[Supplementary Material]
Weakly Supervised Region Proposal Network and Object Detection

Peng Tang[1], Xinggang Wang[1], Angtian Wang[1], Yongluan Yan[1],
Wenyu Liu[1 (✉)], Junzhou Huang[2,3], Alan Yuille[4]

[1] School of EIC, Huazhong University of Science and Technology, Wuhan, China
{pengtang,xgwang,angtianwang,yongluanyan,liuwy}@hust.edu.cn
[2] Tencent AI lab, Shenzhen, China
[3] Department of CSE, University of Texas at Arlington, Arlington, USA
jzhuang75@gmail.com
[4] Department of Computer Science, The Johns Hopkins University, Baltimore, USA
alan.l.yuille@gmail.com

In the supplementary material, we provide our improvement of the state-of-the-art weakly supervised object detection network [1] in Section 1, the per-class results on the PASCAL VOC 2012 dataset [2] in Section 2, some qualitative results in Section 3, more ablation experiments on the PASCAL VOC 2007 dataset in Section 4, and runtime analyses in Section 5.

1 The Improvement of the Weakly Supervised Object Detection Network

The network in [1] has one multiple instance detection stream and M instance classifier refinement streams, and each stream produces proposal classification probabilities. The multiple instance detection stream is supervised by image-level annotations, and the instance classifier refinement streams are supervised by bounding box annotations generated by the network. In particular, for the m-th stream, the supervisions are generated by the outputs of the $\{m-1\}$-th stream (we denote the multiple instance detection stream as the 0-th stream). Given an image $\mathbf{I}$ and its image-level annotation $\mathbf{y} = [y_1, ..., y_K] \in \mathbb{R}^K$, N region proposals $\mathcal{P} = \{p_n\}_{n=1}^N$ are generated, where K is the number of object classes, $y_k = 1/0$ indicates that the image is with/without the k-th object class, and p_n is the n-th proposal. Suppose φ_{kn}^m is the classification probability of the n-th proposal for the k-th object class from the m-th stream, to generate the supervisions for the m-th stream, [1] first selects the top-scoring proposal with index $n_k^m = \mathrm{argmax}_n \, \varphi_{kn}^{m-1}$ if the image has the k-th object class (*i.e.* $y_k = 1$), and then assigns labels to proposals according to the Intersection-over-Union (IoU). The assigning label procedure is performed as follows: if the IoU between p_n and $p_{n_k^m}$ is larger than a threshold I_t, p_n is assigned the k-th object class label; if the IoU between p_n and $p_{n_k^m}$ is not larger than I_t, p_n is assigned the background label. This method is motivated by that the top-scoring proposal can cover at least parts of an object and its adjacent proposals may contain

Table 1. Result comparison (AP and mAP in %) for different methods on the PASCAL VOC 2012 `test` set. Our method obtain the best results. See Section 2 for definitions of the Ours-based methods. Results are available at: $^\flat$`http://host.robots.ox.ac.uk:8080/anonymous/DZEIYA.html` $^\natural$`http://host.robots.ox.ac.uk:8080/anonymous/2ZMSDY.html` $^\sharp$`http://host.robots.ox.ac.uk:8080/anonymous/Z6IGOX.html`. Our method obtains the best mAP

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN+context [4]	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
WCCN [5]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.9
HCP+DSD+OSSH3 [6]	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 [1]	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-VGG16$^\flat$	68.4	33.4	40.0	11.3	26.7	55.0	57.8	30.9	1.5	55.1	**43.6**	44.8	61.8	71.8	**41.8**	26.3	45.7	37.5	20.7	41.4	40.8
Ours-Ens.$^\natural$	69.8	57.8	48.6	21.6	**34.0**	54.8	58.8	28.9	10.2	55.7	31.6	42.4	62.2	70.9	24.1	**28.0**	43.9	42.0	27.5	54.4	43.4
Ours-Ens.+FRCNN$^\sharp$	**72.1**	68.7	51.4	22.1	30.0	**57.0**	**61.6**	39.0	9.1	58.7	27.5	**52.2**	**67.9**	**74.4**	29.7	25.4	52.5	**43.4**	19.1	51.7	**45.7**

Table 2. Result comparison (CorLoc in %) for different methods on the PASCAL VOC 2012 `trainval` set. Our method obtain the best results. See Section 2 for definitions of the Ours-based methods. Our method obtains the best mean of CorLoc

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
WSDDN+context [4]	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 [6]	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 [1]	**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-VGG16	85.5	60.8	62.5	36.6	53.8	82.1	80.1	48.2	14.9	87.7	**68.5**	60.7	85.7	89.2	**62.9**	62.1	87.1	54.0	45.1	70.6	64.9
Ours-Ens.	87.9	79.2	70.5	50.0	**59.8**	82.1	**82.0**	45.2	29.8	84.5	52.4	59.3	87.6	90.1	47.1	61.4	86.5	59.1	49.1	80.3	67.2
Ours-Ens.+FRCNN	88.5	85.3	73.4	53.5	59.4	**84.9**	81.4	51.6	29.7	**89.6**	52.0	**63.8**	**89.4**	**91.6**	49.8	**64.8**	**87.7**	**63.2**	47.5	79.7	**69.3**

larger portion of the object. Thus these proposals can be assigned the object class label. In [1], I_t is set to 0.5. More details can be found in [1].

However, there may exist more than one object with the same object label in the same image. Assigning all proposals that have low IoU with $p_{n_k^m}$ to the background label is not perfect, because these proposals may cover another object that also correspond to the k-th object class. To alleviate this problem, we simply ignore some proposals during training, *i.e.*, setting the loss for these proposals to 0 for the m-th instance classifier refinement stream. The ignored proposals are determined as follows: if the IoU between p_n and $p_{n_k^m}$ is lower than I'_t, p_n is ignored. Here we set I'_t to 0.1. Using this improved method, we observe that the performance is boosted: mAP from 41.2% to 42.2% and CorLoc from 60.6% to 60.9% on the PASCAL VOC 2007 dataset using the selective search proposal [3].

2 Per-class Results on the PASCAL VOC 2012 Dataset

The per-class results of different methods on the PASCAL VOC 2012 dataset are shown in Table 1 and Table 2. The definitions of Ours-VGG16, Ours-Ens., and Ours-Ens.+FRCNN in the tables are using our proposals, ensemble of models from our proposals and the selective search proposals, and training a Fast RCNN (FRCNN) [7] using the top-scoring proposals from Ours-VGG16-Ens.,

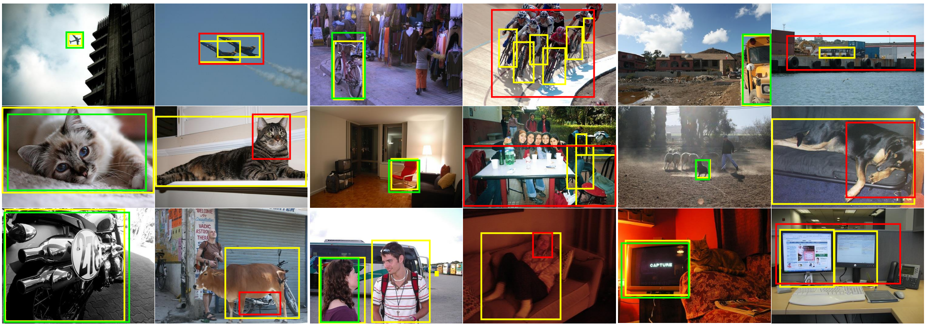

Fig. 1. Example detection results for “aeroplane”, “bike”, “bus”, “cat”, “chair”, “dog”, “motorbike”, “person”, and “monitor”. Yellow, green, and red rectangles indicate groundtruth boxes, success cases (IoU>0.5), and failure cases (IoU<0.5), respectively. Only the highest score detection for each image is shown

respectively. Our method obtains the best results on the PASCAL VOC 2007 and 2012 datasets.

3 Qualitative Results

We show some qualitative results of our method in Fig 1. We observe that our method is robust to the object size and viewpoint, and we can localize objects roughly for almost all cases. The main failures are as follows: for rigid objects like “aeroplane”, “bike”, “chair”, *etc.*, it falsely produces overlarge boxes which contain both the object and its neighboring objects, because the neighboring objects share similar appearances with the object; for non-rigid objects like “cat”, “dog”, and “person”, it falsely produces too small boxes which cover only parts of objects, because parts have less deformations than the objects. We believe that designing objectness score evaluation method for our response maps specifically, instead of using EB directly, to generate better proposals may be helpful.

4 More Ablation Experiments

We conduct more ablation experiments on the PASCAL VOC 2007 dataset to analyze different components of our method.

Using Edge Boxes as $\mathcal{P}^1$. Here we use 10K proposals from Edge Boxes (EB) [8] as $\mathcal{P}^1$ to train the proposal refinement network and the following weakly supervised object detection network. The results are 45.0% mAP and 63.4% CorLoc which are similar to our results (45.3% mAP and 63.8% CorLoc). This is because we generate numerous $\mathcal{P}^1$ to ensure high recall, so the recalls of our method and EB are almost the same. Although the results are similar, our method has its benefits: 1) All object location cues are from a single CNN, so we do not need to train separate edge detectors on other datasets with pixel-level edge labels. 2) The newly proposed 3-stage weakly supervised object detection framework is novel and outperforms all previous methods. 3) Our focus of generating proposals using a weakly supervised CNN has not been explored before. Our state-of-the-art performance shows this promising research direction. Our method can be applied to other deep networks and we can exploit other information in the response maps to search for better results.

Cascade of two OICR-VGG16 [1] model. Here we use 2K Selective Search (SS) [3] proposals as $\mathcal{P}^1$ and an OICR-VGG16 network for proposal refinement in our 3-stage framework. This outperforms the SS+OICR: mAP 43.3% *vs.* 42.2% and CorLoc 62.0% *vs.* 60.9%, confirming the effectiveness of our proposal refinement. These results are worse than ours because the final proposals are still SS which has lower proposal quality than ours, which confirms the effectiveness of our method.

5 Runtime Analyses

Here we give runtime analyses of our method. During testing, our method takes totally 2.3s for each image (0.8s for proposal generation and 1.5s for WSOD), and is faster than the previous best performed method [1] which takes 3.5s for each image (2s for proposal generation and 1.5s for WSOD). This is because we use a single network which is designed for speed: 1) The proposal generation and WSOD share their convolutional computations; 2) We use a small network for proposal refinement. Our proposal generation method is slower than Edge Boxes (0.25s for each image) because our method has an extra proposal refinement stage, but our WSOD results are much better. It is possible to speed up our method through using a smaller network for proposal refinement and compressing fully connected layers by truncated SVD as in [7].

References

1. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: CVPR. (2017) 2843–2851
2. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111**(1) (2015) 98–136
3. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV **104**(2) (2013) 154–171
4. Kantorov, V., Oquab, M., Cho, M., Laptev, I.: Contextlocnet: Context-aware deep network models for weakly supervised localization. In: ECCV. (2016) 350–365
5. Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: CVPR. (2017)
6. Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: CVPR. (2017) 1377–1385
7. Girshick, R.: Fast r-cnn. In: ICCV. (2015) 1440–1448
8. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV. (2014) 391–405