We aimed for a dataset that is challenging and diverse, and has many images where it is not trivial to estimate contact states. Besides, our dataset is challenging and diverse, and has many images where it is not trivial to estimate contact states. For instance, the results from the last two columns show that the model trained on our data has better cross-dataset generalization (by 8% in mAP) when compared to the model trained on the previous dataset [28]. This also shows the benefit of our data.

**Q1:** Are the evaluation results valid since [28] uses FasterRCNN instead of MaskRCNN? The paper should include an evaluation of the proposed method trained on the previous dataset [28].

**A:** In the context of physical contact estimation, there are no conceptual nor empirical differences between MaskRCNN and FasterRCNN. Conceptually, MaskRCNN is FasterRCNN with an additional mask prediction branch. Empirically, they perform similarly as shown in Tab 2. All results are evaluated using bounding boxes. The data from [28] was not available at the time of submission. The experiments requested by the reviewer are now shown in Tab 1. The cross-dataset results from the last two columns show that the model trained on our data has better cross-dataset generalization (by 8% in mAP) when compared to the model trained on the previous dataset [28]. This also shows the benefit of our data.

**Q2:** 4% improvement on average over all the contact states is relatively low. We respect your opinion. But an average improvement of 4% is significant, and a higher experimental standard would not have been satisfied by many previously published NeurIPS papers.

**Q3:** In the ablation study, removing any single part of their network had little change on their results. We respect your concern, and we will clarify it in our revised paper. We adopted the term “weak annotation” from the previously published NeurIPS papers.

**Q4:** Hand joint locations seem more appropriate comparison than human body joints. Following the suggestion, we used OpenPose [5] for hand keypoints, but it failed to detect hands in many unconstrained images, as also reported in [22]. Empirically, we found that the detection AP is only 39.36%, compared to 83.72% of our method. This level of noisy detection results cannot be used for contact state estimation.

**Q5:** For the joint location baselines, why use Mean overlap of the hand with objects, instead of Maximum? We respect your opinion. But Maximum seems to be more intuitive, it does not perform better in practice, yielding an mAP of 33.73%. We originally used Mean because it was thought to be more robust than Max for noisy inputs (i.e., noisy detection results).

**Q6:** For Axis-Parallel, Extended performs better than Exact, but is the opposite case in Quadrilateral. When using Quadrilateral, to crop the polygon into a rectangular image, we first construct a rotated rectangle. Because of this, some surrounding context region is already present. Extending this even more can add a lot of irrelevant regions and leads to a reduction in performance.

**Q7:** For the AP metric, is the IoU computed from the quadrilateral box or the axis-aligned box? We used axis-aligned box, following the standard evaluation protocol for hand detection [1, 22, 28].

**Q8:** Is the dataset balanced? Other-Person-Contact seems to be more difficult? A: We aimed for a dataset of representative images of the real world, so the classes are not balanced. Compared to other classes, the number of hands with Other-Person-Contact labels is much smaller, as reported in lines 236–241.

**Q9:** For the second sentence of Fig 1, does it mean the feature of hand, the feature of the object and feature of union box? A: It means the feature of hand, and the feature of the union box. We will reword this, thanks.