Supplementary Material for "Multiple Instance Detection Network with Online Instance Classifier Refinement"

The document contains the supplementary materials for "Multiple Instance Detection Network with Online Instance Classifier Refinement". In this document we provide detailed per-class results on VOC 2012, and some visualization comparisons among the WSDDN [1], the WSDDN+context [2], and our method (OICR).

1. Per-class results on VOC 2012

The detailed per-class results on VOC 2012 can be viewed in Table 1 and Table 2. Obviously our method outperforms the previous state-of-the-art [2] by a large margin. Similar to results on VOC 2007, our results are better than [2] for rigid objects such as "bicycle", "car", and "motorbike", but worse than [2] for non-rigid objects "cat", "dog", and "person". This is because non-rigid objects are always with great deformation. Our method tends to detect the most discriminative parts of these objects (like head). The [2] considers more context information thus can deal with these classes better.

2. Visualization comparisons

We show some visualization comparisons among the WSDDN [1], the WSDDN+context [2], and our method in Fig. 1. From this visualization and results from tables, we can observe that for classes such as aeroplane, bike, car, *etc*., our method tends to provide more accurate detections, whereas other two methods sometimes fails to produce boxes that are overlarge or only contain parts of objects (the first four rows in Fig. 1). But for some classes such as person, our method always fails to detect only parts of objects (the fifth row in Fig. 1). Because considering context information sometimes help the detection (as in WSDDN+context [2]), we believe our method can be further improved by incorporating context information into our framework.

All these three methods (actually almost all weakly supervised object detection methods) suffers from two problems: producing boxes that not only contain the target object but also include their adjacent similar objects, or only detecting parts of object for objects with deformation (the last row in Fig. 1). An ideal solution for these problems is yet wanted.

References

[1] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 1, 2

[2] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, pages 350–365, 2016. 1, 2

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN+context [2]	64.0	54.9	36.4	8.1	12.6	53.1	40.5	**28.4**	6.6	35.3	**34.4**	**49.1**	42.6	62.4	**19.8**	15.2	27.0	33.1	33.0	50.0	35.3
OICR-VGG_M[†]	64.4	50.6	34.8	16.7	16.5	49.7	44.8	20.4	5.0	39.0	18.2	46.2	50.3	64.3	3.4	15.1	32.4	38.5	36.3	45.1	34.6
OICR-VGG16[‡]	67.7	61.2	41.5	25.6	22.2	54.6	49.7	25.4	**19.9**	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
OICR-Ens.[§]	68.4	58.6	39.9	23.4	21.3	52.8	48.7	23.5	13.5	44.8	22.0	36.5	47.6	68.3	2.6	21.7	39.1	**39.6**	**38.2**	52.7	38.2
OICR-Ens.+FRCNN[♮]	**71.4**	**69.4**	**55.1**	**29.8**	**28.1**	**55.0**	**57.9**	24.4	17.2	**59.1**	21.8	26.6	**57.8**	**71.3**	1.0	**23.1**	**52.7**	37.5	33.5	**56.6**	**42.5**

Table 1. Average precision (in %) for different methods on VOC 2012 test set. [†]http://host.robots.ox.ac.uk:8080/anonymous/PGSNG5.html [‡]http://host.robots.ox.ac.uk:8080/anonymous/6ASJ4I.html [§]http://host.robots.ox.ac.uk:8080/anonymous/UEZKOR.html [♮]http://host.robots.ox.ac.uk:8080/anonymous/XP6BJ7.html

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
WSDDN+context [2]	78.3	70.8	52.5	34.7	36.6	80.0	58.7	**38.6**	27.7	71.2	32.3	48.7	76.2	77.4	**16.0**	48.4	69.9	47.5	**66.9**	62.9	54.8
OICR-VGG_M	84.8	79.7	66.0	44.4	40.2	75.6	72.2	32.3	20.0	82.5	46.2	**64.1**	**84.3**	87.5	12.3	48.6	75.8	**63.9**	62.5	70.3	60.7
OICR-VGG16	86.2	84.2	68.7	55.4	46.5	82.8	74.9	32.2	46.7	82.8	42.9	41.0	68.1	89.6	9.2	53.9	81.0	52.9	59.5	83.2	62.1
OICR-Ens.	86.7	85.1	69.5	57.2	47.4	81.2	76.2	32.0	34.0	84.8	**47.9**	50.4	82.0	88.6	10.4	55.5	77.9	62.6	60.2	80.8	63.5
OICR-Ens.+FRCNN	**89.3**	**86.3**	**75.2**	**57.9**	**53.5**	**84.0**	**79.5**	35.2	**47.2**	**87.4**	43.4	43.8	77.0	**91.0**	10.4	**60.7**	**86.8**	55.7	62.0	**84.7**	**65.6**

Table 2. CorLoc (in %) for different methods on VOC 2012 trainval set.

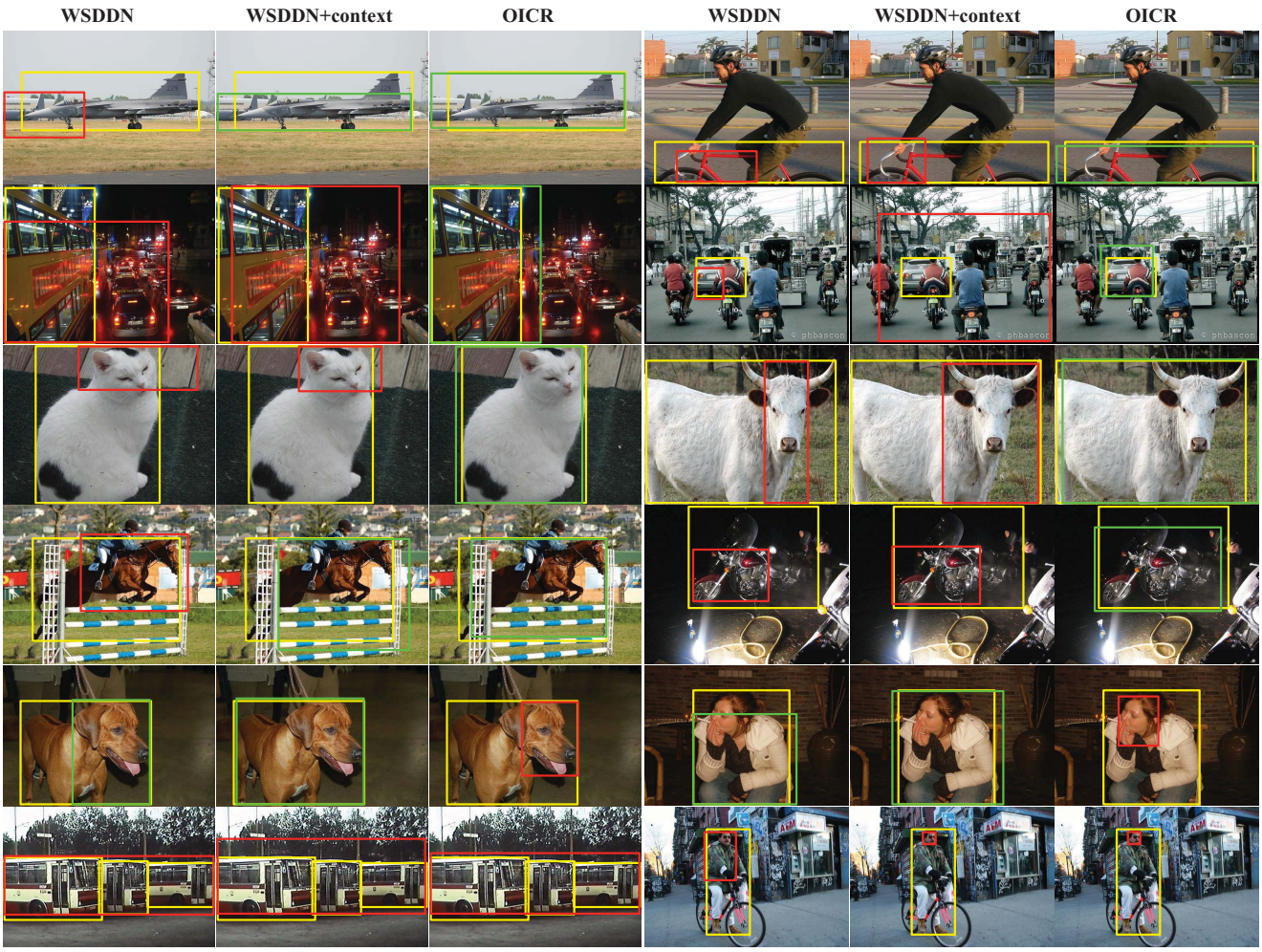

Figure 1. Some visualization comparisons among the WSDDN [1], the WSDDN+context [2], and our method (OICR). Green rectangle indicates success cases (IoU>0.5), red rectangle indicates failure cases (IoU<0.5), and yellow rectangle indicates ground truths. The first four rows show examples that our method outperforms other two methods (with larger IoU). The fifth row shows examples that our method is worse than other two methods (with smaller IoU). The last row shows failure examples for both three methods.