, and Zachary C Lipton. Are perceptually-aligned gradients a general property of robust classifiers? *arXiv preprint arXiv:1910.08640*, 2019.

[50] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. *arXiv:1906.09453*, 2019.

[51] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

[52] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning*, pages 9561–9571. PMLR, 2020.

[53] Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International Conference on Machine Learning*, pages 7502–7511. PMLR, 2019.

[54] Peijie Chen, Chirag Agarwal, and Anh Nguyen. The shape and simplicity biases of adversarially robust ImageNet-trained CNNs. *arXiv preprint arXiv:2006.09373*, 2020.

[55] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

[56] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[57] Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. *Advances in Neural Information Processing Systems*, 2019.

[58] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

[59] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

[60] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations*, 2019.

[61] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *arXiv preprint arXiv:1705.08475*, 2017.

[62] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. *arXiv preprint arXiv:1706.00038*, 2017.

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

[64] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, 32:12805–12816, 2019.

[65] Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, pages 1–13, 2021.

[66] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, 2020.

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

[68] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[69] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural Networks*, 111:47–63, 2019.

[70] Fabian H Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S Tolias. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979, 2019.

[71] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *BioRxiv*, 2020.

[72] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[73] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[74] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The abstract states our core findings, the introduction points out the scope of our work; contributions are stated both in the abstract and at the beginning of Section 5, from where we quote here: "We set out to answer the question: *Are we making progress in closing the gap between human and machine vision?* In order to quantify progress, we performed large-scale psychophysical experiments on 17 out-of-distribution datasets (which will be open-sourced along with Python evaluation code as a benchmark to track future progress). We then investigated models that push the boundaries of traditional deep learning (different objective functions, architectures, and dataset sizes ranging from 1M to 1B), asking how they perform relative to human visual perception. We found that the OOD robustness gap between human and machine vision is closing, as the best models now match or exceed human accuracies. At the same time, an image-level consistency gap remains, however, this gap that is at least in some cases narrowing for models trained on large-scale datasets."
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not foresee any negative societal impacts. Instead, our work aims at providing a better understanding of some of the latest machine learning developments, particularly in comparison to human visual perception, which we believe is an important step towards a comprehensive assessment of what we can expect from current models, and under which circumstances their predictions might fail (e.g. due to a lack of robustness).
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] Our results are of experimental rather than theoretical nature.
   (b) Did you include complete proofs of all theoretical results? [N/A] Our results are of experimental rather than theoretical nature.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] During anonymous submission, they are submitted in the sublementary material; afterwards, we will include a link to our github repository.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section D in the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Our results are based on comparing psychophysical results from 90 observers to 52 models; we tested pre-trained models (see Section A in the Appendix). For those specific models (often trained and released by the authors themselves), the our out-of-distribution testing results are deterministic.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section D in the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Section E in the Appendix.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We collected human data in a psychophysical experiment; additionally we include code to benchmark models on out-of-distribution datasets. The code is included in the supplementary material; a github link will be added after the end of the anonymous review period.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We collected human data in a psychophysical experiment.

Participants were asked to review and sign the following declaration of consent (of which they received a copy):

Psychophysical study
Your task consists of viewing visual stimuli on a computer monitor and evaluating them by pressing a key. Participation in a complete experimental session is remunerated at 10 Euros/hour.
Declaration of consent
Herewith I agree to participate in a behavioural experiment to study visual perception. My participation in the study is voluntary. I am informed that I can stop the experiment at any time and without giving any reason without incurring any disadvantages. I know that I can contact the experimenter at any time with questions about the research project.
Declaration of consent for data processing and data publication
*Herewith I agree that the experimental data obtained in the course of the experiment may be used in semianonymised form for scientific evaluation and publication. I agree that my personal data (e.g. name, phone number, address) may be stored in digital form; they will not be used for any other purpose than for contacting me. This personal data will remain exclusively within XXX (*anonymized for submission*) and will not be passed on to third parties at any time.*

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The participants were asked to provide their full name, date of birth, contact information (either address, email or phone number), gender, handedness and whether they use glasses or contact lenses to correct their vision. This personal information was stored in a handwritten document, only available to the experimenter, and will not be released. Therefore, the psychophysical data we release does not contain any personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Section C.1 in the Appendix.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See Section C.2 in the Appendix.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Section C.3 in the Appendix.