

# Object Instance Mining for Weakly Supervised Object Detection

Chenhao Lin,<sup>1</sup> Siwen Wang,<sup>2\*</sup> Dongqi Xu,<sup>1\*</sup> Yu Lu,<sup>1†</sup> Wayne Zhang<sup>1</sup>

<sup>1</sup>SenseTime Research

<sup>2</sup>Dalian University of Technology, Dalian, China, 116024

{linchenhao, luyu, wayne.zhang}@sensetime.com, {wangsiwendut, isdongqixu}@gmail.com

## Abstract

Weakly supervised object detection (WSOD) using only image-level annotations has attracted growing attention over the past few years. Existing approaches using multiple instance learning easily fall into local optima, because such mechanism tends to learn from the most discriminative object in an image for each category. Therefore, these methods suffer from missing object instances which degrade the performance of WSOD. To address this problem, this paper introduces an end-to-end object instance mining (OIM) framework for weakly supervised object detection. OIM attempts to detect all possible object instances existing in each image by introducing information propagation on the spatial and appearance graphs, without any additional annotations. During the iterative learning process, the less discriminative object instances from the same class can be gradually detected and utilized for training. In addition, we design an object instance reweighted loss to learn larger portion of each object instance to further improve the performance. The experimental results on two publicly available databases, VOC 2007 and 2012, demonstrate the efficacy of proposed approach.

## Introduction

Object detection has always been one of the most essential technologies in computer vision field. Deep learning techniques introduced in recent years have significantly boosted state-of-the-art approaches for object detection (Girshick 2015; Liu et al. 2016; Redmon et al. 2016; Ren et al. 2015). However, these approaches usually require large-scale manually annotated datasets (Russakovsky et al. 2015). The high-cost of time-consuming accurate bounding box annotations, has impeded the wide deployment of CNN-based object detection technologies in real applications.

To address this limitation, weakly supervised object detection (WSOD) technology, which requires only image-level labels for training, has been introduced and explored (Bilen and Vedaldi 2016; Diba et al. 2017; Jie et al. 2017; Oquab et al. 2015; Zhang et al. 2018b; Tang et al. 2017;

\*Work performed as an intern in SenseTime Research  
<https://github.com/bigvideoresearch/OIM>

†Corresponding author

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

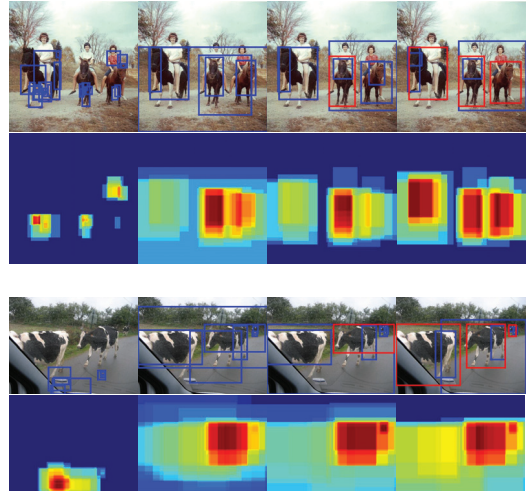


Figure 1: The original images and corresponding objectness maps to show the evolution of object instance mining during learning process (from left to right). The first to fourth columns represent random initialization, epoch1, epoch3, and final epoch, respectively. Blue or red bounding boxes indicate the detected instances (top-scoring proposals after NMS) with detection scores  $< 0.5$  or  $\geq 0.5$ .

Zhang et al. 2018a; Shen et al. 2018; Arun, Jawahar, and Kumar 2019; Pan et al. 2019). Although many approaches have been developed for WSOD and achieved promising results, the lack of object instance level annotations in images leads to huge performance gap between WSOD and fully supervised object detection (FSOD) methods.

Most previous approaches follow the framework of combining multiple instance learning (MIL) with CNN. This framework usually mines the most confident class-specific object proposals for learning CNN-based classifier, regardless of the number of object instances appearing in an image. For the images with multiple object instances from the same class, the object instances (fully annotated with bounding boxes in FSOD) with lower class-specific scores will be probably regarded as background regions. Many images in the challenging VOC datasets contain more than one object

instance from the same class. For example, in VOC2007 *trainval* set the number of image-level object labels and the annotated object instances are 7,913 and 15,662 respectively, which indicates that at least 7,749 instances are NOT selected during training. In this case, the selected object instances with relatively limited scale and appearance variations, may not be sufficient for training a CNN classifier with strong discriminative power. Moreover, the missing instances may be selected as negative samples during training, which may further degrades the discriminative capability of the CNN classifier.

In this paper, an end-to-end object instance mining (OIM) framework is proposed to address the problem of multiple object instances in each image for WSOD. OIM is based on two fundamental assumptions: 1) the highest confidence proposal and its surrounding highly overlapped proposals should probably belong to the same class; 2) the objects from the same class should have high appearance similarity. Formally, spatial and appearance graphs are constructed and utilized to mine all possible object instances present in an image and employ them for training. The spatial graph is designed to model the spatial relationship between the highest confidence proposal and its surrounding proposals, while the appearance graph aims at capturing all possible object instances having high appearance similarities with the most confident proposal. By integrating these two graphs into the iterative training process, an OIM approach that attempts to accurately mine all possible object instances in each image with only image-level supervision is proposed. With more object instances for training, a CNN classifier can have stronger discriminative power and generalization capabilities. The proposed OIM can further prevent the learning process from falling into local optima because more objects per-class with high similarity are employed for training. The original images and the corresponding objectness maps shown in Figure 1 illustrate that with the increasing number of iterations, multiple object instances belonging to the same class can be detected and are employed for training using the proposed approach.

Another observation from existing approaches is that the most confident region proposal is easy to concentrate on the locally distinct part of an object, especially for non-rigid objects such as human and animals. This may lead to the problem of detecting only small part of the object. To alleviate this problem, an object instance reweighted loss using the spatial graph is presented to help the network detect more accurate bounding box. This loss tends to make the network pay less attention on the local distinct parts and focus on learning the larger portion of each object.

Our key contributions can be summarized as follows:

- An object instance mining approach using spatial and appearance graphs is developed to mine all possible object instances with only image-level annotation, and it can significantly improve the discriminative capability of the trained CNN classifier.
- An object instance reweighted loss by adjusting the weight of loss function of different instances is proposed to learn more accurate CNN classifier.

## Related Work

With only image-level annotations, most existing approaches implement weakly supervised object detection (Bilen and Vedaldi 2016; Tang et al. 2017; Jie et al. 2017; Wan et al. 2019) through multiple instance learning (MIL) framework (Dietterich, Lathrop, and Lozano-Pérez 1997). The training images are firstly divided into bag of proposals (instances) containing positive target objects and negative backgrounds and CNN classifier is trained to classify the proposals into different categories. The most discriminative representation of instances is easy to be distinguished by such classifier that may make network trap into local optima.

Recently, Bilen et al. (Bilen and Vedaldi 2016) proposed a weakly supervised deep detection network (WSDDN) to perform object localization and classification simultaneously. Following this work, Tang *et al.* (Tang et al. 2017) introduced an online instance classifier refinement (OICR) strategy to learn larger portion of the objects. Such approach improves the performance of WSOD. However, it is also easy to trap into local optima since only the most discriminative instance is selected for refinement. Wan *et al.* (Wan et al. 2018) developed a min-entropy latent model to classify and locate the objects by minimizing the global and local entropies, which was proved to effectively boost the detection performance. Wan *et al.* (Wan et al. 2019) also attempted to address the local minima problem in MIL using continuation optimization method. In references (Wei et al. 2018; Shen et al. 2019; Li et al. 2019), the authors attempted to integrate segmentation task into weakly supervised object detection to obtain more accurate object bounding boxes. However, these methods require complex training framework with high training and test time complexity.

The authors in (Tang et al. 2018) proposed to use proposal cluster to divide all proposals into different small bags and then classifier refinement was applied. This approach attempted to classify and refine all possible objects in each image. However, many proposals containing part of the object might be ignored using proposal cluster during the training. Gao et al. (Gao et al. 2018) introduced a count-guided weakly supervised localization approach to detect per-class objects in each image. A simple count-based region selection algorithm was proposed and integrated into OICR to improve the performance of WSOD. However, the extra count annotations which needs a certain human labor are introduced and their method requires an alternative training process which can be time-consuming. In this paper, the count annotation is replaced by the proposed OIM algorithm without extra labor cost.

## Proposed Approach

### Overall Framework

The overall architecture of the proposed framework illustrated in Figure 2 mainly consists of two parts. The first part is a multiple instance detector (MID) which is similar to the structure presented in (Bilen and Vedaldi 2016). It performs region selection and classification simultaneously using a weighted MIL pooling. The second part is the proposed object instance mining and the proposed instance reweighted

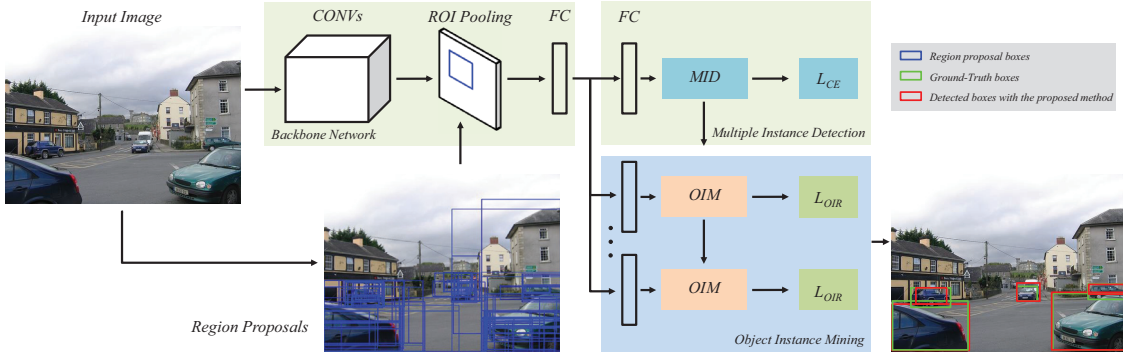


Figure 2: Architecture of the proposed object instance mining framework. *MID* represents multiple instance detector and *OIM* indicates proposed object instance mining.  $L_{CE}$  is multi-class cross entropy loss and  $L_{OIR}$  is proposed instance reweighted loss.

loss. During the training phase, we firstly adopt MID to classify the region proposals into different predicted classes. Then the detection outputs and proposal features are integrated to search all possible object instances from the same class in each image using spatial and appearance graphs. In addition, the instance reweighted loss is designed to learn larger portion of each object. As can be seen from the Figure 2, the multiple object instances belonging to the same class can be accurately detected using the proposed method.

### Object Instance Mining

Previous methods (Tang et al. 2017; Gao et al. 2018; Zhang et al. 2018b; Wei et al. 2018) often select the most confident proposal from each class as the positive sample to refine the multiple instance detector. The performance improvement can be limited using these methods, since only the top-scoring and surrounding proposals are selected for refinement. While in many conditions, there are multiple object instances belonging to the same class in an image. Those ignored object instances may be regarded as negative samples during the training that may degrade the performance of WSOD. Therefore, we propose an object instance mining (OIM) approach by building spatial graphs and appearance graphs to search all possible object instances in each image and integrate them into the training process.

Based on the assumption that the top-scoring and surrounding proposals with large overlaps (spatial similarity) should have the same predicted class, the spatial graphs can be built. We also assume that the objects from the same class should have similar appearance. Based on the similarities between the top-scoring proposal and the other proposals, the appearance graphs are built. Then we search all possible object instances in each image and employ them for training through these graphs.

Given an input image  $I$  with class label  $c$ , a set of region proposals  $P = \{p_1, \dots, p_N\}$  and their corresponding confidence scores  $X = \{x_1, \dots, x_N\}$ , the core instance (proposal)  $p_{i_c}$  with the highest confidence score  $x_{i_c}$  can be selected. Here  $i_c$  donates the index of this core instance (proposal). The core spatial graph can be defined by  $G_{i_c}^s = (V_{i_c}^s, E_{i_c}^s)$ , where each node in  $V_{i_c}^s$  represents a selected proposal which has the overlap, i.e. spatial similarity, with the core instance

larger than a threshold  $T$ . Each edge in  $E_{i_c}^s$  represents such spatial similarity. All the nodes in spatial graph  $G_{i_c}^s$  will be selected and labelled to the same class as  $p_{i_c}$ .

We define feature vectors of each proposal as  $F = \{f_1, \dots, f_N\}$  and it can be generated from the fully connected layer. Each vector encodes a feature representation of a region proposal. Then the appearance graph is defined as  $G^a = (V^a, E^a)$ , where each node in  $V^a$  is a selected proposal which has high appearance similarity with the core instance and each edge in  $E^a$  represents the appearance similarity. This similarity can be calculated from the feature vectors of core instance and one of the other proposals (e.g.  $p_j$ ) using the Euclidean distance, denoted as follows,

$$D_{i_c, j} = \|f_{i_c} - f_j\|_2. \quad (1)$$

Only if the proposal  $p_j$  meets the condition that  $D_{i_c, j} < \alpha D_{avg}$  and  $p_j$  has no overlap with all the proposals previously selected, such proposal will be added into the nodes in  $G^a$ .  $D_{avg}$  represents the average inter-class similarity of the core spatial graph  $G_{i_c}^s$  using average distance of all the nodes in  $G_{i_c}^s$  and it can be defined as follows,

$$D_{avg} = \frac{1}{M} \sum_k D_{i_c, k}, \quad (2)$$

$$\text{s.t. } IoU(p_{i_c}, p_k) > T.$$

where  $p_k$  represents the node meet the constraints above and  $M$  indicates the number of these nodes in  $G_{i_c}^s$ .  $\alpha$  is a hyper parameter which is determined by experiments.

The proposed object instance mining (OIM) approach using spatial and appearance graphs is summarized in Algorithm 1. We also build spatial graph  $G^s$  for each node in appearance graph  $G^a$  and then all these nodes will be included for training. If no proposal has high similarity with the core instance, only the core instance and surrounding proposals, i.e. spatial graph  $G_{i_c}^s$  will be employed. In such a way, more instances from the same class with similar appearance and different poses will be employed for training. It results in that not only more object instances can be detected but also more accurate detected boxes can be learned.

Figure 3 illustrates the process to detect all possible object instances from to the same class using spatial and appearance graphs. Figure 3 (a) is the core spatial graph and

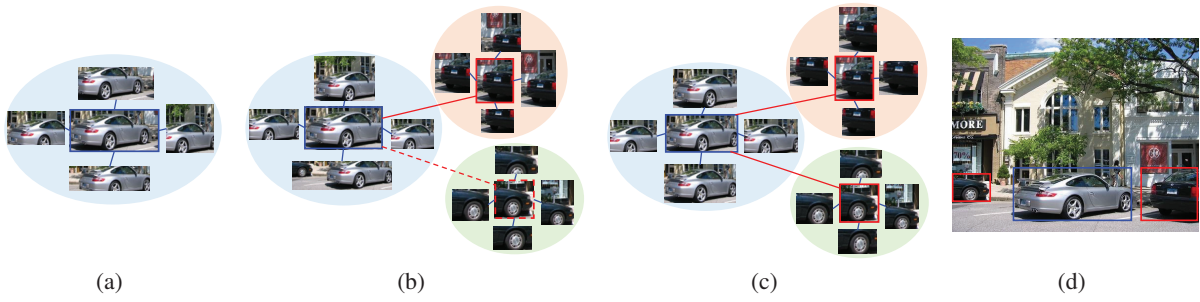


Figure 3: Process to explore all possible object instances from the same class using OIM. (a)-(c) illustrate the spatial and appearance graphs of different epochs and (d) shows all detected instances. Blue bounding boxes represent the detected core instance with the highest confidence score. Red bounding boxes represent the other detected instances which have high appearance similarities with the core instance. Blue and red line represent spatial and appearance graph edge respectively. Red broken line in (b) means the appearance similarities is smaller than the threshold and thus the object instances are not employed in this stage.

---

### Algorithm 1: Object Instance Mining

---

**Input:** Image  $I$ , region proposals  $P = \{p_1, \dots, p_N\}$ , image label  $Y = \{y_1, y_2, \dots, y_c\}$   
**Output:** All the nodes  $V^a$  in appearance graph

- 1 Feed Image  $I$  and its proposals into the network to produce feature vectors  $F = \{f_1, \dots, f_N\}$
- 2 **for**  $c$  in  $C$ ,  $C$  is the list of training data class **do**
- 3   **if**  $y_c == 1$  **then**
- 4      $V^s \leftarrow \emptyset, V^a \leftarrow \emptyset, D \leftarrow 0, D_{avg} \leftarrow 0, M \leftarrow 0,$   
 $flag \leftarrow 0$
- 5     Choose the top-scoring proposal  $i_c$
- 6      $V_{i_c}^s \leftarrow p_{i_c}, V^a \leftarrow p_{i_c}$
- 7     **for**  $j = 1$  to  $N$  **do**
- 8       Compute the appearance similarity  $D_{i_c, j}$  using Eq. 1
- 9       Compute  $IoU(p_{i_c}, p_j)$
- 10       **if**  $IoU(p_{i_c}, p_j) > T$  **then**
- 11           $V_{i_c}^s \leftarrow p_j$
- 12           $D \leftarrow D_{i_c, j} + D, M \leftarrow M + 1$
- 13        $D_{avg} \leftarrow \frac{D}{M}$
- 14       Sort (ascend)  $P$  based on  $D_{i_c, j}$
- 15       **for**  $j = 1$  to  $N$  **do**
- 16          **if**  $D_{i_c, j} < \alpha D_{avg}$  **then**
- 17           **if**  $\exists p_k \in V^a, IoU(p_k, p_j) > 0$  **then**
- 18              $flag \leftarrow 1$
- 19           **if**  $flag == 0$  **then**
- 20              $V^a \leftarrow p_j, V_j^s \leftarrow p_j$

---

figure 3 (b)-(c) describe the spatial and appearance graphs in different epochs. With the increased number of iterations, more instances can be detected using the proposed OIM.

### Instance Reweighted Loss

In addition to exploring all possible object instances in each image, we also design an object instance reweighted loss to learn more accurate detected boxes. During the iterative learning process, the CNN-based classifier is easy to learn the most distinct part of each object instance instead of the whole body, especially for the non-rigid one. We propose to assign different proposal weights to individual proposals

to balance the weight of the top-scoring proposal and surrounding less discriminative ones. Thus the larger portion of each instance is expected to be detected.

Given an image with label  $Y$  and predicted label  $Y_j = [y_{0,j}, y_{1,j}, \dots, y_{C,j}]^T \in \mathbb{R}^{(C+1) \times 1}$  for the  $j$ -th proposal in a spatial graph  $G^s$ , where  $y_{c,j} = 1$  or 0 indicates the proposal belonging to class  $c$  or not, and  $c = 0$  is index of background class. The loss in Eq. 3 is similar to the loss in (Tang et al. 2017), where  $w_j$  is the loss weight of  $j$ -th proposal.  $x_{c,j}^s$  with class label  $c$  in  $G^s$ , are the proposals used for training and  $x_{c,i_c}^s$  is center (core) proposal with the highest score.

$$\mathcal{L} = -\frac{1}{|P|} \sum_{j=1}^{|P|} \sum_{c=1}^{C+1} w_j y_{c,j} \log x_{c,j}^s. \quad (3)$$

It can be seen from Eq. 3 that proposals in each spatial graph contribute equally. Thus, the non-center proposals with relative low scores in each spatial graph are difficult to be learned during training. To address this problem, an instance reweighted loss function is designed as follows,

$$\mathcal{L} = -\frac{1}{|P|} \sum_{j=1}^{|P|} \sum_{c=1}^{C+1} w_j y_{c,j} (1 + z_j^s) \log x_{c,j}^s, \quad (4)$$

where  $z_j^s$  is introduced to balance the proposal weights in spatial graph  $G^s$  as defined in Eq. 5.  $\beta$  is hyper-parameter.

$$z_j^s = \begin{cases} \beta, & j \neq i_c \\ \beta - 1, & j = i_c \end{cases} \quad (5)$$

To guide the network to pay more attention on learning the less discriminative regions of the object instance in each graph  $G^s$ , we balance the weight of the surrounding less discriminative proposals with the center proposal using Eq. 4 and Eq. 5. As a result, gradients of surrounding proposals are scaled up to  $(1 + \beta)$  of its original value, while gradient of the center proposal is scaled to  $\beta$  of its original value during back-propagation. Similar to the implementation in (Gao et al. 2018), we also use the standard multi-class cross entropy loss for the multi-label classification and it is combined with the proposed instance reweighted loss for training.

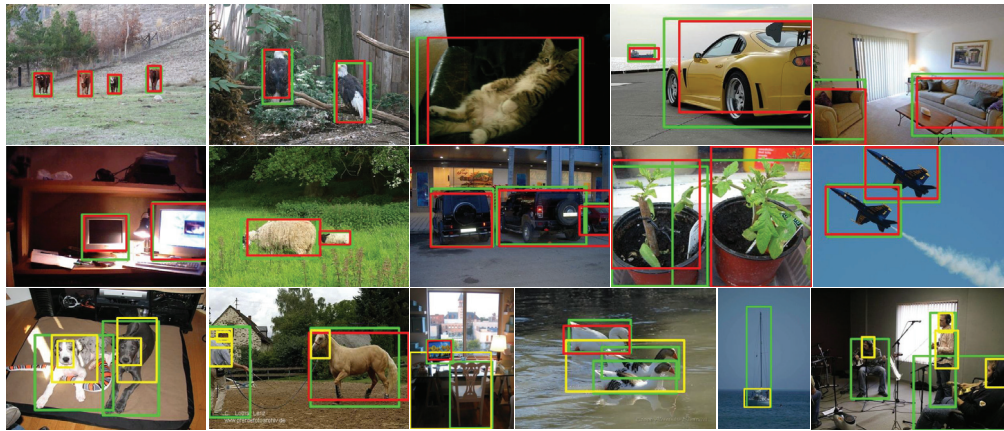


Figure 4: Detections examples on VOC2007 *test* set. The green bounding boxes represent the ground-truth. The successful detections ( $\text{IoU} \geq 0.5$ ) are marked with red bounding boxes, and the failed ones are marked with yellow color. We show all detections with scores  $\geq 0.5$  and NMS is performed to remove duplicate detections.

## Experiments

### Datasets and Evaluation Metrics

Following the previous state-of-the-art methods on WSOD, we also evaluate our approach two datasets, PASCAL VOC2007(Everingham et al. 2010) and VOC2012(Everingham et al. 2015), which both contain 20 object categories. For VOC2007, we train the model on the *trainval* set (5,011 images) and evaluate the performance on the *test* set (4,952 images). For VOC2012, the *trainval* set (11,540 images) and the *test* set (10,991 images) are used for training and evaluation respectively. Additionally, we train our model on the VOC2012 *train* set (5,717 images) and proceed evaluation on the *val* set (5,823 images) to further validate the effectiveness of proposed approach. Following previous work, we use mean average precision (mAP) to evaluate the performance of proposed approach. Correct localization (CorLoc) is applied to evaluate the localization accuracy.

### Implementation Details

To make a fair comparison, VGG16 model pre-trained on the ImageNet dataset (Russakovsky et al. 2015) is adopted as the backbone network to finetune the CNN classifier. The object proposals are generated using Selective Search(Uijlings et al. 2013). The batch size is set to 2, and the learning rates are set to 0.001 and 0.0001 for the first 40K and the following 50K iterations respectively. During training and test, we take five image scales  $\{480, 576, 688, 864, 1200\}$  along with random horizontal flipping for data augmentation. Following (Tang et al. 2017), the threshold  $T$  is set to 0.5. With the increased number of iterations, the network has more stable learning ability, we dynamically set the hyper parameters  $\alpha$  as  $\alpha_1 = 5$  for the first 70K and  $\alpha_2 = 2$  for the following 20K iterations.  $\beta$  are empirically set to 0.2 in our experiments. We also analyze the influence of these parameters in the ablation experiments section. 100 top-scoring region proposals are kept and Non-Maximum Suppression with IoU of 0.3 per class is performed to calculate mAP and CorLoc.

### Comparison with State-of-the-arts

State-of-the-art WSOD methods are used for comparison to validate the effectiveness of the proposed approach. Table 1 shows performance comparison in terms of mAP on VOC2007 *test* set. By only using OIM, better or similar results can be achieved compared with previous SOTA methods such as MELM, SDCN, *etc.* We attribute this improvement to the OIM, which increases the representation capability of the trained CNN by searching more objects from the same class and employing them into training. As the detected bounding boxes and objectness maps shown in Figure 1, the confidence scores of less discriminative objects are gradually improved and more objects from the same class can be detected during the training. It further proves that integrating the less discriminative objects into training improves the performance for WSOD. Further performance improvement can be achieved using the proposed instance reweighted loss. The proposed approach achieves a mAP of 50.1%, which outperforms the PCL,C-WSL\*, SDCN, WSJDS methods, *etc.*, and the performance is similar to the result of C-MIL. We further used the learned objects as pseudo ground-truth to train a Fast-RCNN-based detector, our approach also achieve better or similar performance as compared with previous state-of-the-art methods.

In particular, by only using the proposed OIM strategy, our approach outperforms C-WSL method by 1.4 % without introducing extra per-class count supervision. Our work attempts to include all possible object instances from each class for training since many images contain more than one per-class object instance. Figure 5 illustrates that most classes in two datasets have more than one object instance in an image. Specifically, almost half of categories contain more than two object instances in an image. Especially for class “sheep”, which the average number of sheep appearing in an image is larger than 3, our OIM method (57.9 % mAP) performs better than all the other methods. In addition, for most non-rigid objects (“cat”, “dog”, “horse”, “person”, *etc.*), as can be seen from Table 1, by applying in-

Table 1: Comparison with the state-of-the-arts in terms of mAP (%) on the VOC2007 *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              | 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        |
| PCL               | 54.4        | 69.0        | 39.3        | 19.2        | 15.7        | 62.9        | 64.4        | 30.0        | 25.1        | 52.5        | 44.4        | 19.6        | 39.3        | 67.7        | 17.8        | 22.9        | 46.6        | 57.5        | 58.6        | 63.0        | 43.5        |
| TS <sup>2</sup> 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        |
| C-WSL*            | 62.9        | 64.8        | 39.8        | 28.1        | 16.4        | 69.5        | 68.2        | 47.0        | <b>27.9</b> | 55.8        | 43.7        | 31.2        | 43.8        | 65.0        | 10.9        | 26.1        | 52.7        | 55.3        | 60.2        | 66.6        | 46.8        |
| MELM              | 55.6        | 66.9        | 34.2        | 29.1        | 16.4        | 68.8        | 68.1        | 43.0        | 25.0        | 65.6        | 45.3        | 53.2        | 49.6        | 68.6        | 2.0         | 25.4        | 52.5        | 56.8        | 62.1        | 57.1        | 47.3        |
| OICR+W-RPN        | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | -           | 46.9        |
| SDCN              | 59.8        | 67.1        | 32.0        | <b>34.7</b> | 22.8        | 67.1        | 63.8        | 67.9        | 22.5        | 48.9        | 47.8        | 60.5        | 51.7        | 65.2        | 11.8        | 20.6        | 42.1        | 54.7        | 60.8        | 64.3        | 48.3        |
| WS-JDS            | 52.0        | 64.5        | 45.5        | 26.7        | 27.9        | 60.5        | 47.8        | 59.7        | 13.0        | 50.4        | 46.4        | 56.3        | 49.6        | 60.7        | 25.4        | 28.2        | 50.0        | 51.4        | 66.5        | 29.7        | 45.6        |
| C-MIL             | 62.5        | 58.4        | 49.5        | 32.1        | 19.8        | 70.5        | 66.1        | 63.4        | 20.0        | 60.5        | 52.9        | 53.5        | 57.4        | 68.9        | 8.4         | 24.6        | 51.8        | 58.7        | 66.7        | 63.5        | <b>50.5</b> |
| OIM               | 62.2        | 67.2        | 48.0        | 29.6        | 23.5        | 68.7        | 69.3        | 64.3        | 22.8        | 59.6        | 39.6        | 30.7        | 42.7        | 69.8        | 3.1         | 23.3        | <b>57.9</b> | 55.4        | 63.4        | 63.5        | 48.2        |
| OIM+IR            | 55.6        | 67.0        | 45.8        | 27.9        | 21.1        | 69.0        | 68.3        | 70.5        | 21.3        | 60.2        | 40.3        | 54.5        | 56.5        | 70.1        | 12.5        | 25.0        | 52.9        | 55.2        | 65.0        | 63.7        | <b>50.1</b> |
| C-WSL*+FRCNN      | 62.9        | 68.3        | 52.9        | 25.8        | 16.5        | 71.1        | 69.5        | 48.2        | 26.0        | 58.6        | 44.5        | 28.2        | 49.6        | 66.4        | 10.2        | 26.4        | 55.3        | 59.9        | 61.6        | 62.2        | 48.2        |
| SDCN+FRCNN        | 61.1        | 70.6        | 40.2        | 32.8        | 23.9        | 63.4        | 68.9        | 68.2        | 18.3        | 60.2        | 53.5        | 63.6        | 53.6        | 66.1        | 14.6        | 21.8        | 50.5        | 56.7        | 62.4        | <b>67.9</b> | 51.0        |
| WS-JDS+FRCNN      | 64.8        | 70.7        | 51.5        | 25.1        | <b>29.0</b> | 74.1        | 69.7        | 69.6        | 12.7        | <b>69.5</b> | 43.9        | 54.9        | 39.3        | <b>71.3</b> | <b>32.6</b> | <b>29.8</b> | 57.0        | 61.0        | 66.6        | 57.4        | 52.5        |
| Pred Net (FRCNN)  | <b>66.7</b> | 69.5        | 52.8        | 31.4        | 24.7        | <b>74.5</b> | <b>74.1</b> | 67.3        | 14.6        | 53.0        | 46.1        | 52.9        | <b>69.9</b> | 70.8        | 18.5        | 28.4        | 54.6        | 60.7        | 67.1        | 60.4        | <b>52.9</b> |
| C-MIL+FRCNN       | 61.8        | 60.9        | <b>56.2</b> | 28.9        | 18.9        | 68.2        | 69.6        | 71.4        | 18.5        | 64.3        | <b>57.2</b> | <b>66.9</b> | 65.9        | 65.7        | 13.8        | 22.9        | 54.1        | <b>61.9</b> | 68.2        | 66.1        | <b>53.1</b> |
| OIM+IR+FRCNN      | 53.4        | <b>72.0</b> | 51.4        | 26.0        | 27.7        | 69.8        | 69.7        | <b>74.8</b> | 21.4        | 67.1        | 45.7        | 63.7        | 63.7        | 67.4        | 10.9        | 25.3        | 53.5        | 60.4        | <b>70.8</b> | 58.1        | <b>52.6</b> |

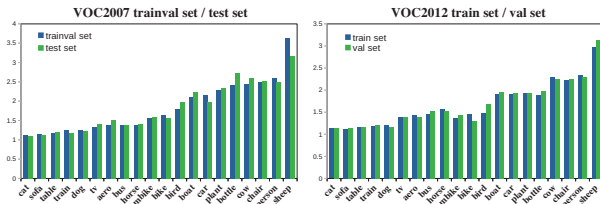


Figure 5: Objects number of each class divided by the number of images which the corresponding class occurs on VOC2007 and VOC2012.

Table 2: Comparison with the state-of-the-arts in terms of CorLoc (%) on the VOC2007 *trainval* set.

| Method                              | Localization (CorLoc) |
|-------------------------------------|-----------------------|
| OICR(Tang et al. 2017)              | 60.6                  |
| PCL(Tang et al. 2018)               | 62.7                  |
| C-WSL* (Gao et al. 2018)            | 63.5                  |
| MELM (Wan et al. 2018)              | 61.4                  |
| WS-JDS (Shen et al. 2019)           | 64.5                  |
| C-MIL (Wan et al. 2019)             | 65.0                  |
| OICR+W-RPN (Singh and Lee 2019)     | 66.5                  |
| SDCN (Li et al. 2019)               | 66.8                  |
| OIM+IR                              | <b>67.2</b>           |
| C-WSL*+FRCNN (Gao et al. 2018)      | 66.1                  |
| WS-JDS+FRCNN (Shen et al. 2019)     | 68.6                  |
| SDCN+FRCNN (Li et al. 2019)         | <b>68.8</b>           |
| Pred Net (FRCNN) (Arun et al. 2019) | <b>70.9</b>           |
| OIM+IR+FRCNN                        | <b>68.8</b>           |

stance reweighted loss more accurate object instance can be detected.

CorLoc is also used as the evaluation metric to ascertain the performance of proposed method. Table 2 shows performance comparison in terms of CorLoc on the VOC2007 *trainval* set. Our result outperforms all existing state-of-the-art methods when Fast-RCNN detector is not used. The proposed OIM framework iteratively explores more object instances and larger portion of the instances from the same class with similar appearance and different poses for training, which makes more accurate detected boxes can be learned. Therefore, the proposed approach not only brings the mAP improvements but also makes the detected boxes

Table 3: Comparison with the state-of-the-arts in terms of mAP (%) on the VOC2012 *test* set.

| Method                              | Dataset       | mAP         |
|-------------------------------------|---------------|-------------|
| MELM (Wan et al. 2018)              | train/val     | 40.2        |
| C-WSL (Gao et al. 2018)             | train/val     | 43.0        |
| OIM+IR                              | train/val     | <b>44.4</b> |
| OICR (Tang et al. 2017)             | trainval/test | 37.9        |
| PCL (Tang et al. 2018)              | trainval/test | 40.6        |
| MELM (Wan et al. 2018)              | trainval/test | 42.4        |
| WS-JDS (Shen et al. 2019)           | trainval/test | 39.1        |
| OICR+W-RPN (Singh and Lee 2019)     | trainval/test | 43.2        |
| SDCN (Li et al. 2019)               | trainval/test | 43.5        |
| OIM+IR                              | trainval/test | <b>45.3</b> |
| Pred Net (FRCNN) (Arun et al. 2019) | trainval/test | 48.4        |
| WS-JDS + FRCNN (Shen et al. 2019)   | trainval/test | 46.1        |
| C-MIL + FRCNN (Wan et al. 2019)     | trainval/test | <b>46.7</b> |
| SDCN + FRCNN (Li et al. 2019)       | trainval/test | <b>46.7</b> |
| OIM+IR + FRCNN                      | trainval/test | <b>46.4</b> |

Table 4: Comparison with the state-of-the-arts in terms of CorLoc (%) on the VOC2012 *trainval* set.

| Method                              | Localization (CorLoc) |
|-------------------------------------|-----------------------|
| OICR(Tang et al. 2017)              | 62.1                  |
| PCL(Tang et al. 2018)               | 63.2                  |
| WS-JDS (Shen et al. 2019)           | 63.5                  |
| OICR+W-RPN (Singh and Lee 2019)     | <b>67.5</b>           |
| SDCN (Li et al. 2019)               | <b>67.9</b>           |
| OIM+IR                              | <b>67.1</b>           |
| Pred Net (FRCNN) (Arun et al. 2019) | 69.5                  |
| WS-JDS + FRCNN (Shen et al. 2019)   | 69.5                  |
| C-MIL + FRCNN (Wan et al. 2019)     | 67.4                  |
| SDCN + FRCNN (Li et al. 2019)       | 69.5                  |
| OIM+IR + FRCNN                      | <b>69.5</b>           |

more accurate which results in better CorLoc.

The proposed approach is also evaluated on VOC2012 dataset. Since some approaches (Gao et al. 2018) only use validation set of VOC2012 for evaluation, we use both *test* and *val* set to evaluate the proposed approach. In Table 3, the detection results in terms of mAP on *test* and *val* set are provided respectively. Table 4 lists the CorLoc results on VOC2012 *trainval* set. The experimental results in Tables 3 and 4 validate the effectiveness of the proposed approach.

Figure 4 visualizes the detection results on the VOC2007

Table 5: Detection performance (mAP%) on the VOC2007 for using different values of parameter  $\alpha$  and parameter  $\beta$ .

|            |             |      |             |      |
|------------|-------------|------|-------------|------|
| $\alpha_1$ | 1           | 2    | 5           | 10   |
| $\alpha_2$ | 1           | 2    | 2           | 5    |
| OIM        | 42.9        | 48.1 | <b>48.2</b> | 46.8 |
| OIM+IR     | 43.4        | 49.3 | <b>50.1</b> | 48.4 |
| $\beta$    | 0.2         | 0.5  | 0.8         |      |
| OIM+IR     | <b>50.1</b> | 48.0 | 46.3        |      |

Table 6: Detection performance comparison of proposed approach on the VOC2007 with various configurations.

| SG | AG | OIM(SG+AG) | IR | mAP (%)     |
|----|----|------------|----|-------------|
|    |    |            |    | 34.8        |
| ✓  |    |            |    | 42.2        |
|    | ✓  |            |    | 46.7        |
| ✓  | ✓  | ✓          |    | 48.2        |
|    |    |            | ✓  | 43.4        |
| ✓  | ✓  | ✓          | ✓  | <b>50.1</b> |

*test* set. The successful ( $\text{IoU} \geq 0.5$ ) and failed ( $\text{IoU} < 0.5$ ) detections are marked with red and yellow bounding boxes respectively. The green bounding boxes are the ground-truths. The first two rows indicate our approach can detect tight boxes even multiple objects from the same class co-occur in an image, e.g. “cow”, “sheep”. The last row shows some failed cases, which are often attribute to localizing the most discriminative parts of non-rigid objects, grouping multiple objects, and background clutter, e.g. “human”.

### Ablation Experiments

We performed ablation experiments to illustrate the effect of parameters introduced in proposed object instance mining ( $\alpha$ ) and instance reweighted loss ( $\beta$ ). Table 5 indicates when parameter  $\alpha$  ( $\alpha_1$  used in the first 70K and  $\alpha_2$  used in the following 20K iterations) becomes smaller or larger, the performance of proposed approach will degrade. If the parameter  $\alpha$  is too small, very less instances will be selected in the appearance graph for training. It results in that in many images, only the most discriminative object is selected and used for training. If the parameter  $\alpha$  is too large, many false instances (background proposals) will be employed for training and it also leads to performance drop. For the proposed instance reweighted loss, as also can be seen from the Table 5, with the increasing of  $\beta$  the performance decreases.

We also studied the WSOD performance by only using appearance graph (AG) or spatial graph (SG) to evaluate their effectiveness separately. The first two columns in Table 6 illustrate the experimental results in terms of mAP on the VOC2007 *test* set. We can see that the performance can be significantly improved for WSOD by only using appearance or spatial graph.

The effectiveness of the proposed instance reweighted loss is also evaluated. We apply the network structure in OICR but just replace the loss with instance reweighted loss. The performance achieved using the proposed instance reweighted loss in terms of mAP on the VOC2007 *test* set is shown in Table 6. It can be seen that the mAP can be im-

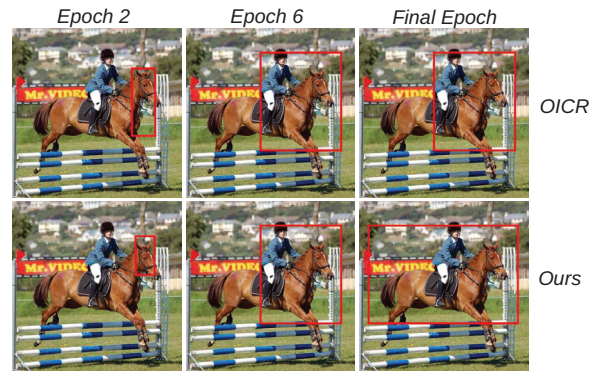


Figure 6: Evolution of object detection during learning process w/o using instance reweighted loss (from left to right). The upper part of each subfigure is the result of OICR (Tang et al. 2017) and the lower part is the result of our method.

proved from 41.2% (Tang et al. 2017) to 43.4% by only using instance reweighted loss. The visual comparison shown in Figure 6 also illustrates that larger portion of the object can be gradually detected using the proposed loss. By incorporating the OIM with instance reweighted loss, the best performance (mAP 50.1%) can be achieved.

### Conclusion

In this paper, an end-to-end object instance mining framework has been presented to address the limitations of existing approaches for WSOD. Object instance mining algorithm is performed using spatial and appearance graphs to make the network learn less discriminative object instances. Thus more possible objects belonging to the same class can be detected accordingly. Without introducing any extra count information, the proposed approach has achieved improved performance comparable to many state-of-the-art results. The object instance reweighted loss is designed to further help the OIM by learning the larger portion of the target object instances in each image. Experimental results on two publicly available datasets illustrate that the proposed approach achieves competitive or superior performance than state-of-the-art methods for WSOD.

### References

Arun, A.; Jawahar, C.; and Kumar, M. P. 2019. Dissimilarity coefficient based weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9432–9441.

Bilen, H., and Vedaldi, A. 2016. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2846–2854.

Diba, A.; Sharma, V.; Pazandeh, A.; Pirsiavash, H.; and Van Gool, L. 2017. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 914–922.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997.

- Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89(1-2):31–71.
- Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2):303–338.
- Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111(1):98–136.
- Gao, M.; Li, A.; Yu, R.; Morariu, V. I.; and Davis, L. S. 2018. C-wsl: Count-guided weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 152–168.
- Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.
- Jie, Z.; Wei, Y.; Jin, X.; Feng, J.; and Liu, W. 2017. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1377–1385.
- Li, X.; Kan, M.; Shan, S.; and Chen, X. 2019. Weakly supervised object detection with segmentation collaboration. *arXiv preprint arXiv:1904.00551*.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.
- Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 685–694.
- Pan, T.; Wang, B.; Ding, G.; Han, J.; and Yong, J. 2019. Low shot box correction for weakly supervised object detection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 890–896.
- Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.
- Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3):211–252.
- Shen, Y.; Ji, R.; Zhang, S.; Zuo, W.; and Wang, Y. 2018. Generative adversarial learning towards fast weakly supervised detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5764–5773.
- Shen, Y.; Ji, R.; Wang, Y.; Wu, Y.; and Cao, L. 2019. Cyclic guidance for weakly supervised joint detection and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 697–707.
- Singh, K. K., and Lee, Y. J. 2019. You reap what you sow: Using videos to generate high precision object proposals for weakly-supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9414–9422.
- Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2843–2851.
- Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. L. 2018. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*.
- Uijlings, J. R.; Van De Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International journal of computer vision* 104(2):154–171.
- Wan, F.; Wei, P.; Jiao, J.; Han, Z.; and Ye, Q. 2018. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1297–1306.
- Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; and Ye, Q. 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2199–2208.
- Wei, Y.; Shen, Z.; Cheng, B.; Shi, H.; Xiong, J.; Feng, J.; and Huang, T. 2018. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 434–450.
- Zhang, X.; Feng, J.; Xiong, H.; and Tian, Q. 2018a. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4262–4270.
- Zhang, Y.; Bai, Y.; Ding, M.; Li, Y.; and Ghanem, B. 2018b. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 928–936.