Natural Adversarial Examples

Dan Hendrycks
UC Berkeley

Kevin Zhao*
University of Washington

Steven Basart*
UChicago

Jacob Steinhardt, Dawn Song
UC Berkeley

Abstract
We introduce two challenging datasets that reliably cause machine learning model performance to substantially degrade. The datasets are collected with a simple adversarial filtration technique to create datasets with limited spurious cues. Our datasets’ real-world, unmodified examples transfer to various unseen models reliably, demonstrating that computer vision models have shared weaknesses. The first dataset is called IMAGENET-A and is like the ImageNet test set, but it is far more challenging for existing models. We also curate an adversarial out-of-distribution detection dataset called IMAGENET-O, which is the first out-of-distribution detection dataset created for ImageNet models. On IMAGENET-A a DenseNet-121 obtains around 2% accuracy, an accuracy drop of approximately 90%, and its out-of-distribution detection performance on IMAGENET-O is near random chance levels. We find that existing data augmentation techniques hardly boost performance, and using other public training datasets provides improvements that are limited. However, we find that improvements to computer vision architectures provide a promising path towards robust models.

1. Introduction
Research on the ImageNet [10] benchmark has led to numerous advances in classification [36], object detection [34], and segmentation [21]. ImageNet classification improvements are broadly applicable and highly predictive of improvements on many tasks [35]. Improvements on ImageNet classification have been so great that some call ImageNet classifiers “superhuman” [23]. However, performance is decidedly subhuman when the test distribution does not match the training distribution [26]. The distribution seen at test-time can include inclement weather conditions and obscured objects, and it can also include objects that are anomalous.

Recht et al., 2019 [42] remind us that ImageNet test examples tend to be simple, clear, close-up images, so that the current test set may be too easy and may not represent harder images encountered in the real world. Geirhos et al., 2020 argue that image classification datasets contain “spurious cues” or “shortcuts” [16, 2]. For instance, models may use an image’s background to predict the foreground object’s class; a cow tends to co-occur with a green pasture, and even though the background is inessential to the object’s identity, models may predict “cow” primarily using the green pasture background cue. When datasets contain

*Equal Contribution.
spurious cues, they can lead to performance estimates that are optimistic and inaccurate.

To counteract this, we curate two hard ImageNet test sets of natural adversarial examples with adversarial filtration. By using adversarial filtration, we can test how well models perform when simple-to-classify examples are removed, which includes examples that are solved with simple spurious cues. Some examples are depicted in Figure 1, which are simple for humans but hard for models. Our examples demonstrate that it is possible to reliably fool many models with clean natural images, while previous attempts at exposing and measuring model fragility rely on synthetic distribution corruptions [18, 26], artistic renditions [24], and adversarial distortions.

We demonstrate that clean examples can reliably degrade and transfer to other unseen classifiers using our first dataset. We call this dataset IMAGE-NET-A, which contains images from a distribution unlike the ImageNet training distribution. IMAGE-NET-A examples belong to ImageNet classes, but the examples are harder and can cause mistakes across various models. They cause consistent classification mistakes due to scene complications encountered in the long tail of scene configurations and by exploiting classifier blind spots (see Section 3.2). Since examples transfer reliably, this dataset shows models have unappreciated shared weaknesses.

The second dataset allows us to test model uncertainty estimates when semantic factors of the data distribution shift. Our second dataset is IMAGE-NET-O, which contains image concepts from outside ImageNet-1K. These out-of-distribution images reliably cause models to mistake the examples as high-confidence in-distribution examples. To our knowledge, this is the first dataset of anomalies or out-of-distribution examples developed to test ImageNet models. While IMAGE-NET-A enables us to test image classification performance when the input data distribution shifts, IMAGE-NET-O enables us to test out-of-distribution detection performance when the label distribution shifts.

We examine methods to improve performance on adversarially filtered examples. However, this is difficult because Figure 2 shows that examples successfully transfer to unseen or black-box models. To improve robustness, numerous techniques have been proposed. We find data augmentation techniques such as adversarial training decrease performance, while others can help by a few percent. We also find that a 10x increase in training data corresponds to a less than a 10% increase in accuracy. Finally, we show that improving model architectures is a promising avenue toward increasing robustness. Even so, current models have substantial room for improvement. Code and our two datasets are available at github.com/hendrycks/natural-adv-examples.

2. Related Work

Adversarial Examples. Real-world images may be chosen adversarially to cause performance decline. Goodfellow
et al. [19] define adversarial examples [49] as “inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.” Most adversarial examples research centers around artificial \( \ell_p \) adversarial examples, which are examples perturbed by nearly worst-case distortions that are small in an \( \ell_p \) sense. Su et al., 2018 [47] remind us that most \( \ell_p \) adversarial examples crafted from one model can only be transferred within the same family of models. However, our adversarially filtered images transfer to all tested model families and move beyond the restrictive \( \ell_p \) threat model.

**Out-of-Distribution Detection.** For out-of-distribution (OOD) detection [27, 28, 29] models learn a distribution, such as the ImageNet-1K distribution, and are tasked with producing quality anomaly scores that distinguish between usual test set examples and examples from held-out anomalous distributions. For instance, Hendrycks et al., 2017 [27] treat CIFAR-10 as the in-distribution and treat Gaussian noise and the SUN scene dataset [52] as out-of-distribution data. They show that the negative of the maximum softmax probability, or the the negative of the classifier prediction probability, is a high-performing anomaly score that can separate in- and out-of-distribution examples, so much so that it remains competitive to this day. Since that time, other work on out-of-distribution detection has continued to use datasets from other research benchmarks as anomaly stand-ins, producing far-from-distribution anomalies. Using visually dissimilar research datasets as anomaly stand-ins is critiqued in Ahmed et al., 2019 [1]. Some previous OOD detection datasets are depicted in the bottom row of Figure 3 [28]. Many of these anomaly sources are unnatural and deviate in numerous ways from the distribution of usual examples. In fact, some of the distributions can be deemed anomalous from local image statistics alone. Next, Meinke et al., 2019 [41] propose studying adversarial out-of-distribution detection by detecting adversarially optimized uniform noise. In contrast, we propose a dataset for more realistic adversarial anomaly detection; our dataset contains hard anomalies generated by shifting the distribution’s labels and keeping non-semantic factors similar to the original training distribution.

**Spurious Cues and Unintended Shortcuts.** Models may learn spurious cues and obtain high accuracy, but for the wrong reasons [38, 16]. Spurious cues are a studied problem in natural language processing [8, 20]. Many recently introduced NLP datasets use adversarial filtration to create “adversarial datasets” by sieving examples solved with simple spurious cues [44, 4, 56, 13, 6, 25]. Like this recent concurrent research, we also use adversarial filtration [48], but the technique of adversarial filtration has not been applied to collecting image datasets until this paper. Additionally, adversarial filtration in NLP removes only the easiest examples, while we use filtration to select only the hardest examples and ignore examples of intermediate difficulty. Adversarially filtered examples for NLP also do not reliably transfer even to weaker models. In Bisk et al., 2019 [5] BERT errors do not reliably transfer to weaker GPT-1 models. This is one reason why it is not obvious \textit{a priori} whether adversarially filtered images should transfer. In this work, we show that adversarial filtration algorithms can find examples that reliably transfer to both weaker and stronger models. Since adversarial filtration can remove examples that are solved by simple spurious cues, models must learn more robust features for our datasets.

**Robustness to Shifted Input Distributions.** Recht et al., 2019 [42] create a new ImageNet test set resembling the original test set as closely as possible. They found evidence that matching the difficulty of the original test set required selecting images deemed the easiest and most obvious by Mechanical Turkers. However, Engstrom et al., 2020 [14] estimate that the accuracy drop from ImageNet to ImageNetV2 is less than 3.6%. In contrast, model accuracy can decrease by over 50% with \textsc{ImageNet-A}. Brendel et al., 2018 [7] show that classifiers that do not know the spatial ordering of image regions can be competitive on the
ImageNet test set, possibly due to the dataset’s lack of difficulty. Judging classifiers by their performance on easier examples has potentially masked many of their shortcomings. For example, Geirhos et al., 2019 [17] artificially overwrite each ImageNet image’s textures and conclude that classifiers learn to rely on textural cues and under-utilize information about object shape. Recent work shows that classifiers are highly susceptible to non-adversarial stochastic corruptions [26]. While they distort images with 75 different algorithmically generated corruptions, our sources of distribution shift tend to be more heterogeneous and varied, and our examples are naturally occurring.

3. IMAGENET-A and IMAGENET-O

3.1. Design

IMAGENET-A is a dataset of real-world adversarially filtered images that fool current ImageNet classifiers. To find adversarially filtered examples, we first download numerous images related to an ImageNet class. Thereafter we delete the images that fixed ResNet-50 [22] classifiers correctly predict. We chose ResNet-50 due to its widespread use. Later we show that examples which fool ResNet-50 reliably transfer to other unseen models. With the remaining incorrectly classified images, we manually select visually clear images.

Next, IMAGENET-O is a dataset of adversarially filtered examples for ImageNet out-of-distribution detectors. To create this dataset, we download ImageNet-22K and delete examples from ImageNet-1K. With the remaining ImageNet-22K examples that do not belong to ImageNet-1K classes, we keep examples that are classified by a ResNet-50 as an ImageNet-1K class with high confidence. Then we manually select visually clear images.

Both datasets were manually constructed by graduate students over several months. This is because a large share of images contain multiple classes per image [46]. Therefore, producing a dataset without multilabel images can be challenging with usual annotation techniques. To ensure images do not fall into more than one of the several hundred classes, we had graduate students memorize the classes in order to build a high-quality test set.

IMAGENET-A Class Restrictions. We select a 200-class subset of ImageNet-1K’s 1,000 classes so that errors among these 200 classes would be considered egregious [10]. For instance, wrongly classifying Norwich terriers as Norfolk terriers does less to demonstrate faults in current classifiers than mistaking a Persian cat for a candle. We additionally avoid rare classes such as “snow leopard,” classes that have changed much since 2012 such as “iPod,” coarse classes such as “spiral,” classes that are often image backdrops such as “valley,” and finally classes that tend to overlap such as “honeycomb,” “bee,” “bee house,” and “bee eater”; “eraser,” “pencil sharpener” and “pencil case”; “sink,” “medicine cabinet,” “pill bottle” and “band-aid”; and so on. The 200 IMAGENET-A classes cover most broad categories spanned by ImageNet-1K; see the Supplementary Materials for the full class list.

IMAGENET-A Data Aggregation. The first step is to download many weakly labeled images. Fortunately, the website iNaturalist has millions of user-labeled images of animals, and Flickr has even more user-tagged images of objects. We download images related to each of the 200 ImageNet classes by leveraging user-provided labels and tags. After exporting or scraping data from sites including iNaturalist, Flickr, and DuckDuckGo, we adversarially select images by removing examples that fail to fool our ResNet-50 models. Of the remaining images, we select low-confidence images and then ensure each image is valid through human review. If we only used the original ImageNet test set as a source rather than iNaturalist, Flickr, and DuckDuckGo, some classes would have zero images after the first round of filtration, as the original ImageNet test set is too small to contain hard adversarially filtered images.

We now describe this process in more detail. We use a small ensemble of ResNet-50s for filtering, one pre-trained on ImageNet-1K then fine-tuned on the 200 class subset, and one pre-trained on ImageNet-1K where 200 of its 1,000 logits are used in classification. Both classifiers have similar accuracy on the 200 clean test set classes from ImageNet-1K. The ResNet-50s perform 10-crop classification for each image, and should any crop be classified correctly by the ResNet-50s, the image is removed. If either ResNet-50 assigns greater than 15% confidence to the correct class, the image is also removed; this is done so that adversarially filtered examples yield misclassifications with low confidence in the correct class, like in untargeted adversarial attacks. Now, some classification confusions are greatly overrepresented, such as Persian cat and lynx. We would like IMAGENET-A to have great variability in its types of errors and cause classifiers to have a dense confusion matrix. Consequently, we perform a second round of filtering to create a shortlist where each confusion only appears at most 15 times. Finally, we manually select images from this shortlist in order to ensure IMAGENET-A images are simultaneously valid, single-class, and high-quality. In all, the IMAGENET-A dataset has 7,500 adversarially filtered images.

As a specific example, we download 81,413 dragonfly images from iNaturalist, and after running the ResNet-50 filter we have 8,925 dragonfly images. In the algorithmically diversified shortlist, 1,452 images remain. From this shortlist, 80 dragonfly images are manually selected, but hundreds more could be selected if time allows.

The resulting images represent a substantial distribution shift, but images are still possible for humans to classify. The Fréchet Inception Distance (FID) [31] enables us to de-
termine whether IMAGE-NET-A and ImageNet are not identically distributed. The FID between ImageNet’s validation and test set is approximately 0.99, indicating that the distributions are highly similar. The FID between IMAGE-NET-A and ImageNet’s validation set is 50.40, and the FID between IMAGE-NET-A and ImageNet’s test set is approximately 50.25, indicating that the distribution shift is large. Despite the shift, we estimate that our graduate students’ IMAGE-NET-A human accuracy rate is approximately 90%.

**IMAGE-NET-O Class Restrictions.** We again select a 200-class subset of ImageNet-1K’s 1,000 classes. These 200 classes determine the in-distribution or the distribution that is considered usual. As before, the 200 classes cover most broad categories spanned by ImageNet-1K; see the Supplementary Materials for the full class list.

**IMAGE-NET-O Data Aggregation.** Our dataset for adversarial out-of-distribution detection is created by fooling ResNet-50 out-of-distribution detectors. The negative of the prediction confidence of a ResNet-50 ImageNet classifier serves as our anomaly score [27]. Usually in-distribution examples produce higher confidence predictions than OOD examples, but we curate OOD examples that have high confidence predictions. To gather candidate adversarially filtered examples, we use the ImageNet-22K dataset with ImageNet-1K classes deleted. We choose the ImageNet-22K dataset since it was collected in the same way as ImageNet-1K. ImageNet-22K allows us to have coverage of numerous visual concepts and vary the distribution’s semantics without unnatural or unwanted non-semantic data shift. After excluding ImageNet-1K images, we process the remaining ImageNet-22K images and keep the images which cause the ResNet-50 to have high confidence, or a low anomaly score. We then manually select a high-quality subset of the remaining images to create IMAGE-NET-O. We suggest only training models with data from the 1,000 ImageNet-1K classes, since the dataset becomes trivial if models train on ImageNet-22K. To our knowledge, this dataset is the first anomalous dataset curated for ImageNet models and enables researchers to study adversarial out-of-distribution detection. The IMAGE-NET-O dataset has 2,000 adversarially filtered examples since anomalies are rarer; this has the same number of examples per class as ImageNetV2 [42]. While we use adversarial filtration to select images that are difficult for a fixed ResNet-50, we will show

---

Figure 4: Additional adversarially filtered examples from the IMAGE-NET-A dataset. Examples are adversarially selected to cause classifier accuracy to degrade. The black text is the actual class, and the red text is a ResNet-50 prediction.

Figure 5: Additional adversarially filtered examples from the IMAGE-NET-O dataset. Examples are adversarially selected to cause out-of-distribution detection performance to degrade. Examples do not belong to ImageNet classes, and they are wrongly assigned highly confident predictions. The black text is the actual class, and the red text is a ResNet-50 prediction and the prediction confidence.
Figure 6: Examples from IMAGE-NET-A demonstrating classifier failure modes. Adjacent to each natural image is its heatmap. Classifiers may use erroneous background cues for prediction. These failure modes are described in Section 3.2.

these examples straightforwardly transfer to unseen models.

3.2. Illustrative Failure Modes

Examples in IMAGE-NET-A uncover numerous failure modes of modern convolutional neural networks. We describe our findings after having viewed tens of thousands of candidate adversarially filtered examples. Some of these failure modes may also explain poor IMAGE-NET-O performance, but for simplicity we describe our observations with IMAGE-NET-A examples.

Consider Figure 6. The first two images suggest models may overgeneralize visual concepts. It may confuse metal with sundials, or thin radiating lines with harvestman bugs. We also observed that networks overgeneralize tricycles to bicycles and circles, digital clocks to keyboards and calculators, and more. We also observe that models may rely too heavily on color and texture, as shown with the dragonfly images. Since classifiers are taught to associate entire images with an object class, frequently appearing background elements may also become associated with a class, such as wood being associated with nails. Other examples include classifiers heavily associating hummingbird feeders with hummingbirds, leaf-covered tree branches being associated with the white-headed capuchin monkey class, snow being associated with shovels, and dumpsters with garbage trucks. Additionally Figure 6 shows an American alligator swimming. With different frames, the classifier prediction varies erratically between classes that are semantically loose and separate. For other images of the swimming alligator, classifiers predict that the alligator is a cliff, lynx, and a fox squirrel. Assessing convolutional networks on IMAGE-NET-A reveals that even state-of-the-art models have diverse and systematic failure modes.

4. Experiments

We show that adversarially filtered examples collected to fool fixed ResNet-50 models reliably transfer to other models, indicating that current convolutional neural networks have shared weaknesses and failure modes. In the following sections, we analyze whether robustness can be improved by using data augmentation, using more real labeled data, and using different architectures. For the first two sections, we analyze performance with a fixed architecture for comparability, and in the final section we observe performance with different architectures. First we define our metrics.

**Metrics.** Our metric for assessing robustness to adversarially filtered examples for classifiers is the top-1 accuracy on IMAGE-NET-A. For reference, the top-1 accuracy on the 200 IMAGE-NET-A classes using usual ImageNet images is usually greater than or equal to 90% for ordinary classifiers.

Our metric for assessing out-of-distribution detection performance of IMAGE-NET-O examples is the area under the precision-recall curve (AUPR). This metric requires anomaly scores. Our anomaly score is the negative of the maximum softmax probabilities from a model that can classify the 200 IMAGE-NET-O classes. The maximum softmax probability detector is a long-standing baseline in OOD detection. We collect anomaly scores with the ImageNet validation examples for the said 200 classes. Then, we collect anomaly scores for the IMAGE-NET-O examples. Higher performing OOD detectors would assign IMAGE-NET-O examples lower confidences, or higher anomaly scores. With these anomaly scores, we can compute the area under the precision-recall curve. Random chance levels for the AUPR is approximately 16.67% with IMAGE-NET-O, and the maximum AUPR is 100%.
**Data Augmentation.** We examine popular data augmentation techniques and note their effect on robustness. In this section we exclude IMAGENET-O results, as the data augmentation techniques hardly help with out-of-distribution detection as well. As a baseline, we train a new ResNet-50 from scratch and obtain 2.17% accuracy on IMAGENET-A. Now, one purported way to increase robustness is through adversarial training, which makes models less sensitive to $\ell_p$ perturbations. We use the adversarially trained model from Wong et al., 2020 [51], but accuracy decreases to 1.68%. Next, Geirhos et al., 2019 [17] propose making networks rely less on texture by training classifiers on images where textures are transferred from art pieces. They accomplish this by applying style transfer to ImageNet training images to create a stylized dataset, and models train on these images. While this technique is able to greatly increase robustness on synthetic corruptions [26], Style Transfer increases IMAGENET-A accuracy only 0.13% over the ResNet-50 baseline. A recent data augmentation technique is AugMix [30], which takes linear combinations of different data augmentations. This technique increases accuracy to 3.8%. Cutout augmentation [11] randomly occludes image regions and corresponds to 4.4% accuracy. Moment Exchange (MoEx) [40] exchanges feature map moments between images, and this increases accuracy to 5.5%. Mixup [57] trains networks on elementwise convex combinations of images and their interpolated labels; this technique increases accuracy to 6.6%. CutMix [55] superimposes images regions within other images and yields 7.3% accuracy. At best these data augmentations techniques improve accuracy by approximately 5% over the baseline. Results are summarized in Figure 7. Although some data augmentation techniques are purported to greatly improve robustness to distribution shifts [30, 54], their lackluster results on IMAGENET-A show they do not improve robustness on some distribution shifts. Hence IMAGENET-A can be used to verify whether techniques actually improve real-world robustness to distribution shift.

**More Labeled Data.** One possible explanation for consistently low IMAGENET-A accuracy is that all models are trained only with ImageNet-1K, and using additional data may resolve the problem. Bau et al., 2017 [3] argue that Places365 classifiers learn qualitatively distinct filters (e.g., they have more object detectors, fewer texture detectors in conv3) compared to ImageNet classifiers, so one may expect an error distribution less correlated with errors on ImageNet-A. To test this hypothesis we pre-train a ResNet-50 on Places365 [58], a large-scale scene recognition dataset. After fine-tuning the Places365 model on ImageNet-1K, we find that accuracy is 1.56%. Consequently, even though scene recognition models are purported to have qualitatively distinct features, this is not enough to improve IMAGENET-A performance. Likewise, Places365 pre-training does not improve IMAGENET-O detection, as its AUPR is 14.88%. Next, we see whether labeled data from IMAGENET-A itself can help. We take baseline ResNet-50 with 2.17% IMAGENET-A accuracy and fine-tune it on 80% of IMAGENET-A. This leads to no clear improvement on the remaining 20% of IMAGENET-A since the top-1 and top-5 accuracies are below 2% and 5%, respectively.

Last, we pre-train using an order of magnitude more training data with ImageNet-21K. This dataset contains approximately 21,000 classes and approximately 14 million images. To our knowledge this is the largest publicly available database of labeled natural images. Using a ResNet-50 pretrained on ImageNet-21K, we fine-tune the model on ImageNet-1K and attain 11.41% accuracy on IMAGENET-A, a 9.24% increase. Likewise, the AUPR for IMAGENET-O improves from 16.20% to 21.86%, although this improvement is less significant since IMAGENET-O images overlap with ImageNet-21K images. Academic researchers rarely use datasets larger than ImageNet due to computational costs, using more data has limitations. An order of magnitude increase in labeled training data can provide some improvements in accuracy, though we now show that
Figure 8: Increasing model size and other architecture changes can greatly improve performance. Note Res2Net and ResNet+SE have a ResNet backbone. Normal model sizes are ResNet-50 and ResNeXt-50 (32 × 4d), Large model sizes are ResNet-101 and ResNeXt-101 (32 × 4d), and XLarge Model sizes are ResNet-152 and (32 × 8d).

Architectural Changes. We find that model architecture can play a large role in ImageNet-A accuracy and ImageNet-O detection performance. Simply increasing the width and number of layers of a network is sufficient to automatically impart more ImageNet-A accuracy and ImageNet-O OOD detection performance. Increasing network capacity has been shown to improve performance on ℓp adversarial examples [37], common corruptions [26], and now also improves performance for adversarially filtered images. For example, a ResNet-50’s top-1 accuracy and AUPR is 2.17% and 16.2%, respectively, while a ResNet-152 obtains 6.1% top-1 accuracy and 18.0% AUPR. Another architecture change that reliably helps is using the grouped convolutions found in ResNeXts [53]. A ResNeXt-50 (32 × 4d) obtains a 4.81% top1 ImageNet-A accuracy and a 17.60% ImageNet-O AUPR.

Another useful architecture change is self-attention. Convolutional neural networks with self-attention [32] are designed to better capture long-range dependencies and interactions across an image. We consider the self-attention technique called Squeeze-and-Excitation (SE) [33], which won the final ImageNet competition in 2017. A ResNet-50 with Squeeze-and-Excitation attains 6.17% accuracy. However, for larger ResNets, self-attention does little to improve ImageNet-O detection.

We consider the ResNet-50 architecture with its residual blocks exchanged with recently introduced Res2Net v1b blocks [15]. This change increases accuracy to 14.50% and the AUPR to 19.5%. A ResNet-152 with Res2Net v1b blocks attains 22.4% accuracy and 23.9% AUPR. Compared to data augmentation or an order of magnitude more labeled training data, some architectural changes can provide far more robustness gains. Consequently future improvements to model architectures is a promising path towards greater robustness.

We now assess performance on a completely different architecture which does not use convolutions, vision Transformers [12]. We evaluate with DeiT [50], a vision Transformer trained on ImageNet-1K with aggressive data augmentation such as Mixup. Even for vision Transformers, we find that ImageNet-A and ImageNet-O examples successfully transfer. In particular, a DeiT-small vision Transformer gets 19.0% on ImageNet-A and has a similar number of parameters to a Res2Net-50, which has 14.6% accuracy. This might be explained by DeiT’s use of Mixup, however, which provided a 4% ImageNet-A accuracy boost for ResNets. The ImageNet-O AUPR for the Transformer is 20.9%, while the Res2Net gets 19.5%. Larger DeiT models do better, as a DeiT-base gets 28.2% accuracy on ImageNet-A and 24.8% AUPR on ImageNet. Consequently, our datasets transfer to vision Transformers and performance for both tasks remains far from the ceiling.

5. Conclusion

We found it is possible to improve performance on our datasets with data augmentation, pretrained data, and architectural changes. We found that our examples transferred to all tested models, including vision Transformers which do not use convolution operations. Results indicate that improving performance on ImageNet-A and ImageNet-O is possible but difficult. Our challenging ImageNet test sets serve as measures of performance under distribution shift—an important research aim as models are deployed in increasingly precarious real-world environments.
References


