Evaluation. Our initial experiments only used the first 500 images from the MS-COCO data set, as was done in [3]. In the tables below we use the full test set with the best performing reduction based on the initial experiments. As **R3** and **R4** suggest, we will use the full test set to update the comparison Table 1 of Sec. 6 toawrds the final manuscript.

Additional Experiments. As **R3** suggested, we provide additional results for different architectures. While the median-smoothed RCNN has better clean performance as measured by F1 score, its certified precision recall is strictly lower highlighting the trade-off between clean accuracy and robustness as discussed in the literature.

	Precision	**Recall**	**F1**	**Certified Precision**	**Certified Recall**
YOLOv3	91.80%	20.88%	34.02%	29.65%	11.43%
Faster RCNN	86.58%	24.63%	38.35%	17.69%	10.85%
Mask RCNN	85.50%	25.14%	38.85%	17.42%	10.85%

Other Performance Metrics. As **R4** suggested, we examine the average precision (AP) for both the plain and certified detectors; we report the results for YOLOv3 below, and plan to include Faster RCNN and Mask RCNN in the final manuscript. Since varying the objectness threshold changes the base detector (f), we reevaluate the smoothed detector at objectness thresholds $\{0.1, 0.2, 0.4, 0.6, 0.8\}$ and calculate the area under the steps to lower bound true AP. As for inference speed, the smoothing paradigm is inherently costly, as we use 2000 perturbations for Monte Carlo estimation. We leave it to future work to improve these important metrics as needed in practice.

	AP@50	Certified AP@50
YOLOv3	32.0%	4.2%

Comparison to Prior Work. We thank **R4** for bringing relevant prior work to our attention. While we cannot claim to be the first adversarial defense for object detection, we maintain that we provide the first certified defense. The performance of the certified defense approach is currently so weak compared to the adversarial training approach that we do not think a meaningful quantitative comparison can be done. We will make sure to discuss the relation to this prior work as follow: "our certified radius is 0.36 in terms of ℓ_2-norm whereas Zhang et al. (ICCV'19) achieved robustness radius of 8/255 in terms of the stronger ℓ_∞-norm threat model." As **R2** suggests, we will distinguish certification and defense as follows: "Several methods of obtaining robustness certificates for classification problems have been proposed [19, 21, XX]. In addition, [9, 11, 24, 25, 29, YY] proposes methods to both defend the model while enabling better certificates;" we will make sure to include the citations suggested by **R2**.

Tailored Empirical Attacks. As **R2** suggested, we implemented a DAG attack against our best performing model. The DAG attack is modified to include Monte Carlo sampling to increase the strength of the attack. We take 20 PGD steps and draw 5 random samples to estimate the gradient of the smoothed model. Surprisingly, the smoothed model is quite robust within the desired radius. The DAG attack was only able to decrease recall by 1.1%. This illustrates that the bound we obtained is likely quite loose with respect to the true robustness of the object detector, and we leave improvements of the robustness certificate as future work.

Certifiable radius. **R2** rightly points out that as of right now the certifiable radius is too low for real-world applications. We emphasize that certified robustness is a challenging domain and none of the existing methods yield practical certificates even for classification problems. For example, while the SOTA certified defense by Salman et al. (NeurIPS'19) achieved 68.2% certified accuracy for $||\epsilon||_\infty < 2/255$ on CIFAR-10, the empirical approach of Xie et al. (CVPR'19) can achieve similar empirical robustness but at the ImageNet scale and with larger radius. That said, our work leverages a principled smoothing approach to provide the first non-trivial certificates for architectures as complex as object detectors.

Generality of the Techniques. **R1** remarks that the sorting and bucketing techniques proposed in Section 5 may be too specific for object detection. We note that the proposed techniques are potentially applicable to certifying other networks through reductions to a regression formulation. This is particularly relevant for tasks that have variable length outputs, such as key points detection, instance segmentation, or image captioning. Viewed in the broader context of adversarial robustness for ML models, computer vision continues to provide exemplar problems and we hope our work on object detection will help advance both the theory and practice of this important field.

Other Clarifications and Corrections. **R2** correctly points out an issue with Equation 3. We already fixed this issue in the supplemental materials, replacing min/max with inf/sup. We will also clarify the definition of the worst-case bounding box as "the box with coordinates satisfying the certified upper and lower bounds which realizes the lowest