TS^{2}C: Tight Box Mining with Surrounding Segmentation Context for Weakly Supervised Object Detection

Yunchao Wei[1], Zhiqiang Shen[1,2*], Bowen Cheng[1*] , Honghui Shi[3], Jinjun Xiong[3], Jiashi Feng[4], and Thomas Huang[1]

[1]University of Illinois at UrbanaChampaign, IL, USA
{yunchao, shen54, bcheng9, t-huang1}@illinois.edu
[2]Fudan University, Shanghai, China
[3]IBM T.J. Watson Research Center, Yorktown Heights, USA
Honghui.Shi@ibm.com jinjun@us.ibm.com
[4]National University of Singapore, Singapore, Singapore
elefjia@nus.edu.sg

Abstract. This work provides a simple approach to discover *tight* object bounding boxes with only image-level supervision, called **T**ight box mining with **S**urrounding **S**egmentation **C**ontext (TS^{2}C). We observe that object candidates mined through current multiple instance learning methods are usually trapped to discriminative object parts, rather than the entire object. TS^{2}C leverages surrounding segmentation context derived from weakly-supervised segmentation to suppress such low-quality distracting candidates and boost the high-quality ones. Specifically, TS^{2}C is developed based on two key properties of desirable bounding boxes: 1) high purity, meaning most pixels in the box are with high object response, and 2) high completeness, meaning the box covers high object response pixels comprehensively. With such novel and computable criteria, more tight candidates can be discovered for learning a better object detector. With TS^{2}C, we obtain 48.0% and 44.4% mAP scores on VOC 2007 and 2012 benchmarks, which are the new state-of-the-arts.

Keywords: weakly-supervised learning, object detection, semantic segmentation

1 Introduction

Weakly Supervised Object Detection (WSOD) [3, 7, 10, 17, 18, 20, 21, 23, 32, 33, 35, 42–44] aims to detect objects only using image-level annotations for supervision. Despite remarkable progress, existing approaches still have difficulties in accurately identifying tight boxes of target objects with only image-level annotations, thus their performance is inferior to the fully supervised counterparts [6, 13, 22, 25, 28–30].

* denotes equal contribution

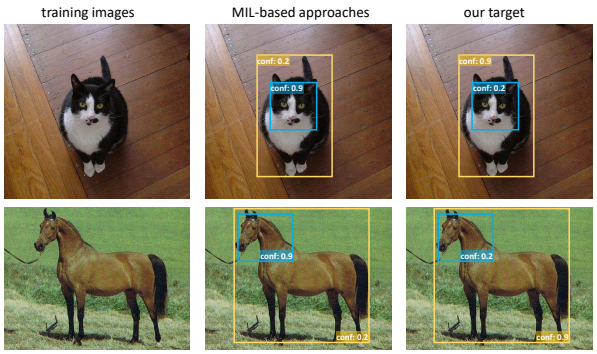

Fig. 1. Comparison of MIL-based approaches and our target. MIL-based approaches tend to assign high confidence to discriminative parts (blue boxes) of target objects. Our target is to alleviate such cases and lift the confidence of the tight ones (yellow boxes). Best viewed in color.

To localize objects with weak supervision information, one popular solution is to apply Multiple Instance Learning (MIL) for mining high-confidence region proposals [34, 47] 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, as shown in Figure 1. This inability of providing the complete object severely limits its effectiveness for WSOD. To address this issue, Li *et al.* [21] exploited the contrastive relationship between a selected region and its mask-out image for proposal selection. Nevertheless, the mask-out strategy fails for multi-instance cases. The selector is easily confused by remained instances with high responses, even though the correct object has been masked out.

Recently, some weakly supervised semantic segmentation approaches [19, 36, 38, 40] have demonstrated promising performance. Utilizing the inferred segmentation confidence maps, Diba *et al.* [10] presented a cascaded approach that leverages segmentation knowledge to filter noisy proposals and achieves competitive detection results. However, we argue that their solution is sub-optimal and insufficient as it only considers the segmentation confidence *inside* the proposal boxes, thus is unable to filter high-response fragments of object parts, as the magenta boxes shown in Figure 2 (b).

In this work, we propose a principled and more effective approach, compared with [10], to mine tight object boxes by exploiting segmentation confidence maps in a creative way, aiming for addressing the challenging WSOD problems. Our approach is motivated by the following observations, as illustrated by two examples in Figure 2 (a). We use blue and yellow to encode two kinds of boxes, which partially and tightly cover objects respectively. Based on the semantic segmentation confidence maps obtained in a weakly supervised manner, many pixels surrounding the blue boxes have high predicted segmentation confidence, while very few high-confidence pixels are included in the surrounding context

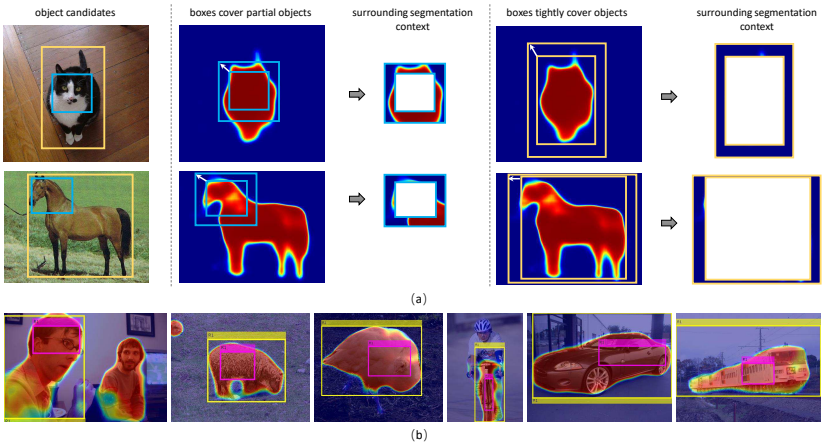

Fig. 2. (a) Motivation of the proposed TS^{2}C: fewer high response pixels on the segmentation confidence map are included by enlarging higher-quality boxes of object candidates (the yellow one) compared with partial bounding boxes (the blue one). (b) Comparison of the rank 1 proposal using the strategy proposed by [10] (magenta boxes) and ours (yellow boxes). Best viewed in color.

for the yellow ones of higher tightness. We find that a desirable tight object box generally needs to satisfy two properties based on segmentation context:

- *Purity*: most pixels inside the box should have high confidence scores, which guarantees that the box is located around the target object;
- *Completeness*: very few pixels are with high confidence scores in the surrounding context of the target box.

Based on these properties, we devise a simple yet effective approach, named Tight box mining with Surrounding Segmentation Context (TS^{2}C), to efficiently select object candidates of high quality from thousands of candidates. Specifically, the proposed TS^{2}C examines two kinds of regions for evaluating the tightness of bounding boxes: 1) the region included in the box and 2) the region surrounding the box. It computes objectness scores of the two regions by averaging the corresponding pixel confidence values on the segmentation maps. Tight boxes are expected to be with high and low objectness values of the two kinds of regions simultaneously. Thus, the difference of two objectness scores is then taken as the quality metric on the final tightness for ranking object candidates. Figure 2 (b) shows the top 1 object candidate inferred by the proposed TS^{2}C. We can see that our approach is more effective for mining tight object boxes than [10]. Moreover, our proposed TS^{2}C is generic and can be easily integrated into any WSOD framework by introducing a parallel semantic segmentation branch for class-specific confidence map prediction. Benefiting from our TS^{2}C, we achieve 48.0% and 44.4% mAP scores on the challenging Pascal VOC 2007 and VOC 2012 benchmarks, which are the new state-of-the-arts in the WSOD community.

2 Related Work

Multiple Instance Learning (MIL) provides a suitable way for formulating and solving WSOD. In specific, if an image is annotated with a specific class, at least one proposal instance from the image is positive for this class; and no proposal instance is positive for unlabeled classes. Previous works on applying MIL to WSOD can be roughly categorized into two-step [7, 17, 21, 35] and end-to-end [3, 10, 18, 20, 32, 33] based approaches.

Two-step approaches first extract proposal representation leveraging handcrafted features or pre-trained CNN models and employ MIL to select the best object candidate for learning the object detector. For instance, Wang *et al.* [35] presented a latent semantic clustering approach to select the most discriminative cluster for each category. Cibis *et al.* [7] learned a multi-fold MIL detector by re-labeling proposals and re-training the object classifier iteratively. Li *et al.* [21] first trained a multi-label classification network on entire images and then selected class-specific proposal candidates using a mask-out strategy, followed by MIL for learning a Fast R-CNN detector. Recently, Jie *et al.* [17] took a similar strategy as Li *et al.* [21] and proposed a more robust self-taught approach to learn a detector by harvesting more accurate supportive proposals in an online manner. However, splitting the WSOD into two steps results in a non-convex optimization problem, making such approaches trapped in local optima.

End-to-end approaches combine CNNs and MIL into a unified framework for addressing WSOD. Oquab *et al.* [27] and Wei *et al.* [39] adopted a similar strategy to learn a multi-label classification network with max-pooling MIL. The learned classification model was then applied to coarse object localization [27]. Bilen *et al.* [3] proposed a novel Weakly Supervised Deep Detection Network (WSDDN) including two key streams, one for classification and the other for object localization. The outputs of these two streams are then combined for better rating the objectness of proposals. Based on WSDDN, Kantorov *et al.* [18] proposed to learn a context-aware CNN with contrast-based contextual modeling. Both [18] and our approach employ proposal context to identify high-quality proposals. However, [18] exploits inside/outside context features of each bounding box for learning to classification, in contrast, we leverage objectness scores obtained by segmentation confidence maps to pick out tight candidates. Recently, Tang *et al.* [32] also employed WSDDN as the basic network and augmented it with several Online Instance Classifier Refinement (OICR) branches, which is the state-of-the-art on the challenging WSOD task. In this work, we employ both WSDDN and OICR to develop our framework where the proposed TS^2C is leveraged to further improve performance. Both [10] and our approach utilizes object segmentation knowledge to benefit WSOD. However, Diba *et al.* [10] only considered the confidence of pixels included in the bounding box for rating the proposal objectness, which is not as effective as ours.

Beyond the above mentioned related works, some fully-supervised object detection approaches [5, 12, 22, 46] also exploit contextual information of bounding boxes for benefiting object detection. Both Chen *et al.* [5] and Li *et al.* [22] leveraged information of enlarged contextual proposals to enhance the accuracy of

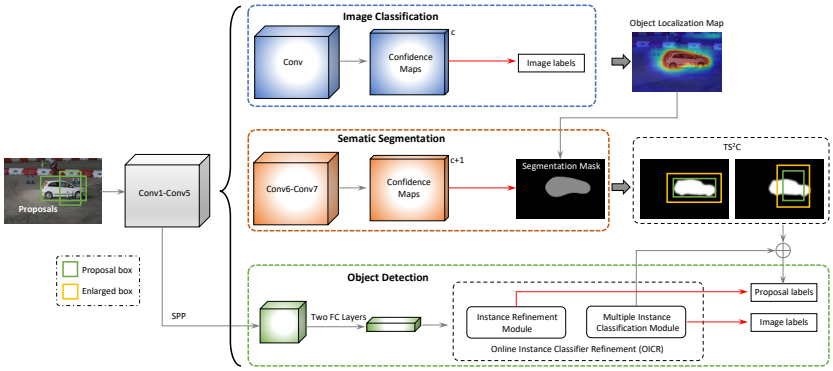

Fig. 3. Overview of the proposed TS^{2}C for weakly supervised object detection. Several convolutional layers are leveraged to extract the intermediate features of an input image. The entire feature maps are firstly fed into a *Classification* branch to produce object localization maps corresponding to image-level labels. We then employ the localization maps to generate the segmentation masks, which serve as supervision to learn the *Segmentation* branch. Based on the segmentation confidence maps, we utilize TS^{2}C to evaluate the objectness scores of proposals according to their purity and completeness, which collaborates with the OICR [32] for training the *Detection* branch.

the classifier. Zhu *et al.* [46] proposed to use a pool of segments obtained in the bottom-up manner to obtain better detection boxes. Our TS^{2}C is totally different from these works in terms of both motivation and methodology. In particular, our motivation is to employ surrounding segmentation context to suppress these false positive objects parts. In addition, our approach can be easily embedded into any WSOD framework to make a further performance improvement.

3 The Proposed Approach

We show the overall architecture of the proposed approach in Figure 3. It consists of three key branches, *i.e.* image classification, semantic segmentation and object detection. In particular, the *Classification* branch is employed to generate class-specific localization maps. Following the previous weakly supervised semantic segmentation approaches [37], we leverage the inferred localization maps to produce pseudo segmentation masks of training images, which are then used as supervision to train the *Segmentation* branch. The segmentation confidence maps from the *Segmentation* branch are then employed to evaluate objectness scores of the proposals according to the proposed TS^{2}C, which finally collaborates with the *Detection* branch for learning an improved object detector. The overall framework is trained by minimizing the following composite loss functions from the three branches using stochastic gradient descent:

$$L = L_{cls} + L_{seg} + L_{det}. \tag{1}$$

We will introduce each branch below and then elaborate on details of TS^{2}C.

3.1 Classification for Object Localization

Inspired by [10, 24, 45], the fully convolutional network along with the Global Average Pooling (GAP) operation is able to generate class-specific activation maps, which can provide coarse object localization prior. We conduct experiments on Pascal VOC benchmarks, in which each training image is annotated with one or several labels. We thus treat the classification task as a separate binary classification problem for each class. Following [27], the loss function L_{cls} is thus defined as a sum of C binary logistic regression losses.

3.2 Weakly Supervised Semantic Segmentation

The *Classification* branch can produce localization cues for foreground objects. We assign the pixels with values on the class-specific confidence map larger than a pre-defined normalized threshold (*i.e.* $\geq$0.78) with the corresponding class label. Beyond the object regions, background localization cues are also needed for training the segmentation branch. Motivated by [19, 36, 38, 40], we leverage the saliency detection technology [41] to produce the saliency map for each training image. Based on the generated saliency map, we choose the pixels with low normalized saliency values (*i.e.* $\leq$0.06) as background. However, both the class-specific confidence map and the saliency map are not accurate enough to guarantee a high-quality segmentation mask. To alleviate the negative effect caused by falsely assigned pixels, we ignore the ambiguous pixels during training the *Segmentation* branch, including 1) pixels that are not assigned semantic labels, 2) foreground pixels of different categories that are in conflict, and 3) low-saliency pixels that fall in the foreground pixels. With the produced pseudo segmentation mask, we train the *Segmentation* branch with pixel-wise cross-entropy loss L_{seg}, which is widely adopted by fully-supervised schemes [4, 26].

3.3 Learning Object Detection with TS^{2}C

For each training or test image, Selective Search [34] is employed to generate object proposals and Spatial Pyramid Pooling (SPP) [15] is leveraged to generate constant size feature maps for different proposals. Our TS^{2}C aims to select high-quality object candidates from thousands of candidates to improve the effectiveness of training, which can be easily implanted into any WSOD framework. We choose the state-of-the-art Online Instance Classifier Refinement (OICR) [32] as the backbone of the *Detection* branch, which collaborates with the proposed TS^{2}C for learning a better object detector. In the following, we will first make a brief introduction of OICR, and then explain how to leverage our TS^{2}C to benefit the learning process of WSOD.

OICR As shown in Figure 3, the OICR mainly includes two modules, *i.e.* multiple instance classification and instance refinement. In particular, the multiple instance classification module is inspired from [3], which includes two branches

to extract parallel data streams from the input features pooled by SPP, as shown in Figure 4 (a). The upper stream conducts softmax operation on each individual proposal for classification. The bottom stream estimates a probability distribution over all candidate proposals using softmax, which indicates the contribution of each proposal to classifier decision for each class. Therefore, these two streams provide classification-based and localization-based features for each proposal. Both inferred scores are then fused with element-wise product operation and finally aggregated into image-level prediction by sum-pooling over all proposals. With the supervision of image-level annotations, the multiple instance classification module can be learned with binary logistic regression losses as detailed in Section 3.1.

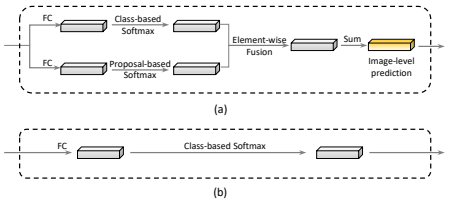

Fig. 4. Details of (a) Multiple Instance Classification Module and (b) Instance Refinement Module in TS^{2}C.

By leveraging multiple instance classification module as a basic classifier for obtaining initial classification scores for each proposal, progressive refinement is then conducted via the instance refinement module, as detailed in Figure 4 (b). In particular, the instance refinement module first selects the top-scoring proposal of each image-level label. Those proposals with high spatial overlap scores over the top-scoring one are then labeled correspondingly. The idea behind such a module is that the top-scoring proposal may only contain part of a target object and its adjacent proposals may cover more object regions. Benefiting from both two modules embedded in the OICR, each proposal is assigned with a pseudo class label, which is then employed as supervision for learning detection with the softmax cross-entropy loss [13, 14, 29]. To address the initialization issue (*i.e.* the classifier cannot well recognize proposals with randomly initialized parameters at the beginning of training), OICR adopts a weighted loss by assigning different weights to different proposals during different training iterations. Thus, the L_{det} is composed of binary logistic regression losses for image-level classification and softmax cross-entropy loss for proposal-level classification. Please refer to [32] for more details.

Problems However, such progressive refinement operation of OICR highly relies on the quality of initial object candidates from the multiple instance classification module. This means without reasonable object candidates received from the multiple instance classification module for initialization, the following progressive refinement strategy of OICR cannot find the correct proposals with high IoU scores over ground-truth bounding boxes. This brings a critical risk: if the multiple instance classification module fails to produce reasonable object candidates then the OICR cannot recall the missed object with any hope. We propose

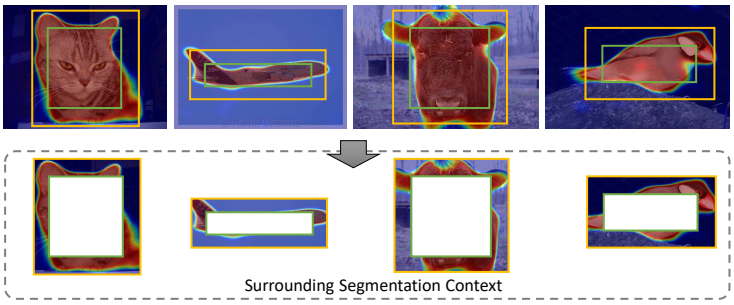

Fig. 5. Motivation of the conditional average strategy: only a small number of pixels belong to objects in the surrounding regions. To promote the objectness score of surrounding context, we only employ pixels with large confidence values (highlighted by red color) for conducting average calculation. Best viewed in color.

to reduce such a risk by designing an objectness rating approach from a totally new perspective. In particular, we detail our proposed TS^{2}C that rates the proposals' objectness from the segmentation view in the following.

TS^{2}C for Learning Detection As shown in Figure 3, TS^{2}C uses the segmentation confidence maps from the *Segmentation* branch to rate the proposal objectness. We consider $x_i (i = 1 \cdots n)$ as one proposal from a given training image annotated by class c. Let H_c denote the confidence map of category c predicted by the semantic *Segmentation* branch. For x_i, we calculate objectness scores of both the region inside the box P_I and the surrounding context P_S between x_i and the corresponding enlarged one. Let $avg(H_c, x_i)$ denote the operation of computing P_I, which takes all pixel values included in x_i into account. P_I of a large value can guarantee that x_i is around the target object. To obtain a robust surrounding objectness score P_S, we adopt a conditional average strategy $a\hat{v}g(H_c, x_i)$. As shown in Figure 5, many surrounding regions of negative candidates include a large number of un-related (*i.e.* background) pixels, which are with low confidence scores. Therefore, the resulted objectness score will be small if we average all the pixel values for computing P_S, in a similar way as for P_I. However, we expect the value of P_S to be large, so that negative candidates of such cases can be suppressed by $P_I - P_S$. To this end, we first rank the pixels in the surrounding region according to their confidence scores and the conditional average strategy only employs the first 50% pixels for calculating the objectness score. Then, the objectness score $O(x_i)$ of the proposed TS^{2}C is finally calculated as

$$O(x_i) = P_I - P_S = avg(H_c, x_i) - a\hat{v}g(H_c, x_i).$$

We rank all the object candidates according to $O(x_i)$ and build a candidate pool by selecting the top two hundred proposals, collaborating with the OICR for learning a better detector. As shown in Figure 3, $\oplus$ means the OICR will

only select object candidates from the pool produced by TS^{2}C for the following training process.

During the testing stage, we ignore the *Classification* and *Segmentation* branches, and leverage the classification outputs from the instance refinement module to obtain the final detection results.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets We conduct experiments on Pascal VOC 2007 and 2012 datasets [11], which are the two most widely used benchmarks for weakly supervised object detection. For VOC 2007, we train the model on the *trainval* set (5,011 images) and evaluate on the *test* set (4,096 images). We also make extensive ablation analysis on VOC 2007 to verify the effectiveness of some settings. For VOC 2012, we train the model on the *trainval* set (11,540 images) and evaluate on *test* set (10,991 images) by submitting the testing result to the evaluation server.
Metrics Following [10, 17, 32], we adopt two metrics for evaluation, *i.e.* mean average precision (mAP) and correct localization (CorLoc) [9], for evaluation on *test* and *trainval* sets, respectively. Both two metrics employ the same threshold of bounding box overlaps with ground-truth boxes, *i.e.* IoU $>= 0.5$.

4.2 Implementation Details

We use the object proposals generated by Selective Search [34], and adopt the VGG16 network [31] pre-trained on ImageNet [8] as the backbone of the proposed framework. We employ the Deeplab-CRF-LargeFOV [4] model to initialize the corresponding layers in the segmentation branch. For the newly added layers, the parameters are randomly initialized with a Gaussian distribution $\mathcal{N}(\mu, \delta)(\mu = 0, \delta = 0.01)$. We take a mini-batch size of 2 images and set the learning rates of the first 40K and the following 30K iterations as 0.001 and 0.0001 respectively. During training, we take five image scales $\{480, 576, 688, 864, 1200\}$ for data augmentation. For TS^{2}C, we adopt an enlarged ratio of 1.2 to obtain the surrounding context, which is further employed for evaluating completeness of object candidates. Our experiments use the OICR [32] code, which is implemented based on the publicly available Caffe [16] deep learning framework. All of our experiments are run on NVIDIA TITAN X PASCAL GPUs.

4.3 Comparison with Other State-of-the-arts

We compare our approach with both two-step [7,17,21,35] and end-to-end [3,10, 18, 20, 32, 33] approaches. Top-3 results are indicated by ***green***, ***red*** and ***blue*** colors. Table 1 shows the comparison in terms of AP on the VOC 2007. It can be observed that the proposed TS^{2}C is effective and outperforms all the other approaches. In particular, we adopt OICR proposed by Tang *et al.* [32] as the

Table 1. Comparison of detection average precision (AP) (%) on PASCAL VOC.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
Comparisons on VOC 2007:																					
Bilen [1]	42.2	43.9	23.1	9.2	12.5	44.9	45.1	24.9	8.3	24.0	13.9	18.6	31.6	43.6	7.6	20.9	26.6	20.6	35.9	29.6	26.4
Bilen [2]	46.2	46.9	24.1	16.4	12.2	42.2	47.1	35.2	7.8	28.3	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
Cinbis [7]	39.3	43.0	28.8	20.4	8.0	45.5	47.9	22.1	8.4	33.5	23.6	29.2	38.5	47.9	20.3	20.0	35.8	30.8	41.0	20.1	30.2
Wang [35]	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
Li [21]	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
Bilen [3]	46.4	58.3	35.5	25.9	14.0	66.7	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	49.6	56.9	50.8	39.3
Teh [33]	48.8	45.9	37.4	26.9	9.2	50.7	43.4	43.6	10.6	35.9	27.0	38.6	48.5	43.8	24.7	12.1	29.0	23.2	48.8	41.9	34.5
Tang [32]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
Jie [17]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	54.3	3.0	53.6	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
Diba [10]	49.5	60.6	38.6	29.2	16.2	70.8	56.9	42.5	10.9	44.1	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	**42.8**
Lai [20]	48.4	61.5	33.3	30.0	15.3	72.4	62.4	59.1	10.9	42.3	34.3	53.1	48.4	65.0	20.5	16.6	40.6	46.5	54.6	55.1	**43.5**
TS^{2}C	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	**44.3**
Comparisons on VOC 2012:																					
Kantorov [18]	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
Tang [32]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**37.9**
Jie [17]	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**
TS^{2}C	67.4	57.0	37.7	23.7	15.2	56.9	49.1	64.8	15.1	39.4	19.3	48.4	44.5	67.2	2.1	23.3	35.1	40.2	46.6	45.8	**40.0**

Table 2. Comparison of detection AP (%) by training FRCNN detectors.

Method	VOC 2007	VOC 2012
TS^{2}C + FRCNN	48.0	44.4
OICR-Ens. + FRCNN [32]	47.0	42.5

detection backbone in the proposed framework. Our approach outperforms OICR by 3.1%. The gains are mainly from using both purity and completeness metrics for filtering noisy object candidates. We also show the comparison between our approach and other state-of-the-arts on PASCAL VOC 2012 in terms of AP. Our result[1] outperforms the baseline (*i.e.* Tang *et al.* [32]) and the state-of-the-art approach (*i.e.* Jie *et al.* [17]) by 2.1% and 1.7%, respectively.

Following [32], we also train a FRCNN [13] detector using top-scoring proposals produced by TS^{2}C as pseudo ground-truth bounding boxes. As shown in Table 2, the performance can be further enhanced to 48.0% and 44.4%[2] on VOC 2007 and 2012, respectively. Our results from a single model are much better than those of [32] obtained by models (*e.g.* VGG16 and VGG-M) fusion. In addition, we conduct additional experiments using CorLoc as the evaluation metric. Table 3 shows the comparison on the VOC 2007 and 2012. Our approach achieves 61.0% and 64.4% in terms of CorLoc score, which are competitive compared with the state-of-the-arts. We visualize some successful detection results (blue boxes) on VOC 2007, as shown in Figure 6. Results from OICR (green boxes) and ground truth (red boxes) are employed for comparison. It can be seen that our approach effectively reduces false positives including partial objects.

[1] http://host.robots.ox.ac.uk:8080/anonymous/GDNUDG.html
[2] http://host.robots.ox.ac.uk:8080/anonymous/ECKWR7.html

Fig. 6. Examples of our object detection results on VOC 2007 test set. Ground-truth annotations, predictions of OICR and ours are indicated by red, green and blue bounding boxes respectively. Best viewed in color.

4.4 Ablation Experiments

We conduct extensive ablation analyses of the proposed TS^{2}C, including the influence of the enlarged scale for obtaining surrounding context and the proposed tightness criteria (*i.e.* purity and completeness). All experiments are based on VOC 2007 benchmark.

Purity and Completeness One of our main contributions is the proposed criteria of purity and completeness for measuring the tightness of object candidates based on the semantic segmentation confidence maps. To validate the effectiveness of our approach (*i.e.* $P_I - P_S$), we test the other popular setting where only the purity (*e.g.* P_I) is taken into account. Specifically, we firstly leverage the two metrics to rank object candidates for annotated class(es). For example, if the image is annotated with two labels, we will produce two rankings according to segmentation confidence maps of the two classes, which are then employed for evaluating recall scores. As shown in Figure 7, we vary the top number of object candidates based on the rankings from two metrics. Since our evaluation method only takes one object candidate for each annotated category in the top-1 case, the upper bound of the recall is 57.9% due to the existence of multi-instance images. Despite the apparent simplicity, the recall scores of our proposed $P_I - P_S$ significantly outperform those of P_I under different settings according to the top number, which demonstrates that the completeness metric is effective for reducing noisy object candidates. More visualizations of rank 1 boxes produced by $P_I - P_S$ and P_I are shown in Figure 8. We can observe that our approach can successfully discover the tight ones from thousands of candidates. To further validate the effectiveness of the proposed TS^{2}C, we also conduct experiments

Table 3. Comparison of correct localization (CorLoc) (%) on PASCAL VOC.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
Comparisons on VOC 2007:																					
Bilen [2]	66.4	59.3	42.7	20.4	21.3	63.4	74.3	59.6	21.1	58.2	14.0	38.5	49.5	60.0	19.8	39.2	41.7	30.1	50.2	44.1	43.7
Cinbis [7]	65.3	55.0	52.4	48.3	18.2	66.4	77.8	35.6	26.5	67.0	46.9	48.4	70.5	69.1	35.2	35.2	69.6	43.4	64.6	43.7	52.0
Wang. [35]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
Li [21]	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
Bilen [3]	65.1	63.4	59.7	45.9	38.5	69.4	77.0	50.7	30.1	68.8	34.0	37.3	61.0	82.9	25.1	42.9	79.2	59.4	68.2	64.1	56.1
Jie [17]	72.7	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	75.5	56.1
Diba [10]	83.9	72.8	64.5	44.1	40.1	65.7	82.5	58.9	33.7	72.5	25.6	53.7	67.4	77.4	26.8	49.1	68.1	27.9	64.5	55.7	56.7
Tang [32]	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
Lai [20]	71.0	76.5	54.9	49.7	54.1	78.0	87.4	68.8	32.4	75.2	29.5	58.0	67.3	84.5	41.5	49.0	78.1	60.3	62.8	78.9	**62.9**
Teh [33]	84.0	64.6	70.0	62.4	25.8	80.7	73.9	71.5	35.7	81.6	46.5	71.3	79.1	78.8	56.7	34.3	69.8	56.7	77.0	72.7	**64.6**
TS^2C	84.2	74.1	61.3	52.1	32.1	76.7	82.9	66.6	42.3	70.6	39.5	57.0	61.2	88.4	9.3	54.6	72.2	60.0	65.0	70.3	**61.0**
Comparisons on VOC 2012:																					
Kantorov [18]	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
Jie [17]	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**
Tang [32]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**62.1**
TS^2C	79.1	83.9	64.6	50.6	37.8	87.4	74.0	74.1	40.4	80.6	42.6	53.6	66.5	88.8	18.8	54.9	80.4	60.4	70.7	79.3	**64.4**

Table 4. Ablation study on PASCAL VOC 2007.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
P_I *vs.* $P_I - P_S$:																					
P_I	54.8	64.3	37.5	28.7	13.9	63.7	62.4	47.3	16.7	45.5	29.6	26.6	41.4	63.1	10.1	23.0	42.5	50.5	63.3	57.9	42.2
$P_I - P_S$	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	**44.3**
Enlarged scales:																					
baseline	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
scale (1.4)	61.3	58.1	44.7	26.2	10.1	65.0	60.5	37.2	28.3	49.8	40.9	24.2	38.9	62.1	9.4	23.9	41.7	51.0	60.8	58.8	42.6
scale (1.3)	61.2	60.2	39.7	29.0	9.8	65.2	59.5	53.3	24.5	48.3	41.0	33.9	40.4	61.4	12.2	22.5	42.1	52.5	59.4	60.9	43.8
scale (1.2)	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	**44.3**
scale (1.1)	59.6	58.1	41.3	29.1	13.3	64.0	60.6	52.9	25.7	49.9	45.6	29.2	40.4	61.4	11.6	22.9	40.8	48.3	60.3	60.7	43.8
Conditional average strategy:																					
top 30%	60.8	58.7	39.7	33.2	11.2	64.3	60.5	52.6	24.8	48.1	37.2	25.6	45.5	63.7	11.4	23.8	40.9	49.1	58.4	59.9	43.5
top 50%	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	**44.3**
top 70%	60.9	61.5	41.8	31.8	12.8	64.8	60.3	46.5	22.8	49.7	38.7	26.3	50.2	63.2	12.7	22.4	41.6	49.4	60.0	60.3	43.9
all pixels	60.8	60.3	38.2	31.2	11.3	63.6	60.1	55.6	20.9	51.9	40.0	33.4	41.2	64.6	11.1	23.2	43.0	47.7	59.6	59.3	43.8

using purity *i.e.* P_I for ranking object candidates as adopted in [10] for proposal selection, which results in 42.2% in mAP. By simultaneously taking purity and completeness into account, *i.e.* $P_I - P_S$, the result surpasses the baseline by 2.1% as shown in Table 4.

Influence of Enlarged Scale To evaluate the completeness of object candidates, we need to enlarge the original box with a specific ratio. As shown in Table 4, we examine four ratios (*i.e.* from 1.1 to 1.4) for obtaining the surrounding context of object candidates, which are then employed to calculate objectness scores with the proposed TS^2C. We can observe that all the models trained with the proposed TS^2C can outperform the baseline by more than 1.4%. In particular, the best result is achieved by adopting the ratio of 1.2. By continually enlarging the ratio, the performance will be decreased. The reason may be that some training images include multiple instances with the same semantics, and

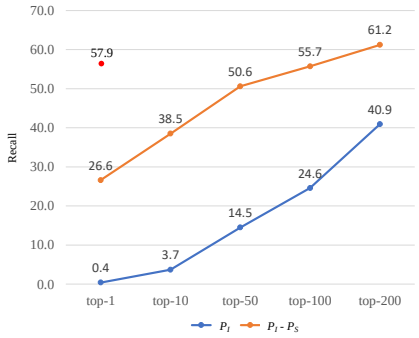

Fig. 7. Comparison of recall scores (%) between the proposed TS^{2}C ($P_I - P_S$) and the purity strategy (P_I).

the completeness score of each object candidate will be influenced by adjacent instances in the case of using larger ratios.

Influence of Conditional Averaging Strategy As shown in Table 4, we also examine the threshold of conditional average strategy. The best result is achieved by employ the first 50% largest pixels to calculate the objectness score of surrounding region.

Discussion Some failure cases are shown in the last row of Figure 8. These samples share some similar characteristics: low-quality segmentation predictions or many semantically identical instances are linked together. For instance (the middle image of the last row), the semantic segmentation branch makes a false prediction for the object under the *bird*, leading to incorrect inference of our approach. It is believed that such a case can be well addressed with the development of weakly supervised semantic segmentation techniques. For other failure samples, although the segmentation branch can provide high quality confidence maps, the overlap between objects results in false prediction of our TS^{2}C. In this case, we may need to develop effective instance-level semantic segmentation approaches in a weakly supervised manner.

However, the limitation of our TS^{2}C to deal with overlapping objects with the same semantics does not affect its good performance on WSOD. We do not employ the top-1 proposal according to the objectness score as the object candidate, but build a candidate pool by selecting the top two hundred proposals. In this case, these tight boxes may still be recalled even without the largest tightness score. The effectiveness of our TS^{2}C can be well proved by the performance gains on VOC 2007 and 2012 compared with [32].

5 Conclusion and Future Work

In this work, we proposed a simple approach, *i.e.* TS^{2}C, for mining tight boxes by exploiting surrounding segmentation context. The TS^{2}C is effective for suppressing low quality object candidates and promoting high quality ones tightly

Fig. 8. Rank 1 object candidates inferred by the proposed TS^{2}C (yellow boxes) and the strategy only using purity metric for ranking (magenta boxes). Some failure cases are given in the last row. Best viewed in color.

covering the target object. Based on the segmentation confidence map, TS^{2}C introduces two simple criteria, *i.e.* purity and completeness, to evaluate objectness scores of object candidates. Despite apparent simplicity, the proposed TS^{2}C can effectively filter thousands of noisy candidates and be easily embedded into any end-to-end weakly supervised framework for performance improvement. In the future, we plan to design more effective metrics for mining tight boxes by further boosting our current approach.

Acknowledgements This work is in part supported by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network, NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133, MOE Tier-II R-263-000-D17-112 and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00341. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

1. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with posterior regularization. In: BMVC. pp. 1–12 (2014)
2. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: IEEE CVPR. pp. 1081–1089 (2015)
3. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: IEEE CVPR. pp. 2846–2854 (2016)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. preprint arXiv:1412.7062 (2014)
5. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: NIPS. pp. 424–432 (2015)
6. Cheng, B., Wei, Y., Shi, H., Feris, R., Xiong, J., Huang, T.: Revisiting rcnn: On awakening the classification power of faster rcnn. In: ECCV (2018)
7. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE TPAMI **39**(1), 189–203 (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE CVPR. pp. 248–255 (2009)
9. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. IJCV **100**(3), 275–293 (2012)
10. Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: IEEE CVPR (2017)
11. 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), 98–136 (2014)
12. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: IEEE ICCV. pp. 1134–1142 (2015)
13. Girshick, R.: Fast r-cnn. In: IEEE ICCV. pp. 1440–1448 (2015)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE CVPR. pp. 580–587 (2014)
15. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European Conference on Computer Vision. pp. 346–361 (2014)
16. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM Multimedia. pp. 675–678 (2014)
17. Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: IEEE CVPR (2017)
18. Kantorov, V., Oquab, M., Cho, M., Laptev, I.: Contextlocnet: Context-aware deep network models for weakly supervised localization. In: ECCV. pp. 350–365 (2016)
19. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: ECCV. pp. 695–711 (2016)
20. Lai, B., Gong, X.: Saliency guided end-to-end learning for weakly supervised object detection. In: IJCAI (2017)
21. Li, D., Huang, J.B., Li, Y., Wang, S., Yang, M.H.: Weakly supervised object localization with progressive domain adaptation. In: IEEE CVPR. pp. 3512–3520 (2016)

22. Li, J., Wei, Y., Liang, X., Dong, J., Xu, T., Feng, J., Yan, S.: Attentive contexts for object detection. IEEE Transactions on Multimedia **19**(5), 944–954 (2017)
23. Liang, X., Liu, S., Wei, Y., Liu, L., Lin, L., Yan, S.: Towards computational baby learning: A weakly-supervised approach for object detection. In: IEEE ICCV. pp. 999–1007 (2015)
24. Lin, M., Chen, Q., Yan, S.: Network in network. ICLR (2013)
25. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37 (2016)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE CVPR (2015)
27. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: IEEE CVPR. pp. 685–694 (2015)
28. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: IEEE CVPR (2017)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
30. Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., Xue, X.: Dsod: Learning deeply supervised object detectors from scratch. In: IEEE ICCV (2017)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (2015)
32. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: IEEE CVPR (2017)
33. Teh, E.W., Rochan, M., Wang, Y.: Attention networks for weakly supervised object localization. In: BMVC (2016)
34. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV **104**(2), 154–171 (2013)
35. Wang, C., Ren, W., Huang, K., Tan, T.: Weakly supervised object localization with latent category learning. In: ECCV. pp. 431–445 (2014)
36. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: IEEE CVPR (2017)
37. Wei, Y., Liang, X., Chen, Y., Jie, Z., Xiao, Y., Zhao, Y., Yan, S.: Learning to segment with image-level annotations. Pattern Recognition (2016)
38. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Feng, J., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE TPAMI (2016)
39. Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: Hcp: A flexible cnn framework for multi-label image classification. IEEE TPAMI **38**(9), 1901–1907 (2016)
40. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: IEEE CVPR. pp. 7268–7277
41. Xiao, H., Feng, J., Wei, Y., Zhang, M., Yan, S.: Deep salient object detection with dense connections and distraction diagnosis. IEEE Transactions on Multimedia (2018)
42. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.: Adversarial complementary learning for weakly supervised object localization. In: IEEE CVPR (2018)
43. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: ECCV (2018)
44. Zhao, F., Li, J., Zhao, J., Feng, J.: Weakly supervised phrase localization with multi-scale anchored transformer network. In: IEEE CVPR. pp. 5696–5705 (2018)

45. Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. IEEE CVPR (2016)
46. Zhu, Y., Urtasun, R., Salakhutdinov, R., Fidler, S.: segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In: IEEE CVPR. pp. 4703–4711 (2015)
47. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV. pp. 391–405 (2014)