We thank all reviewers for their time and helpful comments. We would like to clarify the following points.

References, more background and originality (R #1, #3) We are very sorry that the page with the references was accidentally excluded when uploading the files to CMT. We believe the references give very in-depth background on spatial robustness which we omitted in the page-limited manuscript in order to focus on our own contributions. However, if anything particular would help to understand our concepts better, we'd be glad to hear your thoughts.

As apparent from the citation numbering in the submitted pdf, we in fact cite 50 related papers. Below we list the most closely related works (numbering as in submission) and provide a distilled summary of the originality of our theoretical and empirical contributions compared to these:

[11] L. Engstrom et al. Exploring the landscape of spatial robustness. ICML, 2019.

[20] H. Kannan et al. Adversarial Logit Pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[39] D. Tsipras et al. Robustness may be at odds with accuracy. ICLR, 2019.

[45] H. Zhang et al. Theoretically principled trade-off between robustness and accuracy. ICML, 2019.

To the best of our knowledge we are the first who

1. provide **theoretical** justification why several previously used methods that add regularization on top of augmentation (such as adversarial training, [20] and [45]) can improve the robustness of the solution (our analysis even holds for perturbation sets derived from any kind of group transformation)
2. perform a well-controlled **empirical** comparison of spatial robustness gains between unregularized and regularized augmentation-based procedures, and methods based on architectural modifications to incorporate spatial invariances.
3. Furthermore, papers [39], [45] show for certain examples that predictors with high robust accuracy must have lower than optimal standard/natural accuracy. We provide precise conditions on the perturbation sets for which we can **prove** that there is no such "trade-off" (see below). Notably, it even increases under mild assumptions.

More in-depth analysis, discussion of effectiveness of regularization (R #3) In a revised version we have restructured Sec. 4 to present the main take-aways more transparently. In our opinion, the negligible computational overhead is important to advocate for a more wide-spread use of regularization in practice. Apart from this point, we have in fact done a rather extensive analysis of the effectiveness of regularization for achieving high robust accuracy from various perspectives—including but not limited to: comparison to non-regularized methods, comparison between different choices of `batch`, `def` among regularized methods, comparison to specialized networks.

Regarding regularization vs. vanilla baseline methods (fixing the defense method `def` and batch type `batch`):

1. Adversarial training (`batch`: rob, `def`: Wo-10) without (AT) vs. with regularization (KL, ℓ_2, ALP) [11]: regularization (with all three regularizers) leads to a relative robust error reduction of $\sim 23\%$
2. For data augmentation (`batch`: nat and rob, `def`: rnd) without (std* in Table 1) vs. with regularization (ℓ_2(·,rnd), KL(·,rnd) in Table 2): Relative robust error reductions of 35% (CIFAR-10) to 47.8% (SVHN)
3. The above robustness gains with regularization hold for a large range of λ-values
4. Regularization also improves robustness of VGG-Net (from 74% to 78% on CIFAR10 and 87% to 90.7% on SVHN)

Regarding regularized augmentation-methods vs. handcrafted equivariant networks and compared against one another:

1. Regularized methods outperform representative specialized spatial-equivariant networks
2. For SVHN, adding regularization to samples obtained both via Wo-10 adversarial search or random transformation (rnd) consistently not only helps robust but also standard accuracy
3. The KL regularizer performs better than ℓ_2 for most settings; S-PGD outperforms other defense methods

We would appreciate specific suggestions by the reviewers regarding further analyses.

More complex and larger datasets (R #2, #3) SVHN and CIFAR-10 have been the most common datasets that were used to evaluate handcrafted spatial-equivariant networks. Furthermore, as mentioned in Sec. 4.1., we have performed a subset of the experiments on CIFAR-100 (see Table 9 in the Supp. Mat.). While it doesn't have a higher resolution, it is a much more complex dataset and regularization still improves robust accuracy of unregularized baselines from 33.4% to 52.58% (relative err. red 28.8%). We originally did not run experiments on ImageNet since there are no well-tuned spatial-equivariant networks available for baseline comparison. However we expect regularization to help for spatial robustness on ImageNet as well and will run experiments to confirm that.

Title choice "Trade-off between natural and robust accuracy" and notation in Tab. 1 (R #2) We have improved the clarity of Sec. 2.3 in a revised version. We originally chose the title to help the reader draw the connection to previous papers that use the same expression (e.g. abstract of [39], title of [45]). Regarding Table 1, the reviewer's guess is correct: the rows correspond to the test setting and the columns to the train settings. The precise naming convention for the train settings (columns) are described in the first paragraph of Sec. 3. We will add a comment about the row naming in the caption: "nat" refers to the standard test examples and "grid" to worst-case transformations using grid search as described in Sec. 3.2.