

LAION-C: AN OUT-OF-DISTRIBUTION BENCHMARK FOR WEB-SCALE VISION MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) robustness is a desired property of computer vision models. Improving model robustness requires high-quality signals from robustness benchmarks to quantify progress. While various benchmark datasets such as ImageNet-C were proposed in the ImageNet era, most ImageNet-C corruption types are no longer OOD relative to today’s large datasets scraped from the web, which already contain common corruptions such as blur or JPEG compression artifacts. Consequently, these standard benchmarks are no longer well-suited for evaluating OOD robustness in the era of web-scale datasets. Indeed, recent models show saturating scores on ImageNet-era OOD benchmarks, indicating that it is unclear whether models trained on web-scale datasets truly become better at OOD generalization or whether they have simply been exposed to the test distortions during training. To address this, we here introduce LAION-C as a benchmark alternative for ImageNet-C. LAION-C consists of six novel distortion types across five severity levels designed to be OOD, even for web-scale datasets such as LAION. In a comprehensive evaluation of state-of-the-art models, we find that the LAION-C dataset poses significant challenges to contemporary models. We additionally conducted a psychophysical experiment to evaluate the difficulty of our proposed corruptions for human observers, enabling a comparison of models to lab-quality human robustness data. We observe a paradigm shift in OOD generalization: from humans outperforming models, to the best models now matching or outperforming the best human observers.

1 INTRODUCTION

Vision models have been a cornerstone of modern machine learning, driving breakthroughs in diverse applications. In recent years, large-scale vision models such as vision transformers (Dosovitskiy et al., 2021) and ConvNeXt (Liu et al., 2022), trained on expansive web-scale datasets like LAION (Schuhmann et al., 2022), have pushed the boundaries of performance on standard benchmarks. However, the continued advancement and reliable evaluation of these models depends on the availability of datasets that effectively challenge model robustness and generalization capabilities.

ImageNet-C (Hendrycks & Dietterich, 2019) has long stood as the de facto standard for OOD evaluation, particularly for models trained on ImageNet (Russakovsky et al., 2015). It contains images that are systematically different from those in ImageNet, meaning that models trained on ImageNet must robustly generalize to perform well on ImageNet-C. Previous work (e.g., Hendrycks & Dietterich, 2019) found that OOD generalization is not trivial to achieve: Many vision models do indeed struggle with OOD datasets like ImageNet-C even if they perform well on ImageNet. Hence, these types of unfamiliar inputs are crucial for evaluating the robustness of machine learning models since they are indicative of performance on unexpected input; a challenge that many deployed models face. Modern models trained on much larger web-scale datasets, e.g., CLIP (Radford et al., 2021), exhibit much better performance on classic OOD datasets than IN-trained models, potentially suggesting that they have learned robust representations which better generalize to unseen data.

However, as modern training datasets are scaled well beyond ImageNet, existing OOD benchmarks might not be truly OOD with respect to web-scale datasets anymore. OOD datasets such as ImageNet-C were explicitly created to be OOD with respect to the most popular dataset at that time: ImageNet. ImageNet-C contains images with corruptions potentially relevant for practical ap-

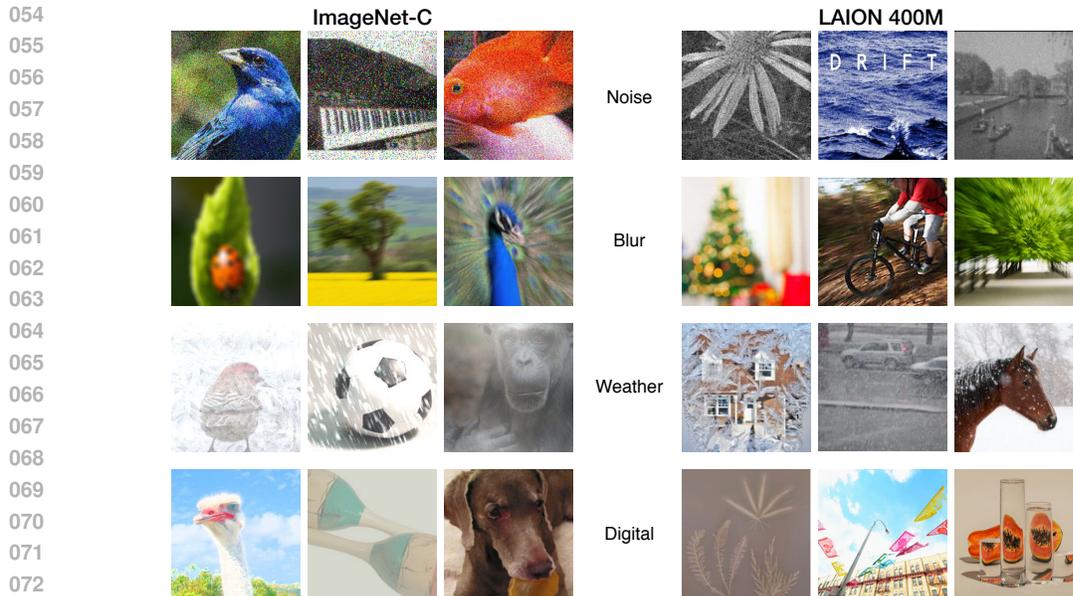


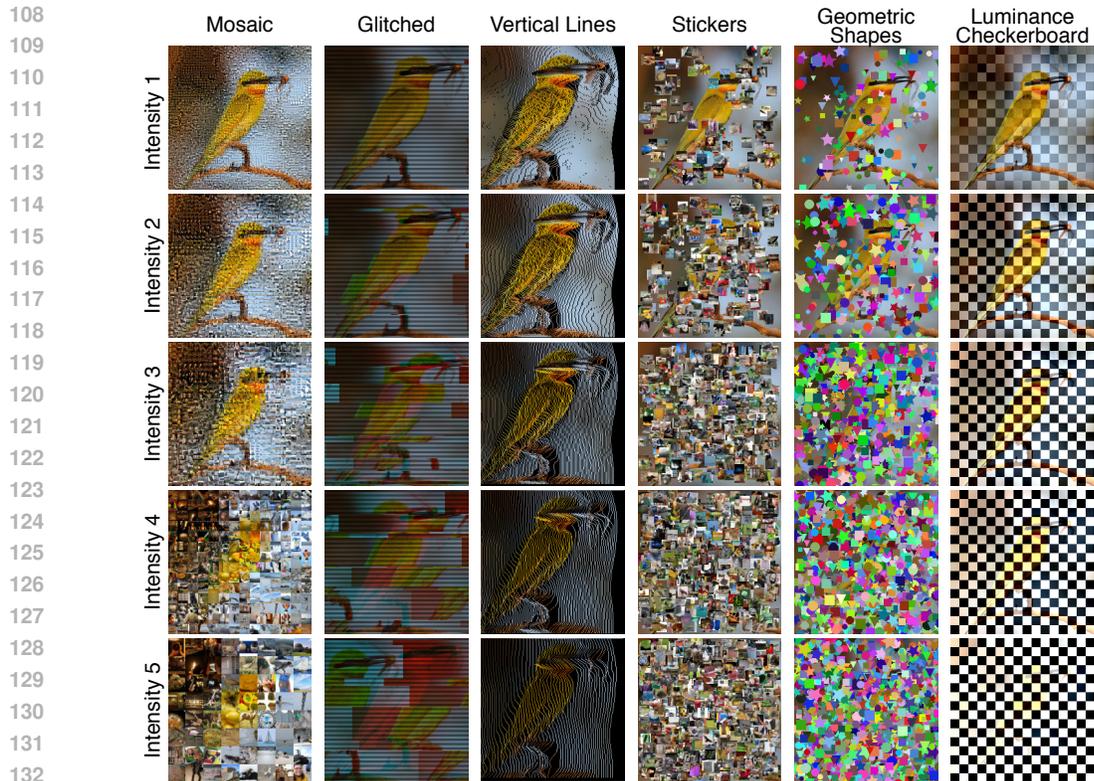
Figure 1: **ImageNet-C corruptions are not out-of-distribution (OOD) for web-scale datasets like LAION-400M.** Exemplary corrupted images from ImageNet-C (left) are similar to LAION-400M samples (right). Each row shows example corruptions and dataset images for one ImageNet-C corruption category (Noise, Blur, Weather, Digital). The presence of these distortions in web-scale datasets indicates the need for an OOD benchmark in the era of web-scale vision models.

lications but (by design) not contained in ImageNet and, thus, OOD. However, with the change of the reference dataset from ImageNet to web-scale datasets such as LAION, these corruptions might no longer be OOD. For example, many images in LAION are blurry—not by deliberate design, but because LAION images were not sampled from a few websites with (implicit) quality standards, like ImageNet images were, but from almost any publicly accessible website online. Simply put, models trained on LAION might have seen the types of corruption on which they are tested with ImageNet-C. For another type of OOD benchmark, namely distribution shifts defined by the style of an image, recent work empirically shows that such datasets are not OOD but overlap with LAION-400M (Mayilvahanan et al., 2023; 2024). This raises a central question: Are modern vision models genuinely improving in terms of OOD generalization, or are they simply trained on datasets that already contain the corruptions, essentially testing in-distribution rather than OOD generalization? This distinction is crucial because if these modern models were not truly more robust than standard models, they might also not perform better on the real OOD test data one might face in practice.

Given the importance of OOD generalization in practice, there is a pressing need for a new benchmark that more effectively evaluates the OOD robustness of state-of-the-art foundation models: an OOD dataset for the era of web-scale vision models. Our **contributions** are as follows:

1. Given that existing OOD datasets are often no longer OOD for models trained on web-scale datasets, we introduce LAION-C, a **novel benchmark** dataset with six manually designed corruption types and 16 superclasses to evaluate the robustness of web-scale vision models.
2. We conduct a comprehensive performance analysis of various models on LAION-C and report a robust human OOD generalization baseline obtained through **psychophysical experiments** with 19 participants, collecting 11,400 trials in a highly controlled laboratory environment.
3. The resulting data serves as an OOD benchmark for current and future models, enabling not only an assessment of their generalization ability on truly OOD data but also providing insights into the **discrepancies between human and machine perception**, observing a paradigm shift in OOD generalization: from humans outperforming models to the best models now matching or outperforming the best human observers.

A detailed related work section can be found in [Appx. A.1](#).



133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

Figure 2: **LAION-C distortions, intended to be OOD even for web-scale datasets.** This figure illustrates the six LAION-C distortions at five intensity levels. Best viewed on screen.

2 METHODS

2.1 CONSTRUCTING NEW OOD DISTORTIONS

As described in the introduction and depicted in Fig. 1, ImageNet-C is not OOD for models trained on large-scale datasets. Given the limitations of existing benchmarks like ImageNet-C, we develop a novel dataset specifically designed to challenge these foundation models more rigorously. Our dataset introduces six carefully designed, fully synthetic distortions that models have not encountered during training. These distortions are designed to be OOD even for web-scale datasets (as supported by quantitative evidence presented later). Hence, models truly need to generalize beyond their training distributions to perform well on this benchmark which we term LAION-C.

Distortions The core idea behind our distortions is to intentionally disrupt visual consistency and perceptual cues that models rely on for image classification, such as texture (Geirhos et al., 2019). Following ImageNet-C, each distortion consists of five different *intensity levels*. The distortions capture a range of visual challenges ranging from disruptions of local image information to more global structural alterations, as described below and illustrated in Fig. 2.

- **Mosaic:** The original image is broken down into smaller tiles, each replaced by a chromatically similar picture. This patchwork creates a mosaic effect that disrupts edges and textures while introducing contextually irrelevant information.
- **Glitched:** The original image undergoes an artistic digital corruption with horizontal lines overlaying shifted image segments and color channel shifts. This dislocates the global contextual structure of the image. While the concept of such glitchy images has been explored in earlier work (Kaufmann et al., 2019), our transformation introduces a more intense corruption.

- **Vertical Lines:** The original image is deconstructed into bent vertical line segments. This distortion retains the original colors but strips away local information, disrupting the contours and edges of the image and introducing visual discontinuity.
- **Geometric Shapes:** The original image is overlaid with overlapping geometric figures such as squares, circles, and stars. This visual clutter introduces local noise that obscures the main object, like the Kaleidoscope corruption from Kaufmann et al. (2019).
- **Stickers:** The original image is augmented with assorted image patches. This addition of visual elements masks features of the primary object.
- **Luminance Checkerboard:** The original image is divided into a grid, with the luminance of each cell altered in a checkerboard pattern. The stark luminance contrast between adjacent tiles and artificial grid boundaries makes this distortion challenging.

We intend to build a challenging dataset that has the potential to guide the future development of vision models. Our dataset incorporates these tougher and less common distortions to simulate the difficulty of OOD scenarios that models might encounter in real-world applications. We tune the intensity levels of each distortion such that either humans or a contemporary vision model (ViT-B) achieve chance performance on the highest intensity level, i.e. no model is expected to perform well on the hardest levels. The other intensity levels are chosen so that we can observe a gradual decline in accuracy, ensuring that the distortions are sufficiently challenging.

These distortions are then applied to a carefully curated subset of images from the ImageNet validation dataset. To contextualize model performance, we later want to compare it to human performance. As human evaluations on datasets with hundreds of classes cannot be scaled to sufficiently many participants, we follow previous work (Geirhos et al., 2018) and simplify the task to a 16-class classification problem. We extract 285 ImageNet-classes to form 16 superclasses, namely ball, bird, boat, bottle, butterfly, car & truck, cat, chair, dog, fish, fruit, instrument, primate, snake, timekeeping, and tool. For robust statistical analysis, our dataset comprises 273 images for each superclass. This data size selection allows us to ensure that a 3% difference in the performance between models is statistically significant. Our dataset serves as a proxy for the unforeseen OOD environments future models must handle, advancing the state of robustness evaluation. Additionally, we manually filter the dataset to ensure that none of the images in one superclass contain objects from another class or require specific cultural knowledge for classification. This process helps to avoid ambiguous ground-truth labels.

2.2 MEASURING MODEL PERFORMANCE

We use the generated datasets to evaluate the performance of a suite of 58 vision models. Our selection includes models trained on large-scale web datasets and fine-tuned on ImageNet-1k, such as Vision Transformers (ViT) (Dosovitskiy et al., 2021), ConvNeXt (Liu et al., 2022), and EVA (Fang et al., 2023; 2024). For comparison, we also evaluate the performance of smaller-scale model families such as ResNet (He et al., 2016) and MobileNet (Howard, 2017) and large-scale models trained only on ImageNet-1k. Additionally, we also evaluate GPT-4o (OpenAI, 2024) and Gemini 1.5 Pro (Team et al., 2024) on a representative subset of LAION images. See Tab. 6 for a complete list of all models we evaluate. To address the imbalance caused by distinct numbers of subclasses within each superclass, we compute the average probability values across subclasses for each superclass, a method first suggested by Geirhos et al. (2018). This adjustment mitigates biases introduced by the varying subclass distributions, ensuring a more accurate model performance evaluation.

2.3 COLLECTING HUMAN PERFORMANCE VIA LAB EXPERIMENTS

To explore the discrepancies between human and machine perception, we design a psychophysical experiment to gather human classification data on the augmented images. This experiment builds on previous paradigms (Geirhos et al., 2018; 2021) to ensure consistency and comparability. In the experiment, 19 human subjects are briefly presented with a distorted image and are asked to classify it into one of 16 classes, reminiscent of how a DNN might be evaluated on a classification task (in contrast to e.g. open response paradigms, where participants could give arbitrary textual responses). Participants were recruited from the university student body, and screened for normal or corrected-to-normal vision and absence of color blindness. The experiment was conducted in a controlled

216 dark environment using a 22" VIEWPixx 3D light LCD monitor, with stimuli presented centrally
 217 at a fixed viewing distance to ensure foveal presentation. The classification task was divided into
 218 two warm-up blocks and ten main experiment blocks, with each block containing images from 16
 219 superclasses. Participants were given 2.5 s to view each image, followed by a 2 s response window
 220 to classify the image by clicking on a set of icons. To motivate high performance, a monetary bonus
 221 was awarded for surpassing fixed, pre-determined performance thresholds for each block. Further
 222 methodological details are provided in [Appx. A.2](#).

224 2.4 QUANTIFYING HUMAN-MACHINE ALIGNMENT VIA ERROR CONSISTENCY.

226 To quantify the alignment between human and machine visual perception, we adopt the error consis-
 227 tency metric proposed in [Geirhos et al. \(2020b\)](#). Error consistency, denoted as $\kappa \in [-1, 1]$, measures
 228 the degree of agreement between the classification mistakes of two observers. In brief, κ takes on
 229 a value of 1.0 if two observers are perfectly consistent, i.e. if they make classification mistakes on
 230 exactly the same images. Two independent binomial observers that agree no more than expected by
 231 chance will result in a κ of 0, while two maximally inconsistent observers will have a κ of -1. See
 232 [Appx. A.3](#) or [Geirhos et al. \(2020b\)](#) for a detailed explanation of the metric.

234 3 RESULTS

237 3.1 How OOD is LAION-C?

238 Now that we have outlined the construction of our
 239 LAION-C dataset, we empirically evaluate whether
 240 it is indeed OOD relative to the large-scale image
 241 datasets used to train modern vision models. Rigor-
 242 ously quantifying how OOD a test dataset is with re-
 243 spect to some training dataset requires a precise defi-
 244 nition of the test and training domain ([Mayilvahanan
 245 et al., 2024](#)). As the distribution shifts introduced
 246 by the distortions of our proposed LAION-C and
 247 ImageNet-C are fuzzy in nature, we use three tools
 248 to compare the OOD-ness of our proposed dataset
 249 to the OOD-ness of ImageNet-C. First, we use a quali-
 250 tative assessment. By searching for the name and re-
 251 lated concepts of ImageNet-C corruptions, we easily
 252 find LAION samples with visual distortions akin to
 253 those of ImageNet-C samples (see [Fig. 1](#)).

254 Second, we use the difficulty of a test dataset (mea-
 255 sured by the performance that models trained on a
 256 reference dataset yield on the test dataset) as a proxy
 257 for how much the test dataset differs from the refer-
 258 ence dataset. Here, the reasoning is that if a test dataset can be solved almost perfectly by a model,
 259 it means that either the model has great generalization skills or the test dataset is not strictly OOD.
 260 If, at the same time, another dataset has much greater difficulty according to the same models, the
 261 second dataset is likely more OOD than the first. For the sake of comparability, we here use a ver-
 262 sion of ImageNet-C restricted to the same 16 superclasses that were used for LAION-C, where we
 263 implemented the ImageNet-C augmentations through the code by [Michaelis et al. \(2019\)](#). Indeed, a
 264 comparison of the performance achieved by our suite of models (see [Fig. 3](#)) suggests that LAION-C
 is more OOD to LAION than ImageNet-C is.

265 Third, we use the FID ([Heusel et al., 2017](#); [Kynkäänniemi et al., 2022](#)) to quantify the difference
 266 between LAION and ImageNet-C and LAION-C, respectively. Specifically, we employ a CLIP-
 267 trained ViT-B as feature encoder and use the implementation by [Parmar et al. \(2022\)](#) to calculate
 268 FID-scores. In line with the previous evidence, we find a FID of ≈ 70 between LAION and LAION-
 269 C, which is substantially higher than that between LAION and ImageNet-C (≈ 40). This means that
 features of LAION are closer to those of ImageNet-C than those of LAION-C, again highlighting the

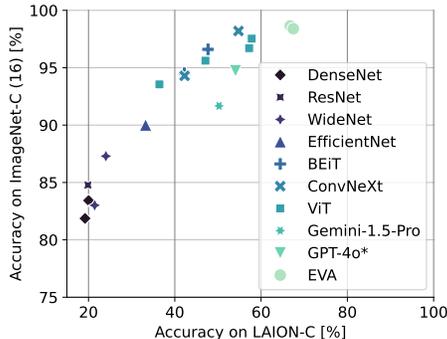


Figure 3: **Performance Divergence of Models on LAION-C and ImageNet-C 16 class.** Evaluating models on the 16-class versions of ImageNet-C and LAION-C produces a plateaued performance on ImageNet-C, while LAION-C still yields a high variance across models.

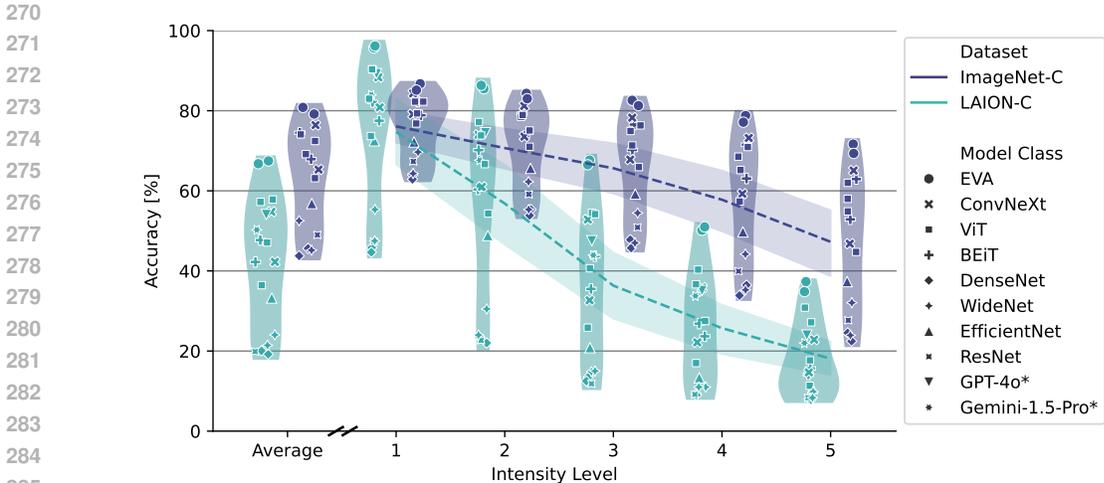


Figure 4: **LAION-C poses a greater challenge to model robustness than ImageNet-C.** We plot distortion intensity against each model’s average accuracy. Visual foundation models evaluated on ImageNet-C maintain high accuracy, with minimal drop across increasing intensity levels. On our LAION-C dataset, the models exhibit a sharper decline in accuracy, highlighting the benchmark’s effectiveness in measuring model robustness.

greater OOD-ness of LAION-C. In summary, we have presented three different kinds of evidence suggesting that LAION-C is more OOD than ImageNet-C to LAION.

3.2 MACHINE PERFORMANCE

In Fig. 4, we compare model performance on ImageNet-C against performance on LAION-C. Evidently, the average model performance on ImageNet-C stays above or close to 60%, indicating that current models are increasingly adept at handling the distortions in ImageNet-C. This observation reinforces our hypothesis that the challenge presented by ImageNet-C may no longer be sufficiently difficult to rigorously test the robustness of modern models.

In contrast, models achieve much lower accuracy on LAION-C on average and exhibit more inter-model variability. This showcases our dataset’s ability to uncover nuances that remain hidden on more saturated benchmarks. These performance differences are particularly obvious at higher intensity levels, illustrating LAION-C’s potential to serve as a more challenging and insightful benchmark for evaluating robustness.

We also provide a detailed breakdown of non-averaged, dataset-specific results in Fig. 10. We observe significant variability in the performance of different vision models across various datasets and distortion levels, highlighting the diversity in model robustness. These results further highlight the effectiveness of our datasets in eliciting different responses from models of different architectures, parameter scales, and training data sizes. This diversity is particularly valuable for understanding which model designs are more robust to specific types of distortions, offering insights that are beneficial for advancing the state-of-the-art model robustness.

In Tab. 1, we present a comprehensive evaluation of our suite of models on LAION-C. We report each model’s top-1 accuracy on the (undistorted) ImageNet validation set as a baseline (*Clean Accuracy*) and the average top-1 accuracy on LAION-C averaged across distortion types and intensity levels (*LAION-C*). We then break the latter down into the six distortion types. This enables a thorough comparison of model performance, highlighting which architectures generalize best.

3.3 IS LAION-C A PROXY FOR MORE REALISTIC DISTRIBUTION SHIFTS?

To demonstrate that model performance on LAION-C is indicative of real-world performance despite the highly synthetic nature of our corruptions, we analyze the correlations between model

Table 1: **LAION-C benchmark results.** Numbers show the top-1 accuracy in percent. *ImageNet* refers to model accuracy on the (uncorrupted) ImageNet validation set, with values sourced from the timm leaderboard (Wightman, 2024). For each corruption, we report the mean top-1 accuracy across all intensity levels, with *LAION-C* as the overall benchmark metric (averaged across corruption types). GPT-4o and Gemini 1.5 Pro were evaluated on 48,000 images, 100 for each class. For full model names and descriptions, see Tab. 6 in the Appendix.

| Model | ImageNet | LAION-C | Mosaic | Vertical | Glitched | Luminance | Geometric | Stickers |
|------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| EVA-G-P14-560-M30M-IN22K | 89.8 | 67.5 | 48.8 | 53.6 | 70.8 | 97.2 | 81.0 | 53.4 |
| EVA02-L-P14-448-MIM-M38M-IN22K | 90.1 | 66.8 | 53.6 | 58.2 | 78.2 | 93.6 | 76.4 | 40.6 |
| ViT-H-P14-336-CLIP-LAION-IN12K | 88.6 | 57.3 | 45.2 | 51.2 | 69.8 | 88.2 | 64.4 | 24.6 |
| ViT-L-P14-224-CLIP-OpenAI-IN12K | 88.3 | 57.8 | 52.6 | 49.8 | 68.2 | 98.6 | 55.4 | 22.4 |
| ViT-B-P32-384-CLIP-LAION-IN12K | 85.4 | 36.4 | 36.8 | 35.2 | 35.8 | 54.0 | 37.6 | 19.2 |
| ViT-B-P16-224-AugReg-IN21K | 85.5 | 47.1 | 46.4 | 42.8 | 62.0 | 71.4 | 42.4 | 17.6 |
| BEiT-v2-L-P16-224-IN1K | 87.4 | 47.7 | 52.4 | 44.8 | 63.2 | 70.2 | 11.8 | 43.8 |
| BEiT-v2-B-P16-224-IN1K | 85.6 | 42.2 | 46.2 | 40.4 | 52.6 | 68.2 | 11.4 | 34.6 |
| ConvNeXt-XXL-CLIP-LAION-IN1K | 88.6 | 54.8 | 53.0 | 53.4 | 71.8 | 77.4 | 52.2 | 20.8 |
| ConvNeXt-B-CLIP-LAION-AugReg-IN12K | 87.6 | 42.3 | 37.6 | 43.8 | 44.4 | 54.2 | 50.4 | 23.2 |
| WRN101-2-TV-IN1K | 78.8 | 21.4 | 30.4 | 28.4 | 22.0 | 22.8 | 18.2 | 6.8 |
| WRN50-2-RACM-IN1K | 81.5 | 24.0 | 26.8 | 21.4 | 17.0 | 45.0 | 24.6 | 9.4 |
| RN50-A1-IN1K | 81.2 | 19.9 | 28.0 | 18.8 | 20.8 | 23.4 | 21.2 | 7.0 |
| EFF-B3-RA2-IN1K | 82.3 | 33.2 | 32.4 | 31.8 | 40.2 | 45.2 | 37.6 | 12.2 |
| DN201-TV-IN1K | 77.3 | 19.2 | 28.6 | 26.2 | 13.2 | 23.2 | 16.8 | 7.2 |
| DN161-TV-IN1K | 77.3 | 20.0 | 31.0 | 26.8 | 15.2 | 25.2 | 15.4 | 6.6 |
| GPT-4o | - | 54.1 | 42.8 | 45.4 | 65.1 | 80.1 | 54.2 | 36.5 |
| Gemini 1.5 Pro | - | 50.2 | 34.9 | 37.0 | 46.2 | 84.4 | 60.9 | 38.1 |
| Best human observer | - | 55.2 | 58.0 | 55.3 | 78.7 | 63.4 | 40.4 | 35.7 |
| Average human observer | - | 47.0 | 50.8 | 43.6 | 71.0 | 53.1 | 34.3 | 29.1 |

Table 2: **LAION-C is challenging but can be solved by fine-tuning on the exact distortions.** We compare the performance of ViT-H-P14-336-CLIP-LAION-IN12K before and after fine-tuning it on ImageNet-1k training images with LAION-C corruptions. As the performance after fine-tuning is high, this means that LAION-C, although challenging, remains solvable as it retains enough signal when applying distortions.

| Accuracy [%] | Mosaic | Vertical Lines | Glitched | Luminance | Geometric | Stickers |
|--------------|--------|----------------|----------|-----------|-----------|----------|
| Before | 45.2 | 51.2 | 69.8 | 88.2 | 64.4 | 24.6 |
| After | 79.0 | 93.5 | 95.8 | 97.7 | 90.2 | 61.0 |

accuracy on LAION-C and on several well-established OOD benchmark datasets such as ImageNet-R, ImageNet-A and ImageNet-Sketch in Tab. 5. Clearly, models that achieve high accuracy on LAION-C are also robust to other distribution shifts. However, our main goal is not to measure real-world performance, but to measure a model’s ability to generalize beyond its training data, which requires a truly OOD test set - a requirement that might even be incompatible with the requirements of a real-world distribution shift.

3.4 CAN LAION-C BE SOLVED?

Given the low performance of current state-of-the-art models on LAION-C, one might wonder whether LAION-C is simply impossible to solve because the distortions destroy all information necessary for correct classification of the images. To disprove this hypothesis and highlight the validity of LAION-C as a benchmark for evaluating model robustness, we conduct a fine-tuning experiment to assess whether the challenges posed by LAION-C are solvable at all. Specifically, we fine-tune a ViT-Huge model, which was originally pretrained with a CLIP-objective on LAION-2B and then fine-tuned on ImageNet-22k and ImageNet-1k. For this experiment, we use a custom dataset sub-sampled from the ImageNet-1K training set and augmented with the distortions introduced in LAION-C. This dataset consists of over 336,000 images uniformly sampled across the 16 superclasses defined for LAION-C.

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

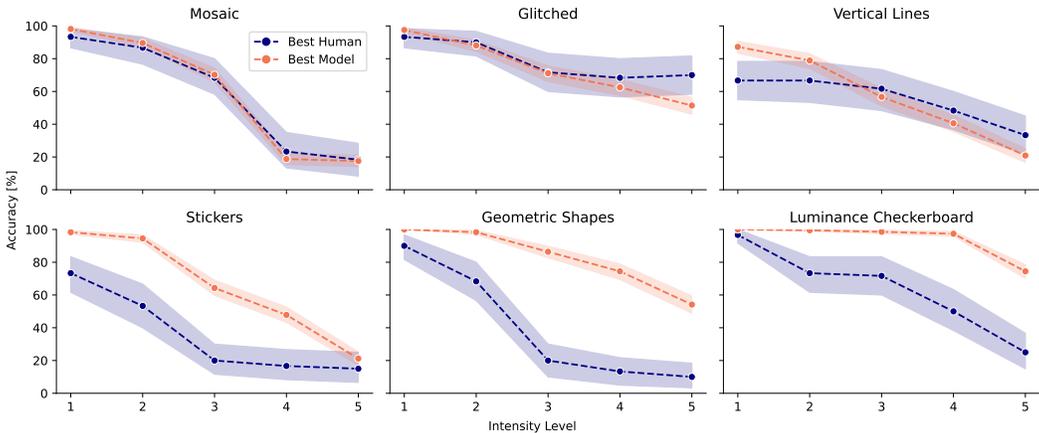


Figure 5: **Human vs. machine accuracy on all distortions.** For each LAION-C distortion, we plot the distortion intensity against the accuracy of the best human and the best model in this condition. The shaded regions indicate the 95% confidence intervals around the means. On the Mosaic, Glitched and Vertical Lines distortions, humans and machines perform similarly, whereas the best model vastly outperforms the best human observer on the Stickers, Geometric Shapes, and Luminance Checkerboard distortions.

As shown in Tab. 2, fine-tuning the model results in substantial accuracy gains, which define an upper bound on LAION-C accuracy that no normal model can be expected to achieve. Notably, these accuracy gains are particularly pronounced on higher-intensity distortions, as detailed in Tab. 4, where accuracy is broken down by distortion intensity. The fine-tuned model likely achieves such good performance by employing un-human-like (or “spurious”) features, but the purpose of this experiment is *not* to suggest that fine-tuning on LAION is a sensible approach to improve OOD robustness, but to quantify how much learnable signal is left. LAION-C provides meaningful robustness tests without being intractable, making it a valuable tool for a more comprehensive evaluation of model performance under difficult conditions.

3.5 HUMAN AND MACHINE VISION DISCREPANCY

Accuracy Differences. In Fig. 5, we summarize how our suite of models performs in terms of classification accuracy, compared to the human participants in our psychophysical experiment. We report the best performances, since averages would be unfairly influenced by some older models we included as points of comparison. In Fig. 10, we provide a more detailed breakdown of performance by model. While human observers still outperform most vision models on images with Mosaic or Glitched distortions, the best models match (or even slightly surpass) human performance. For distortions involving occlusion and luminance manipulations, the vision models typically achieve higher accuracy than humans. Overall, current state-of-the-art vision models now match or even outperform human observers in OOD scenarios, including on our synthetic distortions, which they likely have never encountered during training—a stark contrast to just a few years ago, when humans were still vastly outperforming models (Geirhos et al., 2018; Dodge & Karam, 2019; Taori et al., 2020; Jang & Tong, 2024).

Occlusion and Luminance Manipulations. For distortions involving occlusions, such as Stickers and Geometric Shapes, models usually match or exceed human performance (see second row of Fig. 5). One possible hypothesis is that this can be attributed to the robustness that models develop after e.g., masked image modeling (MIM) (Fang et al., 2023; 2024). The fact that models perform so much better than humans on partially occluded images implies that models use different features than humans. For example, for the Stickers distortion, certain ViT models outperform humans, likely due to their ability to focus on those parts of the image background that remain visible despite the occlusions. As shown in Fig. 1, the stickers occlude nearly the entire image on higher intensity levels, and little to no meaningful object information is retained. Nevertheless, certain models are still able to correctly classify the image based on subtle background cues. This indicates that while mod-

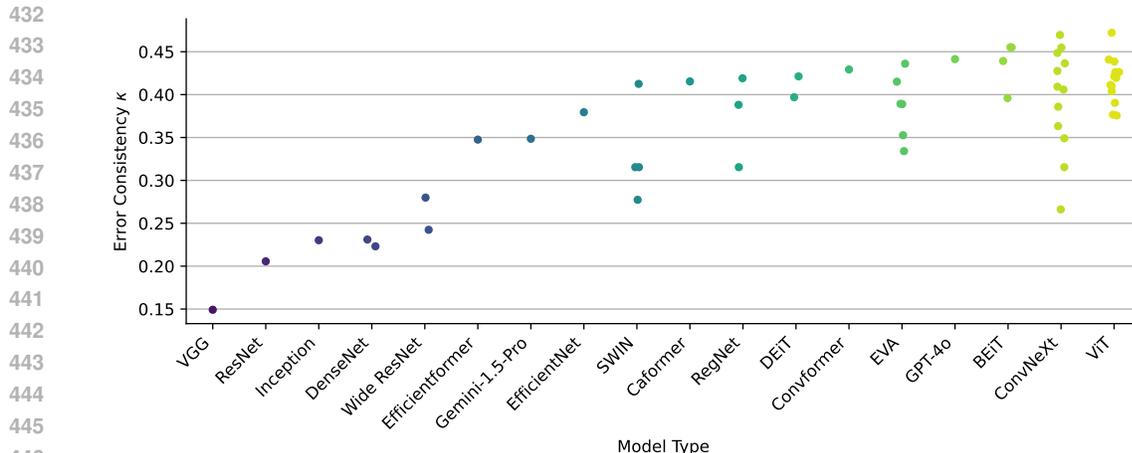


Figure 6: **Humans and models make different mistakes.** We analyze the agreement of error patterns between different families of vision models (see Tab. 6 for a complete list) and human observers. The error consistency (κ) could theoretically achieve a maximum value of 1, but in line with earlier work (Geirhos et al., 2021), the EC values range between 0 and 0.4, indicating that behavioral differences between humans and machines are still quite large. Marker colors encode model families.

els are performing well, they may be doing so by leveraging unintended shortcuts (Geirhos et al., 2020a), such as exploiting background information, when faced with severely occluded images. For the Luminance Checkerboard distortion, we observe that models from the ViT and EVA families outperform humans by a large margin. This advantage could potentially stem from their architectural features, such as self-attention mechanisms and patch-based processing (Fang et al., 2023; Dosovitskiy et al., 2021), which enable them to extract meaningful information from both light and dark regions independently, as well as handle subtle luminance variations. These capabilities give them a clear edge over humans and older models.

Performance on Complex Distortions. When analyzing more complex distortions such as Mosaic, Vertical Lines, and Glitched images (first row of Fig. 5), we observe that human performance is generally on par with the best-performing models. Especially at greater intensity levels, humans perform competitively, e.g., outperforming all models for the strongest Vertical Lines distortions. As we show in Fig. 10, the gap between humans and older models like the ResNet variants is particularly large on these complex distortions. However, modern model classes demonstrate substantial progress, approaching human-level performance even at higher intensity levels. While some margin for improvement remains, the narrowing gap suggests that achieving human-level robustness on classification tasks is no longer the primary challenge for state-of-the-art architectures.

Human-Machine Error Consistency on LAION-C. For a more fine-grained analysis of the behavioral agreement between models and human observers, we calculate error consistency as described in Sec. 2. As illustrated in Fig. 6, there is a high degree of variability in error consistency between human observers and different vision models ranging from 0 to 0.4. This indicates that while model families such as ViT and EVA rival or surpass human performance, they are approaching the task utilizing different strategies than humans, demonstrating less human-like behaviors. The observed value range matches the one found in previous work for older models and different image data (Geirhos et al., 2021). This again suggests that while recent developments have boosted model performance, these models have not become more human-like, as they follow alternative strategies.

4 DISCUSSION

Summary. Given that existing OOD benchmarks are often no longer OOD for models trained on web-scale datasets like LAION since distortions like blur and digital corruptions are commonplace

486 on the web, we here introduce LAION-C. LAION-C is a benchmark designed to evaluate the ro-
487 bustness and generalization capabilities of modern vision models trained on web-scale datasets. Our
488 empirical results demonstrate that LAION-C is more challenging for a representative suite of vision
489 models than the existing ImageNet-C benchmark, particularly at higher distortion intensity levels.
490 Our thorough human evaluation in a highly controlled psychophysical laboratory totaling 11,400
491 trials shows that *the best models often outperform even the best human observers*. While they do not
492 always follow similar strategies (as indicated by error consistency analysis), this reassuring finding
493 indicates that the best models have indeed substantially progressed in their ability to handle unex-
494 pected input and are not just getting better on in-distribution distortions. Given that the LAION-C
495 benchmark dataset, by virtue of its construction, serves as a better proxy for a model’s ability to
496 recognize objects despite an unexpected distortion, we recommend it as an OOD benchmark for
497 current and future web-scale vision models.

498 **Limitations.** While we have shown that LAION-C can effectively reveal shortcomings in model
499 robustness, we have not yet fully explored why certain models underperform on specific distortions.
500 Although our empirical results provide valuable insights, further investigation is required to clarify
501 which visual cues the models rely on under different conditions. Such an analysis could inform the
502 development of new inductive biases or architectural improvements, since a better understanding of
503 these mechanisms could lead to improvements in both model interpretability and robustness. Given
504 our current focus on introducing the dataset, this was not fully addressed here, but could be an area
505 for future exploration. Furthermore, it is an open question what the performance limit on LAION-C
506 looks like. Since fine-tuning models on LAION-C results in significant performance gains, particu-
507 larly at higher distortion levels, there clearly is potential for optimization through advanced training
508 techniques. However, how to further improve generalization across OOD scenarios, especially to
509 enhance the models’ ability to handle the novel distortions presented by LAION-C, remains an open
510 question that warrants further exploration. To retain its value as an OOD benchmark, LAION-C
511 should not be used as a training or fine-tuning dataset (except for analysis purposes).

512 **Conclusion and outlook.** Just a few years ago, early investigations into generalization abilities of
513 deep neural networks showed humans vastly outperforming the best models (Geirhos et al., 2018;
514 Dodge & Karam, 2019). Fast-forwarding to today, LAION-C shows that the best models either
515 match or outperform human performance on challenging OOD distortions. This finding is reassuring
516 in the light of growing concerns about the quality of existing evaluation datasets, including the
517 concern that OOD datasets like ImageNet-C may no longer serve their original purpose in the era
518 of web-scale training datasets. Our findings indicate that the often *super-human performance* of
519 modern models is achieved through *super-human strategies*: Models use a variety of image cues—
520 including, in all likelihood, background pixels to perform well on some of our challenging datasets.
521 Given their high performance across the board, they no longer rely on a single strategy that fails
522 when faced with a test case that distorts a particular image cue. This marks a paradigm shift in OOD
523 generalization: From humans outperforming models to models outperforming humans, from relying
524 on a single strategy to a diverse set of multiple robust strategies, and from OOD benchmarking
525 measuring progress towards human-like object recognition to better performance now indicating
526 super-human (in other words, *less human-like*) vision models.

527 CODE AND DATASET AVAILABILITY

528
529 We will publicly release the dataset and the code to generate distortions and evaluate models af-
530 ter acceptance of the paper. During the anonymous review period, the code is available from the
531 supplementary material.

533 ETHICS STATEMENT

534
535 We confirm that all experimental procedures involving human subjects in our study had IRB ap-
536 proval. In addition, we ensured that all participants gave informed consent prior to their inclusion in
537 the study. Detailed information was provided to each participant beforehand, outlining the study’s
538 purpose, procedures and benefits, ensuring they were fully informed before agreeing to participate.
539 Participants were also informed that they could abort the study at any time, without incurring any
negative consequences. Experimental data and contact information for the participants was stored

540 in accordance with GDPR. Participants were compensated with an hourly base rate of 12 EUR and
 541 received bonus payments based on classification performance, as is customary in psychophysical
 542 experiments, so that the final reimbursements exceeded the local minimum wage.

544 REFERENCES

- 545 Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers.
 546 In *ICLR*, 2022.
- 547 Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are
 548 we done with imagenet? *arXiv preprint*, 2020. URL <https://arxiv.org/abs/2006.07159>.
- 549 Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological mea-*
 550 *surement*, 1960.
- 551 Samuel Dodge and Lina Karam. Human and dnn classification performance on images with quality
 552 distortions: A comparative study. *ACM Transactions on Applied Perception (TAP)*, 2019.
- 553 Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lind-
 554 say, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The
 555 neuroconnectionist research programme. *Nature Reviews Neuroscience*, 2023.
- 556 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
 557 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszko-
 558 reit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at
 559 scale. In *ICLR*, 2021.
- 560 Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong
 561 Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale.
 562 In *CVPR*, 2023.
- 563 Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A
 564 visual representation for neon genesis. *Image and Vision Computing*, 2024.
- 565 Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cyber-*
 566 *netics*, 1975.
- 567 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,
 568 Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64
 569 (12):86–92, 2021.
- 570 Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and
 571 Felix A. Wichmann. Generalisation in humans and deep neural networks. In *NeurIPS*, 2018.
- 572 Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and
 573 Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias im-
 574 proves accuracy and robustness. In *ICLR*, 2019.
- 575 Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel,
 576 Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature*
 577 *Machine Intelligence*, 2020a.
- 578 Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-
 579 trial behaviour of cnns and humans by measuring error consistency. In *NeurIPS*, 2020b.
- 580 Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge,
 581 Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and
 582 machine vision. In *NeurIPS*, 2021.
- 583 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-
 584 nition. In *CVPR*, 2016.

- 594 Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common
595 corruptions and perturbations. In *ICLR (Poster)*, 2019.
- 596
- 597 Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul
598 Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical
599 analysis of out-of-distribution generalization. In *ICCV*, 2021a.
- 600 Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial
601 examples. In *CVPR*, 2021b.
- 602
- 603 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.
604 Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*,
605 2017.
- 606 Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applica-
607 tions. *arXiv preprint*, 2017. URL <https://arxiv.org/abs/1704.04861>.
- 608
- 609 Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected
610 convolutional networks. In *CVPR*, 2017.
- 611
- 612 Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to
613 blur in convolutional neural networks. *Nature Communications*, 15(1):1989, 2024.
- 614 Max Kaufmann, Daniel Kang, Yi Sun, Steven Basart, Xuwang Yin, Mantas Mazeika, Akul Arora,
615 Adam Dziedzić, Franziska Boenisch, Tom Brown, Jacob Steinhardt, and Dan Hendrycks. Testing
616 robustness against unforeseen adversaries. *arXiv preprint*, 2019. URL <https://arxiv.org/abs/1908.08016>.
- 617
- 618 Mario Kleiner, David Brainard, Denis Pelli, Allen Ingling, Richard Murray, and Christopher Brou-
619 sard. What’s new in psychtoolbox-3. *Perception*, 2007.
- 620
- 621 Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of
622 imagenet classes in fréchet inception distance. *arXiv preprint*, 2022. URL <https://arxiv.org/abs/2203.06026>.
- 623
- 624 Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie.
625 A convnet for the 2020s. In *CVPR*, 2022.
- 626
- 627 Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel.
628 Does clip’s generalization performance mainly stem from high train-test similarity? *arXiv*
629 *preprint*, 2023. URL <https://arxiv.org/abs/2310.09562>.
- 630
- 631 Prasanna Mayilvahanan, Roland S. Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila
632 Juhos, Matthias Bethge, and Wieland Brendel. In search of forgotten domain generalization. In
633 *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL <https://openreview.net/forum?id=Bc2p8T4V32>.
- 634
- 635 Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexan-
636 der S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detec-
637 tion: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.
- 638
- 639 Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and cor-
640 ruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*,
641 34:3571–3583, 2021.
- 642
- 643 Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith.
Human alignment of neural network representations. In *ICLR*, 2023.
- 644
- 645 OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed:
646 2024-11-19.
- 647
- Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in
GAN evaluation. In *CVPR*, 2022.

- 648 Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling
649 with vector-quantized visual tokenizers. *arXiv preprint*, 2022. URL <https://arxiv.org/abs/2208.06366>.
650
651
- 652 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
653 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
654 models from natural language supervision. In *ICML*, 2021.
- 655 Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers
656 generalize to imagenet? In *ICML*, 2019.
657
- 658 Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng
659 Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual
660 recognition challenge. *International journal of computer vision*, 2015.
- 661 Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Ko-
662 hitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which
663 artificial neural network for object recognition is most brain-like? *BioRxiv*, 2018.
664
- 665 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi
666 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An
667 open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.
- 668 Chase T. Glitch this, 2020. URL [https://github.com/TotallyNotChase/](https://github.com/TotallyNotChase/glitch-this)
669 [glitch-this](https://github.com/TotallyNotChase/glitch-this). MIT License.
- 670 Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural
671 networks. In *ICML*, 2019.
672
- 673 Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig
674 Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*,
675 2020.
- 676 Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer,
677 Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal under-
678 standing across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
679
- 680 Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of
681 human visual perception? *Annual Review of Vision Science*, 2023.
- 682 Ross Wightman. The timm (pytorch image models) leaderboard. [https://huggingface.co/](https://huggingface.co/spaces/timm/leaderboard)
683 [spaces/timm/leaderboard](https://huggingface.co/spaces/timm/leaderboard), 2024.
684
- 685 Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training
686 procedure in timm. *arXiv preprint*, 2021. URL <https://arxiv.org/abs/2110.00476>.
- 687 Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J
688 DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual
689 cortex. *Proceedings of the national academy of sciences*, 2014.
690
- 691 Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint*, 2016. URL
692 <https://arxiv.org/abs/1605.07146>.
- 693 Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Bayer. Scaling vision transformers.
694 In *CVPR*, 2022.
695
696
697
698
699
700
701

A APPENDIX

A.1 RELATED WORK

OOD generalization ability of vision models. As deep learning has advanced to the point where models can reliably generalize to data that matches their training distribution or even exceed the quality of the original labels (Beyer et al., 2020), OOD-robustness, as measured by specifically designed OOD test sets, has moved to the center stage of computer vision research. In particular, ImageNet-C (Hendrycks & Dietterich, 2019), a dataset containing images from the test-set of ImageNet to which various fairly natural corruptions such as blurring and pixelation were applied, has long been the gold standard for assessing OOD-performance, to the point where data augmentations proposed to increase OOD robustness were found to only work well because they are similar to the ImageNet-C corruptions (Mintun et al., 2021). In contrast, ImageNet-R (Hendrycks et al., 2021a) instead provides a more complex distribution shift by collecting different renditions of the target classes such as sculptures and paintings, instead of photos. A more subtle distribution shift which still caused considerable drops in model performance for ImageNet-trained models, was proposed by Recht et al. (2019). They collected ImageNetV2, a new test set for ImageNet that should theoretically not differ from the ImageNet test set at all, because it was collected with a very similar methodology, but revealed that models do perform slightly worse on ImageNetV2 than on the original test set. Hendrycks et al. (2021b) proposed two other OOD-test sets which do not rely on synthetic image manipulations but instead consist of natural images that are in some sense OOD relative to ImageNet, either by virtue of displaying object classes not present in ImageNet (ImageNet-O) or by showing an object of an ImageNet-class in a scene that is weird enough to fool most models (ImageNet-A). What all of these datasets have in common is that, by design, they provide distribution shifts *relative to ImageNet*. But with the rapid evolution of deep learning, these traditional benchmarks have become less challenging for state-of-the-art vision models trained on expansive web-scale datasets (Radford et al., 2021). While it is to some degree possible to predict a model’s OOD generalization from its training set performance (Taori et al., 2020), the only reliable measurements of this capability stem from empirical evaluations of models on OOD datasets. Our work addresses this need by introducing LAION-C, a dataset that incorporates novel and complex synthetic distortions tailored to challenge even advanced vision systems.

Advancement in visual foundation models The rise of visual foundation models such as Vision Transformers (ViT) (Dosovitskiy et al., 2021), ConvNeXt (Liu et al., 2022) and BeiT (Bao et al., 2022) has redefined what constitutes standard performance across many visual tasks. These improvements in performance partially stem from architectural innovations and parameter optimization, but were mostly powered by the effective leveraging of unprecedented dataset sizes (Zhai et al., 2022). However, because visual foundation models were trained on web-scale datasets, the extent of their generalization capability remains underexplored.

Comparing human vs. machine perception. Deep Neural Networks were originally conceived as models of human vision (Fukushima, 1975) and were found to be the best available models for neuronal activity in the primate visual cortex (Yamins et al., 2014), even if not trained for this task. Today, there is a growing body of research dedicated to evaluating the adequacy of neural networks as behavioral models of human core object recognition (Doerig et al., 2023; Schrimpf et al., 2018; Wichmann & Geirhos, 2023; Muttenthaler et al., 2023). Building upon the findings of Geirhos et al. (2021), who illustrate the narrowing of the behavioral gap between humans and machines in terms of their error consistency, our study further explores this dynamic utilizing LAION-C. We conducted a systematic analysis of differences in perception between human and machine observers, and assessed if the behavioral gap is closing further, as well as highlighting the persistent cognitive differences between humans and machines.

A.2 EXPERIMENT PROCEDURE AND PARTICIPANT INCENTIVES

Participant recruitment and setup. We recruited 20 participants (10 female) from the university student body via mailing lists. All participants were screened to ensure normal or corrected vision and no color blindness, and gave informed consent to participate. One participant was excluded post-hoc due to reporting extreme tiredness. Our experiments were conducted in a darkened cabin,

756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809



Figure 7: **Interface presented to participants.** This figure illustrates the icon layout as displayed to participants during the study. The grid is adapted from (Geirhos et al., 2018), while most of the categories and therefore symbols are different.

using a 22" VIEWPixx 3D light LCD monitor (VPixx Technologies, Saint-Bruno, Canada) at a refresh rate of 120 Hz (scanning backlight mode on). The screen measures 484×302 mm, at a resolution of 1920×1200 pixels. Stimuli were presented foveally in the center of the screen, with a viewing distance of 65 cm, resulting in 5° of visual angle. In line with earlier experiments, the background was set to a grey value of 0.454 in the $[0, 1]$ range. A chin rest was used to maintain a fixed viewing distance and angle. The experiment was implemented using the Psychophysics Toolbox (Kleiner et al., 2007, version 3.0.12) in MATLAB (Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States) using a 12-core desktop computer (AMD HD7970 graphics card "Tahiti" by AMD, Sunnyvale, California, United States) running Kubuntu 14.04 LTS.

The entire classification task, including both the warm-up and main experiment phases, was organized into 12 blocks. In each block, participants were shown images from the 16 superclasses for 2.5 seconds—a duration empirically determined to balance efficient overall experiment length with sufficient exposure time allowing for viewing and consideration time. After each image, the 16 corresponding class icons were displayed on screen, allowing participants 2 seconds to classify each image into one of these categories. The icons were organized in a layout that roughly grouped them by size and general category (e.g., animals and objects), as illustrated in image Fig. 7. To encourage responses rather than leaving selections blank, a message was displayed at the top of the screen 0.75 second before icon display time ended, prompting participants to make a choice. At the end of each block, if a participant surpassed the 90% accuracy threshold calibrated using internal baseline performance data, they received an encouraging on-screen message ("Congratulations! You just earned some extra money!") along with a \$0.50 bonus per block to incentivize higher performance.

Warm-up session and main experiment. The experiment began with a 10-minute warm-up session, allowing participants to familiarize themselves with the icon layouts and the classification task procedure through two practice blocks. Each practice block contained 45 images, with one block consisting of clean images and the other of augmented images. To avoid test-time adaptation, the augmentations used during the warm-up phase differed from those in the actual trials. The images used for the practice trials were also a subsample of the ImageNet validation dataset, but did not overlap with those used in the main experiment.

Following the warm-up, the main experiment proceeded consisting of 10 blocks, each block comprising 60 images. Each set of 5 blocks was augmented using a consistent method, with a balanced distribution across different intensity levels and superclasses. To avoid fatigue, participants were allowed an unlimited amount of time to rest between blocks and encouraged to rest their eyes or accommodate elsewhere.

810 A.3 ERROR CONSISTENCY

811
812 Here, we provide a more detailed explanation of error consistency (EC), summarizing Geirhos et al.
813 (2020b). The EC between two observers which both classified a sequence of samples is obtained
814 by first using the necessary ground-truth labels to assess which images each observer classified cor-
815 rectly. A trial increases the agreement if both observers solved it correctly, or if they both failed
816 (and decreases it if only one of them got the trial correct while the other one failed). One then calcu-
817 lates how much more agreement was observed between the two observers relative to the agreement
818 expected by chance. This is done by calculating Cohen’s Kappa (Cohen, 1960), which is defined
819 as $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the (empirically measured) proportion of agreement-trials and p_e is the
820 (theoretical) expected agreement when modeling both observers as independent binomial observers.
821 κ takes on values between -1 and 1 , with a higher κ signifying higher levels of agreement, and a
822 κ of 0 indicating that a pair of observers does not agree more frequently than one would expect by
823 chance, given their marginal correctness probabilities.

824 In this work, we calculate the error consistency between model responses and human classification
825 data. To do this, we first collect all human responses. Since each human participant saw a fresh set
826 of stimuli, we thus obtain exactly one human response per image. We then calculate each model’s
827 EC to this list of human responses.

828 A.4 AUGMENTATION DESIGNS

- 829
- 830 • **Mosaic:** The image is divided into an $n \times n$ grid, where each tile is replaced by a patch from the
831 ImageNet validation set whose average color best matches that of the tile. The values of n per
832 intensity level are:
 - 833 – Level 1: $n = 4$
 - 834 – Level 2: $n = 6$
 - 835 – Level 3: $n = 8$
 - 836 – Level 4: $n = 16$
 - 837 – Level 5: $n = 28$
 - 838 • **Glitched:** Alternating rows are replaced with black pixels to create a scan line effect. Pixel shifts
839 and color channel offsets are applied to random regions as follows:
 - 840 – Level 1: Shift up to 8% of image width in 4 regions, ± 4 pixel channel offset.
 - 841 – Level 2: Shift up to 32% of image width in 8 regions, ± 8 pixel channel offset.
 - 842 – Level 3: Shift up to 50% of image width in 10 regions, ± 10 pixel channel offset.
 - 843 – Level 4: Shift up to 128% of image width in 16 regions, ± 16 pixel channel offset.
 - 844 – Level 5: Shift up to 200% of image width in 20 regions, ± 20 pixel channel offset.

845 The implementation is inspired by T (2020)

- 846 • **Vertical Lines:** The image is divided into vertical sections, each of which is further subdivided
847 into smaller sections along the y-axis (called y-steps). A vertical line is drawn within each y-step
848 with a slight x-offset based on the intensity level. The line color is determined by the average
849 color of that section. The parameters for each intensity level are:
 - 850 – Level 1: 224 vertical sections, with 1-pixel steps along the y-axis.
 - 851 – Level 2: 178 vertical sections, with 2-pixel steps along the y-axis.
 - 852 – Level 3: 112 vertical sections, with 4-pixel steps along the y-axis.
 - 853 – Level 4: 84 vertical sections, with 6-pixel steps along the y-axis.
 - 854 – Level 5: 60 vertical sections, with 8-pixel steps along the y-axis.
- 855 • **Luminance Checkerboard:** The image is divided into a 14×14 grid, and the luminance of each
856 tile is altered in a checkerboard pattern. The luminance variation per intensity level is:
 - 857 – Level 1: ± 50 .
 - 858 – Level 2: ± 50 – 100 .
 - 859 – Level 3: ± 100 – 125 .
 - 860 – Level 4: ± 125 – 150 .
 - 861 – Level 5: ± 150 – 255 .

Table 3: **Occlusion ratio of objects in Stickers and Geometric Shapes distortions.** We calculated the object occlusion ratio for the Stickers and Geometric Shapes corruptions as an additional quantitative measurement of the distortion strength.

| Intensity Level | Geometric Shapes (%) | Stickers (%) |
|-----------------|----------------------|--------------|
| 1 | 61.88 | 65.83 |
| 2 | 72.51 | 76.52 |
| 3 | 85.35 | 86.19 |
| 4 | 90.16 | 89.54 |
| 5 | 93.21 | 91.63 |

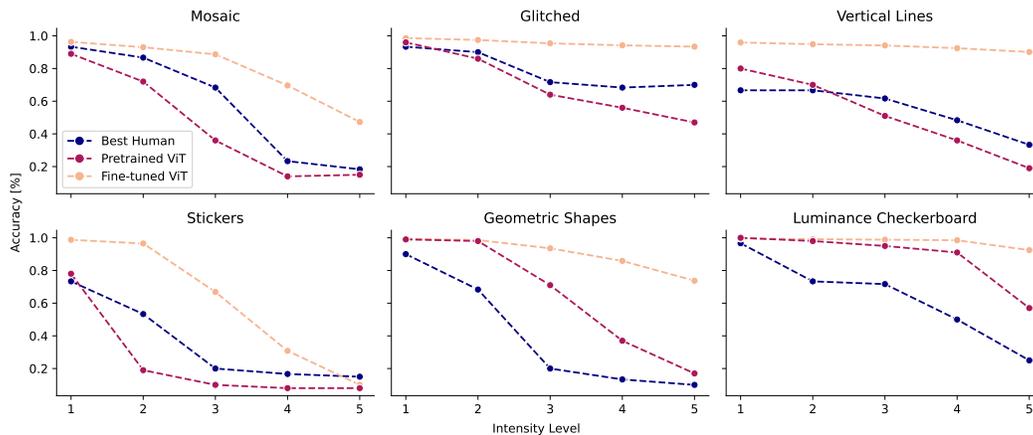


Figure 8: **LAION-C can be solved.** For every distortion, we plot the accuracy of our reference model (ViT-H-P14-336-CLIP-LAION-IN12K) before and after fine-tuning, in comparison to the best human participant for reference. Most distortions can be learned perfectly, only the Stickers and Mosaic distortions might have been too difficult at the highest intensity levels. Further performance gains might be possible with more careful fine-tuning.

- **Stickers:** 16×16 pixel image patches from the ImageNet validation set are randomly placed with uniform probability across the image. The number of patches per intensity level is:

- Level 1: 100 patches.
- Level 2: 200 patches.
- Level 3: 400 patches.
- Level 4: 600 patches.
- Level 5: 1200 patches.

For an estimate of the occlusion ration of the objects per intensity level, see [Tab. 3](#).

- **Geometric Shapes:** Random geometric shapes (triangle, square, star, circle) of varied colors and sizes are introduced. The number of shapes per intensity level is:

- Level 1: 150 shapes.
- Level 2: 300 shapes.
- Level 3: 600 shapes.
- Level 4: 800 shapes.
- Level 5: 1000 shapes.

For an estimate of the occlusion ration of the objects per intensity level, see [Tab. 3](#).

Table 4: **Model (ViT) Accuracy Before and After Fine-Tuning on LAION-C.** The high accuracies after fine-tuning indicate that even though the dataset is challenging, there is, in principle, enough signal left to perform well on LAION-C.

| | Intensity Level | Accuracy Before (%) | Accuracy After (%) |
|------------------------|-----------------|---------------------|--------------------|
| Mosaic | 1 | 89.0 | 96.3 |
| | 2 | 71.9 | 93.0 |
| | 3 | 35.8 | 88.7 |
| | 4 | 14.3 | 69.6 |
| | 5 | 14.7 | 47.7 |
| Vertical Lines | 1 | 79.9 | 95.9 |
| | 2 | 70.1 | 94.9 |
| | 3 | 50.8 | 94.1 |
| | 4 | 36.1 | 92.4 |
| | 5 | 19.4 | 90.0 |
| Glitched | 1 | 95.9 | 98.6 |
| | 2 | 86.2 | 97.5 |
| | 3 | 63.6 | 95.4 |
| | 4 | 55.6 | 94.2 |
| | 5 | 47.1 | 93.4 |
| Luminance Checkerboard | 1 | 99.7 | 99.6 |
| | 2 | 98.4 | 99.2 |
| | 3 | 95.1 | 98.8 |
| | 4 | 90.7 | 98.5 |
| | 5 | 56.6 | 92.5 |
| Geometric Shapes | 1 | 30.9 | 99.4 |
| | 2 | 11.2 | 98.6 |
| | 3 | 6.7 | 93.6 |
| | 4 | 6.6 | 85.9 |
| | 5 | 6.3 | 73.7 |
| Sticker | 1 | 97.3 | 98.8 |
| | 2 | 77.8 | 96.5 |
| | 3 | 28.7 | 63.7 |
| | 4 | 14.9 | 31.8 |
| | 5 | 8.1 | 14.3 |

A.5 ACCURACY

To demonstrate the value of LAION-C as a benchmark for evaluating model robustness, we analyze how model performance on LAION-C correlates with that on ImageNet-C. Grounding our comparison in models that have demonstrated a baseline level of robustness on well-established benchmarks, we apply a threshold to include 40 models that achieved at least 60% accuracy on ImageNet-C.

As shown in Fig. 9, the majority of data points lie above the identity line representing performance alignment on LAION-C and ImageNet-C. The gradual slope of the data points, combined with their positioning, indicates that models generally perform better on ImageNet-C, while their performance on LAION-C is more dispersed and often substantially lower.

This broader distribution of performance highlights that LAION-C introduces more challenging distortions, prompting models to exhibit greater variability in robustness. The moderate Kendall’s tau coefficient ($\tau = 0.66$) between the models’ performances on LAION-C and ImageNet-C further underscores this, indicating notable pairwise differences in how models rank across these two benchmarks, unearthing vulnerabilities that are less pronounced on ImageNet-C. These results demonstrate the necessity of LAION-C as a complementary benchmark for a more comprehensive evaluation of model robustness.

A.6 BREAKDOWN OF MODEL PERFORMANCE

Evaluating VLMs To evaluate GPT-4o (OpenAI, 2024) and Gemini 1.5 Pro (Team et al., 2024) on LAION-C, we decided to test a random subsample of the full dataset, consisting of 100 images per category, which were then tested on all corruptions and intensity levels, resulting in a total of

48,000 images. For ImageNet-C, we limited ourselves to only 10 images per class, to get an initial ballpark estimate of performance.

We employed the following system prompt, in line with our human experiments, during which participants were also shown examples:

System Prompt:

You are an image-recognition API.

You are always asked to classify the main object of images into one of 16 mutually exclusive categories.

In some images, the distortion may be so strong that you might not recognize anything. If you're unsure, provide your best guess - you always have to pick exactly one of the 16 categories.

The 16 categories are: primate, dog, cat, bird, fish, snake, butterfly, fruit, boat, vehicle, chair, ball, bottle, instrument, timekeeper, tool.

Here is a list of characterizations of every such category:

primate: a primate, like e.g. monkeys, chimpanzees, Orang-Utans etc.

dog: a dog, like e.g. german shepherd, labrador, golden retriever etc.

cat: a cat, like e.g. domestic cat, lion, cheetah etc.

bird: a bird, like e.g. songbird, eagle, chicken etc.

fish: a fish, like e.g. trout, shark, whale etc.

snake: a snake, like e.g. viper, cobra, seasnake etc.

butterfly: a butterfly, like e.g. monarch, cabbage butterfly, ringlet etc.

fruit: a fruit, like e.g. apple, orange, pineapple etc.

boat: a boat, like e.g. ship, gondola, fireboat etc.

vehicle: a vehicle, like e.g. truck, van, sports car etc.

chair: a chair, like e.g. bench, throne, couch etc.

ball: a ball (or a person playing with a ball), like e.g. soccer ball, football, tennis ball etc.

bottle: a bottle, like e.g. water bottle, jug, pill bottle etc.

instrument: a musical instrument (or a person playing an instrument), like e.g. sax, flute, harp etc.

timekeeper: a timekeeper, like e.g. clock, watch, sundial etc.

tool: a tool (or a person using a tool), like e.g. hammer, power drill, chainsaw etc.

Since you are an API, you always respond with minimal messages that contain exactly one word, which is the category name.

User Prompt:

What is the main object in this image? Categories are: primate, dog, cat, bird, fish, snake, butterfly, fruit, boat, vehicle, chair, ball, bottle, instrument, timekeeper, tool.

Table 5: **Model performance on LAION-C correlates with other OOD benchmarks.** We evaluated a suite of 18 models (ViT and ConvNeXt variants trained on either LAION-2B or ImageNet) on IN-C, IN-A, IN-R, IN-Sketch and IN-Val. Evidently, the correlations between all of these OOD benchmarks are high, indicating that they measure related quantities.

| | IN-C | LAION-C | IN-A | IN-R | IN-Sketch | IN-val |
|-----------|------|---------|------|------|-----------|--------|
| IN-C | 1.00 | 0.86 | 0.88 | 0.91 | 0.86 | 0.90 |
| LAION-C | 0.86 | 1.00 | 0.69 | 0.70 | 0.81 | 0.72 |
| IN-A | 0.88 | 0.69 | 1.00 | 0.99 | 0.94 | 1.00 |
| IN-R | 0.91 | 0.70 | 0.99 | 1.00 | 0.93 | 0.99 |
| IN-Sketch | 0.86 | 0.81 | 0.94 | 0.93 | 1.00 | 0.95 |
| IN-val | 0.90 | 0.72 | 1.00 | 0.99 | 0.95 | 1.00 |

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

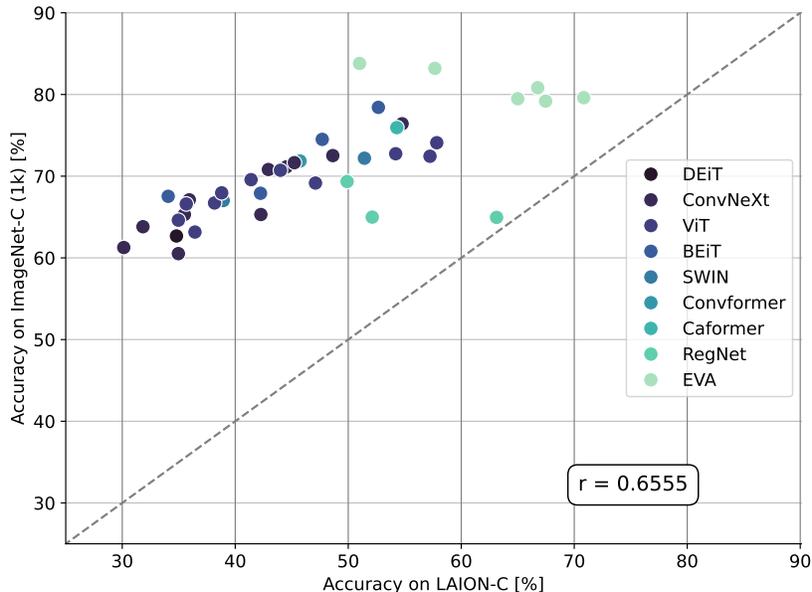


Figure 9: **Performance Divergence of Models on LAION-C and ImageNet-C (1k classes).** The figure illustrates the scattered performance of models across the ImageNet-C and LAION-C dataset, where a Kendall’s tau coefficient (τ) of 0.66 and the shallow slope indicate a dispersed performance on LAION-C. To provide a clearer trend and to better visualize the dispersion, we supplement the suite of models with additional top-performing models sourced from the timm leaderboard (Wightman, 2024), bringing the total number of models to 40 (see Tab. 6 for a complete list).

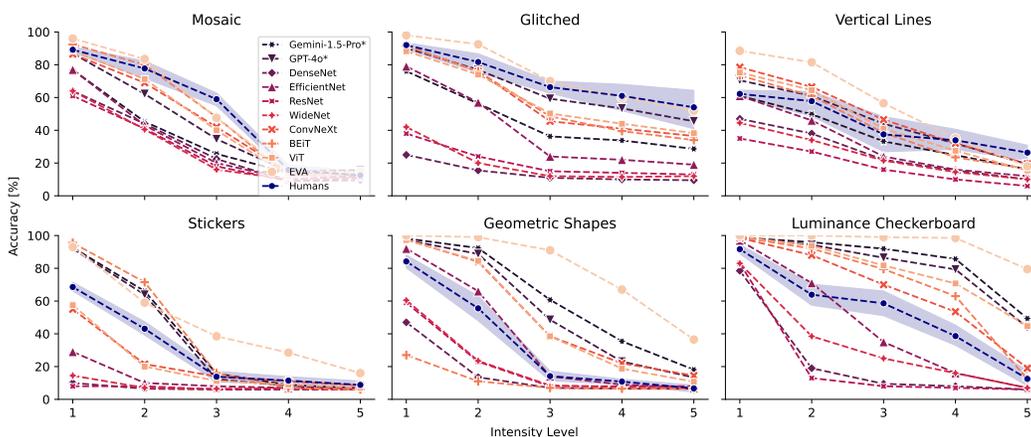


Figure 10: **Model performance on LAION-C.** Analogous to Fig. 5, we relate distortion intensity level to classification accuracy for the different distortions, showing the different models individually. The shaded region around human performance corresponds to the 95% confidence interval, which we omit for the models for better visibility.

A.7 MODELS

Table 6: **Model overview.** For each model used in our evaluation, we show the full model names, as used in timm, an abbreviated name used in the main text and a description of the model. While the first 16 models were used in all analyses and figures, the rest was only used for selective analyses such as Fig. 6.

| Abbreviation | Full Model Name | Description |
|---------------------------------|--|--|
| EVA-G-P14-560-M30M-IN22K | eva_giant_patch14_560.m30m_ft.in22k.in1k | EVA giant model, patch size 14, pre-trained with masked image modeling (MIM) on a Merged-30M dataset, fine-tuned on ImageNet-22k and ImageNet-1k (Fang et al., 2023). |
| EVA02-L-P14-448-MIM-M38M-IN22K | eva02_large_patch14_448.mim.m38m_ft.in22k.in1k | EVA02 large model, patch size 14, pre-trained with masked image modeling (MIM) on a Merged-38M dataset, fine-tuned on ImageNet-22k and ImageNet-1k (Fang et al., 2024). |
| VIT-H-P14-336-CLIP-LAION-IN12K | vit_huge_patch14_clip_336.laion2b_ft.in12k.in1k | Vision Transformer (ViT) huge model, patch size 14, pre-trained on LAION-2B dataset using OpenCLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021). |
| VIT-L-P14-224-CLIP-OPENAI-IN12K | vit_large_patch14_clip_224.openai_ft.in12k.in1k | Vision Transformer large model, patch size 14, pre-trained on WIT-400M using CLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021). |
| VIT-B-P32-384-CLIP-LAION-IN12K | vit_base_patch32_clip_384.laion2b_ft.in12k.in1k | Vision Transformer base model, patch size 32, pretrained on LAION-2B using OpenCLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021). |
| VIT-B-P16-224-AUGREG-IN21K | vit_base_patch16_224.augreg2.in21k_ft.in1k | Vision Transformer base model, patch size 16, trained on ImageNet-21k and fine-tuned on ImageNet-1k (Dosovitskiy et al., 2021). |
| BEITV2-L-P16-224-IN1K | beitv2_large_patch16_224.in1k_ft.in1k | BEiTV2 large model, patch size 16, trained on ImageNet-1k, fine-tuned on ImageNet-22k and ImageNet-1k (Bao et al., 2022; Peng et al., 2022). |
| BEITV2-B-P16-224-IN1K | beitv2_base_patch16_224.in1k_ft.in1k | BEiTV2 base model, patch size 16, trained on ImageNet-1k, fine-tuned on ImageNet-22k and ImageNet-1k (Bao et al., 2022; Peng et al., 2022). |
| CONV-XXL-CLIP-LAION-IN1K | convnext_xlarge_clip_laion2b_soup_ft.in1k | ConvNext xlarge model, pre-trained using OpenCLIP on LAION-2B, fine-tuned on ImageNet-1k (Liu et al., 2022). |
| CONV-B-CLIP-LAION-AUGREG-IN12K | convnext_base_clip_laion2b_augreg_ft.in12k.in1k_384 | ConvNext base model, pre-trained using OpenCLIP on LAION-2B, fine-tuned on ImageNet-12k and ImageNet-1k (Liu et al., 2022). |
| WRN101-2-TV-IN1K | wide_resnet101_2.tv.in1k | Wide ResNet-101 model, trained on ImageNet-1k, with original torchvision model weight (He et al., 2016; Zagoruyko & Komodakis, 2016). |
| WRN50-2-RACM-IN1K | wide_resnet50_2.racm.in1k | Wide ResNet-50 model, trained with RandAugment RACM recipe on ImageNet-1k (He et al., 2016; Zagoruyko & Komodakis, 2016). |
| RN50-A1-IN1K | resnet50.a1.in1k | ResNet-50 model trained on ImageNet-1k (He et al., 2016; Wightman et al., 2021). |
| EFF-B3-RA2-IN1K | efficientnet_b3.ra2.in1k | EfficientNet-B3 model, trained with RandAugment RA2 recipe on ImageNet-1k (Tan & Le, 2019). |
| DN201-TV-IN1K | densenet201.tv.in1k | DenseNet-201, DenseNet pre-trained on ImageNet-1k (Huang et al., 2017). |
| DN161-TV-IN1K | densenet161.tv.in1k | DenseNet-161, DenseNet model pre-trained on ImageNet-1k (Huang et al., 2017). |
| GPT-4o | gpt-4o-2024-08-06 | At the time of writing, the most recent snapshot of OpenAI’s flagship model (OpenAI, 2024). Only evaluated on 48,000 LAION-C samples and 12,000 ImageNet-C samples. |
| Gemini-1.5-Pro | gemini-1.5-pro-002 | At the time of writing, the most recent stable version of Google’s Gemini model (Team et al., 2024). Only evaluated on 48,000 LAION-C samples and 12,000 ImageNet-C samples. |
| | convnextv2_pico.fcmae_ft.in1k | |
| | convnextv2_tiny.fcmae_ft.in22k.in1k | |
| | convnext_base.fb.in22k.ft.in1k | |
| | convnext_large_mlp_clip_laion2b_augreg_ft.in1k_384 | |
| | convnext_large_mlp_clip_laion2b_soup_ft.in12k.in1k_384 | |
| | convnext_tiny.in12k.ft.in1k | |
| | convnext_small.fb.in22k.ft.in1k_384 | |
| | convnext_xlarge.fb.in22k.ft.in1k | |
| | convnext_small.in12k.ft.in1k_384 | |
| | convnextv2_large.fcmae_ft.in22k.in1k_384 | |
| | vit_betwixt_patch16_reg4_gap_256.sbb2.e200.in12k.ft.in1k | |
| | vit_medium_patch16_rope_reg1_gap_256.sbb.in1k | |
| | vit_wce_patch16_reg1_gap_256.sbb.in1k | |
| | vit_medium_patch16_reg4_gap_256.sbb2.e200.in12k.ft.in1k | |
| | vit_medium_patch16_reg4_gap_256.sbb.in12k | |
| | vit_pwee_patch16_reg1_gap_256.sbb.in1k | |
| | vit_betwixt_patch16_rope_reg4_gap_256.sbb.in1k | |
| | vit_betwixt_patch16_reg4_gap_256.sbb.in12k.ft.in1k | |
| | maxxvitv2_rmlp_base_rw_384.sw.in12k.ft.in1k | |
| | vgg19_bn.tv.in1k | |
| | regnety_1280.swag_lc.in1k | |
| | regnety_1280.swag_ft.in1k | |
| | regnety_320.swag_ft.in1k | |
| | inception_v3_ifadv.in1k | |
| | beit_base_patch16_224.in22k.ft.in22k.in1k | |
| | beit_large_patch16_512.in22k.ft.in22k.in1k | |
| | deit3_large_patch16_384.fb.in22k.ft.in1k | |
| | deit_base_distilled_patch16_224.fb.in1k | |
| | swin_base_patch4_window7_224.ms.in22k.ft.in1k | |
| | swin2_base_window12to24_192to384.ms.in22k.ft.in1k | |
| | swin2_large_window12to24_192to384.ms.in22k.ft.in1k | |
| | eva_large_patch14_336.in22k.ft.in1k | |
| | convformer_b36_sail.in22k.ft.in1k_384 | |
| | caformer_b36_sail.in22k.ft.in1k_384 | |
| | efficientformerv2_s2.snap_dist.in1k | |

A.8 DATASHEET FOR LAION-C

As proposed by one of our anonymous reviewers, we here include a Datasheet for LAION-C following the template proposed by Gebru et al. (2021).

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The LAION-C dataset was created to serve as a benchmark for evaluating the robustness and Out-of-Distribution (OOD) generalization of large-scale vision models. It can also be used to study the difference between human and model perception.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Information will be provided upon publication.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Information will be provided upon publication.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the LAION-C dataset represent images grouped into 16 superclasses with various synthetic distortions applied to them at 5 severity levels. Each superclass contains 273 images, and the distortions include mosaic effects, glitched images, vertical lines, geometric shapes, stickers, and luminance checkerboard patterns.

How many instances are there in total (of each type, if appropriate)?

In total, LAION-C consists of 131,040 images. (16 classes \times 273 images \times 6 corruptions \times 5 severity levels.)

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated / verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of the ImageNet validation set and only contains 4,368 of the 50,000 images. As such, LAION-C is not representative of ImageNet, because it only consists of coarse superclasses. This decision was made to facilitate measuring human classification performance on LAION-C, which would not be possible with the fine-grained classes of ImageNet.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of an RGB image, as well as metadata about the ground-truth class, corruption type, and severity level, which are simply part of the filename.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is labeled with its superclass (one of 16) and can be traced back to its original ImageNet class label.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing from individual instances as each image in the dataset is synthetically altered and labeled with the type of distortion and its severity, ensuring comprehensive data for evaluation purposes.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.

1188 The dataset does not contain explicit relation-
 1189 ships between individual instances such as social
 1190 links or ratings since it primarily focuses on
 1191 image recognition and distortion type evaluation
 1192 without any relational context between the im-
 1193 ages.

1194 **Are there recommended data splits (e.g.,
 1195 training, development / validation, test-
 1196 ing)?** If so, please provide a description of
 1197 these splits, explaining the rationale behind
 1198 them.

1199 Since the dataset is primarily used for bench-
 1200 marking purposes, splitting specifics are not pro-
 1201 vided. Essentially, the entire dataset is a valida-
 1202 tion set.

1204 **Are there any errors, sources of noise, or
 1205 redundancies in the dataset?** If so, please
 1206 provide a description.

1207 The dataset is designed to introduce controlled
 1208 noise through synthetic distortions to test model
 1209 robustness. There are no unintentional errors
 1210 or redundancies; all modifications serve the pur-
 1211 pose of benchmark evaluation.

1213 **Is the dataset self-contained, or does
 1214 it link to or otherwise rely on external
 1215 resources (e.g., websites, tweets, other
 1216 datasets)?** If it links to or relies on exter-
 1217 nal resources, a) are there guarantees that
 1218 they will exist, and remain constant, over
 1219 time; b) are there official archival versions of
 1220 the complete dataset (i.e., including the ex-
 1221 ternal resources as they existed at the time
 1222 the dataset was created); c) are there any
 1223 restrictions (e.g., licenses, fees) associated
 1224 with any of the external resources that might
 1225 apply to a future user? Please provide de-
 1226 scriptions of all external resources and any
 1227 restrictions associated with them, as well as
 links or other access points, as appropriate.

1228 The dataset is entirely self-contained.

1230 **Does the dataset contain data that might
 1231 be considered confidential (e.g., data
 1232 that is protected by legal privilege or by
 1233 doctor-patient confidentiality, data that
 1234 includes the content of individuals non-
 1235 public communications)?** If so, please
 1236 provide a description.

1237 The dataset does not contain confidential data as
 1238 it is based on publicly available ImageNet data.

1240 **Does the dataset contain data that, if
 1241 viewed directly, might be offensive, in-**

**sulting, threatening, or might otherwise
 cause anxiety?** If so, please describe why.

The dataset does not contain offensive or dis-
 turbing content as it focuses on visual distortions
 applied to non-sensitive images. Additionally,
 the images sourced from ImageNet are manually
 filtered to exclude any content that could be con-
 sidered disturbing.

Does the dataset relate to people? If not,
 you may skip the remaining questions in this
 section.

Yes, the LAION-C dataset relates to people to
 some extent as it includes images from Image-
 Net, some of which feature human faces and
 figures. While the primary focus of the dataset is
 not on the individuals depicted or on analyzing
 human-specific data, the presence of human im-
 ages means that the dataset does relate to people
 indirectly.

**Does the dataset identify any subpop-
 ulations (e.g., by age, gender)?** If so,
 please describe how these subpopulations
 are identified and provide a description
 of their respective distributions within the
 dataset.

The LAION-C dataset itself does not explic-
 itly identify subpopulations by age, gender, or
 other demographic characteristics as part of its
 core design. However, since it includes im-
 ages from ImageNet, which may contain human
 faces, there is an implicit presence of such de-
 mographic data.

**Is it possible to identify individuals (i.e.,
 one or more natural persons), either di-
 rectly or indirectly (i.e., in combination
 with other data) from the dataset?** If so,
 please describe how.

While the primary intention of the LAION-C
 dataset is not to facilitate the identification of in-
 dividuals, it incorporates images from ImageNet,
 which may include human faces.

**Does the dataset contain data that might
 be considered sensitive in any way (e.g.,
 data that reveals racial or ethnic origins,
 sexual orientations, religious beliefs, po-
 litical opinions or union memberships, or
 locations; financial or health data; bio-
 metric or genetic data; forms of govern-
 ment identification, such as social secu-
 rity numbers; criminal history)?** If so,
 please provide a description.

1242 While the LAION-C dataset primarily features
 1243 synthetic distortions applied to images for technical
 1244 analysis, it includes images sourced from
 1245 ImageNet that may contain human faces. These
 1246 images can indirectly reveal racial or ethnic origins
 1247 due to the diversity of individuals depicted.
 1248 However, there is no explicit focus on collecting
 1249 or analyzing data related to sexual orientations,
 1250 religious beliefs, political opinions, union memberships,
 1251 specific locations, financial or health data, biometric or genetic data,
 1252 government identification numbers, or criminal history. The
 1253 inclusion of human images is incidental and not
 1254 intended for any analysis related to these sensitive
 1255 aspects.

1256 Any other comments?

1257 None.

1261 Collection Process

1262
 1263 **How was the data associated with each instance acquired?** Was the data
 1264 directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey
 1265 responses), or indirectly inferred / derived from other data (e.g., part-of-speech tags, model-
 1266 based guesses for age or language)? If data was reported by subjects or indirectly
 1267 inferred / derived from other data, was the data validated / verified? If so, please describe
 1268 how.

1269 The data for each instance in the LAION-C
 1270 dataset is derived from ImageNet, where images
 1271 are directly observable and not reported by subjects
 1272 or inferred.

1273
 1274 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?

1275 First, 16 sensible high-level classes were selected that the authors deemed suitable for humans to recognize in psychophysical experiments. These classes are: ball, bird, boat, bottle, butterfly, car & truck, cat, chair, dog, fish, fruit, instrument, primate, snake, timekeeping, and tool. Then, 200 classes from the original ImageNet-1k set were selected that can constitute these high-level classes. From the pools of validation set images, 500 images were randomly selected per superclass. These images were then manually filtered to include only images that fall clearly into one of the 16 super-

classes (i.e. an image showing both a ball and a dog would have been filtered out to ensure clean class labels).

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

See previous question. Candidate images from the constituent classes were sampled randomly with uniform probability.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Information will be provided upon publication.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The source dataset for the creation of LAION-C was the 2012 ILSVRC validation set (“ImageNet”) which was collected over several years. The distortions applied in LAION-C were created specifically for benchmarking purposes at the time of dataset development (2023 / 2024), which do not coincide directly with the original image collection periods.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

The original ImageNet dataset underwent various ethical and review processes during its development, details are managed by the original collector for ImageNet.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Only indirectly. LAION-C includes images from ImageNet that feature human faces and figures.

Did you collect the data from the individuals in question directly, or obtain it via

1296 **third parties or other sources (e.g., web-**
1297 **sites)?**

1298 Not applicable.
1299

1300 **Were the individuals in question notified**
1301 **about the data collection?** If so, please
1302 describe (or show with screenshots or other
1303 information) how notice was provided, and
1304 provide a link or other access point to, or oth-
1305 erwise reproduce, the exact language of the
1306 notification itself.

1307 Not applicable.
1308

1309 **Did the individuals in question consent**
1310 **to the collection and use of their data?**
1311 If so, please describe (or show with screen-
1312 shots or other information) how consent was
1313 requested and provided, and provide a link
1314 or other access point to, or otherwise repro-
1315 duce, the exact language to which the indi-
1316 viduals consented.

1317 Not applicable.
1318

1319 **If consent was obtained, were the**
1320 **consenting individuals provided with a**
1321 **mechanism to revoke their consent in the**
1322 **future or for certain uses?** If so, please
1323 provide a description, as well as a link or
1324 other access point to the mechanism (if ap-
1325 appropriate).

1326 Not applicable.
1327

1328 **Has an analysis of the potential impact**
1329 **of the dataset and its use on data sub-**
1330 **jects (e.g., a data protection impact anal-**
1331 **ysis) been conducted?** If so, please pro-
1332 vide a description of this analysis, includ-
1333 ing the outcomes, as well as a link or other
1334 access point to any supporting documenta-
1335 tion.

1336 No specific data protection impact analysis has
1337 been conducted for the LAION-C dataset as
1338 its primary modifications involve applying syn-
1339 thetic distortions like glitches to the images for
1340 technical benchmarking purposes. These alter-
1341 ations do not fundamentally change the nature
1342 of the data regarding privacy or ethical concerns
1343 beyond their original use in ImageNet.

1344 **Any other comments?**

1345 None.
1346
1347
1348

1349 **Preprocessing / cleaning / labeling**

Was any preprocessing / cleaning / la-
beling of the data done (e.g., discretiza-
tion or bucketing, tokenization, part-
of-speech tagging, SIFT feature extrac-
tion, removal of instances, processing of
missing values)? If so, please provide a
description. If not, you may skip the remain-
der of the questions in this section.

Images were resized to 256x256 pixels and
center-cropped to 224x224 pixels, as is common
for ImageNet. Images were filtered manually to
ensure clean labels as described above.

Was the “raw” data saved in addition
to the preprocessed / cleaned / labeled
data (e.g., to support unanticipated fu-
ture uses)? If so, please provide a link or
other access point to the “raw” data.

No, LAION-C only consists of the modified im-
ages, but every filename can be uniquely traced
back to the parent image from the ImageNet val-
idation set, which can be found here: [https://
www.image-net.org/download.php](https://www.image-net.org/download.php)

Is the software used to preprocess /
clean / label the instances available? If
so, please provide a link or other access
point.

Yes, the preprocessing, cleaning, and labeling
of the dataset instances were conducted using
Python. The code used for these processes will
be made available upon publication.

Any other comments?

None.

| |
|-------------|
| Uses |
|-------------|

Has the dataset been used for any tasks
already? If so, please provide a descrip-
tion.

Yes, the LAION-C dataset has been utilized to
evaluate the robustness and out-of-distribution
(OOD) generalization capabilities of large-scale
vision models.

Is there a repository that links to any
or all papers or systems that use the
dataset? If so, please provide a link or other
access point.

Information will be provided upon publication.

What (other) tasks could the dataset be
used for?

Beyond benchmarking vision model robustness, LAION-C could be used in studies investigating the effects of image distortions on human perception.

Is there anything about the composition of the dataset or the way it was collected and preprocessed / cleaned / labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Given that the base images in the LAION-C dataset are sourced from ImageNet, which is already publicly available, the additional risk for harm is negligible.

Are there tasks for which the dataset should not be used? If so, please provide a description.

We would not recommend using the LAION-C dataset for fine-tuning machine learning models, due to dataset size.

Any other comments?

None.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The LAION-C dataset will be made publicly available, allowing for distribution to third parties outside of the originating entity.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

Upon publication, the dataset will be published via Zenodo.

When will the dataset be distributed?

The dataset will be distributed upon publication.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and / or under applicable terms of use (ToU)? If so, please describe this license and / or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

LAION-C will be available under a CC BY-NC 4.0 license, allowing non-commercial use with proper attribution only, to ensure compliance with the original ImageNet license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

The original ImageNet data is subject to terms of access that limit its use to non-commercial research and educational purposes only. The full terms of access can be found here: <https://www.image-net.org/download.php>

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Since the images are modified ImageNet images, the restrictions of the ImageNet license apply.

Any other comments?

None

Maintenance

Who will be supporting / hosting / maintaining the dataset?

Information will be provided upon publication.

How can the owner / curator / manager of the dataset be contacted (e.g., email address)?

Information will be provided upon publication.

1404 **Is there an erratum?** If so, please provide
1405 a link or other access point.

1406 There is not an explicit erratum as for now.

1408 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe
1409 how often, by whom, and how updates will
1410 be communicated to users (e.g., mailing list,
1411 GitHub)?

1412 Information will be provided upon publication.

1417 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so,
1418 please describe these limits and explain how
1419 they will be enforced.

1420 Not applicable (beyond agreements made for
1421 ImageNet).

1427 **Will older versions of the dataset continue to be supported / hosted / main-**

tained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Should newer versions of the dataset be created, older versions will continue to be available via Zenodo.

If others want to extend / augment / build on / contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated / verified? If so, please describe how. If not, why not? Is there a process for communicating / distributing these contributions to other users? If so, please provide a description.

We encourage other researchers to build on LAION-C, for example by contributing their own corruptions. While there is no automatic mechanism (such as publicly accessible version control, e.g. via Github) for this, we encourage interested parties to reach out to the authors.

Any other comments?

None

1452
1453
1454
1455
1456
1457