Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning *(Supplementary Materials)*

Weifeng Ge Sibei Yang Yizhou Yu
Department of Computer Science, The University of Hong Kong

1. Multi-Label Image Classification

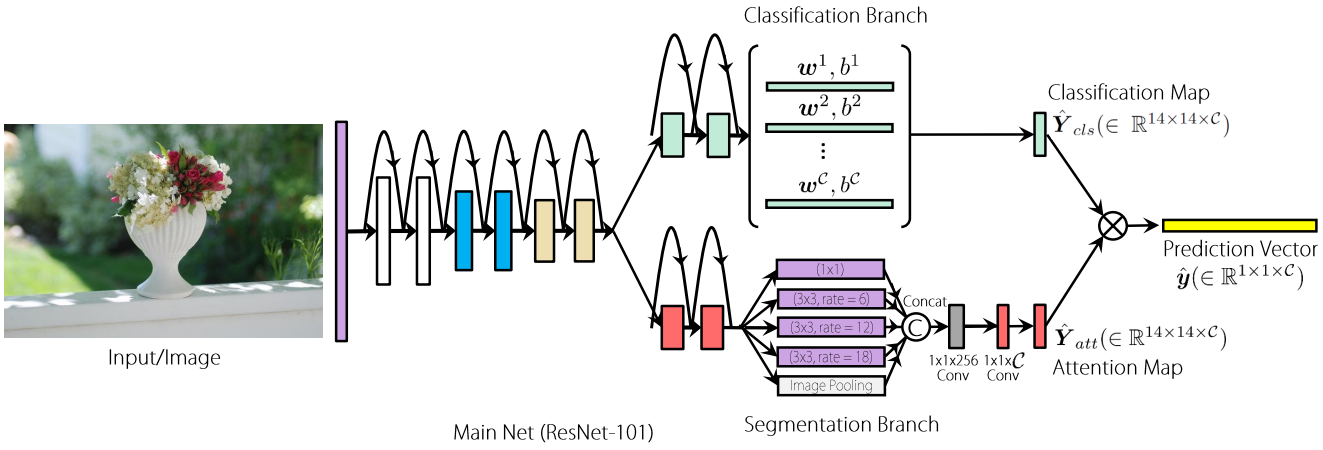

Figure 1. The architecture of our multi-label classification network.

Fig. 1 shows the architecture of our multi-label classification network. Layers before $res4b22\_relu$ of ResNet-101 are shared by the following branches. Both the segmentation and attention branches have the same structure of the $res5$ part of ResNet-101. In the classification branch, the output $\boldsymbol{X}(\in \mathbb{R}^{14\times14\times2048})$ of layer $res5c$ is a $14 \times 14 \times 2048$ tensor. The classification map $\hat{\boldsymbol{Y}}_{cls}(\in \mathbb{R}^{14\times14\times\mathcal{C}})$ is obtained by feeding $\boldsymbol{X}$ directly into a $2048 \times 1 \times 1 \times \mathcal{C}$ convolutional layer. In the segmentation branch, the output of layer $res5c$ is fed into an atrous spatial pyramid pooling layer, and then a $1280 \times 1 \times 1 \times \mathcal{C}$ convolutional layer and a softmax layer to obtain the segmentation map $\hat{\boldsymbol{Y}}_{seg}(\in \mathbb{R}^{14\times14\times\mathcal{C}})$. Normalize the summation of each channel in $\hat{\boldsymbol{Y}}_{seg}$ to obtain the attention map $\hat{\boldsymbol{Y}}_{att}$. In our atrous spatial pyramid pooling layer, we have four dilated convolutional layers and one global convolutional layer. The dilations of the four dilated convolutional layers are [1, 2, 4, 6]. All these convolutional layers have 256 channels.

2. Experimental Results

2.1. Semantic Segmentation

Result comparison. We compare our method with existing state-of-the-art algorithms. Table 1 lists the results of weakly supervised semantic segmentation on the Pascal VOC 2012 validation set. The proposed method achieves 58% mean IoU, and outperforms all existing algorithms by at least 4.9%.

2.2. Object Detection

Result comparison. We compare object detection results from our method with those from OICR-FRCNN (our own implementation) on the Microsoft COCO validation set. Our method achieves 19.3% mAP@.5 and 8.9% mAP@[.5, 0.95], which are 1.9% and 1.2% higher than those achieved by OICR-FRCNN.

method	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
DSCM[9]	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	**38.2**	55.4	32.2	42.6	39.6	44.1
F-B[8]	79.2	60.1	20.4	50.7	41.2	46.3	62.6	49.2	62.3	13.3	49.7	**38.1**	58.4	49.0	57.0	48.2	27.8	55.1	29.6	54.6	26.6	46.6
SEC[4]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.2	62.6	32.1	45.4	45.3	50.7
FCL[7]	85.8	65.2	**29.4**	63.8	31.2	37.2	69.6	64.3	76.2	21.4	56.3	29.8	68.2	60.6	66.2	55.8	30.8	66.1	34.9	48.8	**47.1**	52.8
T-P[3]	82.8	62.2	23.1	65.8	21.1	43.1	71.1	**66.2**	76.1	21.3	59.6	35.1	70.2	58.8	62.3	**66.1**	35.8	69.9	33.4	45.9	45.6	53.1
Ours+CRF	**85.8**	**72.5**	29.1	**66.0**	**55.7**	**49.6**	**73.1**	61.4	**77.5**	**26.6**	**68.5**	31.8	**73.6**	**71.5**	**68.8**	53.1	31.8	**79.8**	**35.7**	**64.9**	41.3	**58.0**

Table 1. Comparison among weakly supervised semantic segmentation methods on the PASCAL VOC 2012 *segmentation val* set.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
OM+MIL+FRCNN[6]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	**21.8**	57.1	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
HCP+DSD+OSSH3[2]	54.2	52.0	35.2	25.9	15.0	59.6	**67.9**	58.7	10.1	**67.4**	27.3	37.8	54.8	**67.3**	5.1	19.7	**52.6**	43.5	56.9	**62.5**	43.7
OICR-Ens+FRCNN[10]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
Ours+FRCNN w/o clustering	66.7	61.8	55.3	41.8	6.7	61.2	62.5	**72.8**	12.7	46.2	40.9	**71.0**	67.3	64.7	30.9	16.7	42.6	56.0	65.0	26.5	48.5
Ours+FRCNN w/o uncertainty	66.8	63.4	54.5	42.2	5.8	60.5	58.3	67.8	7.8	46.1	40.3	71.0	68.2	62.6	30.7	16.5	41.1	55.2	66.8	25.2	47.5
Ours+FRCNN w/o instances	67.7	62.9	53.1	**44.4**	11.2	62.4	58.5	71.2	8.3	45.7	41.5	71.0	68.0	59.2	30.3	15.0	42.4	56.0	**67.2**	26.8	48.1
Ours+FRCNN w/o filtering	**69.0**	67.1	53.8	39.3	13.1	61.4	64.3	72.5	15.3	48.0	42.4	67.2	68.0	65.5	32.4	17.1	42.2	55.6	67.0	23.8	49.3
Ours+FRCNN w/o heatmap	65.9	65.9	**57.6**	40.3	7.6	61.7	62.7	73.4	11.9	49.2	44.3	68.6	**70.8**	64.0	**33.6**	15.2	42.3	54.5	66.1	23.4	49.0
Ours+FRCNN	64.3	**68.0**	56.2	36.4	**23.1**	**68.5**	67.2	64.9	7.1	54.1	**47.0**	57.0	69.3	65.4	20.8	23.2	50.7	**59.6**	65.2	57.0	**51.2**

Table 2. Average precision (in %) of weakly supervised methods on the PASCAL VOC 2007 *detection test* set.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
OICR-VGG16[10]	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
WSDDN+context[1]	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
HCP+DSD+OSSH3+NR[2]	60.8	54.2	34.1	14.9	13.1	54.3	53.4	58.6	3.7	53.1	8.3	43.4	49.8	69.2	4.1	17.5	43.8	25.6	**55.0**	50.1	38.3
OICR-Ens+FRCNN[10]	**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
Ours+FRCNN	71.0	66.9	**55.9**	**33.8**	24.0	**57.6**	**58.0**	**61.4**	**22.5**	58.4	19.2	**58.7**	**61.9**	**75.0**	11.2	**23.9**	50.3	**44.9**	41.3	54.3	**47.5**

Table 3. Average precision (in %) of weakly supervised methods on the PASCAL VOC 2012 *detection test* set.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mCorLoc
OICR-VGG16[10]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	**40.6**	79.5	35.7	33.7	60.5	88.8	21.8	57.9	**76.3**	59.9	75.3	**81.4**	60.6
WSDDN-Ens[1]	68.9	68.7	65.2	42.5	40.6	72.6	75.2	53.7	29.7	68.1	33.5	45.6	65.9	86.1	27.5	44.9	76.0	62.4	66.3	66.8	58.0
OM+MIL+FRCNN[6]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
HCP+DSD+OSSH3[2]	72.2	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	54.3	64.3	88.2	22.2	53.7	72.2	52.6	68.9	74.4	54.9
OICR-Ens+FRCNN[10]	85.8	**82.7**	62.8	45.2	**43.5**	**84.8**	**87.0**	46.8	15.7	**82.2**	**51.0**	45.6	83.7	**91.2**	22.2	**59.7**	75.3	65.1	76.8	78.1	64.3
Ours+FRCNN	**88.3**	77.6	**74.8**	**63.3**	37.8	78.2	83.6	**72.7**	19.4	79.5	46.4	**78.1**	**84.7**	90.4	**28.6**	43.6	76.3	**68.3**	**77.9**	70.6	**67.0**

Table 4. CorLoc (in %) of weakly supervised methods on the PASCAL VOC 2007 *detection trainval* set.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mCorLoc
OICR-VGG16[10]	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
WSDDN+context[1]	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
HCP+DSD+OSSH3+NR[2]	82.4	68.1	54.5	38.9	35.9	84.7	73.1	64.8	17.1	78.3	22.5	57.0	70.8	86.6	18.7	49.7	80.7	45.3	**70.1**	77.3	58.8
OICR-Ens+FRCNN[10]	**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
Ours+FRCNN	88.0	81.6	**75.8**	**60.9**	46.2	**85.3**	75.3	**76.5**	47.2	85.4	**47.7**	**74.3**	**87.8**	**91.4**	**21.6**	55.3	77.9	**68.8**	64.9	75.0	**69.4**

Table 5. CorLoc (in %) of weakly supervised methods on the PASCAL VOC 2012 *detection trainval* set.

method	mAP@.5	mAP@[.5, 0.95]
OICR-Ens+FRCNN[10](impl. in this paper)	17.4	7.7
Ours+FRCNN	**19.3**	**8.9**

Table 6. Average precision (in %) of weakly supervised methods on the Microsfot COCO 2014 *detection validation* set.

2.3. Ablation Study

We perform an ablation study on Pascal VOC 2007 detection test set by replacing or removing a single component in our pipeline every time. First, to verify the importance of object instances, we remove all steps related to object instances, including the entire instance level stage and the operations related to the instance attention map in the pixel level stage. The mAP is decreased by 3.1% as shown in Table 2. Second, the clustering and outlier detection step in the instance level stage is removed. We directly train an instance classifier using the object proposals from the image level stage. The mAP is decreased

by 2.7%. Third, instead of leaving the labels of a subset of pixels uncertain in the pixel level stage, we assign a unique label to every pixel even in the case of low confidence. The mAP drops to 47.5%, 3.7% lower than the performance of the original pipeline. Forth, when removing the clustering and outlier detection step in the instance level stage, we use the original image classifier without retraining the instance classifier to generate the attention map. The mAP is 49.3% which is 1.9% lower than the original pipeline. At last, we remove the object heatmap in the pixel level stage, the mAP becomes 49.0%, which drops by 2.1% compared to the original pipeline.

References

[1] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool. Weakly supervised cascaded convolutional networks. *arXiv preprint arXiv:1611.08258*, 2016. 2

[2] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[3] D. Kim, D. Cho, D. Yoo, and I. So Kweon. Two-phase learning for weakly supervised object localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[4] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016. 2

[5] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 4, 5, 6, 7

[6] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016. 2

[7] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3529–3538, 2017. 2

[8] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *European Conference on Computer Vision*, pages 413–432. Springer, 2016. 2

[9] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, pages 218–234. Springer, 2016. 2

[10] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

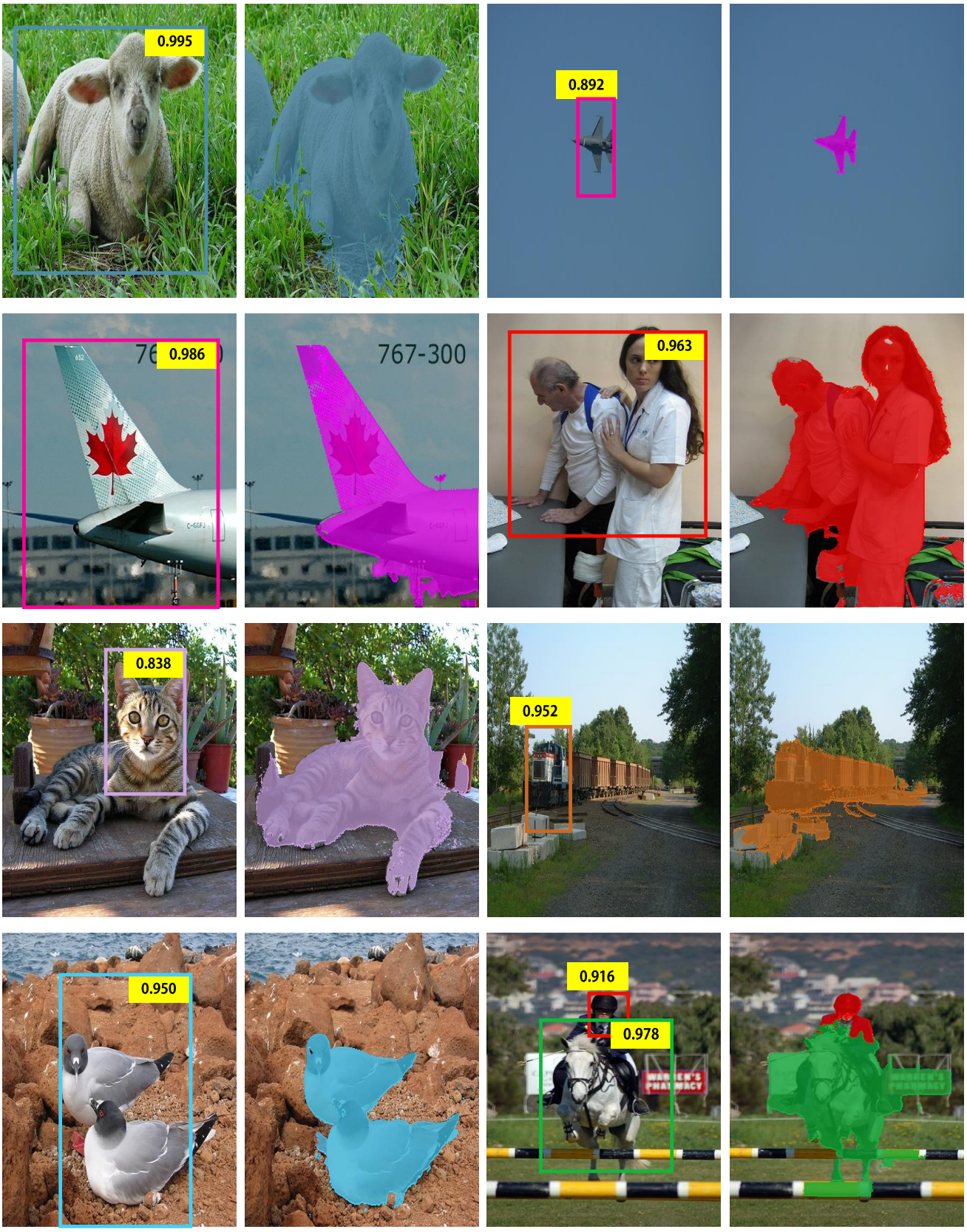

Figure 2. Detection and semantic segmentation results on the Pascal VOC 2007 $test$ set. The detection results are obtained by choosing proposals with the highest confidence within every class. The semantic segmentation results are post-processed by a CRF [5].

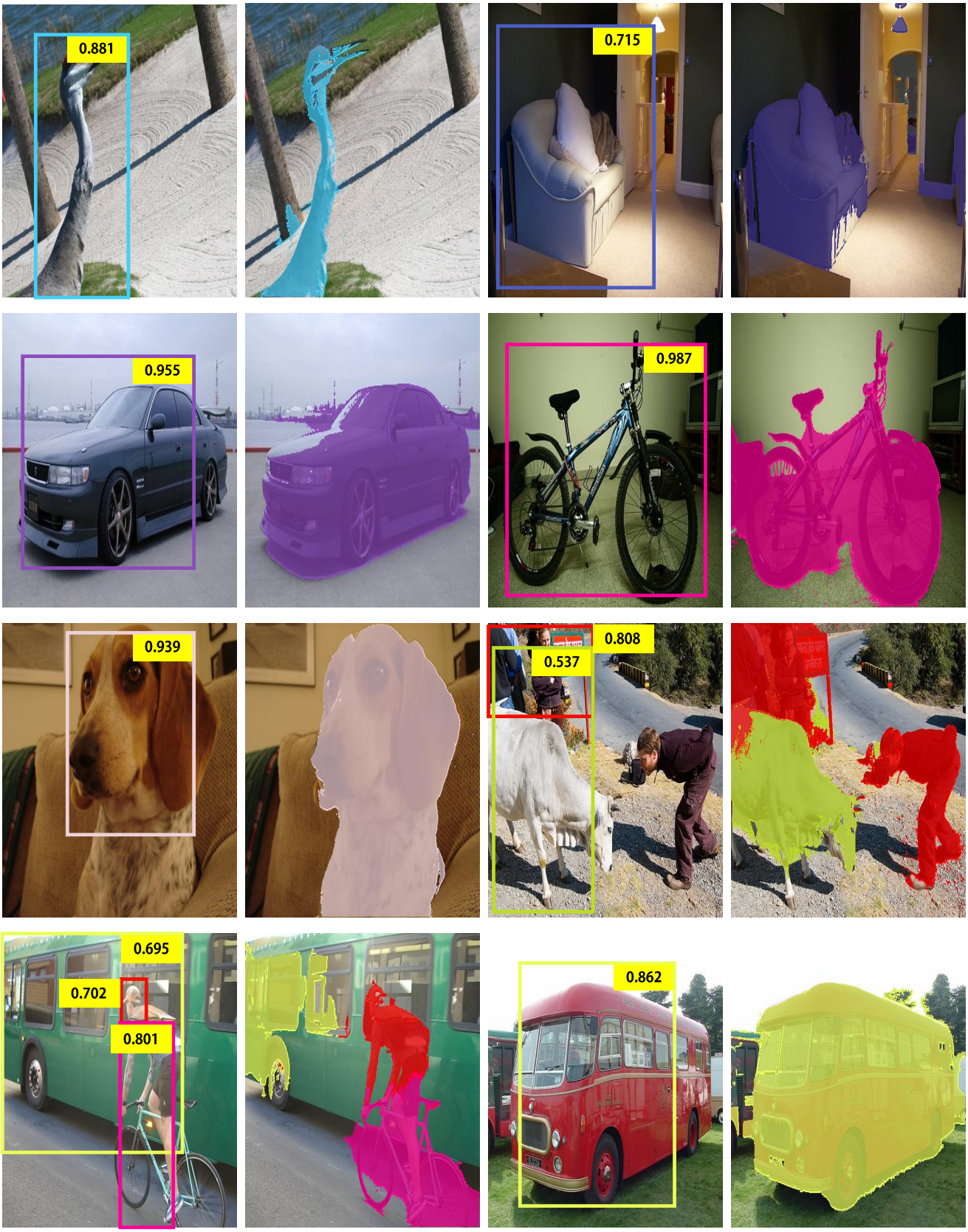

Figure 3. Detection and semantic segmentation results on the Pascal VOC 2007 *test* set. The detection results are obtained by choosing proposals with the highest confidence within every class. The semantic segmentation results are post-processed by a CRF [5].

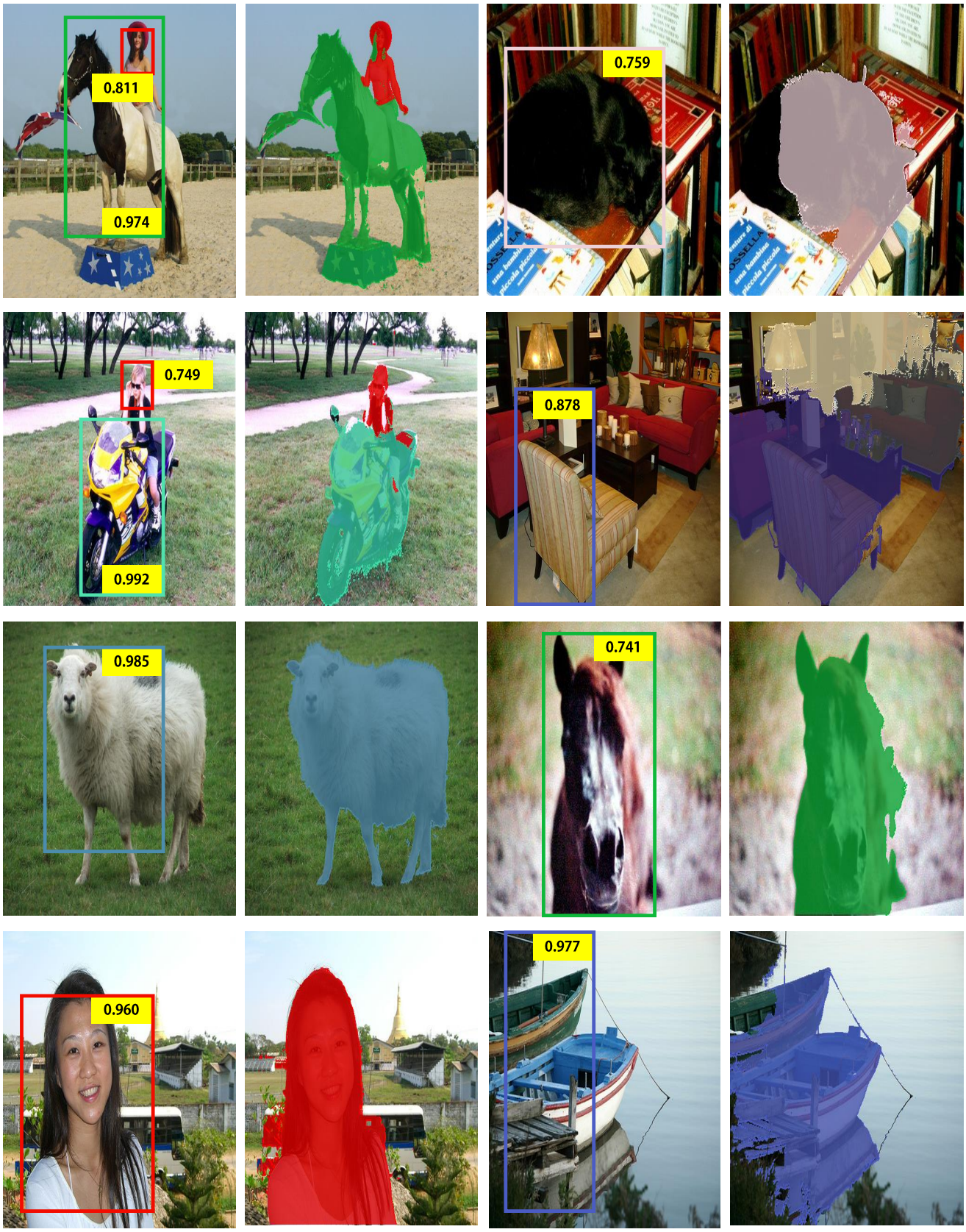

Figure 4. Detection and semantic segmentation results on the Pascal VOC 2012 *test* set. The detection results are obtained by choosing proposals with the highest confidence within every class. The semantic segmentation results are post-processed by a CRF [5].

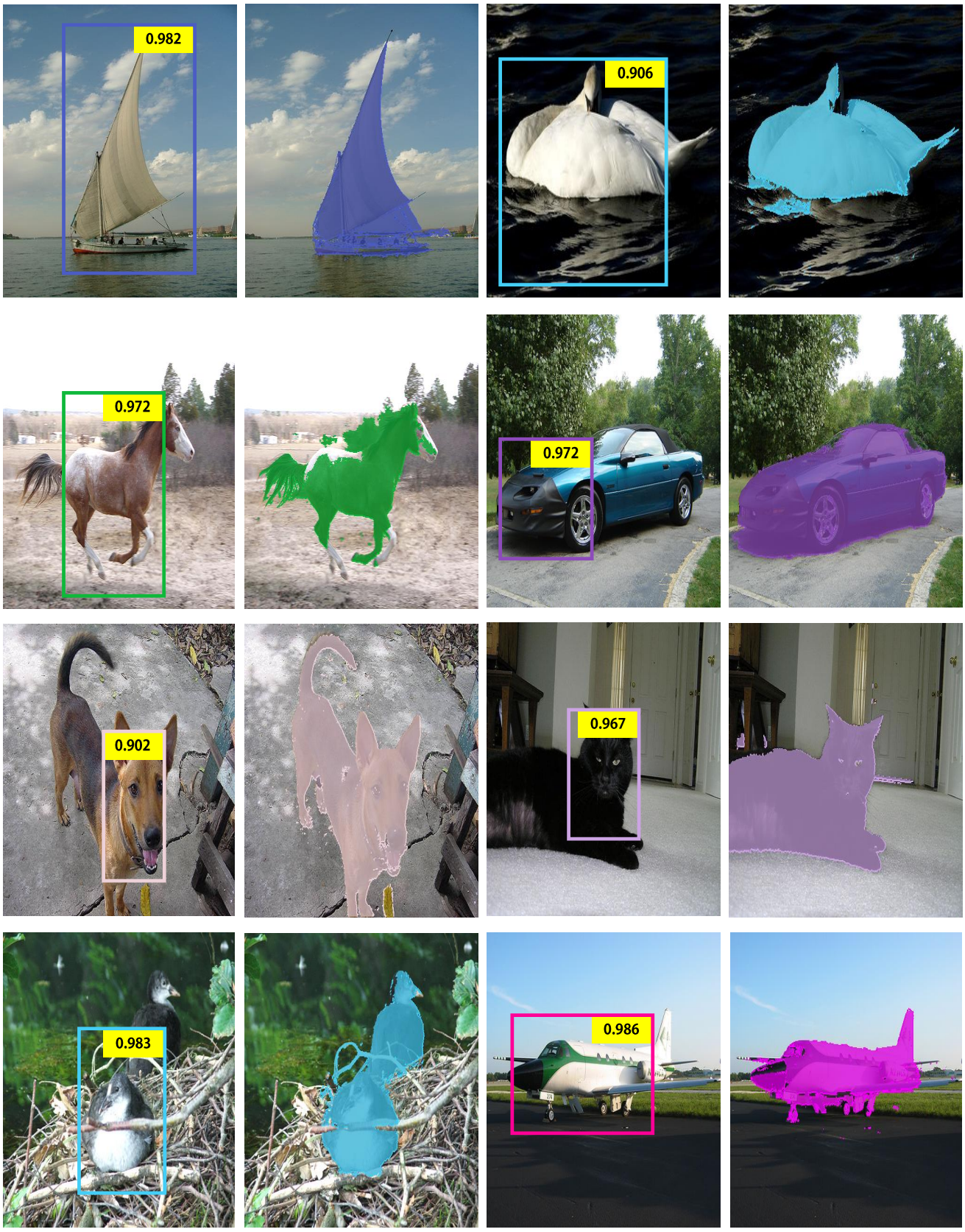

Figure 5. Detection and semantic segmentation results on the Pascal VOC 2012 *test* set. The detection results are obtained by choosing proposals with the highest confidence within every class. The semantic segmentation results are post-processed by a CRF [5].