Appendix

6 Implementation Details for P2BNet

6.1 Sampling Settings.

In CBP sampling, $s \in \{4, 8, 16, 32, 64, 128\} \cdot \delta$, where δ is a factor for dynamic adjustment according to the dataset. $\delta = min(W, H)/100$, where W and H are width and height of the image. $v \in \{1/3, 1/2, 2/3, 1, 3/2, 2, 3\}$ is used as the fixed setting. In PBR sampling, we deem $(v \cdot s)$ and (v/s) as a whole respectively, and set $(v \cdot s) \in \{0.7, 0.8, 1, 1.2, 1.3\}$, $(v/s) \in \{0.7, 0.8, 1, 1.2, 1.3\}$, so we have $5 \times 5 = 25$ options. $(o_x, o_y) \in \{(0, 0), (1, 0), (0, 1), (-1, 0), (-1, -1)\}$ is used to jitter the center position. These settings are simple and fixed, which is beneficial for better generation. In negative sampling, we randomly sample 500 boxes, filter out those which have high IoU with all positive proposals and obtain the final negative sample set $\mathcal{N}$.

6.2 Other Experimental Settings.

ResNet-50 is used as the backbone network unless otherwise specified. The FPN structure is also utilized. The mini-batch is 16 images and all models are trained with 8 GPUs for COCO dataset. In Tab. 2 (a,c), the number of PBR stages has been described in the corresponding section. Loss weights are set as $\alpha_{mil1} = 0.25$ in CBP, $\alpha_{mil2} = 0.25$ and $\alpha_{neg} = 0.75$ in PBR, which follow focal loss [23] and are fixed during training. Unless otherwise specified, we use one CBP stage and one PBR stage as our default configuration, which the experiments of Tab. 2 (b) and 3 (a,b,c) depend on.

7 Other Experiments.

7.1 Different Backbones of P2BNet.

We change the backbone to a larger network, ResNet-101, but the performance decreases. The performance of R-50 is 21.7 AP and 46.1 AP_{50} while that of R-101 is 20.8 AP and 45.8 AP_{50}. We conjecture this is because the larger the network, the more possibility it is to predict the discriminative part rather than the whole object. This phenomenon also happens in WSOD.

7.2 Annotation Time of Quasi-center Point.

Quasi-center point annotation requires annotating objects in center region. Generally, center region of the object is the body part which has more saliency and easy to annotate. For comparison, 200 images were manually annotated by three annotators. The average annotation time per object is 4.27s for quasi-center strategy and 3.84s for the previous random annotation style. Considering the huge performance gain, we believe this is a fair trade-off.

8 The visualization of P2BNet-FR

8.1 Visualization of P2BNet.

In Fig. 7, we illustrate the visualization of P2BNet. With QC point annotation, P2BNet predicts coarse pseudo boxes in the CBP stage, and refine the quality of estimated boxes in the PBR stage. Furthermore, Fig. 7 also shows the performance of P2BNet in complex scenarios, reflecting that the pseudo boxes predicted by P2BNet have the ability to well represent their respective targets. The estimated boxes are used as the pseudo annotation to train the detector, so the high quality of pseudo boxes guarantee high detection performance.

8.2 Visualization of Detection Performance.

Some other detection results of P2BNet-FR are given in Fig. 8. It shows our PSOD detector's performance is comparable with that under box supervision. The performance of WSOD and PSOD is low in complex dataset like COCO. However, our PSOD detector P2BNet-FR has achieved great improvement, especially in complex and dense scenarios.

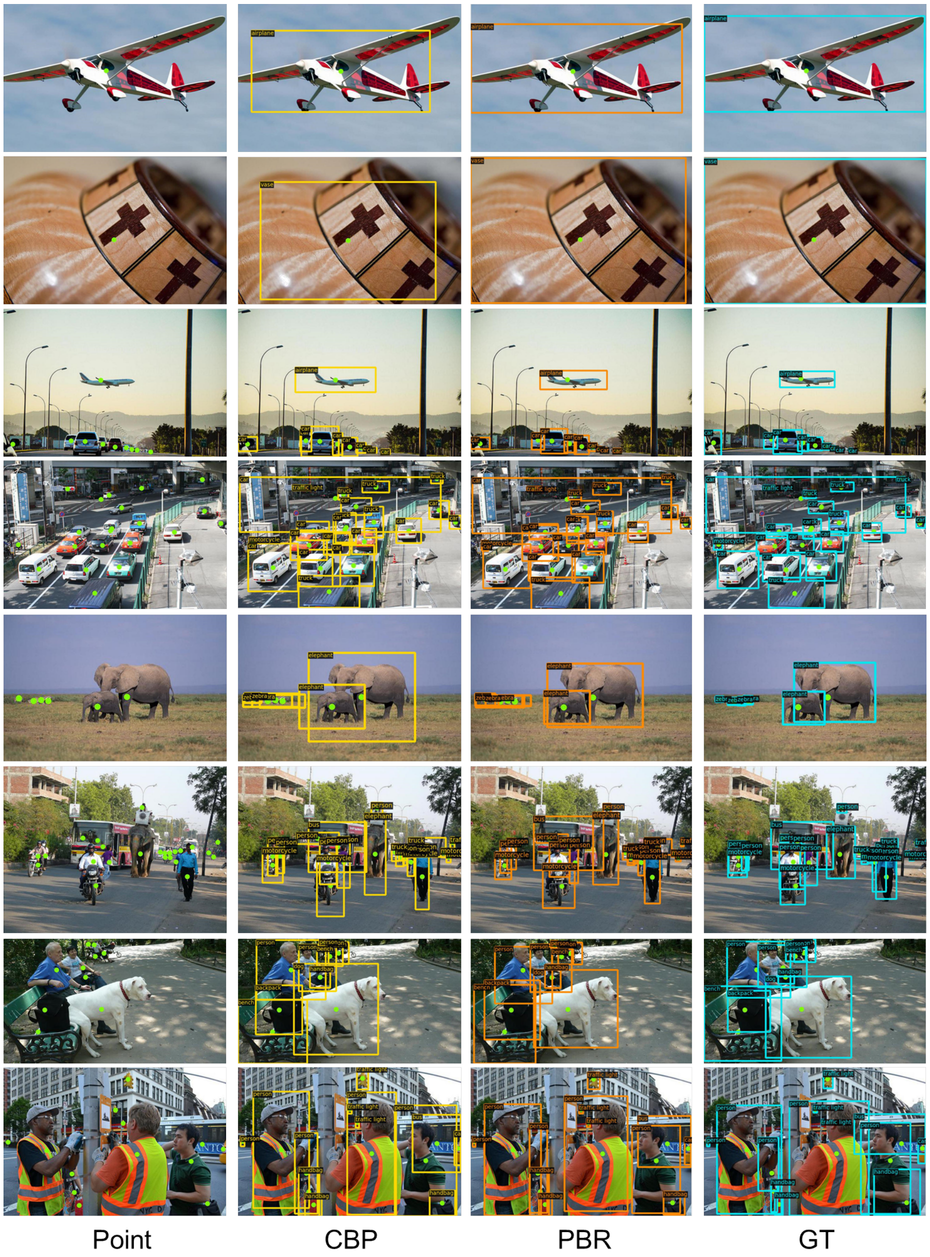

Fig. 7. The Visualization of P2BNet. The green, yellow, orange and blue represent the annotation point, the CBP result, the PBR result and the ground-truth, respectively. With CBP stage and PBR stage, we obtain the high-quality pseudo boxes. The performance in dense scenarios is also great. The images are from COCO-17 train set. (Best viewed in color)

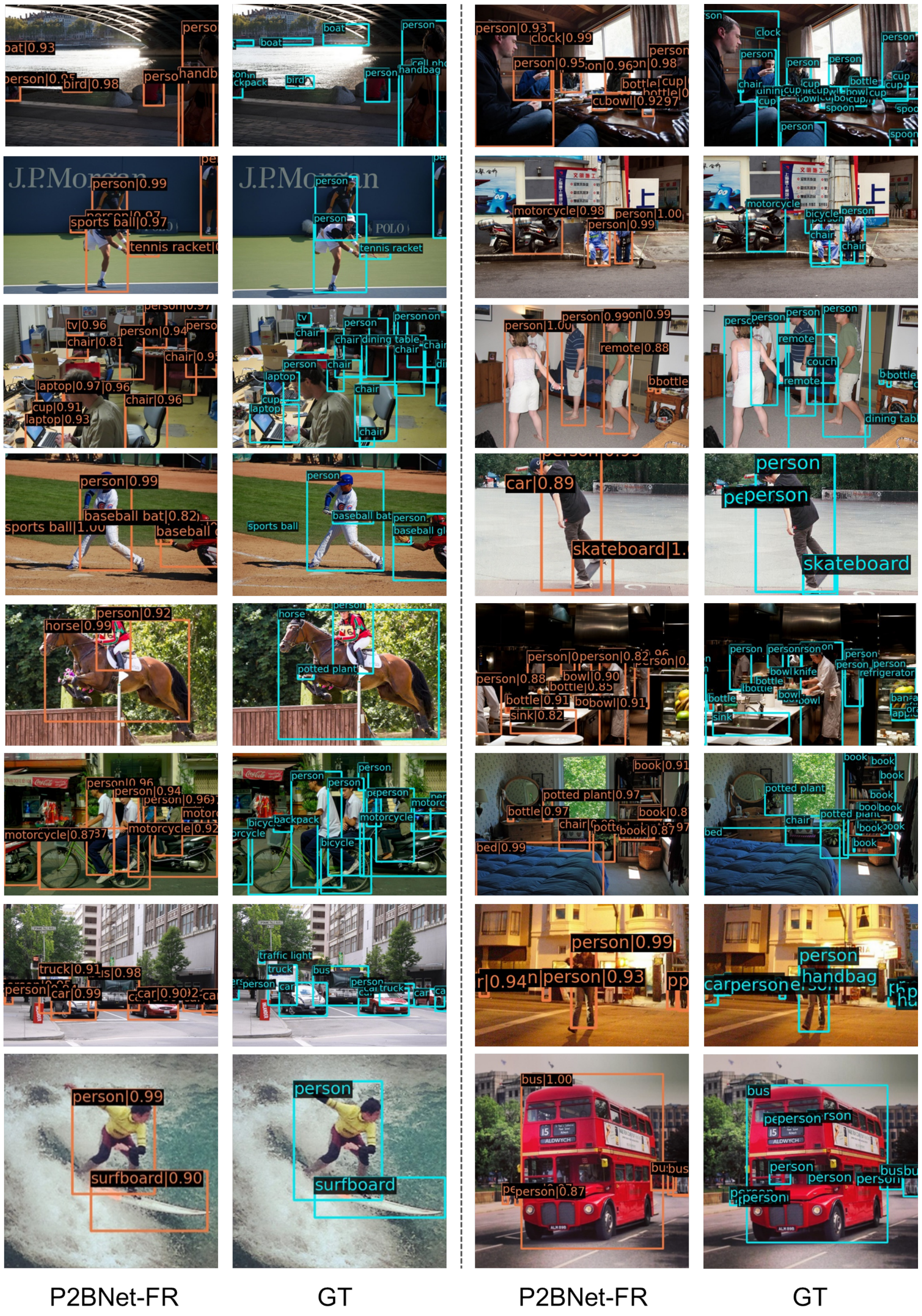

Fig. 8. The detection visualization of P2BNet-FR. Orange and blue represent detection result and ground-truth. Our P2BNet-FR can detect objects in complex scenarios with point annotation. The images are from COCO-17 val set. (Best viewed in color)