Supplementary Materials for Improving Object Detection with Selective Self-supervised Self-training

Yandong Li[1,2,*][0000−0003−2448−1294], Di Huang[2][0000−0001−6021−5175], Danfeng Qin[2][0000−0002−7756−2074], Liqiang Wang[1][0000−0002−1265−4656], and Boqing Gong[2][0000−0003−3915−5977]

[1]University of Central Florida, [2]Google Inc
lyndon.leeseu@outlook.com, lwang@cs.ucf.edu,
{dihuang,qind,bgong}@google.com

We provide the following to support the main text:

Section A: description of the data annotation protocol used in this work (cf. Section 2 in the main paper),
Section B: additional experimental results of the self-training for object detection (SOD) and our selective self-supervised self-training (S^4OD) (cf. Section 5.1 in the main paper), and
Section C: qualitative results of the binary detector (BD) and S^4OD.

A Data annotation protocol

We ask the raters to draw a bounding box around each and every backpack shown in the images, and the boxes should be tight. We underscore two scenarios in the instruction sent to the raters.

1. Please be careful about the labeling because many of the backpacks are very small and hard to find. We want all of them to be labeled.
2. Pay special attention to the backpack straps. Label them even if only the straps are visible.

Additionally, we provide concrete examples as shown in Figure 2 to help the raters understand the annotation instruction. Figure 1 shows the annotations we collected *vs.* COCO annotations.

Fig. 1. Our annotations *vs.* COCO annotations.

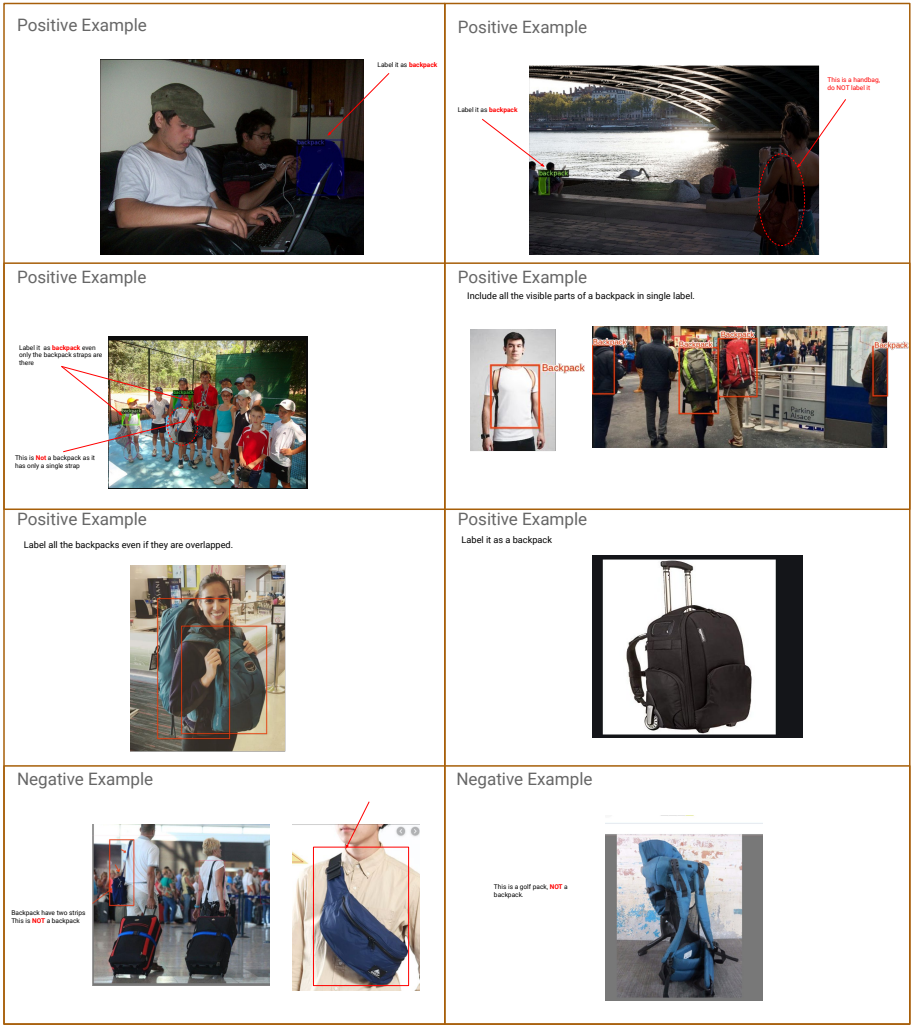

Fig. 2. Some examples we provide to the raters to help them understand our annotation instruction.

B Additional experimental results of SOD and S^{4}OD

B.1 Threshold Selection

The self-training for object detection (SOD) selects pseudo boxes over unlabeled images by thresholding their confidence scores. We show in Table 1 the SOD results with different thresholds: 0.5, 0.7 and 0.9. We find that thresholding the confidence scores can barely improve object detection, and it is sensitive to the threshold. In contrast, we threshold the intersection over union (IoU) of the pseudo and the groundtruth boxes over the training images for learning the selective net in S^{4}OD. The groundtruth boxes lend our selective net robustness to the thresholds γ_h and γ_l (cf. Table 1 which shows S^{4}OD with different choices of thresholding the IoU — S^{4}OD-0.5-0.1 means $\gamma_h = 0.5$ and $\gamma_h = 0.1$). Especially, we can set γ_h by maximizing the detector's performance on the training set (cf. Section 3.2 in the main paper).

Table 1. Results of SOD with different confidence score thresholds and S^{4}OD with different IoU thresholds.

Method	AP@[.5, .95]	Gain Over BD	AP@.5	AP@.75	AP_S	AP_M	AP_L
BD [4]	17.34	–	34.71	15.13	19.03	19.46	21.31
SOD-0.5 [6]	17.21	-0.13	32.66	15.81	18.03	21.03	18.09
SOD-0.7 [6]	17.37	0.03	34.26	14.46	18.77	20.10	17.66
SOD-0.9 [6]	17.22	-0.12	32.76	17.44	19.03	21.05	19.97
S^{4}OD-0.5-0.1	19.41	2.07	36.19	18.67	21.07	22.08	21.02
S^{4}OD-0.6-0.05	19.48	2.14	36.25	19.01	20.31	**23.47**	18.53
S^{4}OD-2_{nd} Iteration	**19.52**	**2.18**	**36.44**	**19.07**	**21.70**	20.90	**24.90**

B.2 "Upper bounds" of S^{4}OD

We provide a "upper bound" for S^{4}OD tighter than the one described in the main text. We train S^{4}OD in two stages: pre-train the student detector on the unlabeled images and then fine-tune it on the labeled set. To obtain a more comparable "upper bound", we also train the binary detector in the two-stage manner except that we reveal labels to the detector in the first stage. Table 2 shows that the two-stage "upper bound" is tighter than the one we provide in the main text. In the main paper, all the results of detecting backpacks using S^{4}OD are obtained using the full Web-backpack. In Table 2, we additionally show the results of S^{4}OD pre-trained using the "Web-backpack labeled" only — but we use the images only and no labels at all.

B.3 More iterations on VOC

We train the second iteration of S^{4}OD on VOC benchmark [2]. Table 3 shows that S^{4}OD-2_{nd} Iteration can improve the single-iteration result by a small margin.

Table 2. Comparison results on the **relabeled** COCO-backpack.

Method	AP@[.5, .95]	AP@.5	AP@.75	AP_S	AP_M	AP_L
BD [4]	18.75	36.03	16.98	11.90	21.62	31.58
Upper bound	22.63	40.89	21.15	15.97	26.53	32.11
Upper bound (two-stage)	22.10	40.35	22.58	16.23	25.56	32.70
S^4OD w/ full Web-backpack	20.27	37.55	20.49	13.51	24.04	31.17
S^4OD w/ subset Web-backpack	19.70	38.33	18.05	14.19	23.51	29.32

This is similar to our observation on COCO-backpack (cf. S^4OD-2_{nd} Iteration in Table 1). We conjecture that it is because the student detector can not obtain more supervision from the teacher in the second iteration.

Table 3. Comparison results on VOC2007 (* the numbers reported in [3])

Method	Labeled data	Unlabeled data	AP	Gain
Supervised [1, 5]	VOC07	–	*73.9/74.1	–
Supervised [1, 5]	VOC07&12	–	*79.4/80.3	–
CSD [3]	VOC07	VOC12	*74.7	*0.8
S^4OD	VOC07	VOC12	76.4	2.3
S^4OD-2_{nd} Iteration	VOC07	VOC12	**76.5**	**2.4**

C Qualitative results

We show some qualitative results of the binary detectors (BDs) and S^4OD. They demonstrate the cases that S^4OD improves over BD and when it does not. They also underscore the challenges in detecting the common, "uninteresting" man-made objects. We apply our most powerful models on the validation set of COCO-backpack and COCO-chair and calculate the per-image AP for each image. We then rank the images according to the per-image AP so that we know which images contain easy-to-detect objects and which images have hard-to-detect objects. Figures 3&4 and 5&6 show the easy-to-detect backpacks and chairs, respectively. Figure 7&8 and 9&10 show the hard backpacks and chairs, respectively. Figure 11 shows some predicted boxes on the images which contain no backpack or chair. Overall, S^4OD can provide more accurate boxes than BD. Especially, BD generates many false positive predictions. The hard-to-detect objects are often small, occluded, or labeled incorrectly.

Fig. 3. Backpacks that are easy to detect. We show the predicted boxes by S^{4}OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

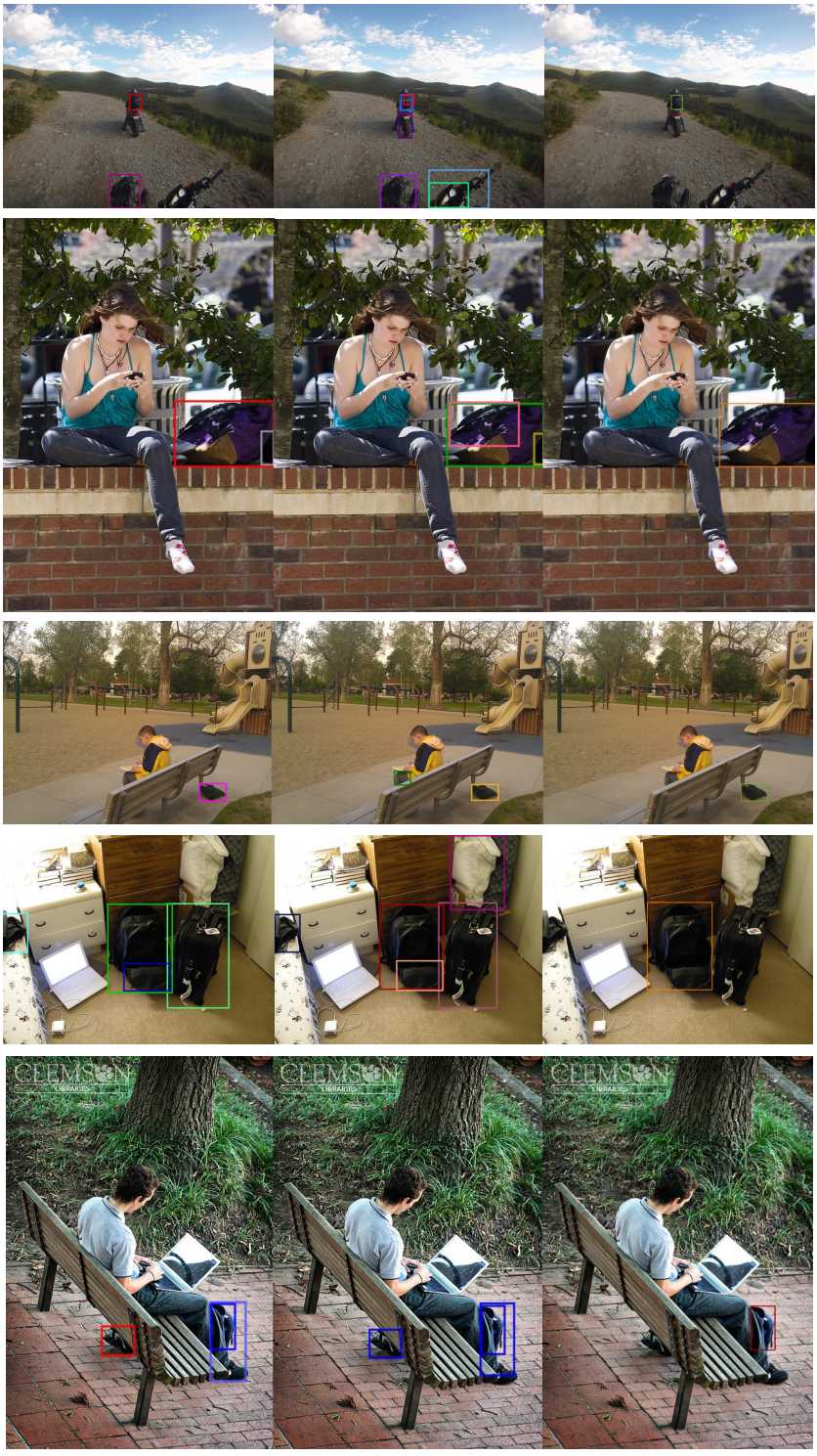

Fig. 4. Backpacks that are easy to detect. We show the predicted boxes by S^4OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

Fig. 5. Chairs that are easy to detect. We show the predicted boxes by S^4OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

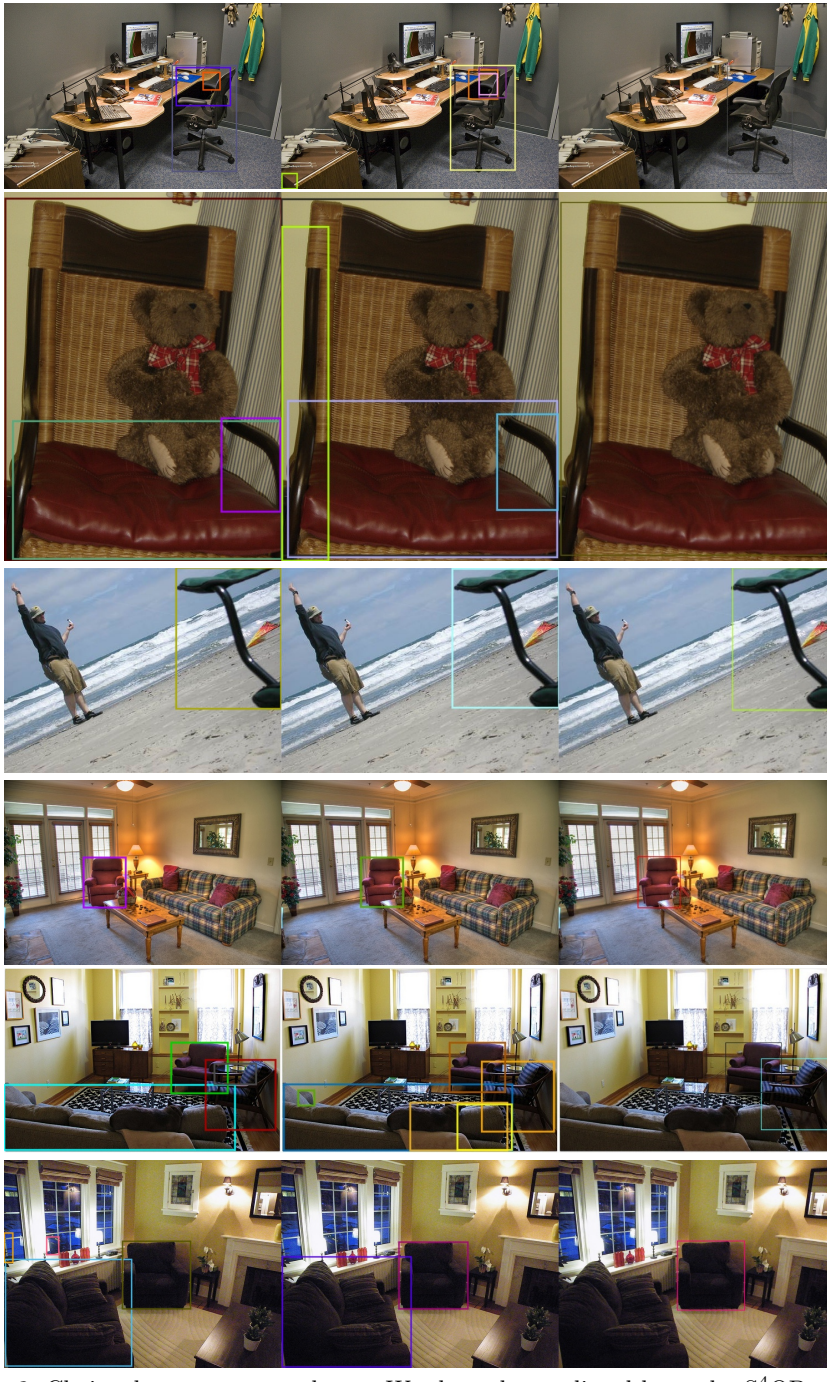

Fig. 6. Chairs that are easy to detect. We show the predicted boxes by S^{4}OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

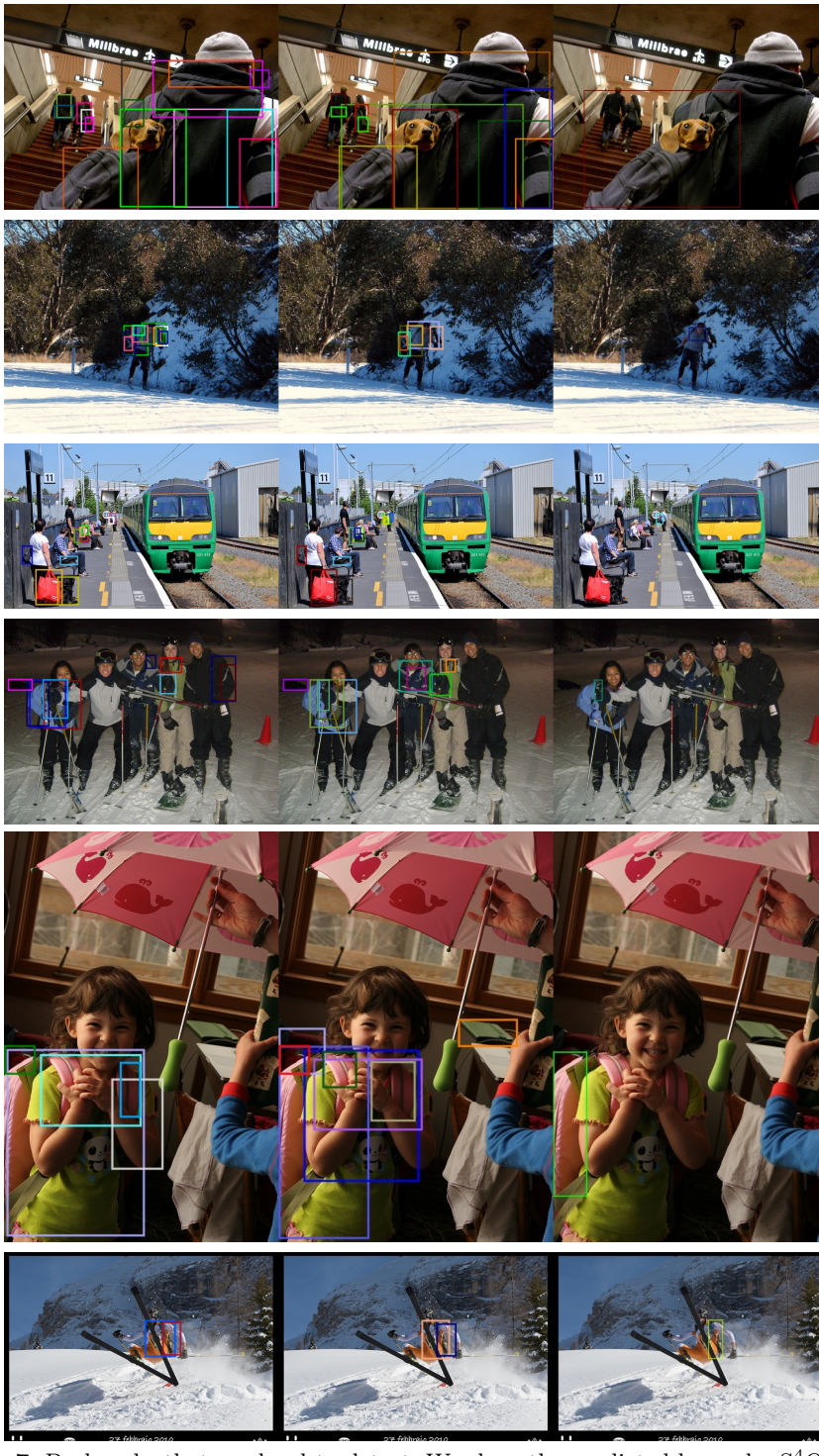

Fig. 7. Backpacks that are hard to detect. We show the predicted boxes by S^4OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

Fig. 8. Backpacks that are hard to detect. We show the predicted boxes by S^4OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

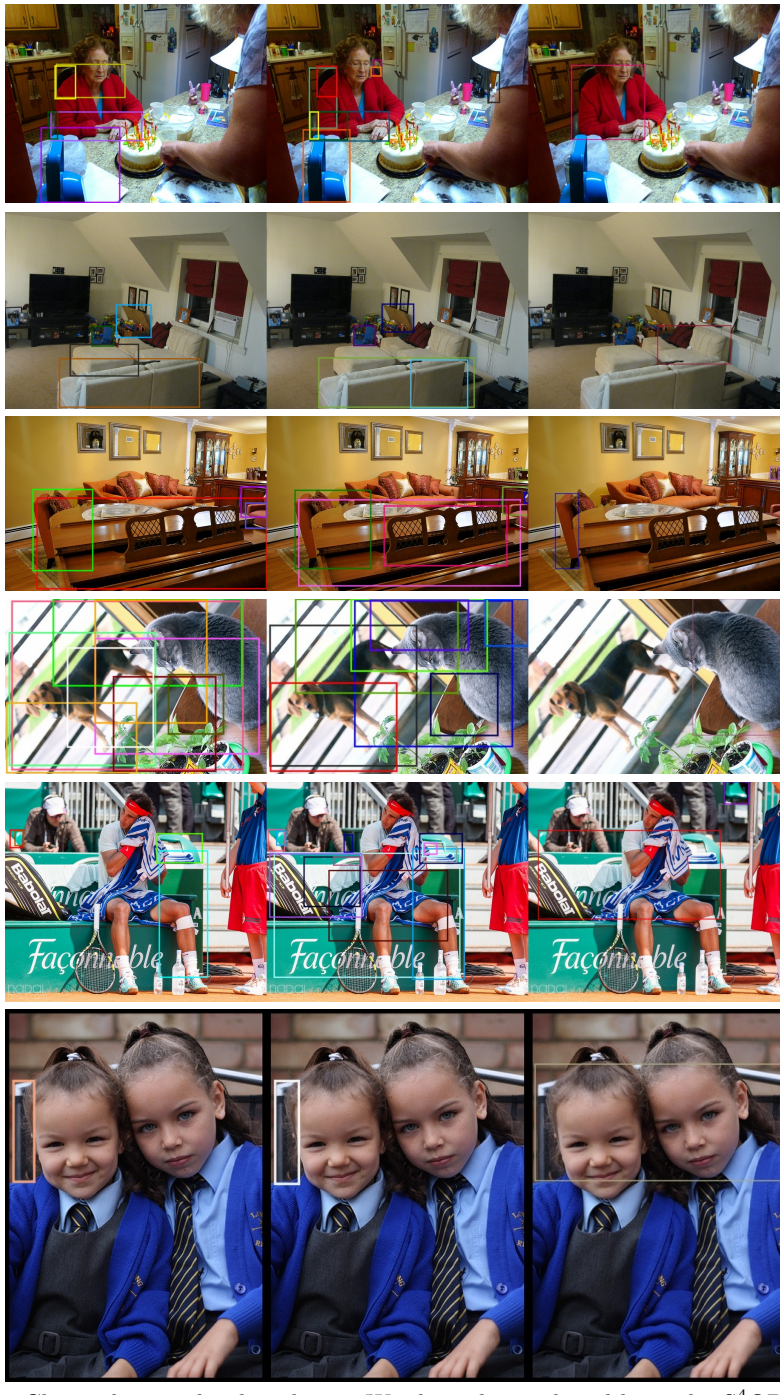

Fig. 9. Chairs that are hard to detect. We show the predicted boxes by S^{4}OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

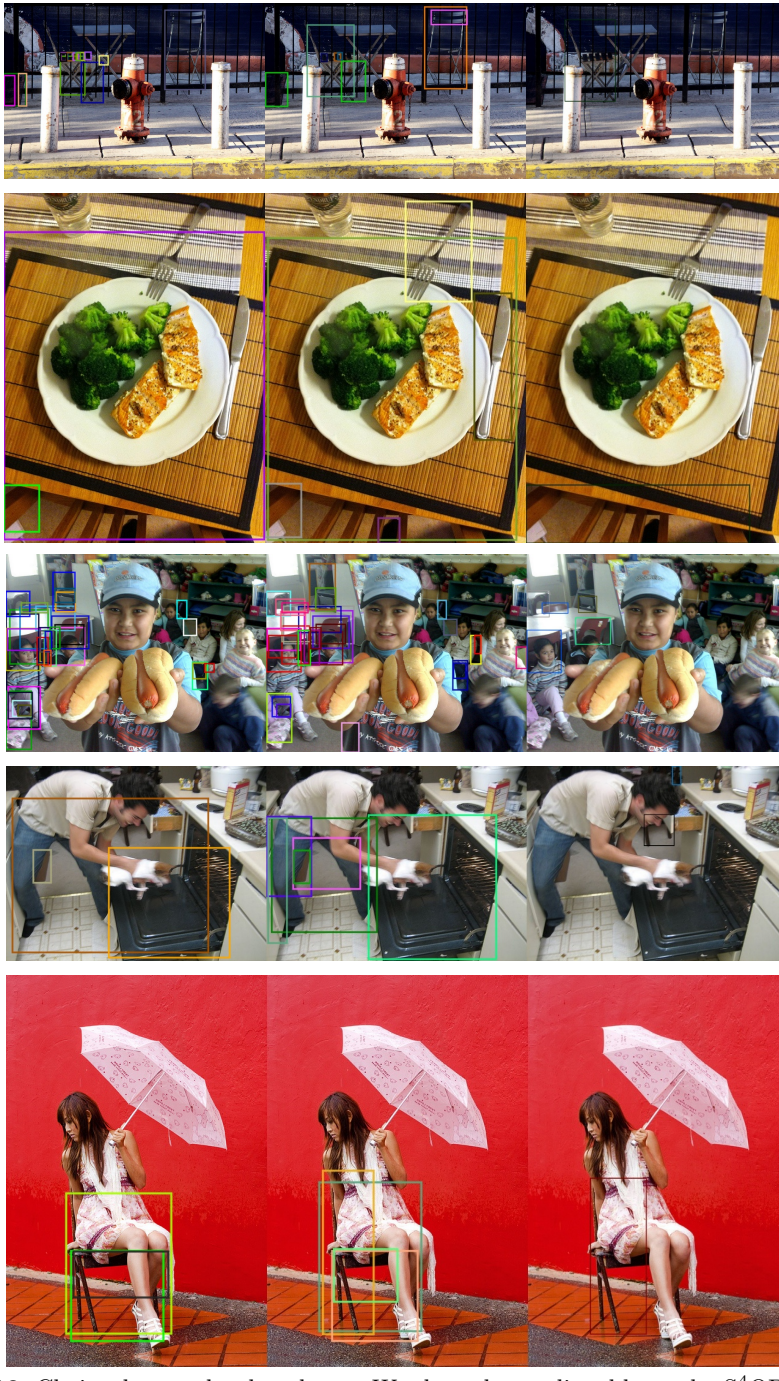

Fig. 10. Chairs that are hard to detect. We show the predicted boxes by S^4OD on the left, the predictions by BD in the middle, and the ground-truth on the right.

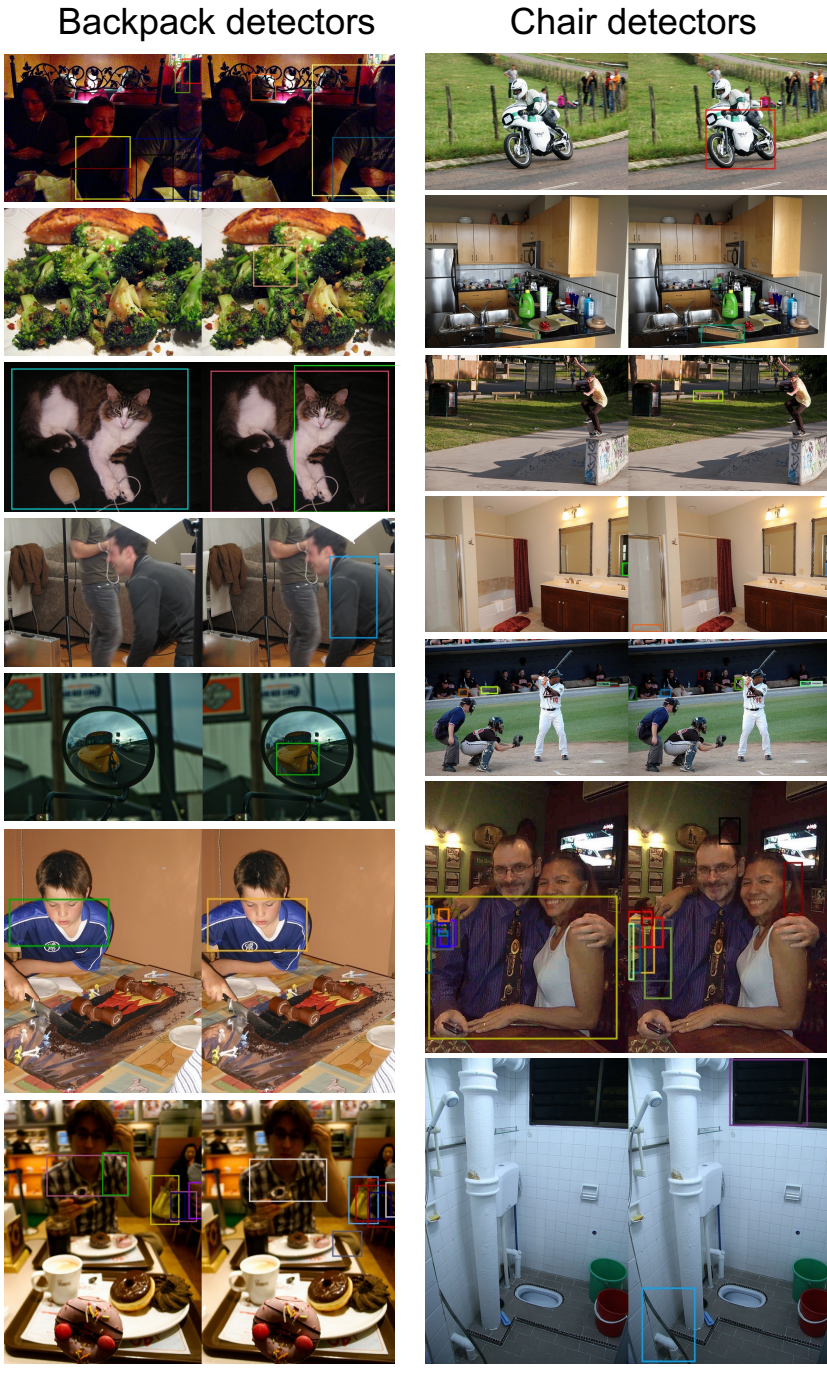

Fig. 11. False positives over the images that contain neither backpacks nor chairs. We show the predicted boxes by S^4OD on the left and the predictions by BD on the right.

References

1. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems. pp. 379–387 (2016)
2. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
3. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. In: Advances in Neural Information Processing Systems. pp. 10758–10767 (2019)
4. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
5. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
6. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory **11**(3), 363–371 (1965)