Learning Globally Optimized Object Detector via Policy Gradient
Supplementary Material

Appendix A: Modifications on Faster R-CNN

We conduct several ablation experiments to show the effectiveness of our modifications on original Faster R-CNN model [10]. Detailed results are presented in Table . These results demonstrate that our modifications are effective and our baseline model greatly outperforms original implementation (5.1 mAP).

Models	mAP
Faster R-CNN [10] (conv5 head [4])	31.2
+ dilated conv (2fc head)	32.5
+ conv-2fc head	33.8
+ ROIAlign	35.0
+ 15 anchors	35.8
+ allow tiny proposals (2 pixels threshold)	36.3

Table 1. Results of ablation experiments on our baseline model, evaluated on the COCO `minival` set. We report the results of COCO-style mAP. All models are trained on `trainval35k`. All results are based on ResNet-101 backbone CNN pre-trained on ImageNet1k dataset and share the same hyper-parameters.

Appendix B: Results on PASCAL VOC

We also evaluate our method on object detection task of PASCAL VOC dataset [2]. Different from COCO, there are 20 object categories in PASCAL VOC dataset. We report the standard evaluation metric mAP (mAP@0.5) over all categories, following [10, 4, 7, 8]. Our baseline model is implemented following the details for COCO described in experiment section except ROIAlign and 15 anchors. Our model is trained on the train and validation subsets of PASCAL VOC 2007 and PASCAL VOC 2012, and is evaluated on the test subset of PASCAL VOC 2007. Results are presented in Table . Our baseline model is better than original model by 3.6 mAP. We can see that our proposed model can consistently improve baseline model on different datasets.

Models	mAP
YOLO [8]	66.4
Fast R-CNN [3]	70.0
Faster R-CNN [10]	73.2
HyperNet [6]	76.3
SSD [7]	76.8
RON [5]	77.6
YOLOv2 [9]	78.6
OHEM+multi-scale+multi-stage [11]	78.9
R-FCN [1]	79.5
R-FCN+multi-scale [1]	80.5
Faster R-CNN (ResNet-101 [4])	76.4
Faster R-CNN (our baseline)	80.01
Faster R-CNN (globally optimized)	81.28

Table 2. Results of object detection, evaluated on the PASCAL VOC 2007 test set. We report the results of mAP@0.5. Our models are trained on train and validation subsets of PASCAL VOC 2007 and PASCAL VOC 2012. Our results are based on ResNet-101 backbone CNN pre-trained on ImageNet1k dataset.

Appendix C: Final Gradient Expression

Using the chain rule, we have:

$$\nabla L_I(\theta, b) = \sum_x \frac{\partial L_I(\theta, b)}{\partial x} \frac{\partial x}{\partial \theta}, \tag{1}$$

where $\frac{\partial L_I(\theta,b)}{\partial x}$ can be computed according to equation 14 as:

$$\frac{\partial L_I(\theta, b)}{\partial x} \approx (c(r(b) - r') + \gamma)(p_b - 1). \tag{2}$$

Appendix D: Visual Results

More examples of our proposed model are presented in Figure 1.

References

[1] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016. 1

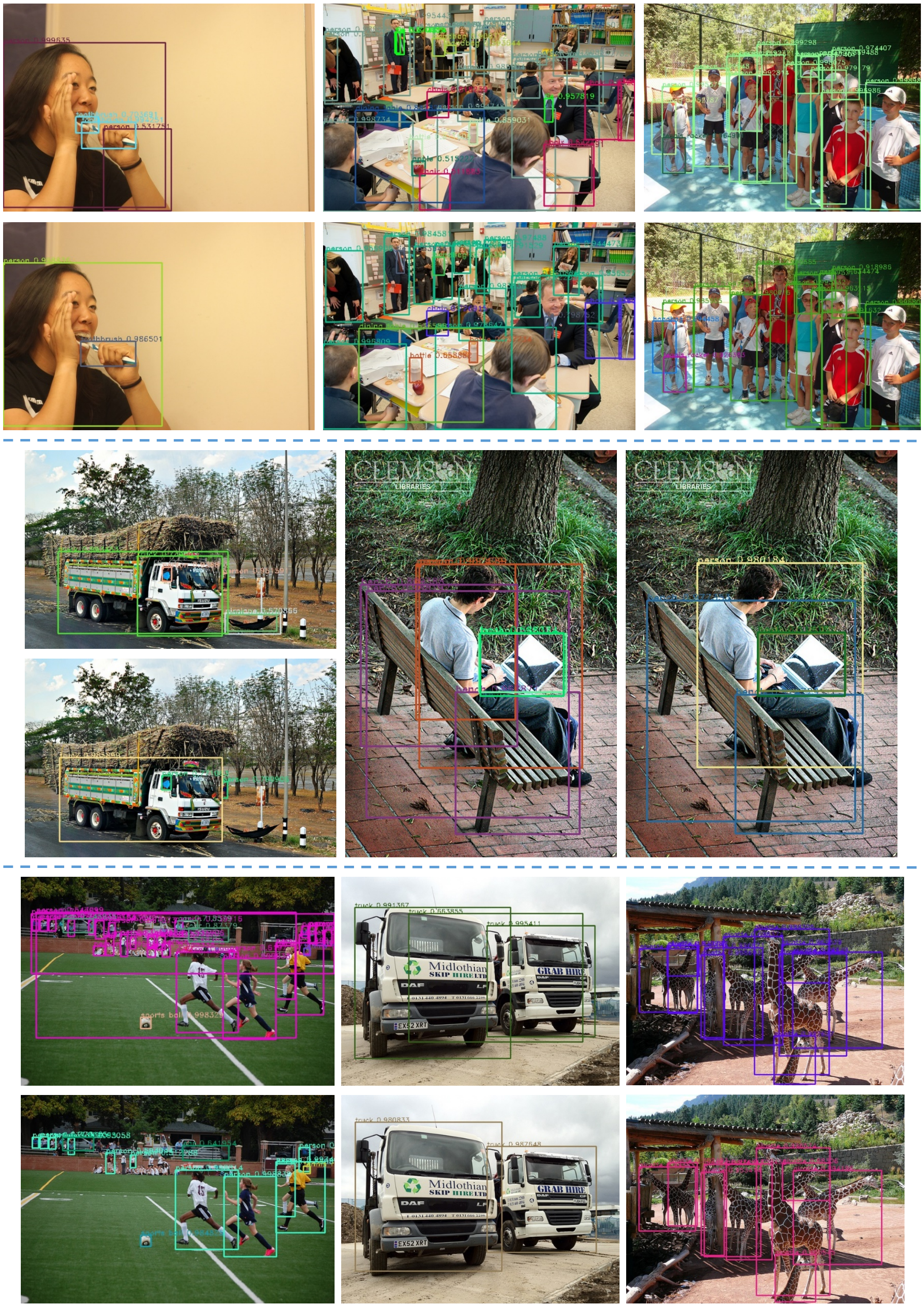

Figure 1. Baseline model (top, left) vs. globally optimized object detector (bottom, right, Faster R-CNN model). We only keep the boxes with the confident scores higher than 0.5. We can see that our proposed model produces more precise results compared to the baseline model. *Wrong detections with high confident scores can be barely found* in results of our model.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1

[3] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 1

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[5] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen. Ron: Reverse connection with objectness prior networks for object detection. *arXiv preprint arXiv:1707.01691*, 2017. 1

[6] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *CVPR*, pages 845–853, 2016. 1

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 1

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1

[9] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016. 1

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1

[11] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016. 1