Single-Shot Refinement Neural Network for Object Detection -Supplementary Material-

Shifeng Zhang[1,2], Longyin Wen[3], Xiao Bian[3], Zhen Lei[1,2], Stan Z. Li[4,1,2]
[1] CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[2] University of Chinese Academy of Sciences, Beijing, China.
[3] GE Global Research, Niskayuna, NY.
[4] Faculty of Information Technology, Macau University of Science and Technology, Macau, China.
{shifeng.zhang,zlei,szli}@nlpr.ia.ac.cn, {longyin.wen,xiao.bian}@ge.com

1. Complete Object Detection Results

We show the complete object detection results of the proposed RefineDet method on the PASCAL VOC 2007 `test` set, PASCAL VOC 2012 `test` set and MS COCO `test-dev` set in Table 1, Table 2 and Table 3, respectively. Among the results of all published methods, our RefineDet achieves the best performance on these three detection datasets, *i.e.*, 85.8% mAP on the PASCAL VOC 2007 `test` set, 86.8% mAP on the PASCAL VOC 2012 `test` set and 41.8% AP on the MS COCO `test-dev` set.

Table 1: Object detection results on the PASCAL VOC 2007 `test` set. All models use VGG-16 as the backbone network.

Method	Data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
RefineDet320	07+12	80.0	83.9	85.4	81.4	75.5	60.2	86.4	88.1	89.1	62.7	83.9	77.0	85.4	87.1	86.7	82.6	55.3	82.7	78.5	88.1	79.4
RefineDet512	07+12	81.8	88.7	87.0	83.2	76.5	68.0	88.5	88.7	89.2	66.5	87.9	75.0	86.8	89.2	87.8	84.7	56.2	83.2	78.7	88.1	82.3
RefineDet320+	07+12	83.1	89.5	87.9	84.9	79.7	70.0	87.5	89.1	89.8	69.8	87.1	76.4	86.6	88.6	88.4	85.3	62.4	83.7	82.3	89.0	83.1
RefineDet512+	07+12	83.8	88.5	89.1	85.5	79.8	72.4	89.5	89.5	89.9	69.9	88.9	75.9	87.4	89.6	89.0	86.2	63.9	86.2	81.0	88.6	84.4
RefineDet320	COCO+07+12	84.0	88.9	88.4	86.2	81.5	71.7	88.4	89.4	89.0	71.0	87.0	80.1	88.5	90.2	88.4	86.7	61.2	85.2	83.8	89.1	85.5
RefineDet512	COCO+07+12	85.2	90.0	89.2	87.9	83.1	78.5	90.0	89.9	89.7	74.7	89.8	79.5	88.7	89.9	89.2	87.8	63.1	86.4	82.3	89.5	84.7
RefineDet320+	COCO+07+12	85.6	90.2	89.0	87.6	84.6	78.0	89.4	89.7	89.9	74.7	89.8	80.5	89.0	89.7	89.6	87.8	65.5	87.9	84.2	88.6	86.3
RefineDet512+	COCO+07+12	85.8	90.4	89.6	88.2	84.9	78.3	89.8	89.9	90.0	75.9	90.0	80.0	89.8	90.3	89.6	88.3	66.2	87.8	83.5	89.3	85.2

Table 2: Object detection results on the PASCAL VOC `2012 test` set. All models use VGG-16 as the backbone network.

Method	Data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
RefineDet320	07++12	78.1	90.4	84.1	79.8	66.8	56.1	83.1	82.7	90.7	61.7	82.4	63.8	89.4	86.9	85.9	85.7	53.3	84.3	73.1	87.4	73.9
RefineDet512	07++12	80.1	90.2	86.8	81.8	68.0	65.6	84.9	85.0	92.2	62.0	84.4	64.9	90.6	88.3	87.2	87.8	58.0	86.3	72.5	88.7	76.6
RefineDet320+	07++12	82.7	92.0	88.4	84.9	74.0	69.5	86.0	88.0	93.3	67.0	86.2	68.3	92.1	89.7	88.9	89.4	62.0	88.5	75.9	90.0	80.0
RefineDet512+	07++12	83.5	92.2	89.4	85.0	74.1	70.8	87.0	88.7	94.0	68.6	87.1	68.2	92.5	90.8	89.4	90.2	64.1	89.8	75.2	90.7	81.1
RefineDet320	COCO+07++12	82.7	93.1	88.2	83.6	74.4	65.1	87.1	87.1	93.7	67.4	86.1	69.4	91.5	90.6	91.4	89.4	59.6	87.9	78.1	91.1	80.0
RefineDet512	COCO+07++12	85.0	94.0	90.0	86.9	76.9	74.1	89.7	89.8	94.2	69.7	90.0	68.5	92.6	92.8	91.5	91.4	66.0	91.2	75.4	91.8	83.0
RefineDet320+	COCO+07++12	86.0	94.2	90.2	87.7	80.4	74.9	90.0	91.7	94.9	71.9	89.8	71.7	93.5	91.9	92.4	91.9	66.5	91.5	79.1	92.8	83.9
RefineDet512+	COCO+07++12	86.8	94.7	91.5	88.8	80.4	77.6	90.4	92.3	95.6	72.5	91.6	69.9	93.9	93.5	92.4	92.6	68.8	92.4	78.5	93.6	85.2

Table 3: Object detection results on the MS COCO `test-dev` set.

Method	Net	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
RefineDet320	VGG-16	29.4	49.2	31.3	10.0	32.0	44.4	26.2	42.2	45.8	18.7	52.1	66.0
RefineDet512	VGG-16	33.0	54.5	35.5	16.3	36.3	44.3	28.3	46.4	50.6	29.3	55.5	66.0
RefineDet320	ResNet-101	32.0	51.4	34.2	10.5	34.7	50.4	28.0	44.0	47.6	20.2	53.0	69.8
RefineDet512	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4	30.6	49.0	53.0	30.0	58.2	70.3
RefineDet320+	VGG-16	35.2	56.1	37.7	19.5	37.2	47.0	30.1	49.6	57.4	36.2	62.3	72.6
RefineDet512+	VGG-16	37.6	58.7	40.8	22.7	40.3	48.3	31.4	52.4	61.3	41.6	65.8	75.4
RefineDet320+	ResNet-101	38.6	59.9	41.7	21.1	41.7	52.3	32.2	52.9	61.1	40.2	66.2	77.1
RefineDet512+	ResNet-101	41.8	62.9	45.7	25.6	45.1	54.1	34.0	56.3	65.5	46.2	70.2	79.8

2. Qualitative Results

We show some qualitative results on the PASCAL VOC 2007 `test` set, the PASCAL VOC 2012 `test` set and the MS COCO `test-dev` in Figure 1, Figure 2, and Figure 3, respectively. We only display the detected bounding boxes with the score larger than 0.6. Different colors of the bounding boxes indicate different object categories. Our method works well with the occlusions, truncations, inter-class interference and clustered background.

Figure 1: Qualitative results of RefineDet512 on the PASCAL VOC 2007 `test` set (corresponding to 85.2% mAP). VGG-16 is used as the backbone network. The training data is 07+12+COCO.

Figure 2: Qualitative results of RefineDet512 on the PASCAL VOC 2012 `test` set (corresponding to 85.0% mAP). VGG-16 is used as the backbone network. The training data is 07++12+COCO.

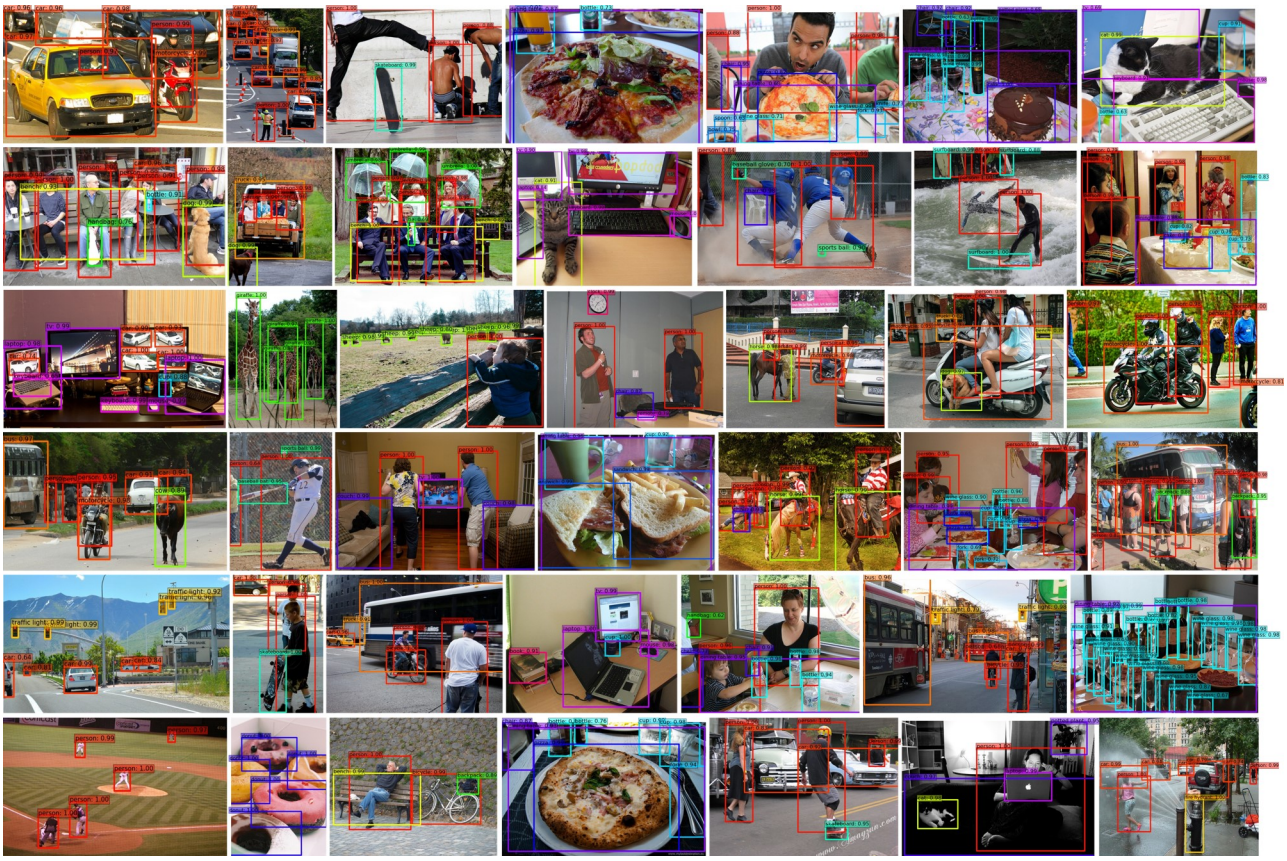

Figure 3: Qualitative results of RefineDet512 on the MS COCO `test-dev` set (corresponding to 36.4% mAP). ResNet-101 is used as the backbone network. The training data is COCO `trainval35k`.

3. Detection Analysis on PASCAL VOC 2007

We use the detection analysis tool[1] to understand the performance of two RefineDet models (*i.e.*, RefineDet320 and RefineDet512) clearly. Figure 4 shows that RefineDet can detect various object categories with high quality (large white area). The majority of its confident detections are correct. The recall is around 95%-98%, and is much higher with "weak" (0.1 jaccard overlap) criteria. Compared to SSD, RefineDet reduces the false positive errors at all aspects: (1) RefineDet has less localization error (Loc), indicating that RefineDet can localize objects better because it uses two-step cascade to regress the objects. (2) RefineDet has less confusion with background (BG), due to the negative anchor filtering mechanism in the anchor refinement module (ARM). (3) RefineDet has less confusion with similar categories (Sim), benefiting from using two-stage features to describe the objects, *i.e.*, the features in the ARM focus on the binary classification (being an object or not), while the features in the object detection module (ODM) focus on the multi-class classification (background or object classes).

Figure 5 demonstrates that RefineDet is robust to different object sizes and aspect ratios. This is not surprising because the object bounding boxes are obtained by the two-step cascade regression, *i.e.*, the ARM diversifies the default scales and aspect ratios of anchor boxes so that the ODM is able to regress tougher objects (*e.g.*, extra-small, extra-large, extra-wide and extra-tall). However, as shown in Figure 5, there is still much room to improve the performance of RefineDet for small objects, especially for the chairs and tables. Increasing the input size (*e.g.*, from 320×320 to 512×512) can improve the performance for small objects , but it is only a temporary solution. Large input will be a burden on running speed in inference. Therefore, detecting small objects is still a challenge task and needs to be further studied.

[1] http://web.engr.illinois.edu/~dhoiem/projects/detectionAnalysis/

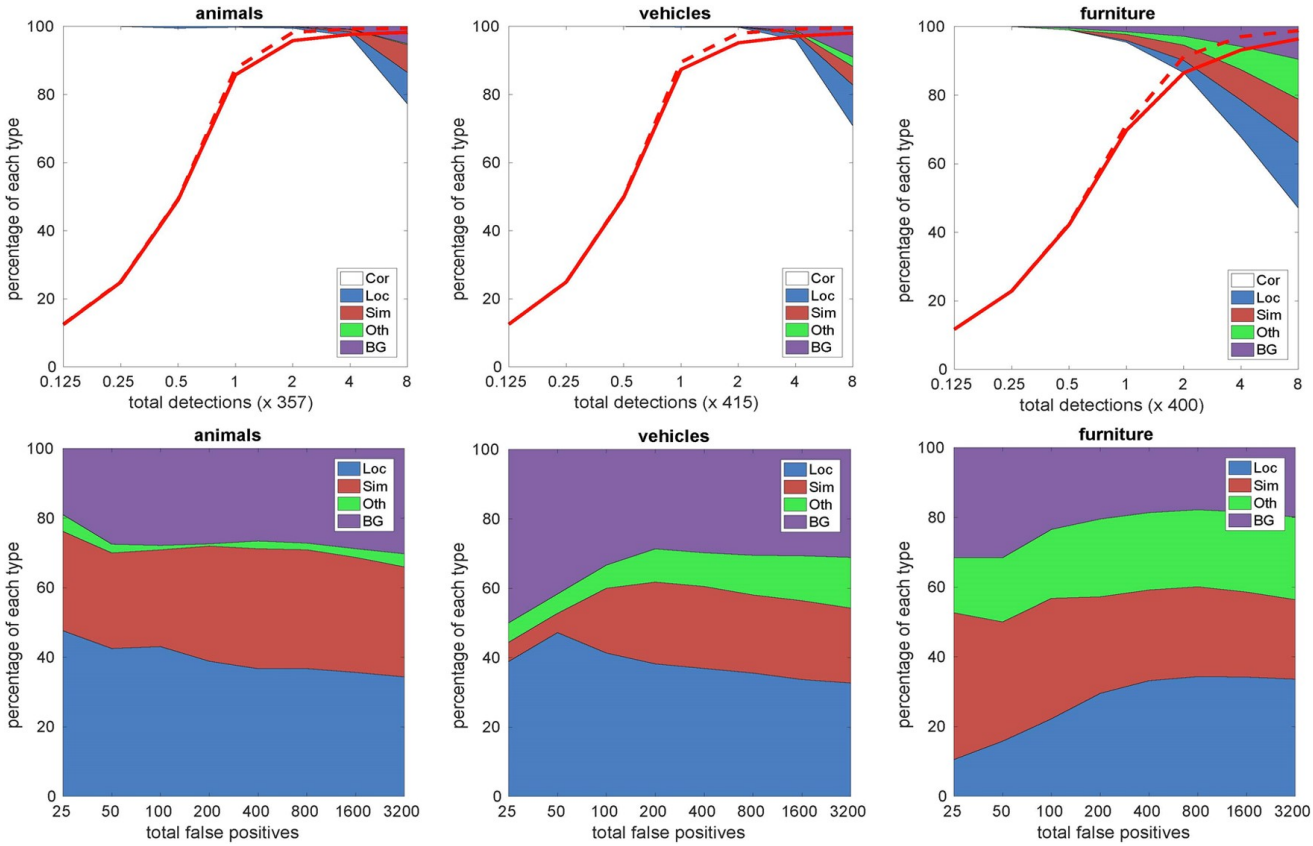

Figure 4: Visualization of the performance of RefineDet512 on animals, vehicles and furniture classes in the VOC 2007 `test` set. The top row shows the cumulative fraction of detections that are correct (Cor) or false positive due to poor localization (Loc), confusion with similar categories (Sim), with others (Oth), or with background (BG). The solid red line reflects the change of recall with strong criteria (0.5 jaccard overlap) as the number of detections increases. The dashed red line is using the "weak" criteria (0.1 jaccard overlap). The bottom row shows the distribution of the top-ranked false positive types.

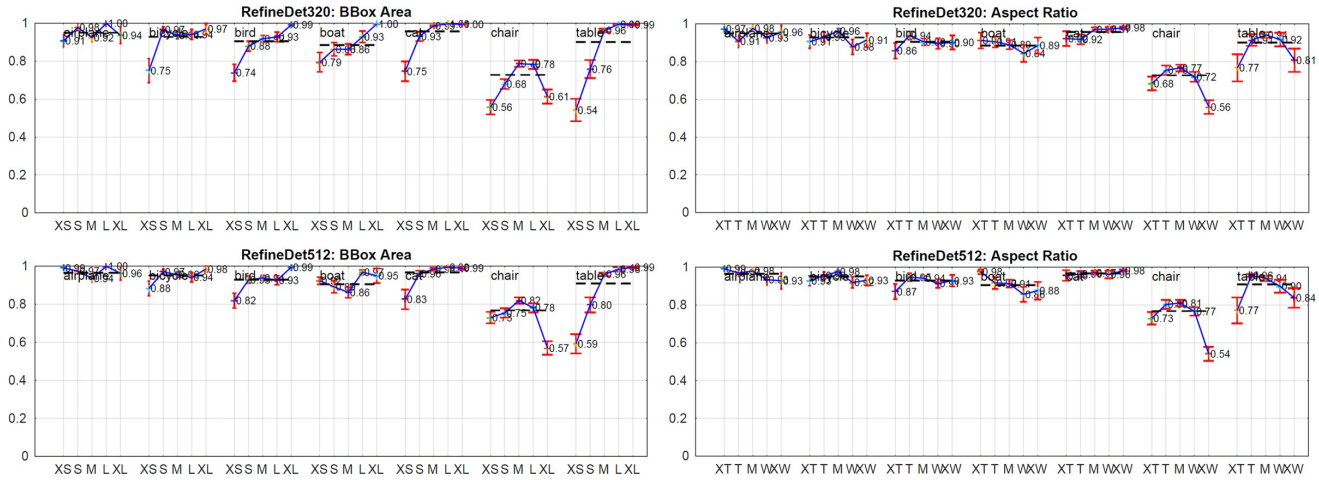

Figure 5: Sensitivity and impact of different object characteristics on the VOC 2007 `test` set. The plot on the left shows the effects of BBox Area per category, and the right plot shows the effect of Aspect Ratio. Key: BBox Area: XS=extra-small; S=small; M=medium; L=large; XL =extra-large. Aspect Ratio: XT=extra-tall/narrow; T=tall; M=medium; W=wide; XW =extra-wide.

4. Recall-IoU Diagram for the Two-Step Regression

We plot the recall *vs.* IoU curves of the initial anchors, anchors after applying the ARM, and anchors after applying the ODM on the VOC07 `test` set in Figure 6. The scores in the legend are the average recall (AR) over the IoU thresholds in $[0.5, 0.95]$. We find that the AR of the initial anchors is only 32.0%. After using the ARM, the AR is improved to 78.0%. Meanwhile, after applying the ODM, the AR is further improved to 82.1%. The results demonstrate the effectiveness of the two-step regression in RefineDet.

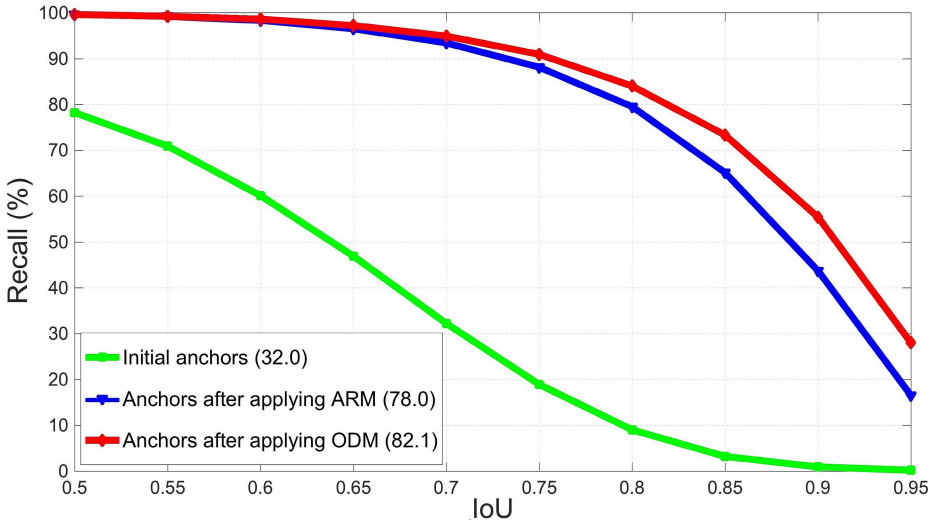

Figure 6: The recall *vs.* IoU curves of the initial anchors, anchors after applying the ARM, and anchors after applying the ODM on the VOC07 `test` set.