Rethinking Segmentation Guidance for Weakly Supervised Object Detection

Ke Yang[*1] Peng Zhang[*2] Peng Qiao[2] Zhiyuan Wang[1] Huadong Dai[1]
Tianlong Shen[1] Dongsheng Li[2] Yong Dou[2]

[1]Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, China
[2]National University of Defense Technology, China

Abstract

Weakly supervised object detection aims at learning object detectors with only image-level category labels. Most existing methods tend to solve this problem by using a multiple instance learning detector which is usually trapped to discriminate object parts, rather than the entire object. In order to select high-quality proposals, recent works leverage objectness scores derived from weakly-supervised segmentation maps to rank the object proposals. Base our observation, this kind of segmentation guided method always fails due to neglect of the fact that objectness of all proposals inside the ground-truth box should be consistent. In this paper, we propose a novel object representation named Objectness Consistent Representation (OCR) to meet the consistency criterion of objectness. Specifically, we project the segmentation confidence scores into two orthogonal directions, namely vertical and horizontal, to get the OCR. With the novel object representation, more high-quality proposals can be mined for learning a much stronger object detector. We obtain 54.6% and 51.1% mAP scores on VOC 2007 and 2012 datasets, significantly outperforming the state-of-the-arts and demonstrating the superiority of OCR for weakly supervised object detection.

1. Introduction

Fully supervised networks need plenty data which provide precise location and category annotations of the objects. However, precise object-level annotations are always expensive in human resource and huge data volume is required by training accurate object detection models. To alleviate this issue, Weakly Supervised Object Detection (WSOD) is a good alternative. WSOD uses only image-level category labels so that significant cost of preparing training data can be saved. Due to the lack of accurate annotations, this problem has not been well handled and the performance is still far from the fully supervised methods.

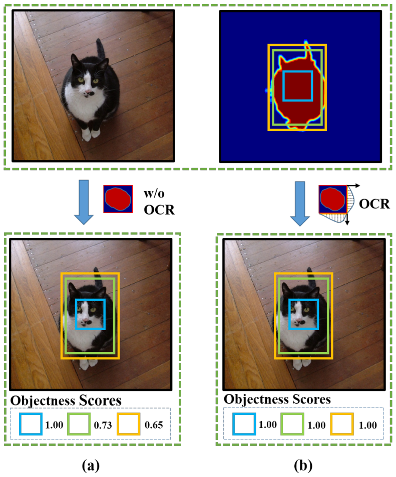

Figure 1. Motivation of our proposed method. (a) For existing segmentation to objectness scoring methods, as the box becomes larger, the classification score decreases quickly even though the box has not exceed the box border of the object, *e.g.* 1.00→0.73→0.65. (b) We project the segmentation score to two orthogonal directions to get our Objectness Consistent Representation (OCR). With the new representation, the score keeps consistent until the box exceeds the real object border, *e.g.* 1.00→1.00→1.00.

To localize objects with weak supervision information, one popular solution is to apply Multiple Instance Learning (MIL) for mining high-confidence region proposals [3, 7, 12] with positive image-level annotations. However, MIL usually discovers the most discriminative part of the target object (*e.g.* the head of a cat) rather than the entire object region. This inability of mining the complete object severely limits the performance of WSOD.

Recently, weakly supervised semantic segmentation methods [8, 15, 1, 2] have demonstrated very promising performance. Diba *et al.* [4] for the first time leveraged semantic segmentation to aid WSOD. They proposed a weak-

[*]Equal contributions. This work is supported by the National Key Research and Development Program of China under Grant No.2018YFB2101100 and the National Natural Science Foundation of China under Grants 91948303, 61902415 and 11801563.

ly cascaded convolutional network that leverages segmentation knowledge to filter noisy proposals with low objectness scores and achieves competitive detection results. Diba *et al*. selected proposals undering the **purity** criterion which means most pixels inside the box should have high confidence scores. High purity can only guarantee that the box is located around the target object, but is unable to filter high-response boxes of *object parts*.

In order to mine boxes of complete objects, in addition to the purity criterion, Wei *et al*. [16] propose a new criteria, *i.e*. **completeness**, to evaluate the objectness scores of object candidates. High completeness requires that very few pixels are with high confidence scores in the surrounding context of the target box.

We argue that their solutions [4, 16] are sub-optimal as they ignore the objectness calculation disparity between pixel-level object representation and box-level representation. They averaged the pixel-level segmentation confidence scores inside the box to estimate the objectness score for that box, which leads to sub-optimal performance. As shown in Figure 1(a), as the box becomes larger, the average confidence score of the box decreases quickly even though the box has not exceed the box border of the object, which is definitely harmful to the selection of high-quality tighter boxes.

We attribute the above problem to the inconsistency of objectness scoring. To solve the above problem, besides the two criteria above, we propose a new criterion, *i.e*. **consistency**, for objectness score calculation of proposals. *Consistency means the scores for each box should be consistent, as long as the box is within the box border of the object.* Considering the consistency criterion, we devise a novel object representation, named Objectness Consistent Representation (OCR), to help select high-quality candidate boxes from large amount object proposals. Specifically, we project the segmentation confidence scores to two orthogonal directions, *i.e*. horizontal and vertical, to get the OCR. With the new representation, the scores keep consistent as long as the boxes do not exceed the border of object, as shown in Figure 1(b). Our proposed OCR is generic and can be easily integrated into any WSOD network by constructing a weakly supervised semantic segmentation branch to produce category-specific segmentation confidence map. We apply object proposal selection with our OCR to the popular baselines of weakly supervised object detection, the experiment results on public datasets show that we significantly outperform the state-of-the-arts.

2. Method

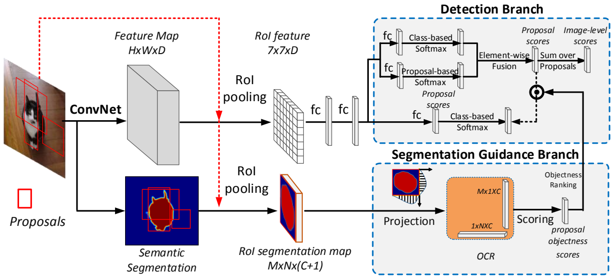

Figure 2. Architecture of our proposed network. (1) Generate semantic segmentation confidence maps using image-level labels. (2) Generate the new object representations from semantic segmentation confidence maps. (3) Calculate objectness scores of proposals and select the proposals with top objectness scores. (4) Feed the extracted RoI features into a MIL network and mining object boxes from the selected top-ranking proposals.

We show the overall architecture of the proposed approach in Figure 2. It consists of three key branches, *i.e*., weakly supervised semantic segmentation branch, segmentation guidance branch and object detection branch. In particular, the weakly supervised semantic segmentation branch is employed to generate class-specific pixel-wise predictions (*i.e*. segmentation confidence maps) with only image-level labels. Then the segmentation confidence maps are passed through the segmentation guidance branch. For segmentation confidence maps inside each proposals, OCRs are calculated. OCRs are then employed to evaluate objectness scores of all proposals. The proposals are ranked by the objectness scores. Finally, we select the top-ranking proposals for object detection branch to further improved object detector. The remainder of this section introduces the three branches in detail.

2.1. Weakly Supervised Semantic Segmentation

In order to verify the individual effect of objectness scoring methods, we fix the segmentation confidence maps to reduce external influences. Specifically, we choose a weakly supervised semantic segmentation network to generate semantic segmentation results in advance. When training the object detector, the network directly takes the pre-computed segmentation results as inputs. *Please note that we use the same segmentation results for all compared segmentation guided methods.* AffinityNet proposed by Ahn *et al*. [2] is used to produce the weakly supervised semantic segmentation results.

2.2. Detection Branch

We only have image-level labels indicating whether an object category appears. To train a standard object detector with regression, it is necessary to mine instance-level supervision such as bounding-box annotations. Therefore, we need to introduce a MIL branch to initialize the pseudo GT annotations. There are a couple of possible choices such as [3, 17, 12]. We choose to adopt OICR network [12] and an enhanced version [17]. For details, please refer to these papers.

Methods	mAP
MIL	41.3
MIL+WCCN[4]	39.7(-1.6)
MIL+TS2C[16]	42.3(+1.0)
MIL+OCR	46.3(+5.0)
MIL$_{reg}$[17]	47.3
MIL$_{reg}$+OCR	50.6(+3.3)

Table 1. Ablation study: The upper part shows AP performance (%) on PASCAL VOC 2007 test. In the brackets are the gaps to the MIL or MIL$_{reg}$ counterpart.

Segmentation Methods	AffinityNet[2]	IRNet[1]
MIL+WCCN[4]	39.9	39.7
MIL+TS2C[16]	42.1	42.3
MIL+OCR	45.3	46.3

Table 2. mAP Performance (%) under different segmentation branches on PASCAL VOC 2007 test.

IoU thresholds	0.5	0.7	ratio(%)
MIL+WCCN[4]	39.7	18.5	46.6
MIL+TS2C[16]	42.3	21.5	50.8
MIL+OCR	46.3	26.3	56.8

Table 3. mAP Performance (%) under different IoU thresholds on PASCAL VOC 2007 test.

Methods	mAP
WSDDN[3]	34.8
ContextLocNet[7]	36.3
OICR[12]	41.2
WCCN[4]	42.8
TS2C[16]	44.3
MIL-OICR+OCR(Ours)	46.3
MIL-OICR$_{reg}$+OCR(Ours)	50.6
WSDDN-Ens.[3]	39.3
OICR-Ens.+FRCNN[12]	47.0
WCCN+FRCNN[4]	43.1
WSRPN-Ens.+FRCNN[13]	50.4
Multi-Evidence[6]	51.2
W2F+RPN+FSD2[18]	52.4
WS-JDS FRCNN[10]	52.5
Ours-Ens.	54.6

Table 4. Comparison of AP performance (%) on PASCAL VOC 2007 test. The upper part shows results by single-phase approaches. The lower part shows results by multi-phase approaches.

Methods	mAP
ContextLocNet[7]	35.3
OICR[12]	37.9
WCCN[4]	37.9
TS2C[16]	40.0
MIL-OICR+OCR(Ours)	44.1
MIL-OICR$_{reg}$+OCR(Ours)	48.3
MELM[6]	42.4
OICR-Ens.+FRCNN[12]	42.5
TS2C+FRCNN[16]	44.0
WSRPN-Ens.+FRCNN[13]	45.7
W2F+RPN+FSD2[18]	47.8
WS-JDS FRCNN[10]	46.1
Ours-Ens.	51.1

Table 5. Comparison of AP performance (%) on PASCAL VOC 2012 test. The upper part shows results by single-phase approaches. The lower part shows results by multi-phase approaches.

2.3. Segmentation Guidance Branch

MIL detector can mining positive boxes from about two thousand proposals and subsequent classifiers and regressor can refine the selection and location of boxes. The refinement operation of OICR highly relies on the quality of initial object candidates from the multiple instance classification module. If the MIL detector fails to retrieve reasonable object proposals candidates as the pseudo GTs, the following refinement will fail too. To reduce the risk of such failure, Wei *et al*. [16] and Diba *et al*. [4] design their objectness rating approachs from the segmentation confidence maps. However, their solutions are sub-optimal and insufficient as they ignore the consistency property of objectness calculation.

Given a segmentation map and the object proposals $\mathcal{R} = (R_1, R_2, ..., R_n)$ of input image $\mathbf{I}$, we can get the segmentation maps of the object proposals $\mathcal{S} = (S_1, S_2, ..., S_n)$ and $S_i \in \mathbb{R}^{M_i \times N_i \times (C+1)}$, where M_i and N_i is the height and width of the i-th object proposal, C is the object category number. What we need to do is computing the objectness scores OS_i of the object proposals from segmentation maps:

$$OS_i = F\left(S_i\right), \tag{1}$$

where F is the objectness calculation function. F is the key factor of a good objectness scoring method. Next we introduce three different types of F.

Average strategy Considering only the *purity* property, Diba *et al*. [4] use a simple average strategy:

$$OS_i = Avg(S_i). \tag{2}$$

It is obvious that simply averaging can not determine whether an object is complete. The objectness score of a small box inside the segmentation maps will get a very high score, but this box is clearly not an ideal candidate. This strategy only considers the *purity* property.

Segmentation Context In order to get proposal candidates with much complete objects, Wei *et al*. [16] select the boxes that have high objectness scores inside the boxes and low objectness scores in the surrounding context regions:

$$OS_i = Avg(S_i) - Avg(\hat{S}_i), \tag{3}$$

where $\hat{S}_i$ is the surrounding context regions of S_i. When the box is close the border of an object, especially for the one with strange or irregular shape, the context confidence (*i.e.* $Avg(\hat{S}_i)$) is negligible and OS_i starts to decrease when the box becomes larger even though the box is still inside the object border. In order to completely solve this problem, we need consistent objectness scoring as long as the box is inside the object boundary. **Consistent Objectness** We propose to project S_i to the directions that are parallel to the bounding boxes to get the objectness consistent representation OCR_i:

$$OCR_i = \{S_i^x, S_i^y\}, \tag{4}$$

where $S_i^y \in \mathbb{R}^{M_i \times 1 \times (C+1)}$ and $S_i^x \in \mathbb{R}^{1 \times N_i \times (C+1)}$ are the projections of S_i in the horizontal and vertical directions, respectively. Specifically, S_i^x and S_i^x are calculated by:

$$\{S_i^x = Max_x(S_i), S_i^y = Max_y(S_i)\}, \tag{5}$$

where Max_x and Max_y means maximize operation in the horizontal and vertical directions, respectively. The OS_i can be calculated by:

$$OS_i = (Avg(S_i^x) + Avg(S_i^y)) - (Avg(\hat{S}_i^x) + Avg(\hat{S}_i^y)), \quad (6)$$

Now the OS_i meets the three criteria. All the object proposals $\mathcal{R} = (R_1, R_2, ..., R_n)$ are ranked according to the objectness scores OS_i. Then following Wei *et al.* [16], the top two hundred proposals are sent to the MIL detector. The MIL detector mines object boxes from these top-ranking proposals. During the testing stage, only detection branch is kept.

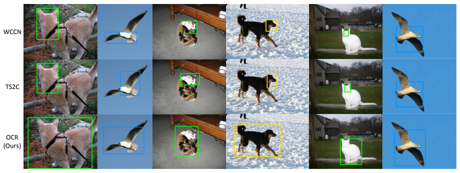

Figure 3. Qualitative detection results of WCCN [4], TS2C [16] and our method.

3. Experiments

In this section, we first introduce the evaluation datasets and the implementation details of our approach. Then we introduce the ablation experiments. Finally, we compare the performance of our method with the-state-of-the-art methods.

Datasets and Evaluation Metrics. We evaluate our method on the popular PASCAL VOC 2007 and 2012 datasets [5]. Average Precision (AP) and the mean of AP (mAP) are taken as the evaluation metrics to test our model on the testing set and are evaluated on the PASCAL criteria, i.e., IoU > 0.5 between ground truths boxes and predicted boxes.

Implementation Details. We use the object proposals generated by selective search windows [14] and adopt VGG16 [11] pre-trained on ImageNet [9] as the backbone of our proposed network. Training details follow OICR [12].

Ablation Studies We conduct ablation experiments on PASCAL VOC 2007 to prove the effectiveness of our proposed OCR.

The baseline is the MIL detector that we introduced in the detection branch section (*i.e.* Section 2.2), which is the same as **OICR** [12], denoted as **MIL**. We also use a detection branch that combine the **MIL** detector (*i.e.* **OICR**) with a box regressor, and we denote this type of detection branch as $\textbf{MIL}_{reg}$[17].

To verify the effect of our proposed new object representation, we conduct ablation studies on the objectness scoring methods. We compare three types of segmentation guidance methods, namely, **WCCN** [4], **TS2C** [16] and our **OCR**. Results are shown in the first four rows of Table 1. As Wei *et al.* [16] did, we adopt only top two hundreds proposals ranked by objectness scores for the detection branch. The performance of **MIL+WCCN** [4] is even lower than the **MIL** baseline. We attribute this to the inferior box objectness scoring method of **WCCN**, thus two hundreds proposals are far from enough. **MIL+TS2C** [16] achieves 1% mAP improvement which is slightly lower than the results reported by Wei *et al.* We attribute this difference to the usage of saliency detection results generated a fully supervised saliency detection model in the original implementation. Our method improves significantly over the baseline by 5.0%. To verify the stability of our OCR under models with different performance, we also test our OCR under a much stronger baseline $\textbf{MIL}_{reg}$, as shown in the last two rows of upper part in Table 1, our method still achieves a large improvement.

To analyze the sensitivity to segmentation quality, we have tried two different segmentation branch, *i.e.* AffinityNet [2] and IRNet [1]. The performance of these two kinds of pseudo semantic segmentation labels in PASCAL VOC 2012 train set is 59.3% and 66.5% mIoU, respectively. The detection results on PASCAL VOC 2007 mAP are shown in Table 2 the absolute decline of OCR is higher than the other two, but still far above their performance. We attribute this to the fact that the improvement brought by their guidances is not obvious, and the impact of switching to lower quality is also small.

Since our method can better handle the situation at the boundary of the object to produce more complete detection region, it is interesting to see whether the performance gap will be larger when raising IOU threshold? When the IoU threshold raised from 0.5 to 0.7, the results change as shown in the Table 3. We can conclude from the results that our OCR indeed helps select high-quality boxes with more precise boundaries.

Comparison with State-of-the-Art To fully compare with other methods, we report the results for both single-phase approaches and multi-phase approaches. The results on VOC 2007 and VOC 2012 are shown in Table 4, Table 5. In Figure 3, we also illustrate some detection results by three segmentation guided methods. It can be concluded from the illustration that our OCR helps the detector detects more complete objects.

4. Conclusion

In this paper, we present a novel object representation named OCR for segmentation guided weakly supervised object detection. We proposed a new criterion, *i.e.* consistency, for evaluating the objectness of object proposals. The proposed OCR can be easily integrated into popular weakly supervised object detection framework.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR 2019*. 1, 3, 4

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR 2018*. 1, 2, 3, 4

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR 2016*. 1, 2, 3

[4] Ali Diba, Vivek Sharma, Ali Mohammad Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *CVPR 2017*. 1, 2, 3, 4

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *IJCV 2010*. 4

[6] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR 2018*. 3

[7] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV 2016*. 1, 3

[8] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV 2016*. 1

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV 2015*. 4

[10] Yunhang Shen, Rongrong Ji, Yan Wang, Yongjian Wu, and Liujuan Cao. Cyclic guidance for weakly supervised joint detection and segmentation. In *CVPR 2019*. 3

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR 2015*. 4

[12] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR 2017*. 1, 2, 3, 4

[13] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *ECCV 2018*. 3

[14] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013. 4

[15] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR 2017*. 1

[16] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: tight box mining with surrounding segmentation context for weakly supervised object detection. In *ECCV 2018*. 2, 3, 4

[17] Ke Yang, Dongsheng Li, and Yong Dou. Towards precise end-to-end weakly supervised object detection network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 4

[18] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *CVPR 2018*. 3