Reviewers, thank you for your careful analyses of our paper.

We would like to clarify the value of our work for self-supervised learning, input corruption robustness, adversarial examples, label corruption robustness, and out-of-distribution detection. A preliminary version of this work has been well-received by the self-supervised community as one of four long oral presentations at a top self-supervised workshop. One reason for its positive reception is, in OOD detection, prior art on natural images for unsupervised techniques such as density estimators and one-class SVMs have performance near chance levels [1]. However, we show that five different self-supervised techniques straightforwardly improve over both unsupervised and one-class methods. We also show that a self-supervised multi-task combination can *even surpass fully supervised techniques* (see Table 4). Another reason our work is valuable to the self-supervised learning community is because we identify self-attention as a useful architectural change; this finding is valuable because self-supervised advancements greatly depend on researchers identifying appropriate architectural choices [2]. For these reasons and more, we believe our OOD detection results have clear value to both the self-supervised learning and OOD detection research communities.

Regarding robustness, training against more data generally does not improve corruption robustness—even training against different corrupted data does not improve robustness [3,4]. These previous works show that training against one type of corruption does not confer robustness to novel corruptions. However, we find self-supervised learning does improve robustness to various novel corruptions. Moreover, we independently experimented with pre-training (R2) on ImageNet and found it did not improve corruption robustness. Further, while self-supervision may be thought of as "a 'data augmentation' method," augmenting the dataset with rotations of multiples of 90 degrees actually *decreases* corruption robustness from 72.3% to 63.7%, but with a rotation prediction loss, it improves to 76.9%. It was not obvious from prior work that combining fully supervised and self-supervised objectives could improve corruption robustness. Hence, this result is surprising and of value.

We agree with R2 that self-supervised learning is a form of bias or regularization, but whether this inductive bias helps is unclear a priori. For adversarial examples, there is much work on training with orders of magnitude more data to increase adversarial robustness [5,6,7]. Rather than training on significantly more data, we show it is possible to extract more predictive information from the training data with self-supervised learning. More, in many domains one does not have external data to train on, such as the medical domain. In these domains, improvements to label corruption robustness and adversarial robustness from self-supervised learning are especially valuable.

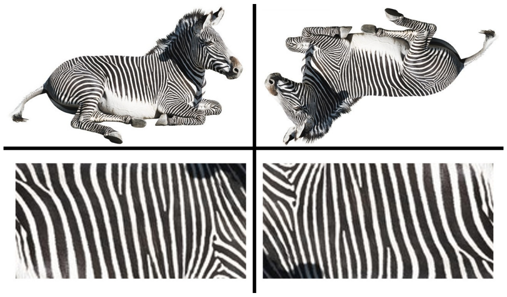

Figure 1: Predicting rotations requires shape, as texture alone is not sufficient for prediction.

R1 and R2 ask why predicting rotations improves robustness. Due to space constraints, we did not speculate on this in the paper, but we think part of the reason is that it requires modeling shape. For example, predicting the zebra's rotation in Figure 1 requires modeling contours and not just texture. This can lead to more robust representations. We will include discussion of this and further analysis in the updated draft.

Individual responses

R1: For OOD detection, our work focuses on the challenging one-class setting, meaning that we fix a single class as in-distribution and the rest as out-of-distribution. Thus, the MSP detector, MC-Dropout, and other techniques suited for multiclass do not apply since we learn with in-distribution data. The performance drop on clean data in Table 1 is a pervasive and a recognized shortcoming of adversarial training itself [8]. Finally, when rotating images, we use the rot90 function from NumPy. This avoids blurriness caused by resampling.

R2: We address many of the concerns in the general comments. On L117, you suggested not attacking the rotation branch, which is a good suggestion. We find that it interestingly performs similarly to attacking the rotation branch and will include this ablation in the updated draft. Thank you.

All our experiments were run with fixed random seeds and hyperparameters chosen as standard values or tuned on validation data. The computational cost of our experiments is high, but we agree that error bars are feasible and informative to add for the common corruption experiments. Due to your suggestion, we have now run these numerous experiments and will add error bars to the updated draft of the paper.

R3: We address many of the concerns in the general comments. In addition to our novel method from the OOD section, our main novelty is in our successful integration of self-supervised learning to four highly researched areas, and our demonstration that robustness and uncertainty can be new dimensions with which to judge self-supervised learning advancements. We will make this clearer in the paper. Thank you.

[1] Implicit Generation and Generalization with Energy Based Models. [2] Revisiting Self-Supervised Visual Representation Learning. [3] Comparing deep neural networks against humans: object recognition when the signal gets weaker. [4] Examining the Impact of Blur on Recognition by Convolutional Networks. [5] Are Labels Required for Improving Adversarial Robustness? [6] Adversarially Robust Generalization Just Requires More Unlabeled Data. [7] Unlabeled Data Improves Adversarial Robustness. [8] Adversarial Training Can Hurt Generalization.