University of Colorado **Boulder**

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

WED-PM-372

Project Webpage: https://vizwiz.org/tasks-and-datasets/image-classification

Reza
Akbarian Bafghi

Danna
Gurari

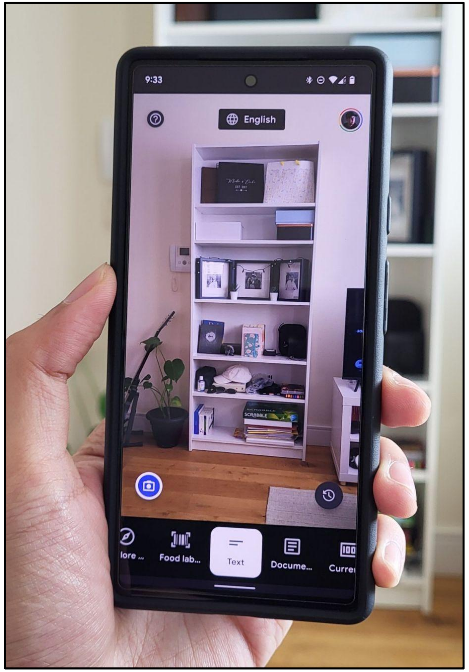
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Overview

Overview

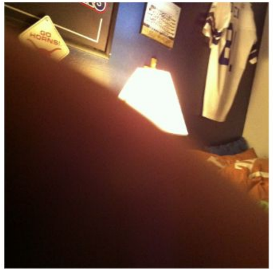
Complementing existing works in robustness testing, we introduce the first dataset for this purpose which comes from **an authentic use case** where blind photographers wanted to learn about the content in their images.

We called our dataset, **VizWiz-Classification** or **VW-C**. It consists of **8,900** images with metadata indicating the presence/absence of **200** ImageNet object categories.



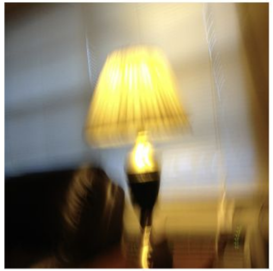
https://phandroid.com/2022/06/24/a-quick-guide-on-how-to-use-androids-lookout-app

Overview



Labels:
Table lamp, lampshade, and T-shirt

Quality issues:
Obscured and Framing



Labels:
Table lamp, lampshade, and studio couch

Quality issues:
Blur

VizWiz-Classification

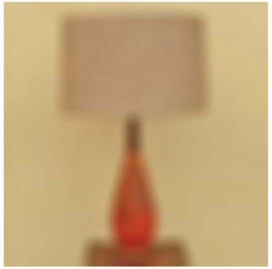
Labels:
Table lamp

Quality issues:
N/A

Labels:
Table lamp

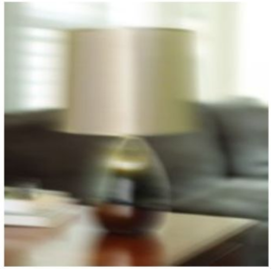
Quality issues:
N/A

ImageNet



Labels:
Table lamp

Quality issues:
Defocus blur

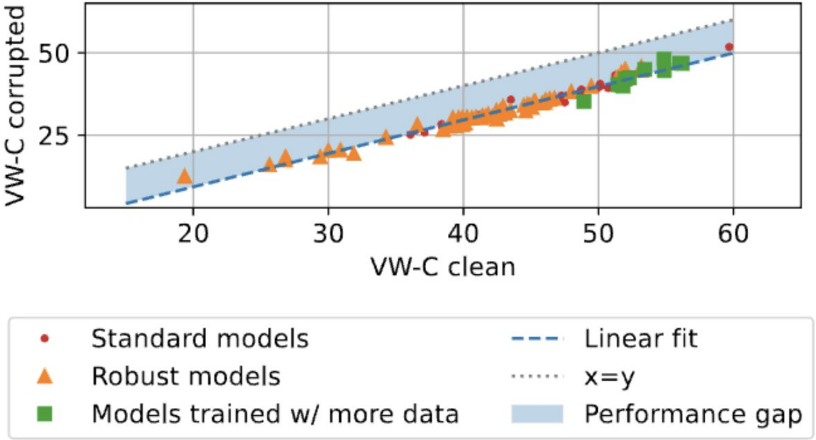


Labels:
Table lamp

Quality issues:
Motion blur

ImageNet-C

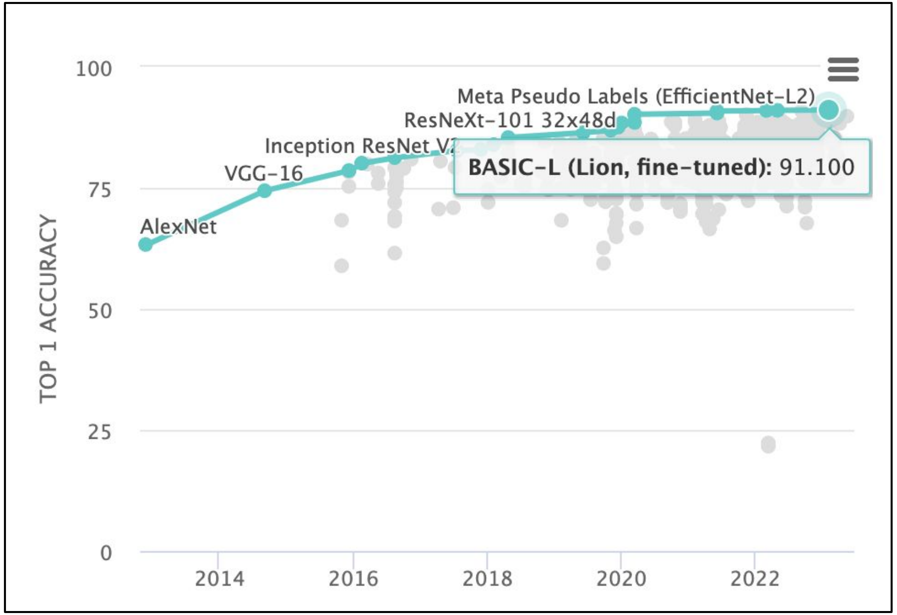
Overview



We analyze the performance of **100** models on our new test dataset. Our fine-grained analysis demonstrates that these models struggle with images with **quality issues**.

Motivation

Motivation



While models exhibit excellent performance on the ImageNet test set, the important question arises regarding their performance in real-world applications and when confronted with distribution shifts.

https://paperswithcode.com/sota/image-classification-on-imagenet

Dataset Creation

Dataset Creation

We use 39,189 images with metadata from the publicly-shared dataset, **VizWiz-Captions.**

A computer screen with a Windows message about Microsoft license terms.

A can of green beans is sitting on a counter in a kitchen.

A photo taken from a residential street in front of some homes with a stormy sky above.

A blue sky with fluffy clouds, taken from a car while driving on the highway.

A hand holds up a can of Coors Light in front of an outdoor scene with a dog on a porch.

A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.

A Winnie The Pooh character high chair with a can of Yoohoo sitting on it in front of a white wall.

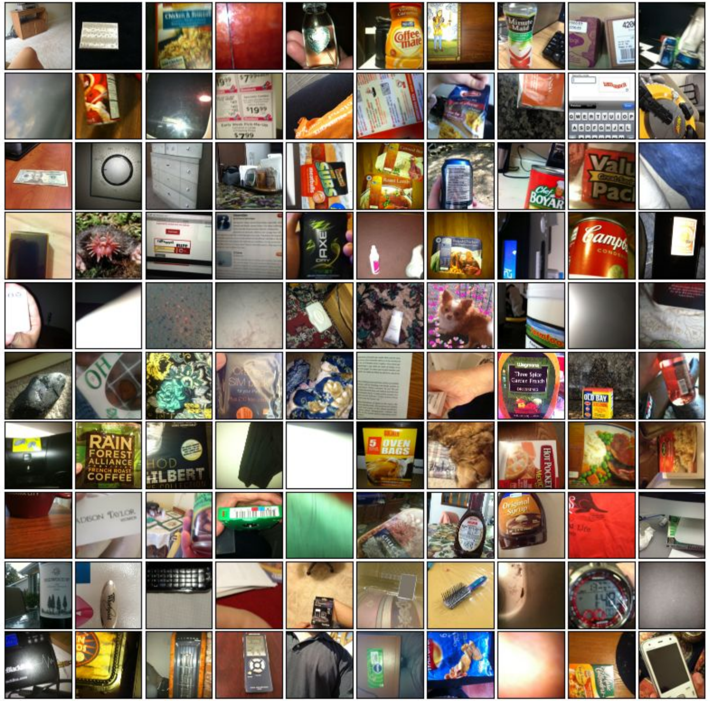
A cup holder in a car holding loose change from Canada.

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

Dataset Creation

Our process for creating the dataset consists of two key parts:

1. Candidate image and category selection
2. Manual data annotation



Number of all images: **39,189**

Dataset Creation

Our process for creating the dataset consists of two key parts:

1. **Candidate image and category selection:**

   We first use automation to identify candidate images that likely contain the ImageNet categories of interest from an initial collection of over 39,000 images.

2. Manual Data Annotation

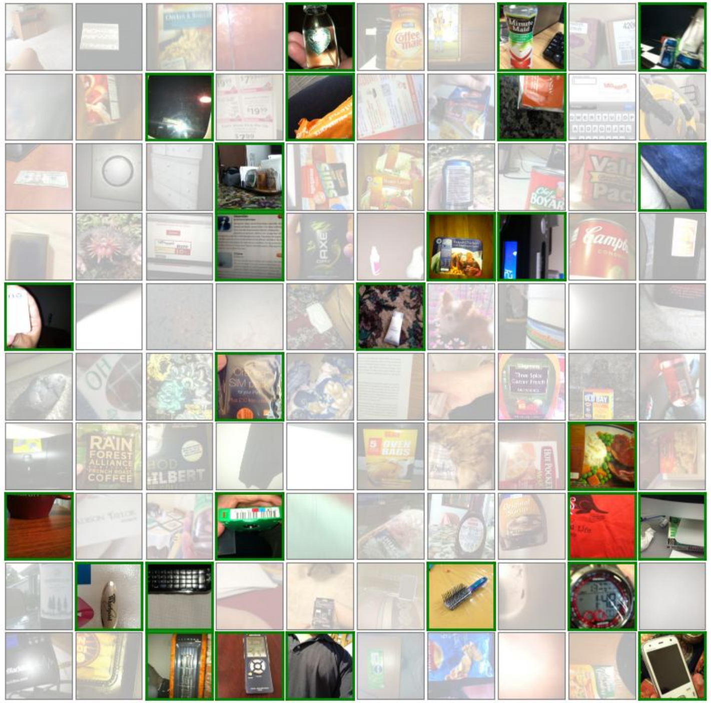
Number of candidate images: **15,567**

Dataset Creation

Our process for creating the dataset consists of two key parts:

1. Candidate image and category selection
2. **Manual Data Annotation:**

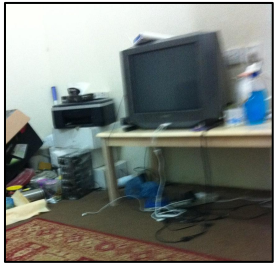
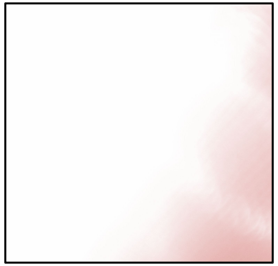
   We leverage human annotation to produce our high-quality labeled dataset.



Number of final images: **8,900**

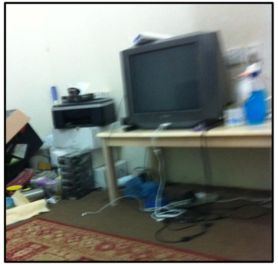
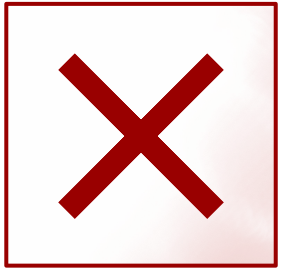
Dataset Creation - Candidate Images and Categories

First, we detect which of the 1000 ImageNet categories are present across all images' captions.

Image				
Caption	A room with a red rug with a TV, and a pile of trash.	A brown young dog sitting on the floor looking up.	A watch with a cartoon character and a shoe in the background.	Quality issues are too severe to recognize visual content.
Candidate Labels	trash can TV prayer rug	boxer	running shoe digital watch	

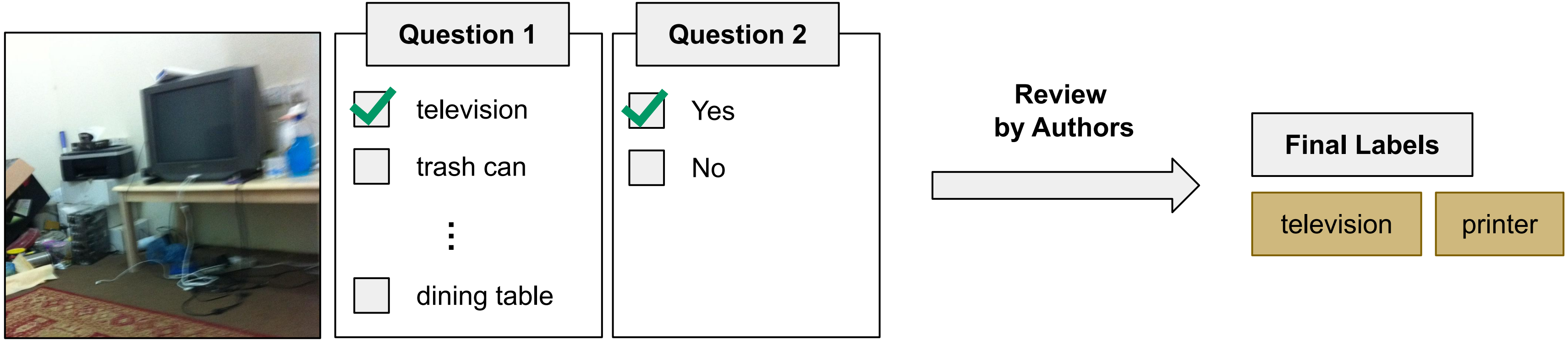
Dataset Creation - Candidate Images and Categories

Then, we selected categories that are obvious to lay audiences. Next, we filtered our initial images based on selected categories.

Image				
Caption	A room with a red rug with a TV, and a pile of trash.	A brown young dog sitting on the floor looking up.	A watch with a cartoon character and a shoe in the background.	Quality issues are too severe to recognize visual content.
Candidate Labels	trash can TV prayer rug	boxer	running shoe digital watch	

Dataset Creation - Manual Data Annotation

For annotation, we asked workers to select observed categories in the image. We then included an additional task of indicating whether additional objects beyond those 10 categories are present in the image.



Question 1

☑ television

□ trash can

⋮

□ dining table

Question 2

☑ Yes

□ No

Review by Authors

Final Labels

television

printer

Dataset Analysis

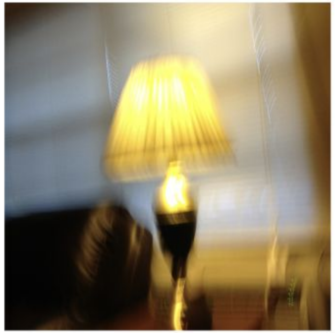
Dataset Analysis - Characteristics of Datasets

Compared to other datasets, ours is the **first** originating from an authentic use case and also containing authentically corrupted images.

Dataset	**#Images**	**#Classes**	**Images/Class**		**Authentic**	**Corrupted**
			#Min	**#Max**		
ImageNet-A	7500	200	3	100	-	-
ImageNet-C	50000	1000	50	50	-	**+**
ImageNetV2	10000	1000	10	10	-	-
ImageNet-O	2000	200	5	30	-	-
ImageNet-R	30000	200	51	430	-	-
Ours	8900	200	4	1311	**+**	**+**

Dataset Analysis - Characteristics of Datasets

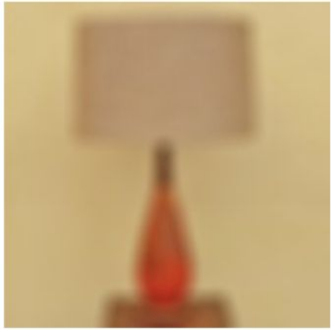
Our dataset and ImageNet-C are only two datasets that support robustness testing with respect to image corruptions. Corruptions in ImageNet-C are generated synthetically and the distribution of corruption labels is artificially chosen.



Corruption

blur

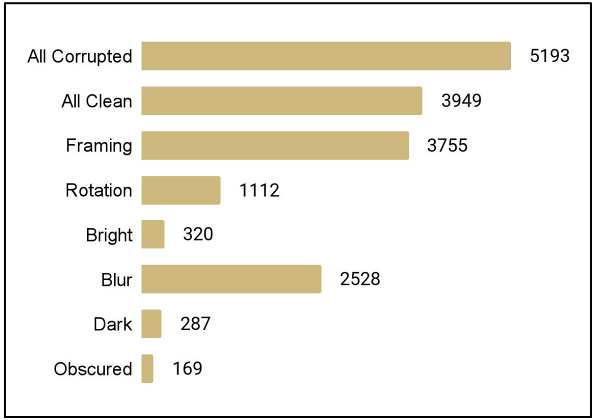
VizWiz-Classification



Corruption

defocus blur

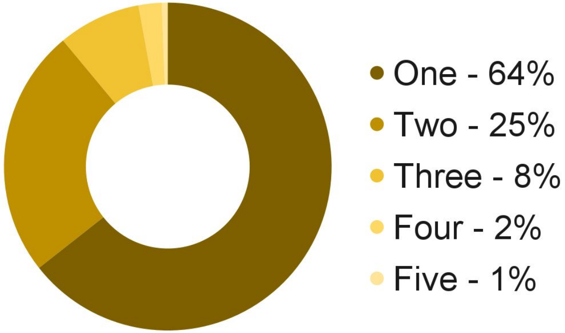
ImageNet-C



Dataset Analysis - Characteristics of Datasets

The dataset allows for multiple labels per image, aiming to address these errors:

- Cluttered images from ImageNet using only a single label.
- Categories with synonymous meanings within ImageNet, leading to model inaccuracies.

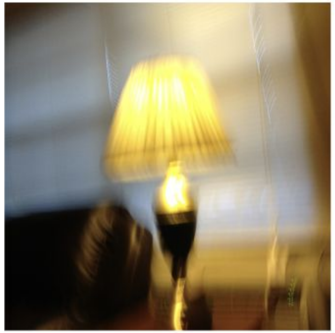


Number of classes assigned per image.
As shown, images in our dataset often have more than one label.

Dataset Analysis - Characteristics of Datasets

The dataset allows for multiple labels per image, aiming to address these errors:

- **Cluttered images from ImageNet using only a single label.**
- Categories with synonymous meanings within ImageNet, leading to model inaccuracies.



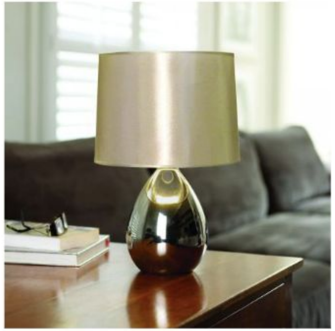
Labels

table lamp

studio couch

lampshade

VizWiz-Classification



Label

table lamp

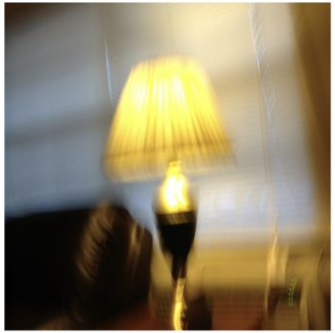
ImageNet

In this example image from ImageNet, there is also a studio couch. If the model detects a studio couch, it would be considered an incorrect prediction.

Dataset Analysis - Characteristics of Datasets

The dataset allows for multiple labels per image, aiming to address these errors:

- Cluttered images from ImageNet using only a single label.
- **Categories with synonymous meanings within ImageNet, leading to model inaccuracies.**



Labels

table lamp

studio couch

lampshade

VizWiz-Classification

Label

table lamp

ImageNet

Categories like "lampshade" and "table lamp" have similar meanings and are frequently found together in images. However, models are typically required to select only one label among them, which can lead to potential inaccuracies.

Algorithm Benchmarking

Algorithm Benchmarking

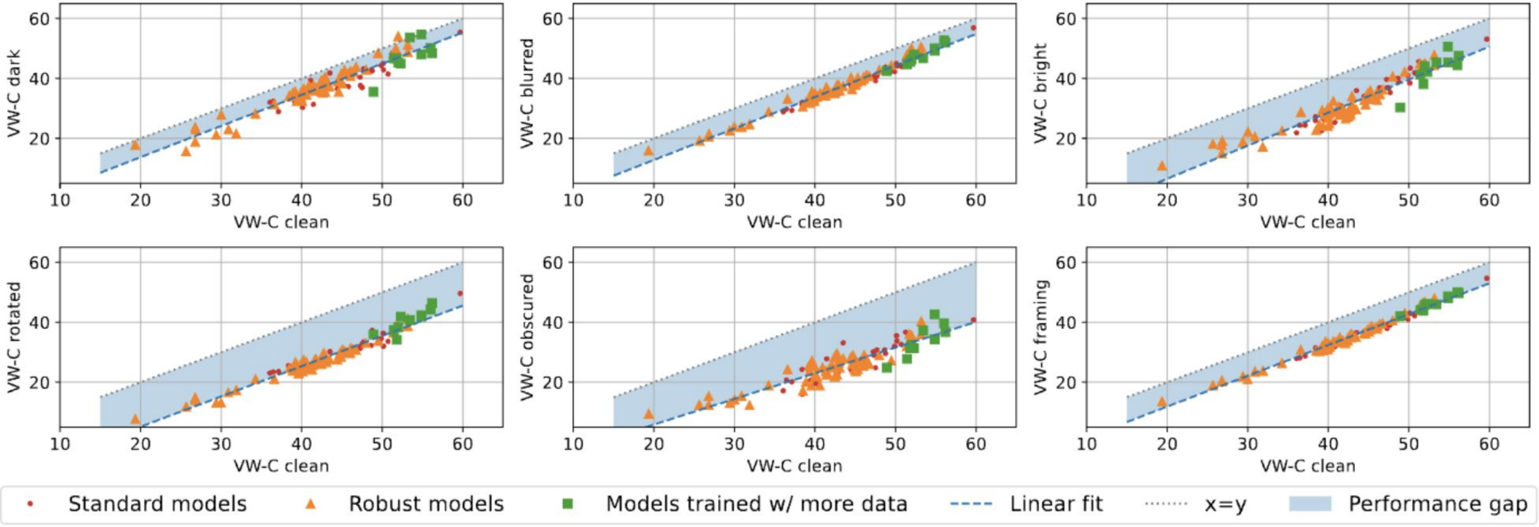
We select 100 models for evaluation on our test dataset. Models are divided into three subclasses:

1. **Standard models** (30 models)
   Trained on ImageNet and do not benefit from any methods for increasing robustness.

2. **Models trained with more data** (10 models)
   Trained on a larger set of training datasets such as ImageNet-21k or IG-1B-Targeted.

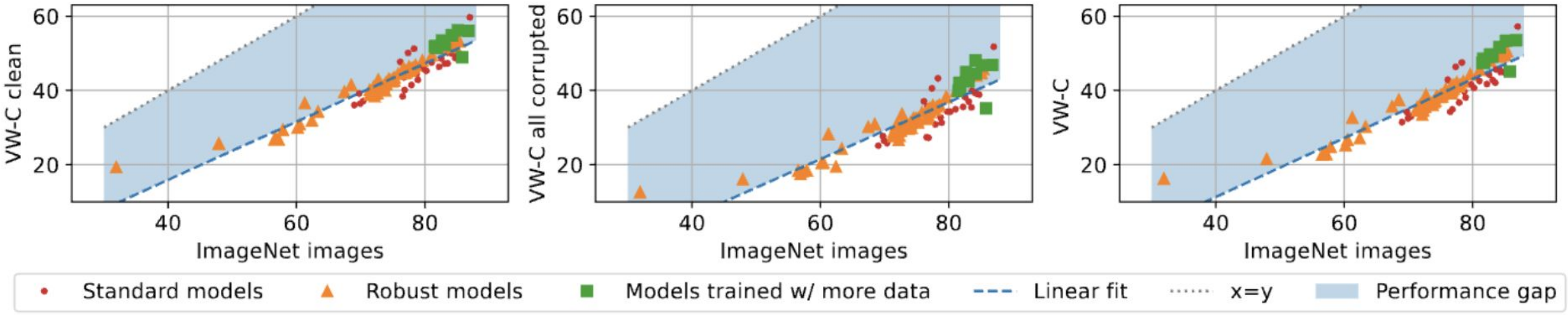
3. **Robust models** (60 models)
   Leverage robustness intervention methods such as data augmentation and adversarial attack methods.

Algorithm Benchmarking -
The Effect of Different Quality Issues



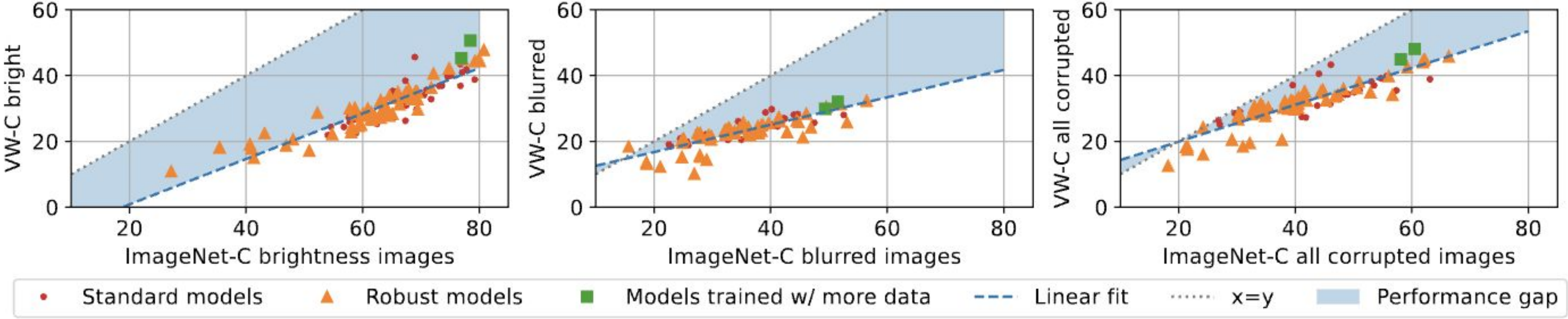
The order of performance gap based on the image quality issues from the **lowest** to **highest** is dark, blurred, framing, bright, rotated, and obscured.

Algorithm Benchmarking -
Measuring Robustness of Models



ImageNet → VizWiz-Classification: Models trained on more data can produce effective robustness. There is a large performance gap even for clean images.

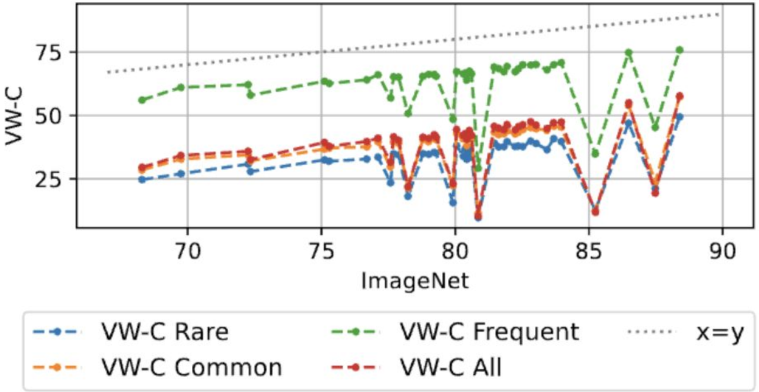
Algorithm Benchmarking -
Can ImageNet-C Track Real-World Corruptions?



ImageNet-C → VizWiz-Classification: Models, which leverage robustness interventions, are over-optimized for synthetic corruptions and cannot simulate real-world quality issues.

Algorithm Benchmarking -
The Effect of The Distribution of Categories

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



Unbalanced 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.

Algorithm Benchmarking - Summary

Distribution Shift	Performance Gap	Effective Robustness		
		Standard	Robust	More Data
VW-C clean → VW-C corrupted	10.3	-0.3	0.04	**0.1**
ImageNet → VW-C	35.8	-1.3	0.1	**3.7**
ImageNet-C → VW-C corrupted	9.5	0.1	-0.2	**4.6**

Performance gap exists in all distribution shifts. The most robust models are models pre-trained on larger datasets.

Conclusions

Conclusions

- Our dataset is the first dataset from **an authentic use case**, specifically from people with visual impairments.

- **Quality issues**, which our images can have, largely affect the performance of models.

- The **performance gap** is major between our new dataset and ImageNet dataset.

- Progress on **ImageNet-C** doesn't ensure similar progress on our dataset.

- The most **robust** models are models **pre-trained** on larger datasets.

- Please join our dataset challenge: https://vizwiz.org/tasks-and-datasets/image-classification

Thank you!