Thanks for your constructive reviews. We tried our best to properly respond to all the inquiries from the reviewers.

1. Common Response

Comparison with SSM: For a practical usage of a model, the stopping criteria must be defined properly. SSM continues the training until no more unlabeled data are included in the training set (Algorithm 1 in [1]). In Fig.4 of [1], the performance improves initially as unlabeled samples are added but it starts to degrade as more samples are added. However, the reported score of the SSM is not from an objective stopping criterion but the peak performance during the entire iterations. With this setting, the influx of the data from out-of-distribution and incorrectly labeled samples cannot be prevented. As we all know that the performance of SSM should not be measured with this setting, we are not sure that the performance of a detector combined with SSM would get better. To check this, we tried to implement SSM in SSD. However, many details are missing in SSM and the learning parameters of single-stage detector and two-stage detector are different. Its intense time-cost and huge hyper-parameter space makes it difficult to implement SSM properly. In our work, we just wanted to present a representative self-learning method.

Dataset: When the unlabeled samples originate from out-of-class distribution, the performance of any semi-supervised learning method usually degrades. This is a well known limitation of SSL [2] and we wanted to show the performance of our method in the in-class distribution setting as other works.
Also, we acknowledge that using a significantly huge and unrefined data would be an ideal setting. However, this paper has its focus on applying the consistency-based semi-supervised approach to the object detection task and the experimental settings have been inspired from those in the paper of SSM. We would like to mention on the dataset-related research direction that some reviewers pointed out in an additional paragraph in the final version.

2. Response to each Reviewer

Reviewer 1: In consistency regularization, various candidates of loss function can be used. We simply chose the Π model, which uses L_2 loss, as a baseline and we will explain this more in detail in the final version.
Particularly in the case of using con-l in SSD300, the performance is quite good even without using BE. We think that it is because the number of samples with flat areas is relatively small for a small input resolution and this helps the regression process while con-c disturbs the effect of con-l.
We have experimented with SSD300 model for the effect of not using CR on labeled samples. SSD300 (CR on unlabeled samples only) scored 72.1mAP in semi-supervised learning using VOC07(labeled) and VOC12(unlabeled) and it is slightly worse than the base CSD (72.3 mAP).
We will search for more related works including the ones the reviewer mentioned and will try to explore a link between the works and Fig. 1 in our paper.

Reviewer 2: Table 2 shows the results of experiments with 20 COCO classes and the entire 80 classes. As the reviewer mentioned, we would like to modify our claim (line 246-248) more clearly as follows: The reason why using 20 class categories as unlabeled data outperforms the other case is because the ratio of *labeled/unlabeled class mismatch* is much smaller. This is similar to the experimental results of Fig. 2 in [2].

Reviewer 3: As mentioned in the paper, the self-training method is an iterative method. So it is time-consuming and computationally intensive. In addition, the mAP score is reported wrongly in SSM, as mentioned above. Meanwhile, our CSD is a new type of semi-supervised learning method for object detection that can be trained with an additional loss reflecting consistency between pairs of images.

Reviewer 4: Both hard negative mining and focal loss require a label in their training. To apply these to the consistency loss of unlabeled data, they should be modified appropriately. Therefore, we propose BE, which is applicable in the semi-supervised learning environment. It seems that the reviewer considered the weighted loss using the background score, which we think can be effective that can be another new contribution. We will apply it in our future research.
It is important to clarify the relationship between boxes by image perturbation to calculate the consistency loss. As the reviewer mentioned, the performance is expected to be improved by diversity if some limited perturbation schemes are applied. Although some of the perturbation methods the reviewer mentioned have minor problems (e.g image wrapping can cut off an object, which is not representable by a single bounding box), we agree that additional perturbation methods will probably improve the performance. However, flipping is the most simple method which always guarantees the one-to-one correspondence of given boxes. In the discussion, we will complement the possibility of using other perturbation methods.

References

[1] Keze Wang, Xiaopeng Yan, Dongyu Zhang, Lei Zhang, and Liang Lin. Towards human-machine cooperation: Self-supervised sample mining for object detection. In *CVPR*, pages 1605–1613, 2018.

[2] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NIPS*, pages 3235–3246, 2018.