A New Dataset Based on Images Taken by Blind People for Testing the Robustness of Image Classification Models Trained for ImageNet Categories

Reza Akbarian Bafghi
University of Colorado, Boulder
reza.akbarianbafghi@colorado.edu

Danna Gurari
University of Colorado, Boulder
danna.gurari@colorado.edu

Abstract

Our goal is to improve upon the status quo for designing image classification models trained in one domain that perform well on images from another domain. Complementing existing work in robustness testing, we introduce the first dataset for this purpose which comes from an authentic use case where photographers wanted to learn about the content in their images. We built a new test set using 8,900 images taken by people who are blind for which we collected metadata to indicate the presence versus absence of 200 ImageNet object categories. We call this dataset VizWiz-Classification. We characterize this dataset and how it compares to the mainstream datasets for evaluating how well ImageNet-trained classification models generalize. Finally, we analyze the performance of 100 ImageNet classification models on our new test dataset. Our fine-grained analysis demonstrates that these models struggle on images with quality issues. To enable future extensions to this work, we share our new dataset with evaluation server at: `https://vizwiz.org/tasks-and-datasets/image-classification`.

1. Introduction

A common approach for designing computer vision solutions is to leverage large-scale datasets to train algorithms. Yet, for many real-world applications, it is not only inefficient to curate such training datasets but also challenging or infeasible. To address this problem, *robustness testing* was recently introduced with the focus of improving the performance of models trained for one domain on a test set in a different domain. In this paper, we focus on robustness testing for the image classification problem.

To date, progress with classification robustness testing has been possibly largely because of numerous publicly-available test datasets with distribution shifts from the original domain. While such datasets have been beneficial in catalyzing progress, they are limited in that they originate from contrived settings. For example, ImageNet-C [15] consists of real images with synthetically generated corruptions to assess model robustness for corrupted images. Yet, as shown in prior work [3], images curated from contrived settings can lack the diversity of challenges that emerge in real-world applications. A consequence of this lack of diversity in test datasets is that algorithm developers do not receive feedback about whether their methods generalize to the range of plausible real-world vision challenges.

We address the above gap for robustness testing by introducing a new test set for image classification. It consists of 8,900 images taken by people who are blind who were authentically trying to learn about images they took with their mobile phone cameras. For each image, we asked crowdworkers to indicate which from 200 object categories were present. We call the resulting dataset VizWiz-Classification. Examples demonstrating how labelled images in our new dataset compare to those in a related robustness testing dataset are shown in Figure 1. We next analyze how our dataset compares to six existing robustness testing datasets and benchmark the performance of 100 modern image classification models on this dataset to highlight challenges and opportunities that emerge for the research community.

Success on our new dataset could benefit real-world applications today. Already, a growing number of blind people are sharing their images with services such as Microsoft's Seeing AI, Google's Lookout, and TapTapSee, which recognize a small number of object categories. Success could broaden such benefits to a longer tail of categories including those underrepresented in the developing world where it can be laborious/infeasible to collect large, labeled datasets, especially from such a specific population as people who are blind. More generally, our new dataset challenge will encourage developing algorithms that handle a larger diversity of real-world challenges. This could benefit applications with similar challenges such as robotics and wearable lifelogging. Finally, image classification is a precursor for many downstream tasks and so we expect progress on our dataset to enable progress on downstream tasks such as object detection, segmentation, and tracking.

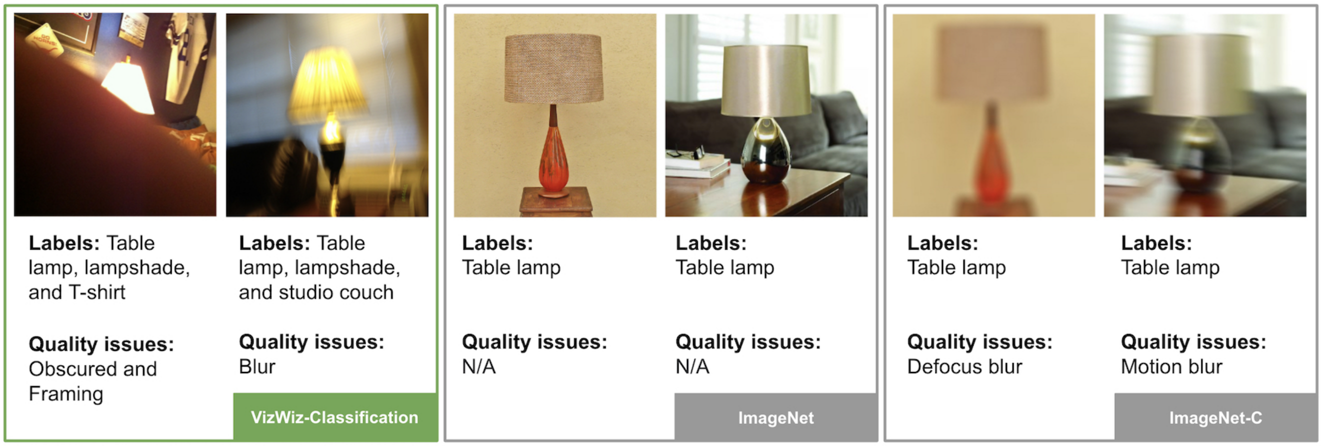
Labels: Table lamp, lampshade, and T-shirt
Quality issues: Obscured and Framing
Labels: Table lamp, lampshade, and studio couch
Quality issues: Blur
VizWiz-Classification
Labels: Table lamp
Quality issues: N/A
Labels: Table lamp
Quality issues: N/A
ImageNet
Labels: Table lamp
Quality issues: Defocus blur
Labels: Table lamp
Quality issues: Motion blur
ImageNet-C

Figure 1. Example labelled images from our new VizWiz-Classification dataset, ImageNet [7], and ImageNet-C [15], where each has the label of "Table lamp". When comparing our dataset to these prior works, (1) our images were taken by blind people who wanted to learn about their environment, whereas ImageNet images were collected from the Internet and ImageNet-C images consist of ImageNet images that were synthetically corrupted and (2) our images can have multiple labels (e.g. also includes a "Lampshade"), while ImageNet and ImageNet-C permit only a single label, which can lead to issues for prediction models when multiple categories are present in an image.

2. Related Works

Domain Adaptation. Research with domain adaptation datasets [15, 21, 26, 27, 35] centers on the development of domain adaptation models [11, 12, 32, 33] that use the data from the target domain (in a supervised or unsupervised fashion) to learn how to generalize to the data of that target domain. Our work, in contrast, is focused on testing models without permitting any training of models on images in our new robustness testing dataset.

Robustness Benchmarks. There has been a recent surge of interest in characterizing model robustness on new test sets [15, 17, 20, 23, 31]. For example, ImageNet-C [15] supports benchmarking model robustness to image corruptions, and was created by synthetically corrupting images in ImageNet's validation set. In ImageNetV2 [23], the authors replicated the data collection process taken by the authors of the original ImageNet dataset in order to create new test datasets. ImageNet-A [17] consists of adversarially filtered images for which classifiers predicted a wrong label with high confidence. ImageNet-O [17] is an out-of-distribution dataset that includes adversarially filtered images by excluding ImageNet-1k images from ImageNet-22k. ObjectNet [2] consists of images taken by crowdworkers who were hired to photograph objects in pre-defined poses, viewpoints, and backgrounds. ImageNet-R [14] consists of renditions of ImageNet categories such as cartoons, paintings, and graffiti. Complementing existing datasets, we introduce the first classification robustness testing benchmark based on pictures taken in real-world scenarios where blind people wanted to know about the content. In addition, our dataset is the first to include images that are naturally corrupted [3].

Model Robustness. Recent work has evaluated the performance of many models for classification robustness, mostly on images sharing the same semantic categories as ImageNet [8, 14, 17, 23, 31, 34]. We similarly benchmark many models on our new dataset that reflect a variety of aspects, including their architectures [19, 39] and training methods such as usage of data augmentation [5, 6, 16], adversarial attacks [4, 10, 28, 29, 36, 38], and model pretraining [25, 38]. For our analysis, we categorize the models into three groups: standard models, robust models, and models trained with more data. We then measure and compare the robustness of each group. We also conduct fine-grained analysis to understand how model performance relates to quality issues in our new dataset.

3. Dataset

We now describe our new dataset for classification robustness testing, which we call "VizWiz-Classification" or "VW-C". We begin in Section 3.1 by describing our process for creating the dataset, which consists of two key parts. Specifically, we first use automation to identify 15,567 candidate images that likely contain the ImageNet categories of interest from an initial collection of over 39,000 images. We describe this *candidate image and category selection* process in Section 3.1.1. Then, we leverage human annotation to produce our high-quality labeled dataset. We describe this *manual data annotation* process in Section 3.1.2. We then characterize our new dataset and how it compares to existing robustness testing datasets in Section 3.2.

3.1. Dataset Creation

3.1.1 Candidate Image and Category Selection

Image Source. We use images taken by blind photographers who shared the pictures with remote humans in order to get assistance in learning about their visual surroundings. We leverage the images with metadata that come from two publicly-shared datasets: VizWiz-Captions [13] and VizWiz-ImageQualityIssues [3]. In total, there are 39,189 images across the train, validation, and test splits. The metadata consists of five captions as well as flags indicating which of the following quality issues are observed for each image: blur (BLR), underexposure/too dark (DRK), overexposure/too bright (BRT), poor framing (FRM), obstructions/obscured (OBS), rotated views (ROT), other flaws (OTH), and no flaws (NON).[1]

Candidate Category Selection. Since we want to benchmark models trained on ImageNet, we need to identify which ImageNet categories to include in our dataset. To do so, we leverage the captions of each image of the VizWiz-Captions dataset [13]. Using string matching, we detect which of the 1000 ImageNet categories are present across all images' captions, which is a total of 505 categories. We then identify which categories are found in captions of more than 7 images, which left us with 250 ImageNet categories. We then removed 58 of these categories, because the authors deemed them to be either ambiguous (e.g., boxer can refer to a dog and athletic) or non-obvious to lay audiences (e.g., specific breeds of dogs and cats). Finally, we added another 23 categories from classes of ImageNet found in captions of less than or equal to 7 images, which the authors deemed to be well-defined (e.g, Christmas stocking and piggy bank). This left us with 218 categories candidate categories for our dataset.

Candidate Image Selection. Having finalized our candidate categories, we then filtered our initial image set to only include those for which our chosen categories appeared in at least two of the five associated captions. This resulted in a total of 15,567 candidate images for our dataset.

3.1.2 Manual Data Annotation

Annotation Tasks. We designed our task interface in Amazon Mechanical Turk (AMT). It showed an image on the left with 10 object categories to the right of the image. The instructions indicated to select all categories that are observed in the image. For each image, we initially provided any categories and nouns that were found in the captions of the image. We then included an additional task of indicating whether additional objects beyond those 10 categories are present in the image in order to decide whether we further review was needed to assess whether additional ImageNet categories are present. Consequently, every task (i.e. human intelligence task, also known as HIT) had 12 possible answers (Question 1 has 10 suggested categories and one checkbox for "None of the above", and question 2 has one option which can be "Yes" or "No").

Annotation Collection. We recruited crowdworkers from AMT who previously had completed over 500 HITs with at least a 99% acceptance rate. Each prospective worker had to pass a qualification test that we provided, which showed five difficult annotation scenarios. The authors established ground truth and then only accepted workers who had an accuracy of more than 90% (i.e. at most 6 wrong answers from 12 × 5 possible answers). From the 100 workers who took our qualification task, 38 workers passed it. Before permitting them to contribute to our dataset annotation, we provided feedback to each worker based on their performance on the qualification task.

We then completed a subsequent round of worker filtering with $1,000$ of our candidate images. We hired 2 workers to annotate each image. We then reviewed all annotated images in this step and only permitted workers (31 workers) to continue on our data annotation task if they more than 90% of their submitted answers were correct (i.e. 12 × all his or her submitted HITs). For workers invited to continue annotating for our task, we again provided feedback to each one individually based on their performance.

For all remaining data collection, we had three different workers annotate each image. We then assigned final category labels using the majority vote; i.e. assign a category to an image only if at least two workers indicate it is present. Additionally, when at least two crowdworkers indicated that additional objects were present in the image beyond the 10 categories, the authors reviewed the image and checked for the presence of additional ImageNet categories. To support ongoing high-quality results throughout data collection, we also conducted ongoing quality control on the workers' results, as described in the Supplementary Materials.

Dataset Finalization. After annotating for the presence of the 218 candidate categories in all 15,567 candidate images, we focused only on categories found in at least 4 images and images with those categories. After conducting this final round of filtering, our final dataset includes 200 categories and 8,900 images.

3.2. Dataset Analysis

In this section, the VizWiz-Classification dataset is analyzed and compared to other popular ImageNet test datasets.

[1] We use quality flaw labels for each image based on if at least 2 of the 5 crowdworkers provided the labels. Since we did not have access to test annotations, we only have quality issues labels for a subset (i.e. 7,147 images) of our dataset.

Dataset	#Images	#Classes	Images/class		Authentic	Corrupted
			#Min	#Max		
ImageNet-A [17]	7500	200	3	100	✗	✗
ImageNet-C [15]	50000	1000	50	50	✗	✓
ImageNetV2 [24]	10000	1000	10	10	✗	✗
ImageNet-O [17]	2000	200	5	30	✗	✗
ImageNet-R [14]	30000	200	51	430	✗	✗
ObjectNet [2]	50000	313 (113)	11	284	✗	✗
Ours	8900	200	4	1311	✓	✓

Table 1. Characterization of our dataset and six related datasets with respect to five factors. Our dataset is the first to come from an authentic use case and so to provide corruptions that are not contrived.

Characteristics of Datasets. We compare our dataset with six mainstream robustness testing datasets for image classification: ImageNet-A [17], ImageNet-C [15], ImageNetV2 [24], ImageNet-O [17], ObjectNet [2], and ImageNet-R [14]. For each dataset, we report how many images are included, how many classes are supported, the number of examples per class (i.e. minimum and maximum values), whether images originate from a contrived versus authentic use case, and whether labels are included with the dataset indicating corruptions/quality issues. Results for each dataset are shown in Table 1.

As shown, our dataset is the first to reflect images which originate from a real-world application. In particular, images come from an authentic use case where blind people wanted to learn about an image's contents. In contrast, other datasets leverage images that originate from contrived settings such as downloading images from the Internet or hiring workers to take pictures to support dataset creation.

Another unique characteristic of our dataset is it is one of only two datasets that support robustness testing with respect to image corruptions. By leveraging annotations from prior work indicating what quality flaws are present for each image in our dataset [3], our test dataset supports fine-grained analysis of model performance with respect to how models handle each. In addition, the distribution of flaws in our dataset reflects a naturally-occurring distribution that arises for an authentic use case. The number of images in our dataset with each quality flaw label is shown in Figure 2. The other dataset with corruption labels is ImageNet-C. However, those image corruptions are generated synthetically and the distribution of corruption labels is artificially chosen.

We also characterize for our dataset how many object labels are associated with each image. We report summative statistics across all images in our dataset in Table 2. This analysis is provided to address the finding from [34] that the performance of image classification models suffers when cluttered images in ImageNet are classified using only a single label. A further observation in [34] is that ImageNet contains categories with synonymous meanings, leading to confusion for models. We aimed to minimize such errors by permitting multiple labels per image.

#Classes	1	2	3	4	5
#Images	5731	2186	727	204	52

Table 2. Number of classes assigned per image. As shown, images in our dataset often have more than one label.

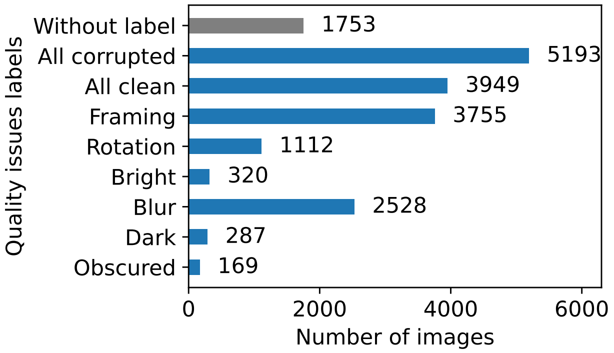

Figure 2. Number of images per each quality issue label. Clean images include images with NON label. Corrupted images consist of images with at least one label from FRM, ROT, BRT, BLR, DRK, and OBS. 1,753 images do not have any labels because we did not have access to their labels.

Dataset Diversity. We now investigate the diversity of images in robustness testing datasets. Inspired by a metric introduced by [40], we claim that dataset A is more diverse than dataset B if categories of dataset A have more diverse images than dataset B. As a consequence, overall, it would mean that images of dataset A that are in the same category would represent a broader distribution. We calculate the difference between a pair of images by calculating the cosine distance between the feature vectors of images. The larger the value of the distance is, the more diverse images are. Then, in each category of a dataset, we calculate the cosine distance between each pair of feature vectors of im-

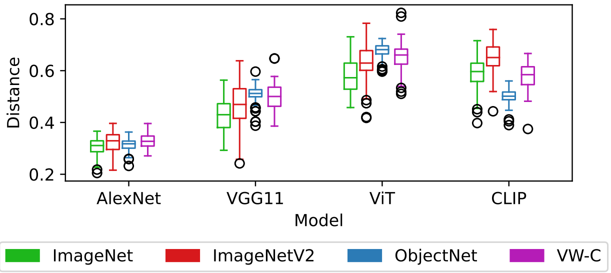

Figure 3. Distribution of mean distance between images of common categories of ImageNet, ImageNetV2, ObjectNet, and VizWiz-Classification. Our dataset has more diverse images than ImageNet and ImageNetV2.

ages. The mean of the distances between all possible image pairs in each category gives us the expected value of the distances of images in the category. We did this to reduce the error that could happen in the annotation process of images (e.g. images that are wrongly annotated for a class).

For our experiment, we choose to compare the diversity of images for the 73 categories which were in common among the following four datasets: ImageNet, ImageNetV2, ObjectNet, and VizWiz-Classification. We excluded ImageNet-O since it is an out-of-distribution dataset. We also excluded datasets with synthesized images, specifically ImageNet-C which includes images with generated corruptions and ImageNet-R which includes renditions of the categories such as paintings and embroidery. We did not include ImageNet-A in our experiment because it only contains 17 common object categories with selected datasets.

We choose AlexNet [18], VGG11 [30], ViT [9], and CLIP [22] for extracting feature vectors. The 'fc7' layer of AlexNet and the 'fc2' layer of VGG11 were selected for extracting the feature vector because we assume that a simpler feature vector enables us to compare images based on attributes such as color and quality rather than semantic information. Also, we use penultimate layer features of ViT [9] and CLIP [22] in order to generate feature vectors that are more robust to occlusions, texture bias, etc [1].

Results are shown in Figure 3. All models show a similar order in terms of diversity in visual description and semantic information. Based on our results, ObjectNet has slightly more diverse images since the mean of all distances is higher than other datasets. Also, ImageNetV2 and ImageNet have similar distributions. The authors of ImageNetV2 tried to reproduce the ImageNet dataset with the same protocols; thus, for those datasets, the distributions of feature vectors match. Further, all models show that our dataset is more diverse than ImageNet and ImageNetV2. The more diversity in our images per category with respect to visual and semantic descriptions can be because the source of images in our dataset is different from ImageNet and ImageNetV2.

4. Algorithm Benchmarking

By benchmarking a wide variety of image classification models on the VizWiz-Classification dataset, we explore their robustness on our new test dataset. We perform analysis of the effect of quality issues and the distribution of classes on the performance of models.

Models. We select 100 models for evaluation on our test dataset. We leverage the test bed provided by [31] in order to calculate accuracy on ImageNet and ImageNet-C and categorize models. Following their work, models are divided into three subclasses: standard models that are trained on ImageNet and do not benefit from any methods for increasing robustness (30 models), models that are trained on a larger set of training datasets such as ImageNet-21k [25] or IG-1B-Targeted [38]) (10 models), and models that leverage robustness intervention methods such as data augmentation and adversarial attack methods (60 models). More details are provided in the Supplementary Materials.

Evaluation Metrics. For benchmarking models on ImageNet and ImageNet-C, we used the standard accuracy top-1 metric. But for calculating model accuracies on our dataset, we consider a prediction for an image correct if it is in the set of labels of the image. Based on top-1 accuracy, the performance of models on multi-label images is much inferior, and top-5 accuracy is optimistic [34]. On this metric, [34] discover that the aforementioned performance reduction vanishes, and models perform comparably on single-object and multi-object images.

Effective Robustness. As described in [31], directly comparing the accuracies of models is not an accurate metric. By contrast, they called a model robust if the accuracy of the model is above the linear trend that models follow. We follow this notation for finding robust models. We fit a linear polynomial to points by minimizing the squared error for learning the linear trend of a set of points, as described in more detail in the Supplementary Materials.

Performance Gap. We expect the performance of models does not change with distribution shifts, and we call the area between the expected accuracy ($y = x$ line) and the actual accuracy (the linear trend) the performance gap. Robust models should be able to fill this gap and not experience an accuracy drop when we test them on a new dataset.

4.1. The effect of different quality issues

We first test whether the models are robust to the real-world corruptions of our dataset. We calculate accuracy on images with corrupted, clean, and all images. Figure 4 shows the results of the 100 models.

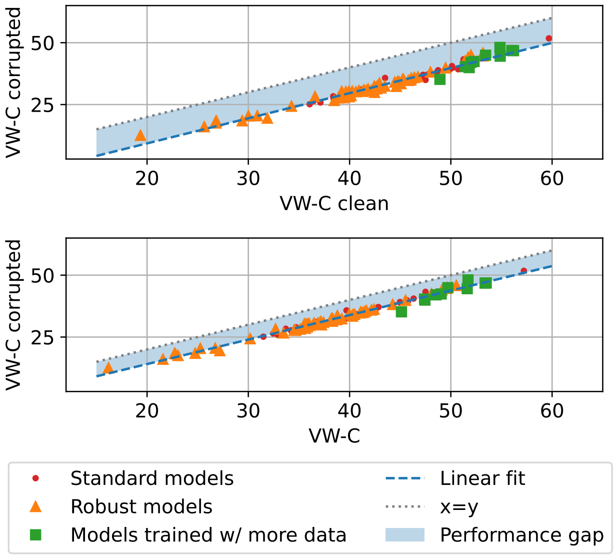

Figure 4. Model accuracies on clean and all images (x-axis) and all corrupted images (y-axis) of VizWiz-Classification. The performance gap increased when we compare corrupted images to clean images rather than all images of our dataset.

Distribution shift	PG	ER		
		S	R	D
VW-C clean → VW-C corr.	10.3	-0.3	0.04	**0.1**
IN → VW-C	35.8	-1.3	0.1	**3.7**
IN-C → VW-C corr.	9.5	0.1	-0.2	**4.6**

Table 3. Describing performance gap (PG) and effective robustness (ER) in different cases. S: Standard models. R: Robust models. D: Models with more data. Corr.: corrupted. IN: ImageNet. IN-C: ImageNet-C. We can observe the performance gap in all cases. Also, the most robust models to distribution shifts are models pre-trained on larger datasets. However, when comparing clean images to corrupted images in our dataset, none of the models show effective robustness. All numbers are percentages.

We observe that none of those models are robust to corruption because the accuracies of all models are placed on the linear fit to points. Only models that are trained on a larger set of training data can make little effective robustness and are slightly above the line. While the performance of a robust model should be the same for corrupted, clean, and all images, instead a performance gap can be seen between these cases. We notice that the range of accuracy drops for comparing all images and corrupted images are from 3.6% to 7.7% and the performance gap is 6.1%. The range of accuracy drops is even larger than in the past case if we compare clean images and corrupted images, which is from 6.7% to 12.8%. The performance gap, in this case, is 10.3%. In conclusion, we find out that the quality of images plays an important role in the performance of models, and, as indicated in Table 3, none of the models can fill this gap.

We also compare accuracies of models with respect to different quality issues and clean images. Figure 5 illustrates how these flaws affect the performance. The order of performance gap based on the quality issue is as dark (5.3%), blurred (6%), framing (7.5%), bright (11.1%), rotated (14.6%), and obscured (17.5%). However, it is worth mentioning that the number of dark, obscured, and bright images in our dataset are 287, 169, and 320 sequentially, which causes making a conclusion about them problematic. Overall, we observe a performance gap for all quality flaws.

Finally, we report about the five models with the best accuracy on our dataset among all 100 models in Table 4. Except for VOLO [39] that leverages a new architecture, all models are trained on more data.

4.2. Measuring robustness of models

First, we aim to evaluate the robustness of 100 models to distribution shifts. As we described in Section 3.2, shifts can occur in the distribution of images and labels. We compare the accuracy of the models on ImageNet to the accuracy of the models on corrupted, clean, and all images of our dataset. Figure 6 shows our results in detail. Based on our result, the range of accuracy drops for clean, corrupted, and all images of our dataset are from 12.6% to 38.1%, from 19.3% to 49.6%, and from 15.7% to 42.3%. Also, the performance gap for clean, corrupted, and all images of our dataset are 31.6%, 41.9%, and 35.9%, which shows there is a large performance gap even for clean images. In this experiment, models trained on more data can produce effective robustness and are located above the fitted line. But, models that use robust interventions provide no robustness when they are tested on our dataset.

4.3. Can ImageNet-C track real-world corruptions?

ImageNet-C consists of 15 varieties of synthetically generated corruptions with five degrees of severity. We compare this dataset to our dataset in three cases (two common corruption types and all images). In this experiment, we compare the performance of 90 models from our 100 models on the ImageNet-C and VizWiz-Classification datasets.

The first aspect is bright images. We use the brightness subset of ImageNet-C and bright images of VizWiz-Classification for our evaluation. The second case is blurred images. ImageNet-C provided six forms of blur corruptions (zoom blur, motion blur, glass blur, gaussian blur, and defocus blur) with five levels of severity. We calculate the accuracy of models for all 30 mentioned sub-sets and then use the mean of them to compare with blurred images of VizWiz-Classification. In the end, we compare accuracies for all images in ImageNet-C with all corrupted images in VizWiz-Classification. Figure 7 depicts our results in detail.

The performance gap between these datasets exists. We

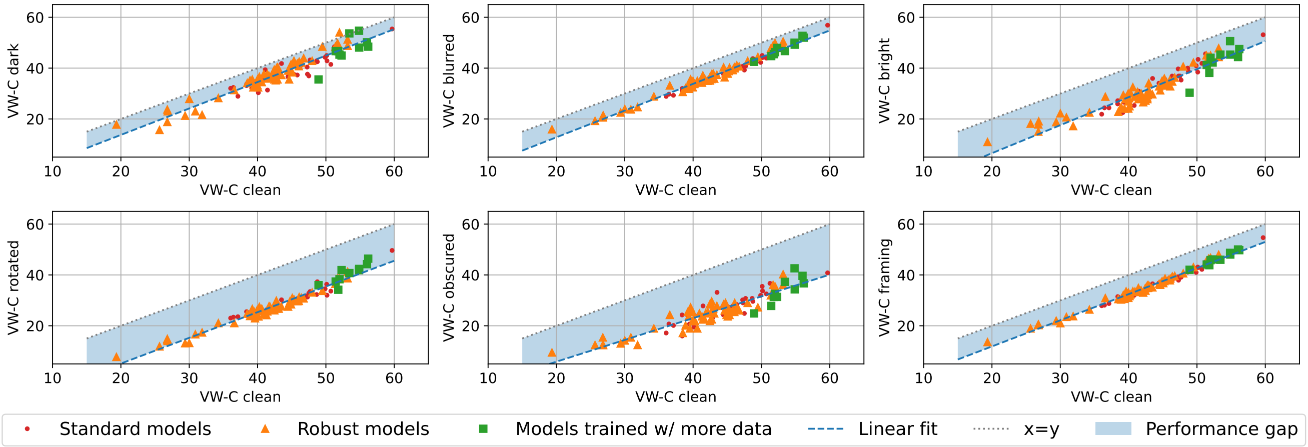

Figure 5. Model accuracies on images with different quality issues (y-axis) and clean images (x-axis) of VizWiz-Classification. The order of performance gap based on the image quality issues from the lowest to highest is dark, blurred, framing, bright, rotated, and obscured.

Network	IN	VizWiz-Classification								
		OBS	DRK	BLR	BRT	ROT	FRM	Corr.	Clean	All
VOLO-d5-512 [39]	87.0	40.8	**55.4**	**56.9**	**53.1**	**49.6**	**54.7**	**51.8**	**59.7**	**57.2**
ConvNeXt-b-IN22k [19]	85.8	39.6	50.2	52.7	44.4	44.2	50.0	46.9	56.0	53.5
ViT-large-p16-384 [9]	**87.1**	36.7	48.4	52.0	47.5	46.4	49.7	46.8	56.2	53.4
ResNeXt101-IG [37]	84.2	**42.6**	54.7	50.0	50.6	41.9	48.7	48.1	54.8	51.7
ViT-base-p16-384 [9]	86.0	34.3	48.1	49.2	45.3	42.4	48.0	44.6	54.9	51.6

Table 4. Top five models based on accuracy for all images in our dataset. IN shows the accuracy of models on ImageNet. Corr. shows the accuracy of models on all corrupted images of our dataset. All models, except for VOLO [39] which uses a different architecture, have been trained on a larger dataset.

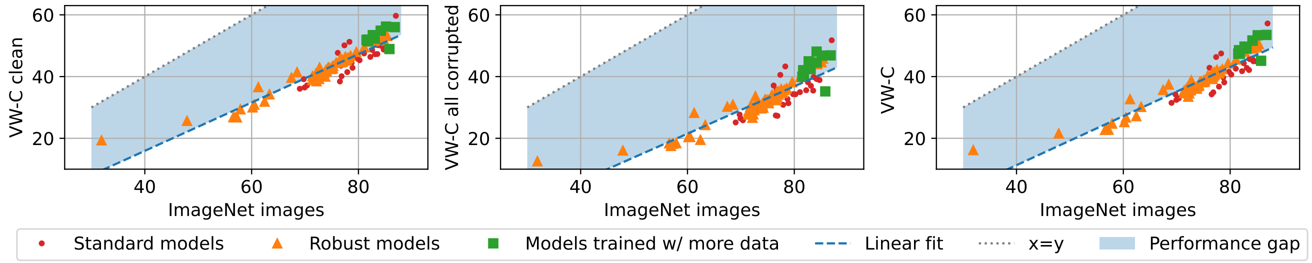

Figure 6. Model accuracies on clean, corrupted, and all images of VizWiz-Classification (y-axis) and ImageNet (x-axis). As shown, the range of accuracy drops for clean, corrupted, and all images from 12.6% to 38.1%, 19.3% to 49.6%, and 15.7% to 42.3% respectively.

notice that the range of accuracy drops for bright images, blurred images, and all images are from 16.2% to 41.1%, from -2.8% to 27.3%, and from -0.1% to 24.2% respectively. Also, the performance gap for bright images, blurred images, and all images are 32%, 12%, and 9.5%. The performance gap is reduced when we compare all images of our dataset to all images of ImageNet-C. In addition, the slope of the linear fit for bright images, blurred images, and all corrupted images is 0.69, 0.42, and 0.56, which means that the progress of the performance of models in ImageNet-C does not guarantee a comparable increase in accuracy for real images of VizWiz-Classification and it underestimates quality issues that images may have in reality. Overall, although models with lower accuracy on ImageNet-C have similar performance on all corrupted images of ImageNet-C and VizWiz-Classification datasets, the gap increases when we compare models with higher accuracy on ImageNet-C. In addition, based on Table 3, we observe that only models trained with more data could have effective robustness. We hypothesize that because models that leverage robust-

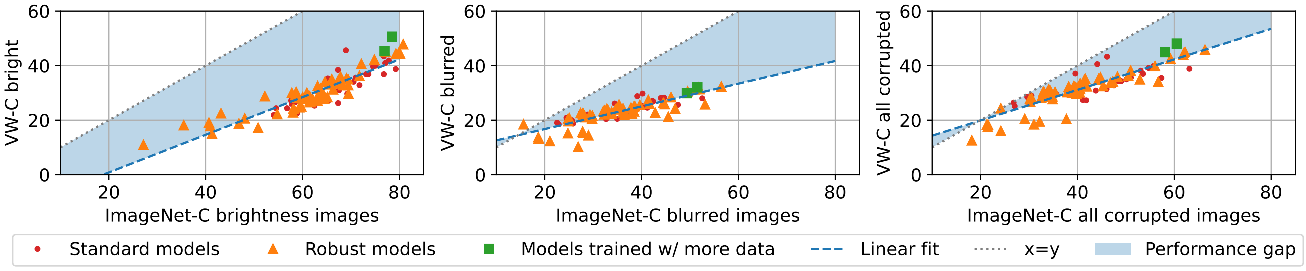

Figure 7. Model accuracies on bright, blurred, and all corrupted images of VizWiz-Classification (y-axis) and brightness, blurred, and all corrupted images of ImageNet-C (x-axis) respectively. The performance of models with lower accuracy on ImageNet-C resembles their performance on VizWiz-Classification, but improvements of models on ImageNet-C does not always yield a similar impact on our dataset.

Dataset	#Images	#Classes	Images/class	
			#Min	#Max
VW-C Rare	896	100	4	21
VW-C Common	5265	90	21	278
VW-C Frequent	5005	10	303	1311
VW-C All	8900	200	4	1311

Table 5. We split VizWiz-Classification into three groups. Images of each group can overlap, but classes are distinct.

ness interventions are designed specifically based on the ImageNet-C synthetic corruptions, they are over-optimized for synthetic corruptions and cannot simulate real-world quality issues.

4.4. The effect of the distribution of categories

In this section, we aimed to examine to which extent the natural distribution of categories of our dataset affects the performance of models. To do so, we split our dataset into three groups with respect to the number of images per category for frequent, common, and rare objects, as shown in Table 5. The results for 40 tested standard models are shown in Figure 8. We observe that predicting the label of images with frequent objects was easier for models than other splits. However, model accuracies are similar on all images of our dataset, images with common objects, and images with rare objects. Because we split our dataset based on the number of images per class, each split is more balanced than our dataset. We also infer that the unbalanced distribution of labels in our dataset is not an important factor for finding the robustness of models because the performance of models follows the same trend in each group. Thus, robust models should be the same in each case.

5. Conclusion

We introduce a new test dataset for measuring the robustness of models to a new distribution shift. Our dataset is the first such dataset with images gathered from a real-world, authentic use case, specifically from people with visual impairments. By benchmarking 100 models on our dataset, we find out our dataset is challenging for current models. Quality issues, which our images can have, affect largely the performance of models and can decrease the accuracy of models by 12.8% when compared to the accuracy of models obtained on clean images of our dataset. Also, we examine the effective robustness of models to distribution shift from ImageNet to our dataset. Results indicate that the performance gap is major between the two datasets and models can experience even by 42.3% drop in accuracy, however, models that are trained on a larger dataset are more robust than other models because all of them are positioned above the linear fit. Finally, we find that although the performance of models with lower accuracy on ImageNet-C is comparable to the performance of them on VizWiz-Classification, the progress of models on ImageNet-C does not guarantee similar progress on our dataset.

Our work contributes to ethics by providing a dataset that acknowledges a population often marginalized in society: people who are blind. As a result, it supports progress in designing more inclusive technology.

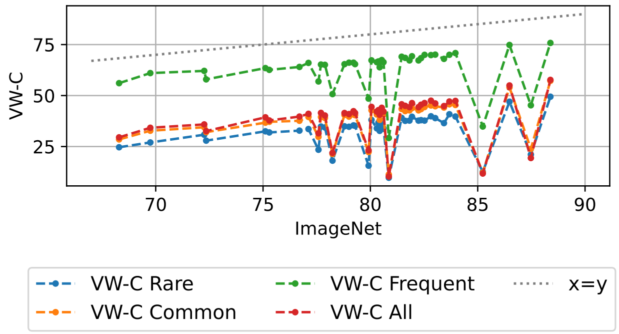

Figure 8. Comparing model accuracies on different splits of our dataset. Model accuracies are similar on all images, images with common objects, and images with rare objects of our dataset.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 5

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua B. Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Neural Information Processing Systems*, 2019. 2, 4

[3] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3643–3653, 2020. 1, 2, 3, 4

[4] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019. 2

[5] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 113–123. Computer Vision Foundation / IEEE, 2019. 2

[6] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[8] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan I. Moldovan, Sylvan Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16453–16463, 2021. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021. 5, 7

[10] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *ICML*, 2019. 2

[11] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *J. Mach. Learn. Res.*, 2016. 2

[12] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2019. 2

[13] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*, 2020. 3

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021. 2, 4

[15] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1903.12261, 2019. 1, 2, 4

[16] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *ArXiv*, abs/1912.02781, 2020. 2

[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Xiaodong Song. Natural adversarial examples. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15257–15266, 2021. 2, 4

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 5

[19] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, 2022. 2, 7

[20] Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *ArXiv*, abs/2102.11273, 2021. 2

[21] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *ArXiv*, abs/1710.06924, 2017. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 5

[23] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *ArXiv*, abs/1902.10811, 2019. 2

[24] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *ArXiv*, abs/1902.10811, 2019. 4

[25] T. Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *ArXiv*, abs/2104.10972, 2021. 2, 5

[26] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2

[27] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 2

[28] Hadi Salman, Greg Yang, Jungshian Li, Pengchuan Zhang, Huan Zhang, Ilya P. Razenshteyn, and Sébastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *ArXiv*, abs/1906.04584, 2019. 2

[29] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019. 2

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[31] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 5

[32] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1378–1388, 2021. 2

[33] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 2

[34] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *ICML*, 2020. 2, 4, 5

[35] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394, 2017. 2

[36] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Loddon Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 501–509, 2019. 2

[37] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. 7

[38] Ismet Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Kumar Mahajan. Billion-scale semi-supervised learning for image classification. *ArXiv*, abs/1905.00546, 2019. 2, 5

[39] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 2, 6, 7

[40] Bolei Zhou, Àgata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 4