The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization

Dan Hendrycks[1] Steven Basart[2*] Norman Mu[1*] Saurav Kadavath[1] Frank Wang[3]
Evan Dorundo[3] Rahul Desai[1] Tyler Zhu[1] Samyak Parajuli[1] Mike Guo[1]
Dawn Song[1] Jacob Steinhardt[1] Justin Gilmer[3]

Abstract

We introduce four new real-world distribution shift datasets consisting of changes in image style, image blurriness, geographic location, camera operation, and more. With our new datasets, we take stock of previously proposed methods for improving out-of-distribution robustness and put them to the test. We find that using larger models and artificial data augmentations can improve robustness on real-world distribution shifts, contrary to claims in prior work. We find improvements in artificial robustness benchmarks can transfer to real-world distribution shifts, contrary to claims in prior work. Motivated by our observation that data augmentations can help with real-world distribution shifts, we also introduce a new data augmentation method which advances the state-of-the-art and outperforms models pretrained with $1000\times$ *more labeled data. Overall we find that some methods consistently help with distribution shifts in texture and local image statistics, but these methods do not help with some other distribution shifts like geographic changes. Our results show that future research must study multiple distribution shifts simultaneously, as we demonstrate that no evaluated method consistently improves robustness.*

1. Introduction

While the research community must create robust models that generalize to new scenarios, the robustness literature [6, 9] lacks consensus on evaluation benchmarks and contains many dissonant hypotheses. Hendrycks et al., 2020 [14] find that many recent language models are already robust to many forms of distribution shift, while others [37, 10] find that vision models are largely fragile and argue that data augmentation offers one solution. In contrast, other researchers [30] provide results suggesting that using pretraining and improving in-distribution test set accuracy improves natural robustness, whereas other methods do not.

Prior works have also offered various interpretations of empirical results, such as the *Texture Bias* hypothesis that convolutional networks are biased towards texture, harming robustness [10]. Additionally, some authors posit a fundamental distinction between robustness on *synthetic* benchmarks vs. *real-world* distribution shifts, casting doubt on the generality of conclusions drawn from experiments conducted on synthetic benchmarks [30].

It has been difficult to arbitrate these hypotheses because existing robustness datasets vary multiple factors (e.g., time, camera, location, etc.) simultaneously in unspecified ways [26, 16]. Existing datasets also lack diversity such that it is hard to extrapolate which methods will improve robustness more broadly. To address these issues and test the methods outlined above, we introduce four new robustness datasets and a new data augmentation method.

First we introduce ImageNet-Renditions (ImageNet-R), a 30,000 image test set containing various renditions (e.g., paintings, embroidery, etc.) of ImageNet object classes. These renditions are naturally occurring, with textures and local image statistics unlike those of ImageNet images, allowing us to compare against gains on synthetic robustness benchmarks.

Next, we investigate the effect of changes in the image capture process with StreetView StoreFronts (SVSF) and DeepFashion Remixed (DFR). SVSF contains business storefront images collected from Google StreetView, along with metadata allowing us to vary location, year, and even the camera type. DFR leverages the metadata from DeepFashion2 [8] to systematically shift object occlusion, orientation, zoom, and scale at test time. Both SVSF and DFR provide distribution shift controls and do not alter texture, which remove possible confounding variables affecting prior benchmarks.

Additionally, we collect Real Blurry Images, which consists of 1,000 blurry natural images from a 100-class subset of the ImageNet classes. This benchmark serves as a real-world analog for the synthetic blur corruptions of the ImageNet-C benchmark [12]. With it we find that synthetic corruptions correlate with corruptions that appear in the wild,

*Equal contribution. [1]UC Berkeley, [2]UChicago, [3]Google. Code is available at https://github.com/hendrycks/imagenet-r.

Figure 1: Images from three of our four new datasets ImageNet-Renditions (ImageNet-R), DeepFashion Remixed (DFR), and StreetView StoreFronts (SVSF). The SVSF images are recreated from the public Google StreetView. Our datasets test robustness to various naturally occurring distribution shifts including rendition style, camera viewpoint, and geography.

contradicting speculations from previous work [30].

Finally, we contribute DeepAugment to increase robustness to some new types of distribution shift. This augmentation technique uses image-to-image neural networks for data augmentation. DeepAugment improves robustness on our newly introduced ImageNet-R benchmark and can also be combined with other augmentation methods to outperform a model pretrained on $1000\times$ more labeled data.

We use these new datasets to test four overarching classes of methods for improving robustness:

- *Larger Models*: increasing model size improves robustness to distribution shift [12, 35].
- *Self-Attention*: adding self-attention layers to models improves robustness [16].
- *Diverse Data Augmentation*: robustness can increase through data augmentation [37].
- *Pretraining*: pretraining on larger and more diverse datasets improves robustness [25, 13].

After examining our results on these four new datasets as well as prior benchmarks, we can rule out several previous hypotheses while strengthening support for others. As one example, we find that synthetic data augmentation robustness interventions improve accuracy on ImageNet-R and real-world image blur distribution shifts, which lends credence to the use of synthetic robustness benchmarks and also reinforces the *Texture Bias* hypothesis. In the conclusion, we summarize the various strands of evidence for and against each hypothesis. Across our many experiments, we do not find a general method that consistently improves robustness, and some hypotheses require additional qualifications. While robustness is often spoken of and measured as a single scalar property like accuracy, our investigations show that robustness is not so simple. Our results show that future robustness research requires more thorough evaluation using more robustness datasets.

2. Related Work

Robustness Benchmarks. Recent works [12, 26, 14] have begun to characterize model performance on out-of-distribution (OOD) data with various new test sets, with dissonant findings. For instance, prior work [14] demonstrates that modern language processing models are moderately robust to numerous naturally occurring distribution shifts, and that IID accuracy is not straightforwardly predictive of OOD accuracy for natural language tasks. For image recognition, other work [12] analyzes image models and shows that they are sensitive to various simulated image corruptions (e.g., noise, blur, weather, JPEG compression, etc.) from their ImageNet-C benchmark.

Recht et al., 2019 [26] reproduce the ImageNet [28] validation set for use as a benchmark of naturally occurring distribution shift in computer vision. Their evaluations show a 11-14% drop in accuracy from ImageNet to the new validation set, named ImageNetV2, across a wide range of architectures. [30] use ImageNetV2 to measure natural robustness and conclude that methods such as data augmentation do not significantly improve robustness. Recently, [7] identify statistical biases in ImageNetV2's construction, and they estimate that re-weighting ImageNetV2 to correct for these biases results in a less substantial 3.6% drop.

Data Augmentation. Recent works [10, 37, 15] demonstrate that data augmentation can improve robustness on ImageNet-C. The space of augmentations that help robustness includes various types of noise [22, 27, 21], highly unnatural image transformations [10, 38, 39], or compositions of simple image transformations such as Python Imaging Library operations [4, 15]. Some of these augmentations can

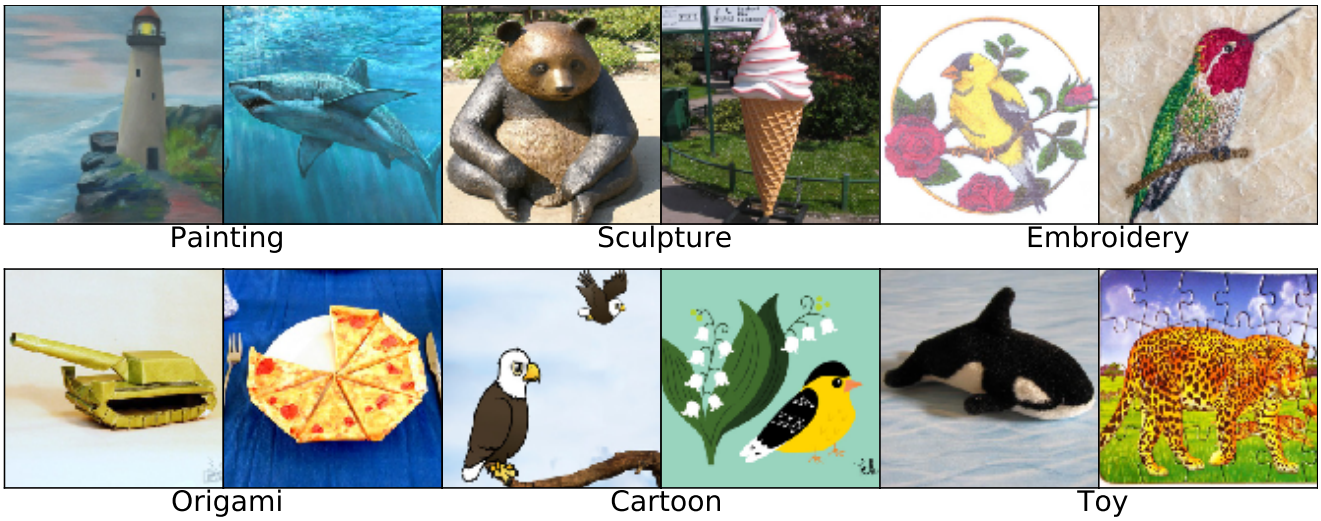

Figure 2: ImageNet-Renditions (ImageNet-R) contains 30,000 images of ImageNet objects with different textures and styles. This figure shows only a portion of ImageNet-R's numerous rendition styles. The rendition styles (e.g., "Toy") are for clarity and are *not* ImageNet-R's classes; ImageNet-R's classes are a subset of 200 ImageNet classes.

improve accuracy on in-distribution examples as well as on out-of-distribution (OOD) examples.

3. New Datasets

In order to evaluate the four robustness methods, we introduce four new benchmarks that capture new types of naturally occurring distribution shifts. ImageNet-Renditions (ImageNet-R) and Real Blurry Images are both newly collected test sets intended for ImageNet classifiers, whereas StreetView StoreFronts (SVSF) and DeepFashion Remixed (DFR) each contain their own training sets and multiple test sets. SVSF and DFR split data into a training and test sets based on various image attributes stored in the metadata. For example, we can select a test set with images produced by a camera different from the training set camera. We now describe the structure and collection of each dataset.

3.1. ImageNet-Renditions (ImageNet-R)

While current classifiers can learn some aspects of an object's shape [24], they nonetheless rely heavily on natural textural cues [10]. In contrast, human vision can process abstract visual renditions. For example, humans can recognize visual scenes from line drawings as quickly and accurately as they can from photographs [3]. Even some primates species have demonstrated the ability to recognize shape through line drawings [18, 29].

To measure generalization to various abstract visual renditions, we create the ImageNet-Rendition (ImageNet-R) dataset. ImageNet-R contains various artistic renditions of object classes from the original ImageNet dataset. Note the original ImageNet dataset discouraged such images since annotators were instructed to collect "photos only, no painting, no drawings, etc." [5]. We do the opposite.

Data Collection. ImageNet-R contains 30,000 image renditions for 200 ImageNet classes. We choose a subset of the ImageNet-1K classes, following [16], for several reasons. A handful ImageNet classes already have many renditions, such as "triceratops." We also choose a subset so that model misclassifications are egregious and to reduce label noise. The 200 class subset was also chosen based on rendition prevalence, as "strawberry" renditions were easier to obtain than "radiator" renditions. Were we to use all 1,000 ImageNet classes, annotators would be pressed to distinguish between Norwich terrier renditions as Norfolk terrier renditions, which is difficult. We collect images primarily from Flickr and use queries such as "art," "cartoon," "graffiti," "embroidery," "graphics," "origami," "painting," "pattern," "plastic object," "plush object," "sculpture," "line drawing," "tattoo," "toy," "video game," and so on. Images are filtered by Amazon MTurk annotators using a modified collection interface from ImageNetV2 [26]. For instance, after scraping Flickr images with the query "lighthouse cartoon," we have MTurk annotators select true positive lighthouse renditions. Finally, as a second round of quality control, graduate students manually filter the resulting images and ensure that individual images have correct labels and do not contain multiple labels. Examples are depicted in Figure 2. ImageNet-R also includes the line drawings from [32], excluding horizontally mirrored duplicate images, pitch black images, and images from the incorrectly collected "pirate ship" class.

3.2. StreetView StoreFronts (SVSF)

Computer vision applications often rely on data from complex pipelines that span different hardware, times, and

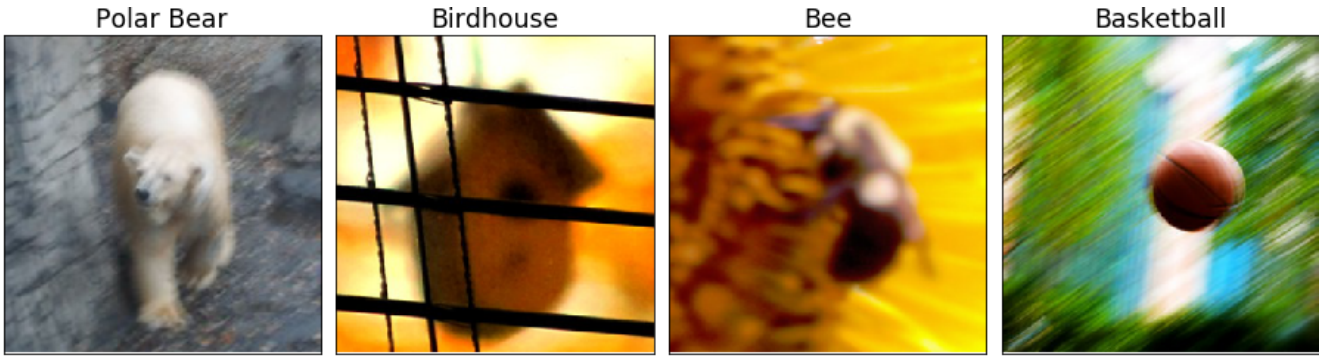

Figure 3: Examples of images from Real Blurry Images. This dataset allows us to test whether model performance on ImageNet-C's synthetic blur corruptions track performance on real-world blur corruptions.

geographies. Ambient variations in this pipeline may result in unexpected performance degradation, such as degradations experienced by health care providers in Thailand deploying laboratory-tuned diabetic retinopathy classifiers in the field [2]. In order to study the effects of shifts in the image capture process we collect the StreetView StoreFronts (SVSF) dataset, a new image classification dataset sampled from Google StreetView imagery [1] focusing on three distribution shift sources: country, year, and camera.

Data Collection. SVSF consists of cropped images of business store fronts extracted from StreetView images by an object detection model. Each store front image is assigned the class label of the associated Google Maps business listing through a combination of machine learning models and human annotators. We combine several visually similar business types (e.g. drugstores and pharmacies) for a total of 20 classes, listed in the Supplementary Materials.

Splitting the data along the three metadata attributes of country, year, and camera, we create one training set and five test sets. We sample a training set and an in-distribution test set (200K and 10K images, respectively) from images taken in US/Mexico/Canada during 2019 using a "new" camera system. We then sample four OOD test sets (10K images each) which alter one attribute at a time while keeping the other two attributes consistent with the training distribution. Our test sets are year: 2017, 2018; country: France; and camera: "old."

3.3. DeepFashion Remixed

Changes in day-to-day camera operation can cause shifts in attributes such as object size, object occlusion, camera viewpoint, and camera zoom. To measure this, we repurpose DeepFashion2 [8] to create the DeepFashion Remixed (DFR) dataset. We designate a training set with 48K images and create eight out-of-distribution test sets to measure performance under shifts in object size, object occlusion, camera viewpoint, and camera zoom-in. DeepFashion Remixed is a multi-label classification task since images may contain more than one clothing item per image.

Data Collection. Similar to SVSF, we fix one value for each of the four metadata attributes in the training distribution. Specifically, the DFR training set contains images with medium scale, medium occlusion, side/back viewpoint, and no zoom-in. After sampling an IID test set, we construct eight OOD test distributions by altering one attribute at a time, obtaining test sets with minimal and heavy occlusion; small and large scale; frontal and not-worn viewpoints; and medium and large zoom-in. See the Supplementary Materials for details on test set sizes.

3.4. Real Blurry Images

We collect a small dataset of 1,000 real-world blurry images to capture real-world corruptions and validate synthetic image corruption benchmarks such as ImageNet-C. We collect the "Real Blurry Images" dataset from Flickr and query ImageNet object class names concatenated with the word "blurry." Examples are in Figure 3. Each image belongs to one of 100 ImageNet classes.

4. DeepAugment

In order to further explore effects of data augmentation, we introduce a new data augmentation technique. Whereas most previous data augmentations techniques use simple augmentation primitives applied to the raw image itself, we introduce DeepAugment, which distorts images by perturbing internal representations of deep networks.

DeepAugment works by passing a clean image through an image-to-image network and introducing several perturbations during the forward pass. These perturbations are randomly sampled from a set of manually designed functions and applied to the network weights and to the feed-forward signal at random layers. For example, our set of perturbations includes zeroing, negating, convolving, transposing, applying activation functions, and more. This setup

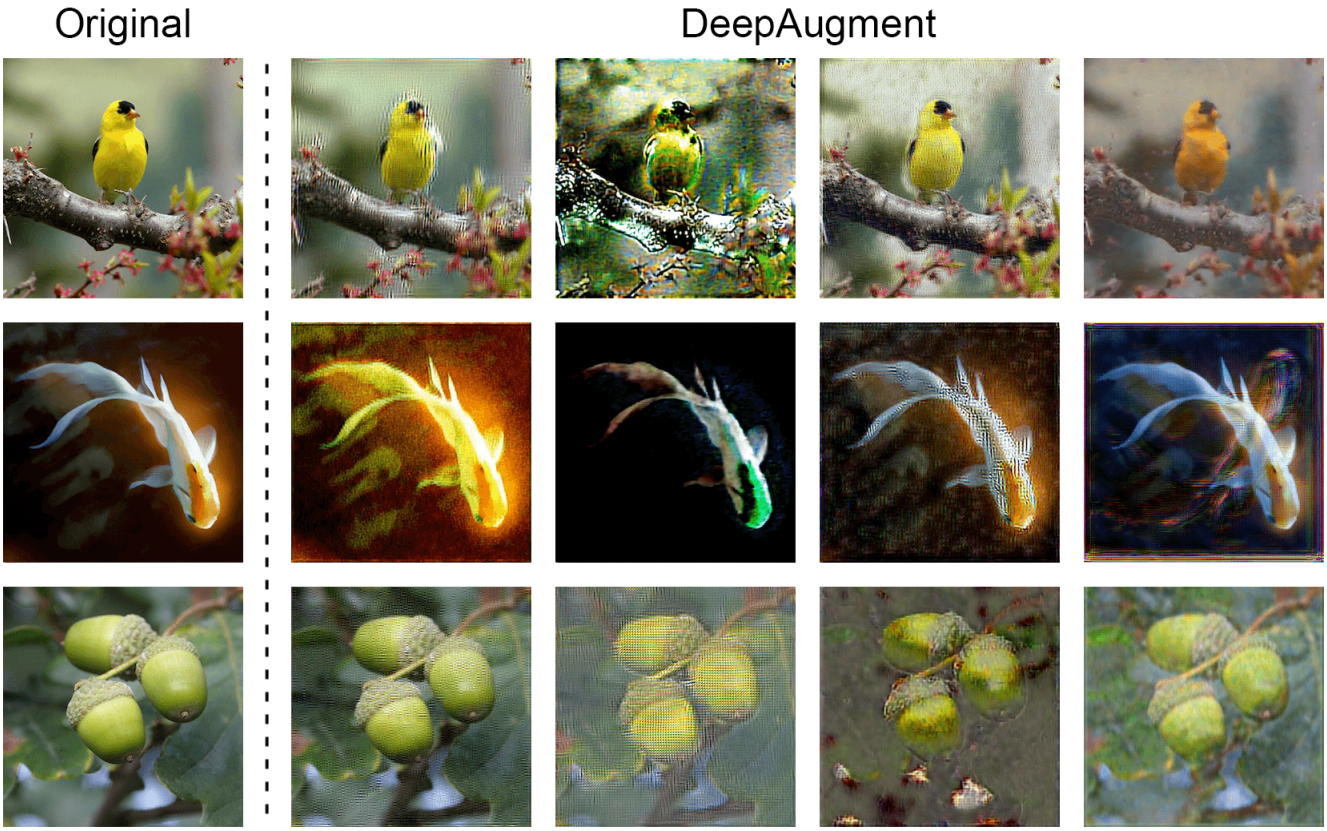

Figure 4: DeepAugment examples preserve semantics, are data-dependent, and are far more visually diverse than, say, rotations.

generates semantically consistent images with unique and diverse distortions as shown in Figure 4. Although our set of perturbations is designed with random operations, we show that DeepAugment still outperforms other methods on benchmarks such as ImageNet-C and ImageNet-R. We provide the pseudocode in the Supplementary Materials.

For our experiments, we specifically use the CAE [31] and EDSR [20] architectures as the basis for DeepAugment. CAE is an autoencoder architecture, and EDSR is a super-resolution architecture. These two architectures show the DeepAugment approach works with different architectures. Each clean image in the original dataset and passed through the network and is thereby stochastically distored, resulting in two distorted versions of the clean dataset (one for CAE and one for EDSR). We then train on the augmented and clean data simultaneously and call this approach DeepAugment. The EDSR and CAE architectures are arbitrary. We show that the DeepAugment approach also works for untrained, randomly sampled architectures in the Supplementary Materials.

5. Experiments

5.1. Setup

In this section we briefly describe the evaluated models, pretraining techniques, self-attention mechanisms, data augmentation methods, and note various implementation details.

Model Architectures and Sizes. Most experiments are evaluated on a standard ResNet-50 model [11]. Model size evaluations use ResNets or ResNeXts [36] of varying sizes.

Pretraining. For pretraining we use ImageNet-21K which contains approximately 21,000 classes and approximately 14 million labeled training images, or around $10\times$ more labeled training data than ImageNet-1K. We also tune an ImageNet-21K model [19]. We also use a large pre-trained ResNeXt-101 model [23]. This was pre-trained on on approximately 1 billion Instagram images with hashtag labels and fine-tuned on ImageNet-1K. This Weakly Supervised Learning (WSL) pretraining strategy uses approximately $1000\times$ more labeled data.

Self-Attention. When studying self-attention, we employ CBAM [34] and SE [17] modules, two forms of self-attention that help models learn spatially distant dependencies.

Data Augmentation. We use Style Transfer, AugMix, and DeepAugment to evaluate the benefits of data augmentation, and we contrast their performance with simpler noise augmentations such as Speckle Noise and adversarial noise. Style transfer [10] uses a style transfer network to apply artwork styles to training images. We use AugMix [15] which randomly composes simple augmentation operations (e.g., translate, posterize, solarize). DeepAugment, introduced

	ImageNet-200 (%)	ImageNet-R (%)	Gap
ResNet-50	7.9	63.9	56.0
+ ImageNet-21K *Pretraining* (10× labeled data)	7.0	62.8	55.8
+ CBAM (*Self-Attention*)	7.0	63.2	56.2
+ ℓ_∞ Adversarial Training	25.1	68.6	43.5
+ Speckle Noise	8.1	62.1	54.0
+ Style Transfer Augmentation	8.9	58.5	49.6
+ AugMix	7.1	58.9	51.8
+ DeepAugment	7.5	57.8	50.3
+ DeepAugment + AugMix	8.0	53.2	45.2
ResNet-152 (*Larger Models*)	6.8	58.7	51.9

Table 1: ImageNet-200 and ImageNet-R top-1 error rates. ImageNet-200 uses the same 200 classes as ImageNet-R. DeepAugment+AugMix improves over the baseline by over 10 percentage points. We take ImageNet-21K Pretraining and CBAM as representatives of pretraining and self-attention, respectively. Style Transfer, AugMix, and DeepAugment are all instances of more complex data augmentation, in contrast to simpler noise-based augmentations such as ℓ_∞ Adversarial Noise and Speckle Noise. While there remains much room for improvement, results indicate that progress on ImageNet-R is tractable.

above, distorts the weights and feedforward passes of image-to-image models to generate image augmentations. Speckle Noise data augmentation muliplies each pixel by $(1 + x)$ with x sampled from a normal distribution [27, 12]. We also consider adversarial training as a form of adaptive data augmentation and use the model from [33] trained against ℓ_∞ perturbations of size $\varepsilon = 4/255$.

5.2. Results

We now perform experiments on ImageNet-R, StreetView StoreFronts, DeepFashion Remixed, and Real Blurry Images. We also evaluate on ImageNet-C and compare and contrast it with real distribution shifts.

ImageNet-R. Table 1 shows performance on ImageNet-R as well as on ImageNet-200 (the original ImageNet data restricted to ImageNet-R's 200 classes). This has several implications regarding the four method-specific hypotheses. Pretraining with ImageNet-21K (approximately 10× labeled data) hardly helps. The Supplementary Materials shows WSL pretraining can help, but Instagram has renditions, while ImageNet excludes them; hence we conclude comparable pretraining was ineffective. Notice self-attention increases the IID/OOD gap. Compared to simpler data augmentation techniques such as Speckle Noise, the data augmentation techniques of Style Transfer, AugMix, and DeepAugment improve generalization. Note AugMix and DeepAugment improve in-distribution performance whereas Style transfer hurts it. Also, our new DeepAugment technique is the best standalone method with an error rate of 57.8%. Last, larger models reduce the IID/OOD gap.

As for prior hypothesis in the literature regarding model robustness, we find that biasing networks away from natural textures through diverse data augmentation improved performance. The IID/OOD generalization gap varies greatly by method, demonstrating that it is possible to significantly outperform the trendline of models optimized solely for the IID setting. Finally, as ImageNet-R contains real-world examples, and since data augmentation helps on ImageNet-R, we now have clear evidence against the hypothesis that robustness interventions cannot help with natural distribution shifts [30].

StreetView StoreFronts. In Table 2, we evaluate data augmentation methods on SVSF and find that all of the tested methods have mostly similar performance and that no method helps much on country shift, where error rates roughly double across the board. Here evaluation is limited to augmentations due to a 30 day retention window for each instantiation of the dataset. Images captured in France contain noticeably different architectural styles and storefront designs than those captured in US/Mexico/Canada; meanwhile, we are unable to find conspicuous and consistent indicators of the camera and year. This may explain the relative insensitivity of evaluated methods to the camera and year shifts. Overall data augmentation here shows limited benefit, suggesting either that data augmentation primarily helps combat texture bias as with ImageNet-R, or that existing augmentations are not diverse enough to capture high-level semantic shifts such as building architecture.

DeepFashion Remixed. Table 3 shows our experimental findings on DFR, in which all evaluated methods have an average OOD mAP that is close to the baseline. In fact, most OOD mAP increases track IID mAP increases. In general, DFR's size and occlusion shifts hurt performance the most. We also evaluate with Random Erasure augmentation, which deletes rectangles within the image, to simulate occlusion [40]. Random Erasure improved occlusion performance, but Style Transfer helped even more. Nothing substantially

		Hardware	Year		Location
Network	IID	Old	2017	2018	France
ResNet-50	27.2	28.6	27.7	28.3	56.7
+ Speckle Noise	28.5	29.5	29.2	29.5	57.4
+ Style Transfer	29.9	31.3	30.2	31.2	59.3
+ DeepAugment	30.5	31.2	30.2	31.3	59.1
+ AugMix	26.6	28.0	26.5	27.7	55.4

Table 2: SVSF classification error rates. Networks are robust to some natural distribution shifts but are substantially more sensitive than the geographic shift. Here data augmentation hardly helps.

			Size		Occlusion		Viewpoint		Zoom	
Network	IID	OOD	Small	Large	Slight/None	Heavy	No Wear	Side/Back	Medium	Large
ResNet-50	77.6	55.1	39.4	73.0	51.5	41.2	50.5	63.2	48.7	73.3
+ ImageNet-21K *Pretraining*	80.8	58.3	40.0	73.6	55.2	43.0	63.0	67.3	50.5	73.9
+ SE (*Self-Attention*)	77.4	55.3	38.9	72.7	52.1	40.9	52.9	64.2	47.8	72.8
+ Random Erasure	78.9	56.4	39.9	75.0	52.5	42.6	53.4	66.0	48.8	73.4
+ Speckle Noise	78.9	55.8	38.4	74.0	52.6	40.8	55.7	63.8	47.8	73.6
+ Style Transfer	80.2	57.1	37.6	76.5	54.6	43.2	58.4	65.1	49.2	72.5
+ DeepAugment	79.7	56.3	38.3	74.5	52.6	42.8	54.6	65.5	49.5	72.7
+ AugMix	80.4	57.3	39.4	74.8	55.3	42.8	57.3	66.6	49.0	73.1
ResNet-152 (*Larger Models*)	80.0	57.1	40.0	75.6	52.3	42.0	57.7	65.6	48.9	74.4

Table 3: DeepFashion Remixed results. Unlike the previous tables, higher is better since all values are mAP scores for this multi-label classification benchmark. The "OOD" column is the average of the row's rightmost eight OOD values. All techniques do little to close the IID/OOD generalization gap.

improved OOD performance beyond what is explained by IID performance, so here it would appear that in this setting, only IID performance matters. Our results suggest that while some methods may improve robustness to certain forms of distribution shift, no method substantially raises performance across all shifts.

Real Blurry Images and ImageNet-C. We now consider a previous robustness benchmark to evaluate the four major methods. We use the ImageNet-C dataset [12] which applies 15 common image corruptions (e.g., Gaussian noise, defocus blur, simulated fog, JPEG compression, etc.) across 5 severities to ImageNet-1K validation images. We find that DeepAugment improves robustness on ImageNet-C. Figure 5 shows that when models are trained with both AugMix and DeepAugment they set a new state-of-the-art, breaking the trendline and exceeding the corruption robustness provided by training on $1000\times$ more labeled training data. Note the augmentations from AugMix and DeepAugment are disjoint from ImageNet-C's corruptions. Full results are shown in the Supplementary Materials. IID accuracy alone is clearly unable to capture the full story of model robustness. Instead, larger models, self-attention, data augmentation, and pretraining all improve robustness far beyond the degree predicted by their influence on IID accuracy.

A recent work [30] reminds us that ImageNet-C uses various *synthetic* corruptions and suggest that they are decoupled from real-world robustness. Real-world robustness requires generalizing to naturally occurring corruptions such as snow, fog, blur, low-lighting noise, and so on, but it is an open question whether ImageNet-C's simulated corruptions meaningfully approximate real-world corruptions.

We evaluate various models on Real Blurry Images and find that *all* the robustness interventions that help with ImageNet-C also help with real-world blurry images. Hence ImageNet-C can track performance on real-world corruptions. Moreover, DeepAugment+AugMix has the lowest error rate on Real Blurry Images, which again contradicts the synthetic vs natural dichotomy. The upshot is that ImageNet-C is a controlled and systematic proxy for real-world robustness.

Our results, which are expanded on in the Supplementary Materials, show that larger models, self-attention, data augmentation, and pretraining all help, just like on ImageNet-C. Here DeepAugment+AugMix attains state-of-the-art. These results suggest ImageNet-C's simulated corruptions track real-world corruptions. In hindsight, this is expected since various computer vision problems have used synthetic corruptions as proxies for real-world corruptions, for decades. In short, ImageNet-C is a diverse and systematic benchmark that is correlated with improvements on real-world corruptions.

6. Conclusion

In this paper we introduced four real-world datasets for evaluating the robustness of computer vision models: ImageNet-Renditions, DeepFashion Remixed, StreetView

Method	ImageNet-C	Real Blurry Images	ImageNet-R	DFR
Larger Models	+	+	+	−
Self-Attention	+	+	−	−
Diverse Data Augmentation	+	+	+	−
Pretraining	+	+	−	−

Table 4: A highly simplified account of each method when tested against different datasets. Evidence for is denoted "+", and "−" denotes an absence of evidence or evidence against.

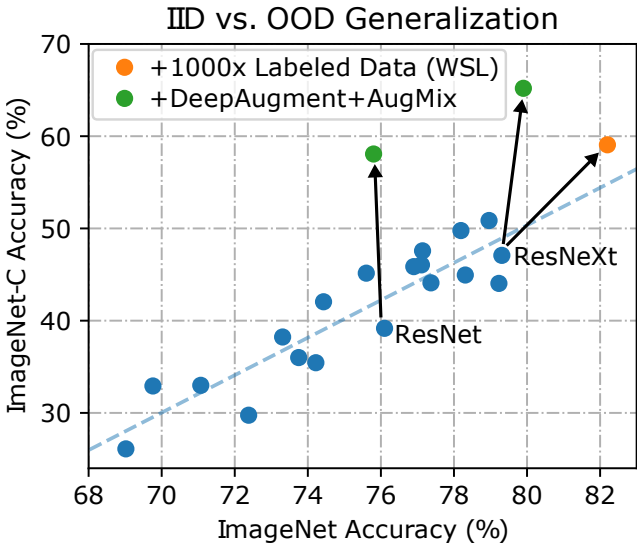

Figure 5: ImageNet accuracy and ImageNet-C accuracy. Previous architectural advances slowly translate to ImageNet-C performance improvements, but DeepAugment+AugMix on a ResNet-50 yields approximately a 19% accuracy increase. This shows IID accuracy and OOD accuracy are not coupled, contra [30].

StoreFronts, and Real Blurry Images. With our new datasets, we re-evaluate previous robustness interventions and determine whether various robustness hypotheses are correct or incorrect in view of our new findings.

Our main results for different robustness interventions are as follows. Larger models improved robustness on Real Blurry Images, ImageNet-C, and ImageNet-R, but not with DFR. While self-attention noticeably helped Real Blurry Images and ImageNet-C, it did not help with ImageNet-R and DFR. Diverse data augmentation was ineffective for SVSF and DFR, but it greatly improved accuracy on Real Blurry Images, ImageNet-C, and ImageNet-R. Pretraining greatly helped with Real Blurry Images and ImageNet-C but hardly helped with DFR and ImageNet-R. It was not obvious *a priori* that synthetic data augmentation could improve accuracy on a real-world distribution shift such as ImageNet-R, nor had pretraining ever failed to improve performance in earlier research [30]. Table 4 shows that many methods improve robustness across multiple distribution shifts. While no single method consistently helped across all distribution shifts, some helped more than others.

Our analysis also has implications for the three robustness hypotheses. In support of the *Texture Bias* hypothesis, ImageNet-R shows that standard networks do not generalize well to renditions (which have different textures), but that diverse data augmentation (which often distorts textures) can recover accuracy. More generally, larger models and diverse data augmentation consistently helped on ImageNet-R, ImageNet-C, and Real Blurry Images, suggesting that these two interventions reduce texture bias. However, these methods helped little for geographic shifts, showing that there is more to robustness than texture bias alone. Regarding more general trends across the last several years of progress in deep learning, while IID accuracy is a strong predictor of OOD accuracy, it is not decisive, contrary to some prior works [30]. Again contrary to a hypothesis from prior work [30], our findings show that the gains from data augmentation on ImageNet-C generalize to both ImageNet-R and Real Blurry Images serve as a resounding validation of using synthetic benchmarks to measure model robustness.

The existing literature presents several conflicting accounts of robustness. What led to this conflict? We suspect that this is due in large part to inconsistent notions of how to best evaluate robustness, and in particular a desire to simplify the problem by establishing the primacy of a single benchmark over others. In response, we collected several additional datasets which each capture new dimensions of distribution shift and degradations in model performance not well studied before. These new datasets demonstrate the importance of conducting multi-faceted evaluations of robustness as well as the general complexity of the landscape of robustness research, where it seems that so far nothing consistently helps in all settings. Hence the research community may consider prioritizing the study of new robustness methods, and we encourage the research community to evaluate future methods on multiple distribution shifts. For example, ImageNet models should at least be tested against ImageNet-C and ImageNet-R. By heightening experimental standards for robustness research, we facilitate future work towards developing systems that can robustly generalize in safety-critical settings.

References

[1] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010.

[2] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[3] Irving Biederman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1):38–64, 1988.

[4] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning augmentation policies from data. *CVPR*, 2018.

[5] Jia Deng. Large scale visual recognition. Technical report, Princeton, 2012.

[6] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE, 2017.

[7] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. Identifying statistical bias in dataset replication. *ICML*, 2020.

[8] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019.

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2019.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

[13] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019.

[14] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *ACL*, 2020.

[15] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *ICLR*, 2020.

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *ArXiv*, abs/1907.07174, 2019.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Shoji Itakura. Recognition of line-drawing representations by a chimpanzee (pan troglodytes). *The Journal of General Psychology*, 121(3):189–197, July 1994.

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Large scale learning of general visual representations for transfer. *arXiv preprint arXiv:1912.11370*, 2019.

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[21] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin Dogus Cubuk. Improving robustness without sacrificing accuracy with patch Gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[23] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri abd Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *ECCV*, 2018.

[24] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. *arXiv*, 2015.

[25] A. Emin Orhan. Robustness properties of facebook's ResNeXt WSL models. *ArXiv*, abs/1907.07640, 2019.

[26] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? *ArXiv*, abs/1902.10811, 2019.

[27] Evgenia Rusak, Lukas Schott, Roland Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. Increasing the robustness of dnns against image corruptions by playing the game of noise. *arXiv preprint arXiv:2001.06057*, 2020.

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

[29] Masayuki Tanaka. Recognition of pictorial representations by chimpanzees (pan troglodytes). *Animal Cognition*, 10(2):169–179, Dec. 2006.

[30] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. When robustness doesn't promote robustness: Synthetic vs. natural distribution shifts on imagenet, 2020.

[31] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017.

[32] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power, 2019.

[33] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[35] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations*, 2020.

[36] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. 2016. *arXiv preprint arXiv:1611.05431*, 2016.

[37] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D Cubuk, and Justin Gilmer. A Fourier perspective on model robustness in computer vision. *arXiv preprint arXiv:1906.08988*, 2019.

[38] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *ICCV*, 2019.

[39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[40] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.