IMAGENET-X: UNDERSTANDING MODEL MISTAKES WITH FACTOR OF VARIATION ANNOTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning vision systems are widely deployed across applications where reliability is critical. However, even today's best models can fail to recognize an object when its pose, lighting, or background varies. While existing benchmarks surface examples that are challenging for models, they do not explain why such mistakes arise. To address this need, we introduce ImageNet-X–a set of sixteen *human* annotations of factors such as pose, background, or lighting for the entire ImageNet-1k validation set as well as a random subset of 12k training images. Equipped with ImageNet-X, we investigate 2,200 current recognition models and study the types of mistakes as a function of model's (1) architecture – e.g. transformer vs. convolutional –, (2) learning paradigm – e.g. supervised vs. self-supervised –, and (3) training procedures – e.g. data augmentation. Regardless of these choices, we find models have consistent failure modes across ImageNet-X categories. We also find that while data augmentation can improve robustness to certain factors, they induce spill-over effects to other factors. For example, color-jitter augmentation improves robustness to color and brightness, but surprisingly hurts robustness to pose. Together, these insights suggests that to advance the robustness of modern vision models, future research should focus on collecting additional diverse data and understanding data augmentation schemes. Along with these insights, we release a toolkit based on ImageNet-X to spur further study into the mistakes the image recognition systems make: https://github.com/ANONYMISED.

1 INTRODUCTION

Despite deep learning surpassing human performance on ImageNet (Russakovsky et al., 2015; He et al., 2015), even today's best vision systems can fail in spectacular ways. Models are brittle to variation in object pose (Alcorn et al., 2019), background (Beery et al., 2018), texture (Geirhos et al., 2018), and lighting (Michaelis et al., 2019).

Model failures are of increasing importance as deep learning is deployed in critical systems spanning fields across medical imaging (Lundervold and Lundervold, 2019), autonomous driving (Grigorescu et al., 2020), and satellite imagery (Zhu et al., 2017). One example from the medical domain raises reasonable worry, as "recent deep learning systems to detect COVID-19 rely on confounding factors rather than medical pathology, creating an alarming situation in which the systems appear accurate, but fail when tested in new hospitals" (DeGrave et al., 2021). Just as worrisome is evidence that model failures are pronounced for socially disadvantaged groups (Chasalow and Levy, 2021; Buolamwini and Gebru, 2018; DeVries et al., 2019; Idrissi et al., 2021).

Existing benchmarks such as ImageNet-A,-O, and -V2 surface more challenging classification examples, but do not reveal why models make such mistakes. Benchmarks don't indicate whether a model's failure is due to an unusual pose or an unseen color or dark lighting conditions. Researchers, instead, often measure robustness with respect to these examples' average accuracy. Average accuracy captures a model's mistakes, but does not reveal directions to reduce those mistakes. A hurdle to research progress is understanding not just *that*, but also *why* model failures occur.

To meet this need, we introduce ImageNet-X, a set of *human* annotations pinpointing failure types for the popular ImageNet dataset. ImageNet-X labels distinguishing object factors such as pose, size, color, lighting, occlusions, co-occurences, and so on for each image in the validation set and

a random subset of 12,000 training samples. Along with explaining how images in ImageNet vary, these annotations surface factors associated with models' mistakes (depicted in Figure 1).

a) `ImageNet-X` annotation form

Given the group of `Prototypical` images to the left and `Sample` image to the right, please answer the following questions in blue:

1. Select all factors that make the `Prototypical` images different from the `Sample` image: ☑ pose / positioning; ☑ object is partially present; ☑ object partially blocked by another object; ☐ object partially blocked by a person; ☐ another object is present; ☑ object is small relative to the image frame; ☐ object is large relative to the image frame; ☐ lightning is brighter; ☐ lightning is darker; ☑ background; ☑ color; ☐ shape; ☐ texture; ☐ pattern; ☐ media style; ☐ subcategory.
2. Describe more about your selections in question 2: cow in water at the beach
3. Describe in one word what the primary difference is between the left images and right image: beach, far-away

b) Robustness analysis enabled by `ImageNet-X`

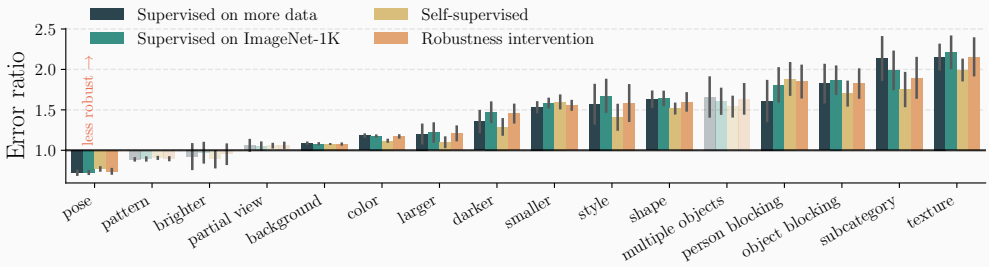

Figure 1: **Models, regardless of architecture, training dataset size, and even robustness interventions all share similar failure types.** `ImageNet-X` annotations allow us to group images into *Factors of Variation* such as pose, pattern or texture (subfigure **a** and full definitions in Appendix A.2). A model can be evaluated on each of these factors, revealing where it makes the most mistakes. We compare error ratios $= \frac{1-\text{acc}(factor)}{1-\text{acc}(overall)}$ on each factor for 4 wide groups of models. Subfigure **b** shows that differences in texture, subcategories (e.g., breeds), and occlusion are most associated with models' mistakes. Transparent bars show the factors where there is no significant difference between the 4 groups (p value > 0.05 with Alexander Govern test).

By analyzing the `ImageNet-X` labels, in section 2, we find that in ImageNet pose and background commonly vary, that classes can have distinct factors (such as dogs more often varying in pose compared to other classes), and that ImageNet's training and validation sets share similar distributions of factors. We then analyze, in section 3, the failure types of more than 2,200 models. We find that models, regardless of architecture, training dataset size, and even robustness interventions all share similar failure types in section 3.1. Additionally, differences in texture, subcategories (e.g., breeds), and occlusion are most associated with models' mistakes (See Figure 1 and section 3.3.1). Among modeling choices such as architecture, supervision, data augmentations, and regularization methods, we find data augmentations can boost models' robustness. Common augmentations such as cropping

and color-jittering however, can have unintended consequences by affecting unrelated factors (see section 3.3.2). For example, cropping improves robustness to pose and partial views, as expected, all the while affecting unrelated factors such as pattern, background, and texture. Together these findings suggests that to advance the robustness of modern vision models, future research should focus on improving training data – by collecting additional data and improving data augmentation schemes – and deemphasize the importance of other aspects such as choice of architecture and learning paradigm.

We release all the ImageNet-X annotations along with an open-source toolkit to probe existing or new models' failure types. The data and code are available at

https://github.com/ANONYMISED.

With ImageNet-X we equip the research community with a tool to pin-point a models' failure types. We hope this spurs new research directions to improve the reliability of deep learning vision systems.

2 IMAGENET-X: ANNOTATING IMAGENET WITH VARIATION LABELS

ImageNet-X contains human annotations for each of the 50,000 images in the validation set of the ImageNet dataset and 12,000 random sample from the training set. Since it's difficult to annotate factors of variations by looking at a single image in isolation, we obtain the annotation by comparing a validation set image to the three class-prototypical images and ask the annotators to describe the image by contrasting it with the prototypical images. We define the prototypical images as the most likely images under ResNet-50 model[1] (He et al., 2015). Trained annotators select among sixteen distinguishing factors, possibly multiple, and write a text description as well as one-word summaries of key differences. The form is illustrated in Figure 1. The factors span pose, various forms of occlusion, styles, and include a subcategory factor capturing whether the image is of a distinct type or breed from the same class (full definitions in Appendix A.2). The text descriptions account for factors outside the sixteen we provide. After training the annotators and verifying quality with multi-review on a subset, each image is annotated by one trained human annotator. For example, the annotator marks whether the object depicted in the Sample image is larger or more occluded than Prototypical images. We provide a datasheet following Gebru et al. (2021) in Appendix A.1.

One word summaries confirm the list of factors considered. Since our pre-selected list of 16 factors may not encompass every type of variation needed to account for model bias, we also asked annotators to provide one-word summaries to best distinguish a given image from its prototypes. We assess whether these free-form responses are encompassed within the pre-defined categories. We find the top-20 one-word annotation summaries are: pattern, close-up, top-view, front-view, grass, black, angle, white, color, background, brown, blue, red, position, facing-left, trees, person, side-view, low-angle, all falling within the 16 categories defined. For example top-view, front-view, facing-left, low-angle, side-view are captured by the *pose* factor.

2.1 WHAT IMAGENET-X REVEALS ABOUT IMAGENET

To better understand the proposed annotations, we explore the distribution of the different factors among ImageNet images. We identify the most common varying factors in ImageNet, confirm factor training and validation set distributions match, and find factors can vary in ways that are class-specific.

Pose and background are commonly selected factors. As we can see in Figure 2, when aggregating all the annotations, pose, background, pattern, and color factors are commonly selected. For instance, pose and background are active for around 80% of the validation images. Since images are not likely to share a background and the objects within are unlikely to share the same pose, annotators selected these factors for many images. Pattern and color, which are the next most common, are also unlikely to be identical across images.

[1]The prototypical images for a class are those correctly classifed with the highest predicted probability by ResNet-50. For 115 models with accuracies better than ResNet-50's, the average accuracy on the prototypical examples is 99.4%. Therefore all good models seem to agree on this set of prototypes.

Figure 2: **Some factors are selected for most images; choosing the top factor allows to focus on the main change in the image.** Figure shows the distribution of each factor in the training and validation set both with all factors selected and only top factor selected.

Selecting the top factor per image. While the annotators could select multiple factors, we select a unique, *top factor* per image. To do so, we use the free-from text justification that the annotator provided. We embed the text justification using a pretrained language model provided by Spacy (Honnibal and Montani, 2017), and we compare it to that of selected factors' embeddings. This selection allows us to extract the *main* change in each image, thus avoiding over-representing factors that are triggered by small changes, such as pose or background. We see a clear reduction of those factors in Figure 2 (all vs top factors).

Training and validation sets have similar distributions of factors To see if there is a distribution shift between training and validation, we annotate a subset of 12k training samples in a similar fashion as the validation[2]. Figure 2 reports the counts of each factor in this subset (denoted *train all* and *train top*). Comparing with the validation dataset, we see most represented factors are very similar to the validation set. We confirm this statistically by performing a χ^2-test on the factor distribution counts, confirming we cannot reject the null hypothesis that the two distributions differ (with $p < 0.01$). Similarly, in Appendix A.3 we see that the distribution of factors ticked by the human annotator (referred to as *active*) per image are close.

To ease our analysis, we consider for each image its meta-label. A meta-label regroups many ImageNet classes into bigger classes. We consider 17 meta-labels that we extract using the hierarchical nature of ImageNet classes: we select 16 common ancestors in the wordnet tree of the original ImageNet 1000 classes. These chosen meta-labels are : device, dog, commodity, bird, structure, covering, wheeled vehicle, food, equipment, insect, vehicle, furniture, primate, vessel, snake, natural object, and other for classes that don't belong to any of the first 16.

Classes can have distinct variation factors. In Appendix A.4 (right) we also observed statistically significant correlations between factors and meta-labels. For instance, the dog meta-label is negatively correlated with pattern, color, smaller, shape, and subcategory while being positively correlated with pose and background. This suggests that the images of dogs in the validation set have more variation for pose and background while having less variation for pattern, color, smaller, shape, and subcategory. The commodity meta-label, which contains clothing and home appliances, is positively correlated with pattern and color and negatively correlated with pose.

In addition to revealing rich information about ImageNet, `ImageNet-X` can be used to probe at varying levels of granularity a model's robustness (see Figure 3).

3 Probing Model Robustness

By measuring when factors appear among a model's mistakes relative to the overall dataset, we can characterize any ImageNet model's robustness across the 16 `ImageNet-X` factors. Robustness via `ImageNet-X` goes beyond revealing model mistakes to pinpointing the underlying factors associated with each mistake.

[2]In the training annotation scheme, we separated "pose" into two factors: "pose" and "subject location", where the latter covers cases where the subject in the same pose but a different part of the image. We reconcile the two here into a unique "pose" to compare with the validation annotations.

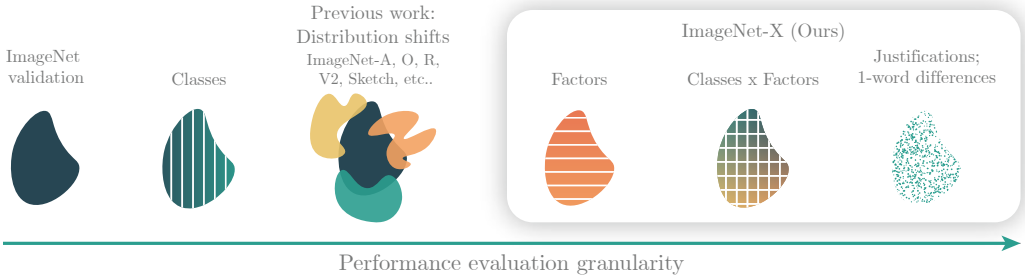

Figure 3: Performance can be evaluated at different levels of granularity. `ImageNet-X` provides factors of variation – aspects that makes an image different from typical images of its class. `ImageNet-X` allows to more precisely pinpoint models' weaknesses and compare them.

Here we systematically study the robustness of 2,200 trained models – including many models from the ImageNet test bed Taori et al. (2020a) and additional self-supervised architectures such as DINO and SimCLR – to reveal the limitations of average accuracy, understand model biases, and characterize which choices in architecture, learning paradigm, data augmentation, and regularization affect robustness. For our analysis we focus on the most salient factor for each image by ranking the selected factors by their similarity to the text-justification.

3.1 MANY DEEP LEARNING MODELS HAVE THE SAME WEAKNESSES AND STRENGTHS

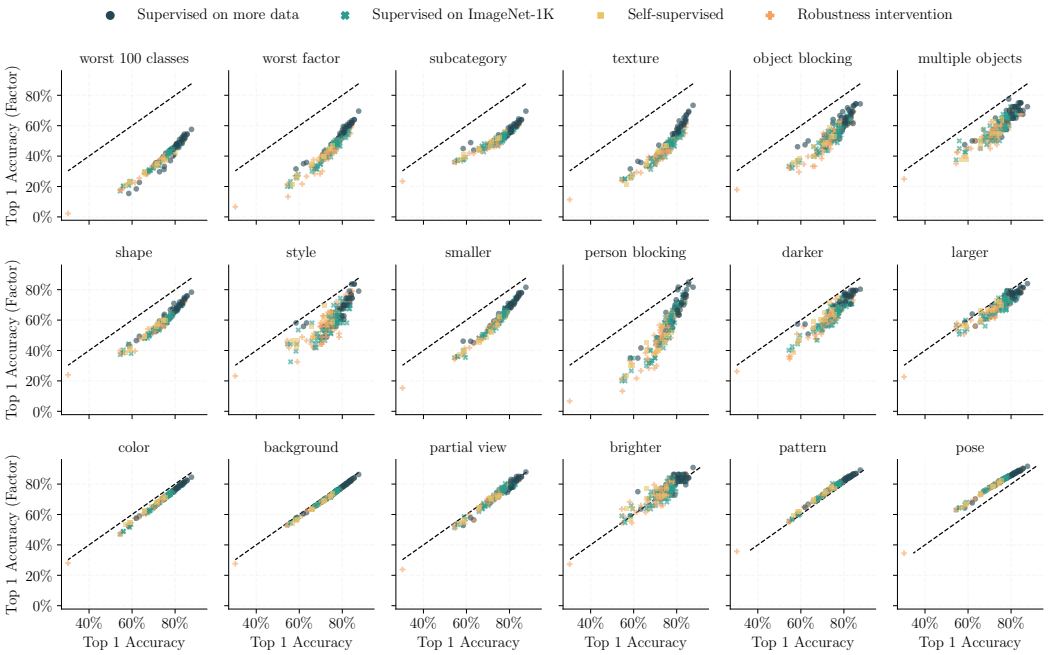

Figure 4: **Most deep learning models, when trained, finetuned or evaluated on ImageNet, have the same biases**. We plot top-1 accuracy for the subset of images labeled with the given factor (y-axis) relative to overall top-1 accuracy (x-axis). The dashed line is an ideal robust models' performance, i.e. performance on each factor is the same as the overall performance. We show the performance of 209 models. We also show the accuracy for the worst factor, and for the images of the worst 100 classes.

To what extent does the commonly reported average accuracy capture a model's robustness? To answer this, we inspect 209 models from the ImageNet testbed Taori et al. (2020a) which includes many architectures (most of which are convolutional, with a few vision transformers), training

procedures (losses, optimizers, hyperparameters), and pretraining data. We include additional self supervised models for completeness.

Models with similar overall accuracies have very similar per factor accuracies. With all this variety, the scatter plots in Figure 4 exhibit surprising consistency; overall accuracy is a good predictor of per factor accuracies. Most models, even with improving overall accuracy, seem to struggle with the same set of factors: subcategory, texture, object blocking, multiple objects and shape. Conversely, they seem to do well on the same set of factors: pose, pattern, brighter, partial view, background, color, larger, and darker. There are some factors where state-of-the-art models seem to have closed the robustness gap such as person blocking or style.

More training data helps, but robustness interventions do not. Models trained with larger datasets (blue circles in Figure 4) exhibit higher accuracy across the factors suggesting larger training datasets do help as others have shown Taori et al. (2020b). Surprisingly, models trained with robustness interventions (such as CutMix, AdvProp, AutoAugment, etc...), which are directly aimed at improving robustness don't show a significant improvement in per factor accuracy as prior work also shows Taori et al. (2020b).

Model weaknesses coincide with labeling errors. The ImageNet labels are known to contain labeling errors Beyer et al. (2020b). With new labels from previous work, we assess the accuracy of the original ImageNet labels and study the distribution of errors across factors. Interestingly, we find that the original label accuracies on all factors coincide with the state of the art models. This offers a potential explanation for the model biases. We leave it to future work to investigate if training on more accurate labels leads to more robust models. We provide details in appendix A.8.

Is accuracy enough? Since our annotations are on the ImageNet validation set, a 100% overall accuracy necessarily means a 100% per factor accuracy. So as models get better overall, they necessarily get better per factor accuracies. To disentangle performance from robustness we need to investigate the distribution of errors across factors for a given model.

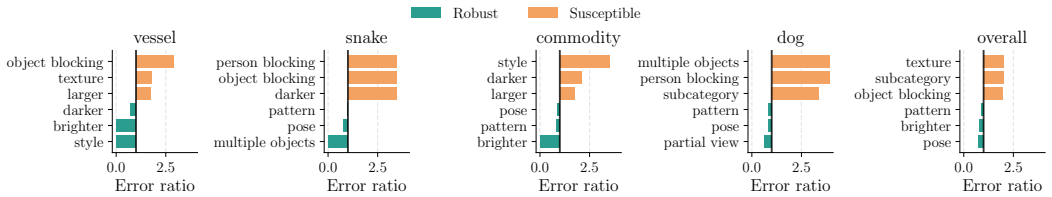

Figure 5: **An illustration of how `ImageNet-X` can identify robustness weaknesses and strengths for Vision Transformers (ViT).** Here we visualize the 3 most susceptible, and 3 most robust factors for the three worst-performing metalabels along with dog. We also include the overall validation set for comparison. We see ViT is susceptible to texture, occlusion, and subcategory, but is robust to pattern, brightness, and pose. While those overall types also show up across metalabels, we see robustness can be distinct by metalabel.

3.2 Moving beyond average accuracy: assessing failure types with ImageNet-X

With `ImageNet-X` we can go beyond average accuracy, to identify the types of mistakes a model makes. To do so we measure the error ratio across each of the 16 `ImageNet-X` factors: Specifically,

$$\text{Error ratio}(\text{factor}, \text{model}) = \frac{1 - \text{accuracy}(\text{factor}, \text{model})}{1 - \text{accuracy}(\text{model})} = \frac{\hat{P}(\text{factor}|\text{correct}(\text{model}) = 0)}{\hat{P}(\text{factor})}.$$

This quantifies how many more mistakes a model makes on a given factor relative to its overall performance. It also measures the increase or decrease in likelihood of a factor when selecting a model's errors vs the overall validation set. A perfectly robust model would have the same error rate across all factors, yielding error ratios of 1 across all factors.

3.2.1 AN EXAMPLE OF VISION TRANSFORMER'S MISTAKES

We illustrate how the error ratio can be used to identify the types of failures and strengths for the popular Vision Transformer (ViT) model. Despite impressive 84% average top-1 accuracy, we find ViT's mistakes are associated with texture, occlusion, and subcategory (appearing 2.02-2.11x more times among misclassified samples than overall) as shown in Figure 5. On the other hand we find ViT are robust to pose, brightness, pattern, and partial views. We also see that these strenghts or weaknesses can vary by metalabel. For example, ViT is susceptible to occlusion for vessel and snake, but not the commodity metalables where mistakes are associated with style, darkness, and texture. For the dog metaclass, ViT is quite robust to different poses. Instead, ViT's mistakes for dogs are associated with the presence of multiple objects and differences among dog breeds (subcategory). The full list of failure types across meta-labels is in Appendix A.9.

3.3 WHICH LEVERS CAN AFFECT MODEL ROBUSTNESS?

In practice model developers have many choices from architecture, learning paradigm, and training regularization to data augmentations. What impact do each of these choices have on model robustness? We systematically examine how each choice affects robustness.

3.3.1 ROLE OF SUPERVISION

We first group models into supervised (1k and with more data), self-supervised, and trained with robustness interventions in Figure 1. We measure the error ratio for each factor across in `ImageNet-X`. We find all model types have comparable error ratios, meaning models make similar types of mistakes. There are a few minor differences. For instance, self supervised models seem to be slightly more robust to the factors: color, larger, darker, style, object blocking, subcategory and texture. Supervised models trained on more data are more robust to the person blocking factor. We isolate whether some of the effects may be due to difference in data augmentation next.

3.3.2 DATA AUGMENTATION

Data-augmentation (DA) is an ubiquitous technique providing significant average test performance gains (Shorten and Khoshgoftaar, 2019). In short, DA leverages a priori knowledge to artificially augment a training set size. The choice of DA policies e.g. image rotations, color jittering, translations is rarely questioned as it is given by our a priori understanding of which transformations will produce nontrivial new training samples whilst preserving their original semantic content. Although intuitive at first glance, DA as recently seen much attention as it often unfairly treats different classes (Balestriero et al., 2022). Equipped with `ImageNet-X`, we now propose a more quantitative and principled understanding of the impact of DA.

To that extend, we propose the following controlled experiment. We employ a ResNet50 —which is one of the most popular architecture employed by practitioners— and perform multiple independent training runs with varying DA policy. Each run across all policies share the exact same optimizer (SGD), weight-decay (1e-5), mini-batch size (512), number of epochs (80), and data ordering through training. For each DA setting, multiple runs are performed to allow for statistical testing. Only the strength of the DAs and the random seed vary within those runs. In all scenarios, left-right flip is used both during training and testing.

Data augmentations can improve robustness, but with spill-over effects to unrelated factors. We report in Fig. 6 the statistically significant effects on error ratio due to three data augmentations: random crop, color jittering and Gaussian blur. For each setting, we measure prevalence shifts i.e. how much more or less likely a factor is to appear among a models' misclassifications.

For random crop, we vary the random crop area lower-bound i.e. how small of a region can the augmented sample be resized from the original image (varying from 0.08 which is the default value, to 1 which is synonym of no DA). We find the prevalence shift decreases for pose and partial view as expected. We also find that pattern, which is unrelated to random cropping, improves as well. However, we also observe a decrease in robustness for subcategory, an unrelated factor.

For color jittering, we vary both the probability to apply some standard color transformations and their strength. In particular, for any given value $t \in (0, 1)$ we employ the composition of ColorJitter

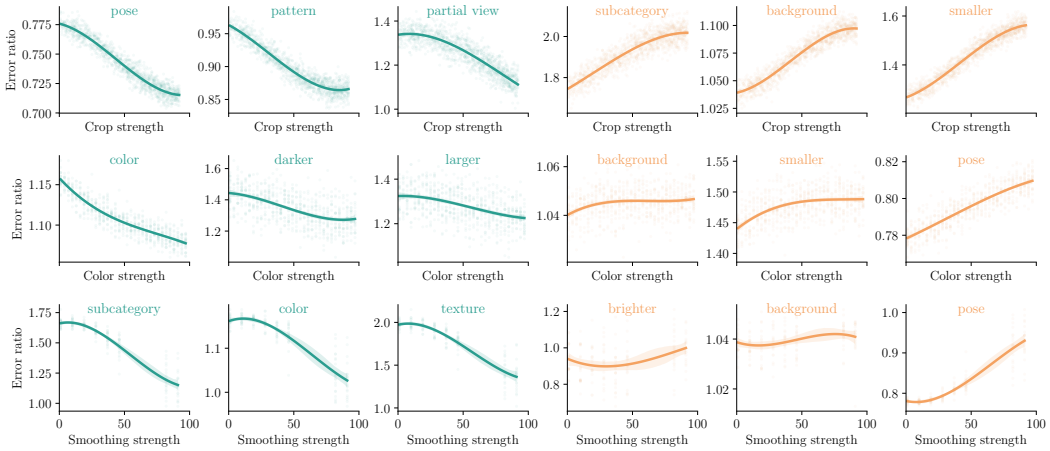

Figure 6: Evaluation of the impact of DA towards robustness of the trained model. We experiment with random crop (top row), color jittering (middle row) and Gaussian blur (bottom row); in all cases, we vary the strength of the respective DA (x-axis) and train multiple independent models for each setting. We observe that decreasing the random crop lower-bound (i.e. increasing the DA strength going from left to right) leads to more robust performance when pose varies, or partial views of the objects are present. Interestingly this also reduce the model robustness to objects that appear smaller and to background variations i.e. **the random crop augmented models become robust towards background and texture.** Color jittering improves robustness to darker objects and to objects with varying color. Surprisingly, color jittering also increases robustness to larger objects. Gaussian blur also presents a dual effect i.e. benefits towards subcategory and color changes and increases sensitivity to change in pose. In short, it seems that regardless of the DA, the benefits are always accompanied by detrimental impacts on unexpected factors hinting at a possible limitation of DA to produce improved robustness throughout all factors.

with probability $0.8t$ with brightness strength $0.4t$, contrast strength $0.4t$, saturation strength $0.1t$, and hue strength $0.1t$, followed by random grayscale with probability $0.5t$, and finally followed by random solarization with probability $0.3t$. Those parameters are found so that when t nears 1, the DA is similar to most aggressive DA pipelines used in recent methods. We naturally observe that the predictions become more robust to change in color and brightness. Interestingly, the model becomes more sensitive to pose variations.

For Gaussian blur, we vary the standard deviation of the Gaussian filter between 0.1 and $0.1 + 3.5t$ for a filter size of 13×13. Note that when t approaches 1 the strength of this DA goes beyond the standard setting used, for example, in SSL. In supervised training, GaussianBlur DA is rarely employed on ImageNet. We observe that the model becomes more robust to texture which might come from blurring removing most of the texture information and forcing the model to rely on other features. Surprisingly, subcategory is also much less of an impact on performances while change in pose becomes more problematic for the Gaussian blur trained model.

4 Related work

The research community has developed approaches for testing models' robustness on extended versions of the ImageNet dataset – see Figure 3 for a visual on different levels of evaluation granularity and Table 1 for an overview of datasets created to test the failure modes of the models trained on the ImageNet data. One common approach is to introduce artificial augmentations , e.g. image corruptions and perturbations (Hendrycks and Dietterich, 2019), renditions (Hendrycks et al., 2021), sketches (Wang et al., 2019; Bordes et al., 2021), etc. These artificial variations capture changes arising from corruptions, but are unlikely to capture changing arising from natural distribution shifts or variation such as changes in pose, lighting, scale, background etc. Consequently, researchers also collected additional natural images to study the performance of the classification models under moderate to drastic distribution shifts Hendrycks and Dietterich (2019); Recht et al. (2019). However,

Dataset	Description	ImageNet images	Natural images	Entire val. set	Human FoV	Ref.
ImageNet-C	algorithmic corruptions	✗	✗	✓	✗	Hendrycks and Dietterich (2019)
ImageNet-P	animated perturbations	✗	✗	✓	✗	Hendrycks and Dietterich (2019)
ImageNet-R	artistic renditions	✗	✗	✗	✗	Hendrycks et al. (2021)
ImageNet-A	natural adversarial examples	✗	✓	✗	✗	Hendrycks et al. (2021)
ImageNet-O	out-of-distribution examples	✗	✓	✗	✗	Hendrycks et al. (2021)
ImageNet-V2	new validation set	✗	✓	✗	✗	Recht et al. (2019)
ImageNet-Sketch	drawn sketches	✗	✗	✗	✗	Wang et al. (2019)
ImageNet-ML	human multi-labels	✗	✓	✗	✗	Shankar et al. (2020)
ImageNet-ReaL	human multi-labels	✓	✓	✓	✗	Beyer et al. (2020a)
ImageNet-ReLabel	machine pixelwise multilabels	✓	✓	✓	✗	Yun et al. (2021)
ImageNet-Stylized	randomly-textured images	✓	✗	✓	✗	Geirhos et al. (2018)
`ImageNet-X`	human FoV annotations	✓	✓	✓	✓	ours

Table 1: Extensions of the ImageNet benchmark extensions designed to inspect failure modes of the ImageNet trained models. We characterize each dataset by looking whether : (1) the dataset images are only coming from the ImageNet validation set — **ImageNet images** —; (2) the dataset images are natural images or are created with algorithmic and artistic perturbations of natural images — **Natural images** —; (3) the dataset annotates the entire ImageNet validation set — **Entire val. set** —; and (4) whether the dataset contains human annotations of image factors of variations — **Human FoV** —. Our proposed `ImageNet-X` is the first dataset based on ImageNet to include human annotations of multiple factors of variation for the entire ImageNet validation set.

most of these datasets are built to assess the model performance when going away from training data distribution and, thus, provide almost no understanding about the nature of the *in-distribution* errors. Currently, the only ImageNet extensions that help analyzing the in-distribution model errors are the multiclass relabelling or saliency of the validation set (Shankar et al., 2020; Beyer et al., 2020a; Yun et al., 2021; Singla and Feizi, 2021). However, this relabelling only explains one type of model error that is caused by the co-occurrences of other objects in the scene. Our contribution, `ImageNet-X`, builds on this line of work to provide granular labels for naturally occurring factors such as changes in pose, background, lighting, scale, etc. to pinpoint the underlying modes of failure.

5 Conclusion

We introduced `ImageNet-X`, an annotation of the validation set and 12,000 training samples of the ImageNet dataset across 16 factors including color, shape, pattern, texture, size, lightning, and occlusion. We showed how `ImageNet-X` labels can reveal how images in popular ImageNet dataset differ. We found that images commonly vary in pose and background, that classes can have distinct factors (such as dogs more often varying in pose compared to other classes), and that ImageNet's training and validation sets share similar distributions of factors. Next, we showed how models mistakes are surprisingly consistent across architectures, learning paradigms, training data size, and common robustness interventions. We identified data augmentation as a promising lever to improve models' robustness to related factors, however, it can also affect unrelated factors. These findings suggest a need for a deeper understanding of data augmentations on model robustness. We hope that `ImageNet-X` serves as an useful resource to build a deeper understanding of the failure modes of computer vision models, and as a tool to measure their robustness across different environments.

Reproducibility Statement The results and figures in the paper can be reproduced using the open-source code and the `ImageNet-X` annotations, which we also release. The annotations were collected by training annotators to contrast three prototypical images from the same class. This setup that can be replicated using the questionnaire we provide here as well as the freely available ImageNet dataset. For a detailed description of the `ImageNet-X` dataset, please see A.1.

REFERENCES

Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *CVPR*, 2019. URL https://arxiv.org/abs/1811.11553.

Randall Balestriero, Leon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. *arXiv preprint arXiv:2204.03632*, 2022.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. URL https://arxiv.org/abs/1807.04975.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv*, 2020a. URL https://arxiv.org/abs/2006.07159.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? (arXiv:2006.07159), Jun 2020b. doi: 10.48550/arXiv.2006.07159. URL http://arxiv.org/abs/2006.07159. arXiv:2006.07159 [cs].

Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021.

Diane Bouchacourt, Mark Ibrahim, and Ari Morcos. Grounding inductive biases in natural images: invariance stems from variations in data. *Advances in Neural Information Processing Systems*, 34: 19566–19579, 2021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Kyla Chasalow and Karen Levy. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021. URL https://arxiv.org/abs/2101.03827.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 2021. URL https://www.nature.com/articles/s42256-021-00338-7.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? download pdf. *arXiv*, 2019. URL https://arxiv.org/abs/1906.02659.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

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. *arXiv preprint arXiv:1811.12231*, 2018.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 2020. URL https://arxiv.org/abs/1910.07738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. URL https://arxiv.org/abs/1502.01852.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv*, 2019. URL https://arxiv.org/abs/1903.12261.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. URL https://arxiv.org/abs/2006.16241.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. *CLeaR*, 2021. URL https://arxiv.org/abs/2110.14503.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 2019. URL https://link.springer.com/article/10.1007/s12194-017-0406-5.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv*, 2019. URL https://arxiv.org/abs/1907.07484.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. URL https://arxiv.org/abs/1902.10811.

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. *IJCV*, 2015. URL https://arxiv.org/abs/1409.0575.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. In *ICML*, 2020. URL https://proceedings.mlr.press/v119/shankar20c.html.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2021.

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*, volume 33, page 18583–18599. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020b.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. URL https://arxiv.org/abs/1905.13549.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *CVPR*, 2021. URL https://arxiv.org/abs/2101.05022.

Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 2017. URL https://arxiv.org/abs/1710.03959.