Collaborative Learning for Weakly Supervised Object Detection

Jiajie Wang, Jiangchao Yao, Ya Zhang*, Rui Zhang
Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, China
{ww1024,sunarker,ya_zhang,zhang_rui}@sjtu.edu.cn

Abstract

Weakly supervised object detection has recently received much attention, since it only requires image-level labels instead of the bounding-box labels consumed in strongly supervised learning. Nevertheless, the save in labeling expense is usually at the cost of model accuracy. In this paper, we propose a simple but effective weakly supervised *collaborative* learning framework to resolve this problem, which trains a weakly supervised learner and a strongly supervised learner jointly by enforcing partial feature sharing and prediction consistency. For object detection, taking WSDDN-like architecture as weakly supervised detector sub-network and Faster-RCNN-like architecture as strongly supervised detector sub-network, we propose an end-to-end Weakly Supervised Collaborative Detection Network. As there is no strong supervision available to train the Faster-RCNN-like sub-network, a new *prediction consistency loss* is defined to enforce consistency of predictions between the two sub-networks as well as within the Faster-RCNN-like sub-networks. At the same time, the two detectors are designed to partially share features to further guarantee the model consistency at perceptual level. Extensive experiments on PASCAL VOC 2007 and 2012 data sets have demonstrated the effectiveness of the proposed framework.

1 Introduction

Learning frameworks with Convolutional Neural Network (CNN) [Girshick, 2015; Ren *et al.*, 2015; Redmon and Farhadi, 2016] have persistently improved the accuracy and efficiency of object detection over the recent years. However, most existing learning-based object detection methods require strong supervisions in the form of instance-level annotations (e.g. object bounding boxes) which are labor extensive to obtain. As an alternative, weakly supervised object detection explores image-level annotations that are more accessible from rich media data [Thomee *et al.*, 2015].

A common practice for weakly supervised object detection is to model it as a multiple instance learning (MIL) problem, treating each image as a bag and the target proposals as instances. Therefore, the learning procedure is alternating between training an object classifier and selecting most confident positive instances [Bilen *et al.*, 2015; Cinbis *et al.*, 2017; Zhang *et al.*, 2006]. Recently, CNNs are leveraged for the feature extraction and classification [Wang *et al.*, 2014]. Some methods further integrate the instance selection step in deep architectures by aggregating proposal scores to image-level predictions [Wu *et al.*, 2015; Bilen and Vedaldi, 2016; Tang *et al.*, 2017] and build an efficient end-to-end network.

While the above end-to-end weakly supervised networks have shown great promise for weakly supervised object detection, there is still a large gap in accuracy compared to their strongly supervised counterparts. Several studies have attempted to combine these two detectors in a cascaded manner, aiming to further refine coarse detection results by leveraging powerful strongly supervised detectors[Tang *et al.*, 2017; Dong *et al.*, 2017]. Generally, instance-level predictions from a trained weakly supervised detector are used as pseudo labels to train strongly supervised detectors. However, these methods only consider a one-off unidirectional connection between two kind of detectors, making the prediction accuracy of the strongly supervised detectors depend heavily on that of the corresponding weakly supervised detectors.

In this paper, we propose a novel weakly supervised collaborative learning (WSCL) framework which bridges weakly supervised and strongly supervised learners in a unified learning process. The consistency of two learners, for both shared features and model predictions, is enforced under the WSCL framework. Focusing on object detection, we further develop an end-to-end weakly supervised collaborative detection network, as illustrated in Fig. 1. A WSDDN-like architecture is chosen for weakly supervised detector sub-network and a Faster-RCNN-like architecture is chosen for strongly supervised detector sub-network. During each learning iteration, the entire detection network takes *only* image-level labels as the weak supervision and the strongly supervised detector sub-network is optimized in parallel to the weakly supervised detector sub-network by a carefully designed prediction consistency loss, which enforces the consistency of instance-level predictions between and within the two detectors. At the same time, the two detectors are designed to partially

*The corresponding author

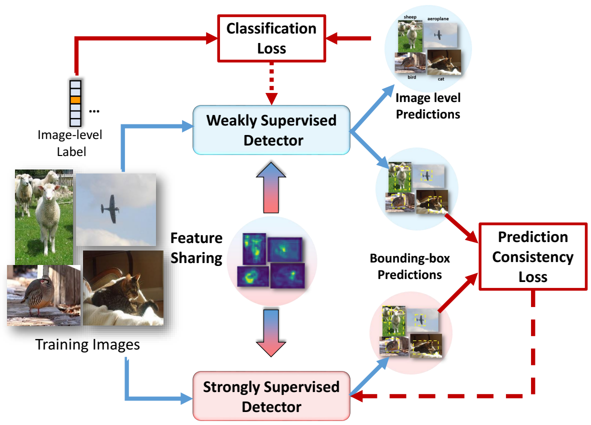

Figure 1: The proposed weakly supervised collaborative learning framework. A weakly supervised detector and a strongly supervised detector are integrated into a unified architecture and trained jointly.

share features to further guarantee the model consistency at perceptual level. Experimental results on the PASCAL VOC 2007 and 2012 data sets have demonstrated that the two detectors mutually enhance each other through the collaborative learning process. The resulting strongly supervised detector manages to outperform several state-of-the-art methods. The main contributions of the paper are summarized as follows.

- We propose a new collaborative learning framework for weakly supervised object detection, in which two types of detectors are trained jointly and mutually enhanced.
- To optimize the strongly supervised detector sub-network without strong supervisions, a prediction consistency loss is defined between the two sub-networks as well as within the strongly supervised detector sub-network.
- We experiment with the widely used PASCAL VOC 2007 and 2012 data sets and show that the proposed approach outperforms several state-of-the-art methods.

2 Weakly Supervised Collaborative Learning Framework

Given two related learners, one weakly supervised learner D_W and one strongly supervised learner D_S, we propose a weakly supervised collaborative learning (WSCL) framework to jointly train the two learners, leveraging the task similarity between the two learners. As shown in Fig. 2(a), D_W learns from weak supervisions and generates fine-grained predictions such as object locations. Due to lack of strong supervisions, D_S cannot be directly trained. But it is expected that D_S and D_W shall output similar predictions for the same image if trained properly. Hence, D_S learns by keeping its predictions consistent with that of D_W. Meanwhile, D_S and D_W are also expected to partially share feature representations as their tasks are the same. The WSCL framework thus enforces D_S and D_W to partially share network structures and parameters. Intuitively, D_S with reasonable amount of strong supervisions is expected to learn better feature representation than D_W. By bridging the two learners under this collaborative learning framework, we enable them to mutual

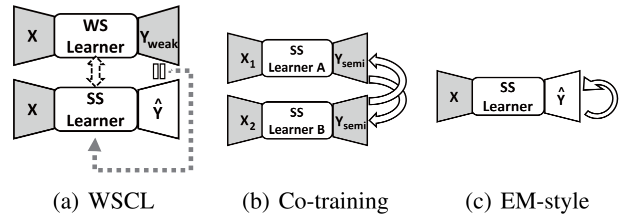

Figure 2: Comparison of WSCL with co-training and EM-style frameworks. SS denotes a strongly-supervsed learning style and WS denotes a weakly-supervised learning style. See text for details.

reinforcement each other through the joint learning process.

WSCL is similar to several learning frameworks such as co-training and the EM-style learning as shown in Fig. 2. *Co-training* framework [Blum and Mitchell, 1998] is designed for semi-supervised settings, where two parallel learners are optimized with distinct views of data. Whenever the labels in either learner are unavailable, its partner's prediction can be used for auxiliary training. Compared with the homogeneous collaboration in co-training, the WSCL framework is heterogeneous, i.e. the two learners have different types of supervisions. Moreover, two learners in WSCL are trained jointly rather than iteratively. *EM-style* framework for weakly supervised object detection task [Jie *et al.*, 2017; Yan *et al.*, 2017] usually utilizes a strongly supervised learner to iteratively select training samples according to its own predictions. However, the strongly supervised learner in this framework may not get stable training samples since it is sensitive to the initialization. By contrast, WSCL trains a weakly supervised and a strongly supervised learner jointly and enables them to mutually enhance each other.

3 Weakly Supervised Collaborative Detection

In this section, we focus on the object detection applications. Given a training set $\{(\mathbf{x_n}, \mathbf{y_n}), n = 1, \cdots, N\}$, where N is the size of training set, $\mathbf{x_n}$ is an image, and the image's label $\mathbf{y_n} \in \mathbb{R}^C$ is a C-dimensional binary vector indicating the presence or absence of each category. The task is to learn an object detector which predicts the locations of objects in an image as $\{(\mathbf{p}_i, \mathbf{t}_i), i = 1, \cdots, B\}$, where B is the number of proposal regions. And for the i-th proposal region $x^{(i)}$, $\mathbf{p}_i$ is a vector of category probability, and $\mathbf{t}_i$ is a vector of four parameterized bounding box coordinates. The image-level annotation $\mathbf{y}$ is considered as a form of weak supervisions, because the detector is also expected to predict object categories and locations in terms of bounding boxes.

Under the weakly supervised collaborative learning framework, we propose a Weakly Supervised Collaborative Detection Network (WSCDN). A two-stream CNN similar to WSDDN [Bilen and Vedaldi, 2016] is chosen as the weakly supervised learner D_W and Faster-RCNN [Ren *et al.*, 2015] is chosen as the strongly supervised learner D_S. The two learners are integrated into an end-to-end collaborative learning network. The overall architecture is illustrated in Fig. 3.

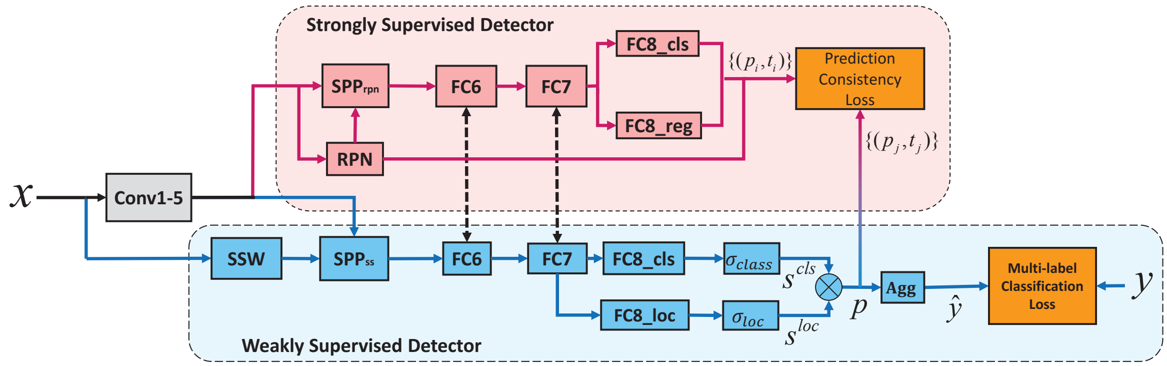

Figure 3: The architecture of our WSCDN model built based on VGG16. Red and blue lines are the forward paths for the strongly and weakly supervised detectors respectively, while black solid and dashed lines indicate the shared parts of two detectors.

3.1 Base Detectors

As shown in the blue area of Fig. 3, the weakly supervised detector D_W is composed of three parts. The first part (up to FC7) takes pre-generated proposal regions and extracts features for each proposal. The middle part consists of two parallel streams, one to compute classification score s_{jc}^{cls} and the other to compute location score s_{jc}^{loc} of each proposal region $x^{(j)}$ for category c. The last part computes product over the two scores to get a proposal's detection score p_{jc}, and then aggregates the detection scores over all proposals to generate the image-level prediction $\hat{y}_c$. Suppose the weakly supervised detector D_W has a total number of B_W proposal regions, the aggregation of prediction scores from the instance-level to the image-level can be represented as

$$\hat{y}_c = \sum_{j=1}^{B_W} p_{jc}\,, \quad \text{where} \quad p_{jc} = s_{jc}^{cls} \cdot s_{jc}^{loc}. \tag{1}$$

With the above aggregation layer, D_W can be trained in an end-to-end manner given the image-level annotations $\mathbf{y}$ and is able to give coordinate predictions directly from $x^{(j)}$ and category predictions from p_{jc}.

The network architecture of the strongly supervised detector D_S is shown in the red area of Fig. 3. Region proposal network (RPN) is used to extract proposals online. Then bounding box predictions $\{(\mathbf{p}, \mathbf{t})\}$ are made through classifying the proposals and refining their coordinates.

3.2 Collaborative Learning Network

For collaborative learning, the two learners are integrated into an end-to-end architecture as two sub-networks and trained jointly in each forward-backward iteration. Because the training data only have weak supervision in forms of classification labels, we design the following two sets of losses for model training. The first one is similar to WSDDN and many other weakly supervised detectors and the second one focuses on checking the prediction consistency, both between the two detectors and within the strongly supervised detector itself.

For the weakly supervised detector sub-network D_W, it outputs category predictions on the image level as well as location predictions on the object level. Given weak supervision $\mathbf{y}$ at the image level, we define a classification loss in the form of a multi-label binary cross-entropy loss between $\mathbf{y}$ and the image-level prediction $\hat{y}_c$ from D_W:

$$L(D_W) = -\sum_{c=1}^{C} (y_c \log \hat{y}_c + (1 - y_c) \log(1 - \hat{y}_c)). \tag{2}$$

$L(D_W)$ itself can be used to train a weakly supervised detector, as has been demonstrated in WSDDN. Under the proposed collaborative learning framework, $L(D_W)$ is also adopted to train the weakly supervised sub-network D_W.

Training the strongly supervised detector sub-network D_S independently usually involves losses consisting of a category classification term and a coordinate regression term, which requires instance-level bounding box annotations. However, the strong supervisions in terms of instance-level labels are not available in the weak settings. The major challenge for training the weakly supervised collaborative detector network is how to define loss to optimize D_S without requiring instance-level supervisions at all. Considering both D_W and D_S are designed to predict object bounding boxes, we propose to leverage the prediction consistency in order to train the strongly supervised sub-network D_S. The prediction consistency consists of two parts: between both D_W and D_S and only within D_S. The former one enforces that the two detectors give similar predictions both in object classification and object locations when converged. The latter one is included because the output of D_W is expected to be quite noisy, especially at the initial rounds of the training. Combining these above two kinds of prediction consistency, we define the loss function for training D_S as

$$L(D_S) = -\sum_{j=1}^{B_W} \sum_{i=1}^{B_S} \sum_{c=1}^{C}$$
$$I_{ij}(\beta \underbrace{p_{jc} \log p_{ic}}_{C_{inter}^{P}} + (1-\beta) \underbrace{p_{ic} \log p_{ic}}_{C_{inner}^{P}} + \underbrace{p_{jc} \mathrm{R}(\mathbf{t}_{jc} - \mathbf{t}_{ic})}_{C_{inter}^{L}}) \tag{3}$$

where the first two cross-entropy terms C_{inter}^{P} and C_{inner}^{P} consider the consistency of category predictions both on the inter and inner level; p_{jc} and p_{ic} are the category predictions from D_W and D_S respectively; the last one C_{inner}^{L} is a regression term promising the consistency of only inter-networks' localization predictions, which measures the coordinate difference between proposals from D_S and D_W. Here, $\mathrm{R}(\cdot)$ denotes a smooth L_1 loss [Girshick, 2015] and is weighted by p_j; B_W and B_S are the numbers of proposal regions for D_W and D_S in a mini-batch respectively; I_{ij} is a binary indicator with the value of 1 if the two proposal regions $x^{(i)}$ and $x^{(j)}$ are closet and have a overlap ratio (IoU) more than 0.5, and 0 otherwise; $\beta \in (0, 1)$ is a hyper parameter which balances two terms of consistency loss for category predictions. If β is larger than 0.5, D_S will trust predictions from D_W more than from itself.

Max-out Strategy. The predictions of D_S and D_W could be inaccurate, especially in the initial rounds of training. For measuring the prediction consistency, it is important to select only the most confident predictions. We thus apply a Max-out strategy to filter out most predictions. For each positive category, only the region with highest prediction score by D_W is chosen. That is, if $y_c = 1$, we have:

$$\hat{p}_{j_c^* c} = 1,\ \ s.j. \sum_j \hat{p}_{jc} = 1,\ \ \text{where}\, j_c^* = \arg\max_j\, p_{jc}. \quad (4)$$

If $y_c = 0$, we have $\hat{p}_{jc} = 0, \forall j, c$. The category prediction $\hat{p}_{jc}$ is then used to replace p_{jc} when calculating the consistency loss in $L(D_S)$. The Max-out strategy can also reduce the region numbers of D_W used to calculate the prediction consistency loss and thus can save much training time.

Feature Sharing. As the two detectors in WSCDN are designed to learn under different forms of supervision but for the same prediction task, the feature representations learned through the collaboration process are expected to be similar to each other. We thus enforce the partial feature sharing between two sub-networks so as to ensure the perceptual consistency of the two detectors. Specifically, the weights of convolutional (conv) layers and part of bottom fully-connected (fc) layers are shared between D_W and D_S.

Network Training. With the image-level classification loss $L(D_W)$ and instance-level prediction consistency loss $L(D_S)$, the parameters of two detectors can be updated jointly with only image-level labels by the stochastic gradient descent (SGD) algorithm. The gradients for individual layers of D_S and D_W are computed only respect to $L(D_S)$ and $L(D_W)$ respectively, while the shared layers' gradients are produced by both loss functions.

4 Experimental Results

4.1 Data Sets and Metrics

We experiment with two widely used benchmark data sets: PASCAL VOC 2007 and 2012 [Everingham *et al.*, 2010], both containing 20 common object categories with a total of 9,962 and 22,531 images respectively. We follow the standard splits of the data sets and use the *trainval* set with only image-level labels for training and the *test* set with ground-truth bounding boxes for testing.

Two standard metrics, Mean average precision (mAP) and Correct localization (CorLoc) are adopted to evaluate different weakly supervised object detection methods. The mAP measures the quality of bounding box predictions in test set.

Methods	I_W	CL_W	CL_S	CS_S
mAP(%)	28.5	40.0	**48.3**	39.4
CorLoc(%)	45.6	58.4	**64.7**	59.3

Table 1: Comparison of detectors built with the WSCL framework to their baselines and counterparts in terms of mAP and CorLoc on PASCAL VOC 2007 data set.

Following [Everingham *et al.*, 2010], a prediction is considered as true positive if its IoU with the target ground-truth is larger than 0.5. CorLoc of one category is computed as the ratio of images with at least one object being localized correctly. It is usually used to measure the localization ability in localization tasks where image labels are given. Therefore, it is a common practice to validate the model's CorLoc on training set [Deselaers *et al.*, 2012].

4.2 Implementation Details

Both the weakly and strongly supervised detectors in the WSCDN model are built on VGG16 [Simonyan and Zisserman, 2014], which is pre-trained on a large scale image classification data set, ImageNet [Russakovsky *et al.*, 2015]. We replace Pool5 layer with SPP layer [He *et al.*, 2014] to extract region features. Two detectors share weights for convolutional (conv) layers and two fully-connected (fc) layers, i.e., *fc6*, *fc7*. For the weakly supervised detector, we use SelectiveSearch [Uijlings *et al.*, 2013] to generate proposals and build network similar with WSDDN: the last fc layer in VGG16 is replaced with a two-stream structure in 3.1, as each stream consists a fc layer followed by a softmax layer focusing on classification and localization respectively. For the strongly supervised detector Faster-RCNN, we follow the model structure and setting of its original implementation.

At training time, we apply image multi-scaling and random horizontal flipping for data augmentation, with the same parameters in [Ren *et al.*, 2015]. We empirically set the hyper parameter β to 0.8. RPN and the following region-based detector in Faster-RCNN are trained simultaneously. We train our networks for total 20 epochs, setting the learning rate of the first 12 epochs to 1e-3, and the last 8 epochs to 1e-4. At test time, we obtain two sets of predictions for each image from the weakly and strongly supervised detectors, respectively. We apply non-maximum suppression to all predicted bounding boxes, with the IoU threshold set to 0.6.

4.3 Influence of Collaborative Learning

To investigate the effectiveness of the collaborative learning framework for weakly supervised object detection, we compare the following detectors: 1) the weakly and strongly supervised detectors built with the collaborative learning framework, denoted as CL_W and CL_S, respectively; 2) The initial weakly supervised detector built above, denoted as I_W; 3) The same weakly supervised and strongly supervised detector networks trained in cascaded manner similar to [Tang *et al.*, 2017; Yan *et al.*, 2017]. The resulting strongly supervised detector is denoted as CS_S.

The results on PASCAL VOC 2007 data set are presented in Table 1. Among the four detectors under comparison, CL_S achieves the best performance in terms of mAP and CorLoc

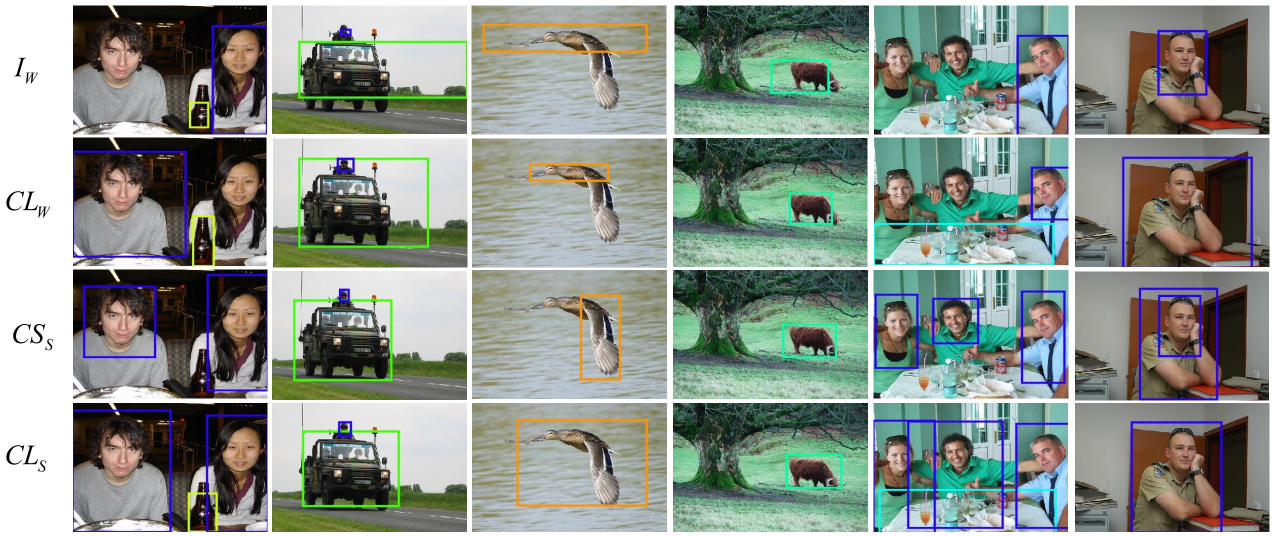

Figure 4: Visualization of the detection results of four detectors in Table 1. Images from the 1st to 4th row are results from the I_W, CL_W, CS_S and CL_S respectively.

and outperforms the baseline I_W, its collaborator CL_W, and its cascade counterpart, CS_S. Compared to CS_S, the mAP and CorLoc are improved from 39.4% to 48.3% and from 59.3% to 64.7%, respectively, suggesting the effectiveness of the proposed collaborative learning framework. Furthermore, CL_W outperforms I_W in terms of mAP and CorLoc by a large margin of 11.5% and 12.8%, respectively, showing that the parameters sharing between the two detectors enables a better feature representation and thus leading to significant improvement of the weakly supervised detector.

We also qualitatively compare the detection results of I_W, CL_W, CS_S and CL_S. As shown in Fig. 4, the strongly supervised detector CL_S clearly outperforms the other three detectors, with more objects correctly detected. For example, in the first column and fifth column where there are multiple objects in one images, only CL_S is able to correctly detect all of them, while the other three detectors missed one or more objects. Moreover, CL_S generates more accurate bounding boxes. Weakly supervised detectors are known for often generating bounding boxes that only cover the most discriminate part of an object (e.g. face of a person or wings/head of a bird). CL_S can generate more bounding boxes that tightly cover the entire objects as shown in the third and sixth column of Fig. 4, indicating the collaborative learning framework is able to learn a better feature representation for objects. Compared to I_W, CL_W generates tighter object bounding box in the second and fourth columns, i.e. the performance of the weakly supervised detector is improved after collaborative learning, suggesting that feature sharing between the two detectors helps optimizing the weakly supervised detector.

To show how CL_W and CL_S improve during the collaborative learning, we plot mAPs of the two detectors for different training iterations. As shown in Fig. 5, both detectors get improved with increasing training iterations. Initially, the strongly supervised detector CL_S has a smaller mAP than the weakly supervised detector CL_W. But in a dozen thousands iterations, CL_S surpasses CL_W and further outperforms CL_W by a large margin in the end, suggesting the effectiveness of the prediction consistency loss we proposed.

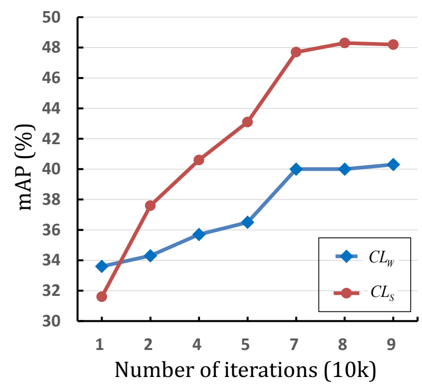

Figure 5: The changes of mAP for CL_S and CL_W on PASCAL VOC 2007 data set during the process of collaborative learning.

4.4 Comparison with State-of-the-Arts

In general, two types of weakly supervised object detection methods are compared. The first includes the MIL methods [Cinbis *et al.*, 2017; Wang *et al.*, 2014] and various end-to-end MIL-CNN models [Bilen and Vedaldi, 2016; Kantorov *et al.*, 2016; Tang *et al.*, 2017] following the two-stream structure of WSDDN [Bilen and Vedaldi, 2016]. The second type of methods builds a curriculum pipeline to find confident regions online, and trains an instance-level modern detector in a strongly supervised manner [Li *et al.*, 2016; Jie *et al.*, 2017]. So the detectors they used to report the results share a similar structure and characteristics with our strongly supervised detector.

For the PASCAL VOC 2007 dataset, the mAP and CorLoc results are shown in Table 2 and Table 3, respectively. The

Methods	**aer**	**bik**	**brd**	**boa**	**btl**	**bus**	**car**	**cat**	**cha**	**cow**	**tbl**	**dog**	**hrs**	**mbk**	**prs**	**plt**	**shp**	**sfa**	**trn**	**tv**	**Avg.**
[Cinbis *et al.*, 2017]	38.1	47.6	28.2	13.9	13.2	45.2	48.0	19.3	17.1	27.7	17.3	19.0	30.1	45.4	13.5	17.0	28.8	24.8	38.2	15.0	27.4
[Wang *et al.*, 2014]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	52.7	**19.1**	17.4	35.9	33.3	34.8	46.5	31.6
[Bilen and Vedaldi, 2016]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
[Kantorov *et al.*, 2016]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	**49.2**	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
[Tang *et al.*, 2017]	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
[Li *et al.*, 2016]	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
[Jie *et al.*, 2017]	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
CL_W	59.7	54.7	31.6	24.1	13.2	59.6	53.2	39.0	19.3	49.9	35.8	45.0	38.2	63.6	7.1	16.9	36.6	47.9	54.9	50.0	40.0
CL_S	**61.2**	**66.6**	**48.3**	26.0	15.8	**66.5**	65.4	53.9	24.7	**61.2**	46.2	**53.5**	48.5	**66.1**	12.1	22.0	49.2	**53.2**	66.2	59.4	**48.3**

Table 2: Comparison of WSCDN to the state-of-the-art on PASCAL VOC 2007 *test* set in terms of average precision (AP) (%).

Methods	**aer**	**bik**	**brd**	**boa**	**btl**	**bus**	**car**	**cat**	**cha**	**cow**	**tbl**	**dog**	**hrs**	**mbk**	**prs**	**plt**	**shp**	**sfa**	**trn**	**tv**	**Avg.**
[Cinbis *et al.*, 2017]	57.2	62.2	50.9	37.9	23.9	64.8	74.4	24.8	29.7	64.1	40.8	37.3	55.6	68.1	25.5	38.5	65.2	35.8	56.6	33.5	47.3
[Wang *et al.*, 2014]	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
[Bilen and Vedaldi, 2016]	65.1	58.8	58.5	33.1	**39.8**	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
[Kantorov *et al.*, 2016]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	**47.1**	59.5	67.0	83.5	**35.3**	39.9	67.0	49.7	63.5	65.2	55.1
[Tang *et al.*, 2017]	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
[Li *et al.*, 2016]	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
[Jie *et al.*, 2017]	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
CL_W	82.5	75.7	63.1	44.1	32.4	72.1	76.7	50.3	35.0	74.0	30.8	57.9	57.5	82.3	19.1	47.6	76.3	50.0	71.1	69.5	58.4
CL_S	**85.8**	**80.4**	**73.0**	42.6	36.6	79.7	82.8	**66.0**	34.1	78.1	36.9	**68.6**	**72.4**	**91.6**	22.2	51.3	**79.4**	**63.7**	74.5	74.6	**64.7**

Table 3: Comparison of WSCDN to the state-of-the-art on PASCAL VOC 2007 *trainval* set in terms of Correct Localization (CorLoc) (%).

Methods	**aer**	**bik**	**brd**	**boa**	**btl**	**bus**	**car**	**cat**	**cha**	**cow**	**tbl**	**dog**	**hrs**	**mbk**	**prs**	**plt**	**shp**	**sfa**	**trn**	**tv**	**Avg.**
[Kantorov *et al.*, 2016]	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 *et al.*, 2017]	67.7	61.2	41.5	**25.6**	**22.2**	54.6	49.7	25.4	19.9	47.0	18.1	26.0	38.9	67.7	2.0	**22.6**	41.1	34.3	37.9	55.3	37.9
[Jie *et al.*, 2017]	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
CL_W	64.0	60.3	40.1	18.5	15.0	57.4	38.3	25.3	17.3	32.4	16.5	33.1	28.6	64.8	6.9	16.6	34.3	**41.4**	52.4	51.2	35.7
CL_S	**70.5**	**67.8**	**49.6**	20.8	22.1	**61.4**	51.7	34.7	**20.3**	50.3	19.0	43.5	49.3	**70.8**	10.2	20.8	**48.1**	41.0	**56.5**	**56.7**	**43.3**

Table 4: Comparison of WSCDN to the state-of-the-art on PASCAL VOC 2012 *test* set in terms of average precision (AP) (%).

Methods	**aer**	**bik**	**brd**	**boa**	**btl**	**bus**	**car**	**cat**	**cha**	**cow**	**tbl**	**dog**	**hrs**	**mbk**	**prs**	**plt**	**shp**	**sfa**	**trn**	**tv**	**Avg.**
[Kantorov *et al.*, 2016]	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
[Tang *et al.*, 2017]	86.2	84.2	68.7	**55.4**	**46.5**	82.8	**74.9**	32.2	46.7	**82.8**	42.9	41.0	68.1	**89.6**	9.2	**53.9**	81.0	52.9	59.5	**83.2**	62.1
[Jie *et al.*, 2017]	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
CL_W	88.0	79.7	66.4	51.0	40.9	84.0	65.4	35.6	46.5	69.9	46.6	49.7	52.4	89.2	21.2	47.2	73.3	54.8	70.5	75.5	60.4
CL_S	**89.2**	**86.0**	**72.8**	50.4	40.1	**87.7**	72.6	37.0	**48.2**	80.3	**49.3**	54.4	72.7	88.8	**21.6**	48.9	**85.6**	**61.0**	**74.5**	82.2	**65.2**

Table 5: Comparison of WSCDN to the state-of-the-art on PASCAL VOC 2012 *trainval* set in terms of Correct Localization (CorLoc) (%).

proposed model gets 39.4% and 49.4% in terms of map for the weakly supervised detector and the strongly supervised detector respectively. On CorLoc, our two detectors also perform well, get 61.1% and 67.5%. In particular, the strongly supervised detector CL_S in our model receives best results among those methods by both mAP and CorLoc.

Compared to the first type of methods, CL_S improves detection performance by more than 7.1% on mAP and 4.1% on CorLoc. Our CL_W that has a similar but the simplest structure, also gets comparable results with regard to other modified models, revealing the mutual enhancement of two kinds of detectors with collaborative learning. With respect to the second set of methods under comparison, we use a weakly supervised detector to achieve confident region selection in a collaboration learning process, instead of those complicated schemes. The collaborative learning framework enables the strongly supervised detector CL_S to outperform [Jie *et al.*, 2017] by 6.6% and 8.6% on mAP and CorLoc respectively.

Similar results are obtained on PASCAL VOC 2012 dataset as shown in Table 4 and Table 5. CL_S achieved 43.3% on mAP and 65.2% on CorLoc, both of which outperform the other state-of-the-art methods, indicating the effectiveness of the collaborative learning framework.

5 Conclusion

In this paper, we propose a simple but effective WSCL framework for weakly supervised object detection, in which two detectors with different mechanics and characteristics are integrated in a unified architecture. In particular, we propose an end-to-end Weakly Supervised Collaborative Detection Network (WSCDN). The weakly supervised learner, WSDDN-like sub-network, is trained with the image-level classification loss. To train the strongly supervised learner, Faster-RCNN-like sub-network, a new *prediction consistency loss* is defined to enforce the prediction consistency of the two networks. Moreover, the two learners are required to partially share parameters to achieve feature sharing. Extensive experiments on benchmark data sets have shown that WSCDN outperforms the state-of-the-arts. The weakly supervised detector and the strongly supervised detector are also shown to benefit each other in the collaborative learning process.

Acknowledgments

This work is supported by The High Technology Research and Development Program of China (2015AA015801), NSFC (61521062), and STCSM (18DZ2270700).

References

[Bilen and Vedaldi, 2016] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016.

[Bilen *et al.*, 2015] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with convex clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1081–1089, 2015.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[Cinbis *et al.*, 2017] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2017.

[Deselaers *et al.*, 2012] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100(3):275–293, 2012.

[Dong *et al.*, 2017] Xuanyi Dong, Deyu Meng, Fan Ma, and Yi Yang. A dual-network progressive approach to weakly supervised object detection. pages 279–287, 2017.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[He *et al.*, 2014] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361. Springer, 2014.

[Jie *et al.*, 2017] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. *arXiv preprint arXiv:1704.05188*, 2017.

[Kantorov *et al.*, 2016] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, pages 350–365. Springer, 2016.

[Li *et al.*, 2016] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016.

[Redmon and Farhadi, 2016] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Tang *et al.*, 2017] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. *arXiv preprint arXiv:1704.00138*, 2017.

[Thomee *et al.*, 2015] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li Jia Li. The new data and new challenges in multimedia research. *Communications of the Acm*, 59(2):64–73, 2015.

[Uijlings *et al.*, 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[Wang *et al.*, 2014] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *European Conference on Computer Vision*, pages 431–445. Springer, 2014.

[Wu *et al.*, 2015] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2015.

[Yan *et al.*, 2017] Ziang Yan, Jian Liang, Weishen Pan, Jin Li, and Changshui Zhang. Weakly-and semi-supervised object detection with expectation-maximization algorithm. *arXiv preprint arXiv:1702.08740*, 2017.

[Zhang *et al.*, 2006] Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424, 2006.