We thank the reviewers for their feedback and reply to the major points raised by each reviewer individually.

Reviewer 1

Our paper focuses on ImageNet classification because this is what almost all prior work on robustness has studied. While we agree that considering robustness across a wider range of tasks would be an improvement, most prior publications have focused on ImageNet exclusively as well. ImageNet thus provides a standard benchmark, which is important since our paper is a meta-study of prior work. Further, since many of the papers that study only ImageNet were publishable results, we believe this indicates our ImageNet analysis is also sufficient.

We also agree that evaluating on uncurated datasets is an extremely interesting area for future work, especially for the real-world deployment of machine learning. However, even in the simpler setting when both training and test sets are well curated, we find that most techniques still do not improve robustness beyond standard accuracy.

We agree that transfer learning is an interesting setting for future work, but note transfer learning cannot always be the solution, e.g., when the shift can not be reliably quantified or it is hard to collect data from the shifted distribution. To reiterate our core argument: studying the robustness of models to small changes in distribution is the main focus of our work. We hope that future work (e.g., transfer learning research) can build on our testbed.

Reviewer 3

Our results are substantially more nuanced than "more data helps": (i) We show that *only* more data currently helps robustness on ImageNetV2 and ObjectNet. Many robustness interventions have been proposed over the last few years, but they do not help on these two distribution shifts. This is a strong negative result. (ii) Appendix B demonstrates that additional data *from the same distribution* does *not* help robustness. (iii) The effect of off-distribution data varies. For instance, one model trained on JFT-300 [57] does not have effective robustness on ImageNetV2 or ObjectNet, while another model trained on JFT-300 [74] does. Beyond the effect of data, we highlight that ℓ_p-robust models provide effective robustness on consistency shifts. Appendix D contains additional results for more granular trends.

Regarding domain adaptation / transfer learning, please refer to the third paragraph in our response to Reviewer 1.

We agree with the concern about the choice of baseline function and will clarify this point. The baseline function depends on the trend given by the models without robustness intervention; any regression method that fits standard models well can be used. In practice for the datasets we considered, we found the (piecewise) log-linear fit to be best.

Reviewer 4

We greatly appreciate R4's detailed feedback. We concur that the pace of work in this area is a key challenge: our work is the result of multiple iterations of peer review, each which led to additional models, datasets, and analyses in the manuscript. One side effect, as R4 points out, is that the manuscript is dense, despite our best attempts at prioritizing the key analyses for the main paper and leaving details to the supplementary material.

ImageNet-v2 as primary example. Another concurrent work, Shankar et. al., ICML '20, shows that humans can classify ImageNet-v2 images as well as ImageNet: that is, humans do not suffer from a drop in accuracy due to the distribution shift in ImageNet-v2, while current vision models do. Thus, even if Engstrom et. al.'s analysis explains the accuracy drop due to selection frequency bias, we would like models that match the robustness of human labelers.

The definition of "distribution shifts arising in real data": We included a diverse array of natural distribution shifts available at the time of submission. As R4 notes, ImageNet-R was released after submission and is a welcome response to our call for more evaluations on natural distribution shifts. We have added ImageNet-R to our testbed (see Fig. 1) and will include it in the camera-ready. We would also be happy to include any other natural shifts that we may have missed.

BiT-L models: As the BiT-L model is private, we have reached out to the authors to include it in our test bed, and are simultaneously working on incorporating the smaller, publicly released BiT-M model.

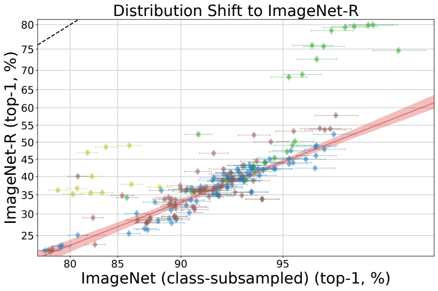

Figure 1: Model accuracies on ImageNet-R, a dataset with renditions such as art, sketches, and graphics of 200 ImageNet classes. Evaluations on this distribution shift follow the high-level trends of the other natural distribution dataset shifts in our testbed, with models trained on extra data providing the most robustness (though the effect is not uniform). ImageNet classes were subsampled to match the class distribution of ImageNet-R. Confidence intervals, linear fit, and axis scaling are computed similarly to Figure 2 in the main text. The color coding is identical to Figure 3 of the main text.