We thank the reviewers for their comments. We address individual concerns below.

Reviewer 1: *It would be good to verify the main findings (e.g. Table 1) for different model types.* We agree that this would be interesting, the main reason we did not include a full analysis of more architectures is computational constraints, since each architecture requires its own grid search over the various hyperparameters outlined in the paper. We believe that the results in G.2 are enough to demonstrate that the increased performance is a robust phenomenon, and we hope to inspire future works that test transfer learning performance across more architectures and datasets.

Width experiments from S4.2. We agree that running with even wider architectures would lead to a better understanding of the trend. Unfortunately, the number of distinct architectures we can train is again bounded by computational constraints. We would like to note that, particularly in the fixed-feature setting, the trend of "eps=0 network increases then plateaus/decreases" holds robustly across datasets, which we believe gives some indication of the generality of the phenomenon. Also, ResNet-50, WRN-50x2, and WRN-50x4 are all ResNet-50 models with varying width (only ResNet-18 differs in architecture).

Dataset granularity hypothesis S4.3. We thank the reviewer for the suggestion. Though we agree that resolution is a coarse proxy, we wanted to focus S4.3 on a quantitative notion of granularity. It would be very interesting future work to test the same relationship with respect to other more complex quantitative notions of dataset granularity.

Reviewer 2: *Lack of novelty.* We believe the reviewer is conflating "transfer learning" (wherein one uses a pre-trained classifier on one dataset to perform better on another dataset) with "adversarial transfer" (the phenomenon where adversarial attacks that fool one architecture tend to also fool another architecture). The two fields are entirely unrelated—our work is on the former, whereas [Mad+18] and others discuss the latter.

The improvement is marginal. While the improvement is sometimes small, note that (a) robust models consistently outperform standard models, which adds significance to the result, (b) on many datasets the improvement given by robust models is outside of error bars, and (c) that robust models have much worse accuracy than standard models, making even modest improvements somewhat surprising.

Lack of clarity #1 (technical details). We are not sure what the reviewer means by this comment. The equation given is fairly standard, and in Appendix F we give a detailed primer on adversarial robustness which introduces each symbol in the first equation explicitly, and also provides other background technical knowledge.

Lack of clarity #2 and lack of reproducibility (experimental details). We are again confused by the reviewer's comment here, since Appendix A in the supplementary materials provides all of the details necessary to reproduce the experiments. Furthermore, we provide a full code release (the link is in the paper) with all of our pre-trained ImageNet models and easy-to-run code for reproducing any of the numbers reported in our paper.

Related work. We hope that the reviewer's concern is alleviated by the clarification above (i.e., the difference between "transfer learning" and "adversarial transfer.") In both Section 5 and Appendix E we outline and discuss all of the related work of which we are aware.

Reviewer 5: Thank you for your comments!

Reviewer 6: *Clarity: Figure 5 a bit confusing.* Thanks for pointing this out. Figure 5 summarizes the results of our fixed-feature transfer learning experiment on various datasets and architectures (dataset names are given above each plot and architecture name below each plot group). Each data point corresponds to an ImageNet model pre-trained with a given robustness level denoted by one of the markers (the legend at the top relates the marker to the robustness level). The x coordinate is the clean accuracy of this model, and the y coordinate is corresponding transfer accuracy on the relevant dataset.

Note that due to a formatting error the y axis legend (which should read "Transfer Accuracy") got cut off, we have fixed this in the updated manuscript. We will also make sure to clarify the figure in the updated version of the paper.

Experiments: comparison with texture robust models. We train only on Stylized ImageNet.

Open questions: insights on how to select the robustness level for a new dataset. This is a great question. From our analysis in section 4.3, it seems that the robustness level is correlated with the scale of the datasets; as the scale of the dataset increases, the "best" corresponding robustness level decreases. One might be able to fit a function mapping dataset scale to robustness level using the results of our experiments on various datasets (the have various scales)

We believe more analysis and experiments are required before reaching conclusions on how to select the best robustness level.