Table 1: Comparing FasterRCNN, MaskRCNN and proposed method, and cross-dataset experiments.

Method	FasterRCNN [28]	MaskRCNN	Proposed	FasterRCNN [28]	MaskRCNN	Proposed	Proposed	Proposed
Train data	[28]	[28]	[28]	ours	ours	ours	[28]	ours
Test data	[28]	[28]	[28]	ours	ours	ours	ours	[28]
No-Contact	66.55 %	67.30 %	68.23 %	59.22%	60.52 %	62.48 %	44.45%	47.13 %
Self-Contact	53.45 %	54.94 %	58.52 %	50.96 %	51.62 %	54.31 %	32.03 %	38.85 %
Other-Person	6.46 %	6.56 %	12.94 %	32.00 %	33.79 %	39.51 %	7.32 %	6.77 %
Object-Contact	89.70 %	90.34 %	92.70 %	70.75 %	67.43 %	73.34%	49.68 %	74.27 %
mAP	54.04 %	54.78 %	58.10 %	53.23 %	53.31 %	57.41 %	33.37 %	41.76 %

Reviewer 1: 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 1. 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. **A:** We respect your opinion. But an average improvement of 4% is significant, and a higher experimental standard would have not 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. **A:** Removing any single component reduces mAP roughly by at least 1.5%, while removing both components reduces mAP by 2.5%.

Q4: Hand joint locations seem more appropriate comparison than human body joints. **A:** 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? **A:** While 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).

Reviewer 3: Q1: Why are two attention modules connected this way? **A:** First, the region between hand and object can have plausible regions of contact and we want to predict contact scores directly by spatially attending such regions using the spatial attention module. Second, the appearance of the hand and its affinity between surrounding objects provides strong cues in determining its contact state. We encode these information using cross-feature affinity attention module and predict another set of scores. The Contact Estimation branch is designed to combine two sets of scores from two attention models.

Q2: Why are hand instances annotated with quadrilateral boxes instead of any number of vertices? **A:** Due to many small and blurry hands in our data, it would be ambiguous and prohibitively laborious to use free form polygons.

Q3: Does the the 1^{st} column of Tab 2 represents the results of [28]? Why not adding the ref [28] to that column instead of calling it Mask-RCNN? Retrain [28] with new data **A:** Yes, this will be fixed. See Tab 1 for requested experiments.

Q4: By using "weak annotation", the paper implicitly assumes that predicting the object is part of the task. **A:** We understand your concern, and we will clarify it in our revised paper. We adopted the term "weak annotation" from the field of multiple instance learning, where detection is not necessarily the main task.

Q5: The dataset of 22K images appears to be small. **A:** We respect your opinion. But 22K images is not small. 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 of Tab 1 shows that our dataset performs much better when compared to the previous larger dataset in the task of cross dataset evaluation.

Q6: For Axis-Parallel, Extended performs better than Exact, but is the opposite case in Quadrilateral. **A:** 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? **A:** We used axis-aligned box, following the standard evaluation protocol for hand detection [1, 22, 28].

Reviewer 4: Q1: 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.

Q2: 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.