JCST
Vol.34 No.6 Nov. 2019

ISSN 1000-9000(Print)
/1860-4749(Online)
CODEN JCTEEM

Journal of Computer Science & Technology

SPONSORED BY INSTITUTE OF COMPUTING TECHNOLOGY
THE CHINESE ACADEMY OF SCIENCES &
CHINA COMPUTER FEDERATION
SUPPORTED BY NSFC
CO-PUBLISHED BY SCIENCE PRESS &
SPRINGER

Wang XG, Wang JS, Tang P *et al.* Weakly- and semi-supervised fast region-based CNN for object detection. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 34(6): 1269–1278 Nov. 2019. DOI 10.1007/s11390-019-1975-z

Weakly- and Semi-Supervised Fast Region-Based CNN for Object Detection

Xing-Gang Wang, *Member*, *CCF*, Jia-Si Wang, Peng Tang, and Wen-Yu Liu*, *Member*, *CCF*

School of Electronic Information and Communications, Huazhong University of Science and Technology Wuhan 430074, China

E-mail: {xgwang, wangjiasi, pengtang, liuwy}@hust.edu.cn

Received March 23, 2019; revised July 26, 2019.

Abstract Learning an effective object detector with little supervision is an essential but challenging problem in computer vision applications. In this paper, we consider the problem of learning a deep convolutional neural network (CNN) based object detector using weakly-supervised and semi-supervised information in the framework of fast region-based CNN (Fast R-CNN). The target is to obtain an object detector as accurate as the fully-supervised Fast R-CNN, but it requires less image annotation effort. To solve this problem, we use weakly-supervised training images (i.e., only the image-level annotation is given) and a few proportions of fully-supervised training images (i.e., the bounding box level annotation is given), that is a weakly- and semi-supervised (WASS) object detection setting. The proposed solution is termed as WASS R-CNN, in which there are two main components. At first, a weakly-supervised R-CNN is firstly trained; after that semi-supervised data are used for finetuning the weakly-supervised detector. We perform object detection experiments on the PASCAL VOC 2007 dataset. The proposed WASS R-CNN achieves more than 85% of a fully-supervised Fast R-CNN's performance (measured using mean average precision) with only 10% of fully-supervised annotations together with weak supervision for all training images. The results show that the proposed learning framework can significantly reduce the labeling efforts for obtaining reliable object detectors.

Keywords object detection, weakly-supervised learning, semi-supervised learning, fast region-based convolutional neural network (Fast R-CNN)

1 Introduction

Significant progress in object detection[1−6] has been witnessed, thanks to deep learning methods and the availability of large-scale datasets with accurate bounding-box-level annotations. However, collecting such accurate annotations can be very labor-intensive and time-consuming, whereas achieving only image-level annotations (i.e., image tags) is much easier, as these annotations are often available at the Internet (e.g., image search queries or user tags).

Using image-level annotations to train object detectors is called weakly-supervised object detection[7,8]. Recently, most of the weakly-supervised object detectors are built upon the fast region-based CNN detector (Fast R-CNN)[1]. They train the R-CNN framework given image-level supervision and consider region scores as latent variables; the latent variables are placed in the detection network and various kinds of region score normalization strategies are adopted. The principle behind this weakly-supervised detector learning problem is multi-instance learning (MIL)[9], where an input image with multiple region proposals is considered as a bag and a proposal in the weakly-supervised detection network is considered as an instance.

The online instance classifier refinement (OICR) scheme[8] is a widely applied framework for weakly-supervised object detection, which improves [7] by proposing a novel online proposal score updating strategy. The main crux of OICR is to propagate the most confident detections to their spatial neighboring boxes. This procedure can significantly reduce the ambiguity of instance/region proposal scores to obtain a better

Regular Paper

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61876212, 61733007, and 61572207, and the National Key Research and Development Program of China under Grant No. 2018YFB1402604.

*Corresponding Author

weakly-supervised detection network. However, the detection performance of OICR is still far from real application requirements. For example, as reported in [8], the detection performance of Fast R-CNN using VGG16 as a base network measured by mAP is 0.69, while the mAP of OICR is 0.41, trained and tested on the PASCAL VOC 2007 dataset.

To bridge the performance gap, we propose to incorporate semi-supervised information by adding a small number of training images that are fully labeled with the bounding box annotations. The semi-supervised information can remedy the flaw in weakly-supervised detector learning and therefore increase the detection accuracy. We call this setting as weakly- and semi-supervised (WASS) object detection.

In this paper, we solve the WASS object detection problem in a principled way by decomposing it into two sub-problems, which are the existing weakly-supervised detector learning problem and a semi-supervised detector refinement problem. The proposed detector named WASS R-CNN entails the following contributions.

1) It is the first work that proposes to learn an end-to-end deep CNN based object detector in a WASS setting and provides useful insights towards creating an effective object detector while greatly reducing the human labeling efforts.

2) It sheds light on a general method to utilize weak supervision and partial labeling for most deep learning models.

3) Tested on a challenging object detection benchmark (PASCAL VOC 2007), it outperforms the state-of-the-art by a significant margin with very few full annotations.

2 Related Work

Reducing supervision in learning effective object detectors is an important and practical topic in computer vision and machine learning. Plenty of studies have been proposed to address this problem from different aspects. In this section, we provide a brief review of some representative examples.

From the aspect of weakly-supervised object detection, one of the earliest researches is from Viola *et al.*[10] which proposed a multi-instance boosting framework to learn an object detector from imprecise bounding box annotations. Then, Zhu *et al.*[11] proposed to use the bottom-up salient object proposals to guide the weakly-supervised detector learning. Following the popularity of object proposals in weakly-supervised object detection, a series of research papers integrate object proposal, multi-instance learning and deep CNNs for weakly-supervised object detection[7,8,12−15]. These papers have witnessed a gradual improvement in the performance of weakly-supervised object detection. Among these methods, the most related one is the multi-fold MIL method[15] in which a multi-fold training scheme is proposed to address the problem that MIL algorithms are easy to converge to poor local optima. It also studies the problem of learning object detector using a mixed weakly- and semi-supervised setting. However, the multi-fold MIL method[15] cannot achieve end-to-end learning and feature sharing for object proposals, leading to limited performance and speed. Besides, our study shows that by end-to-end stochastic gradient descent (SGD) training and finetuning from a model pre-trained on ImageNet, the problem of converging to poor local optima becomes less important.

Besides only using image-level supervision for object detection, there are some other interesting papers toward reducing supervision in training effective object detectors. For example, Redmon and Farhadi[2] proposed to train a 9 000-class object detector with some object classes having full supervision and the other object classes only containing image-level supervision. Note that the WASS setting in [2] is different from ours. Weak supervision from human input has also been considered to improve the training. Papadopoulos *et al.*[16] proposed to use human verification in the process of weakly-supervised detector training, which requires manual inspection in the loop. Further, Papadopoulos *et al.*[17] proposed to use click supervision for training object detectors, which requires to click all bounding boxes. However, in this paper, we exclude any manual supervision and focus on the combination of image-level and bounding-box-level annotations. The above methods focus on learning robust models to extract high-level semantics given the weak annotations.

3 Method

In this section, we present details about how to train a WASS R-CNN using weakly- and semi-supervised training images for object detection. The weakly supervised training images have image-level annotations, i.e., the object category annotations. The semi-supervised information means that a proportion of training images are fully labeled for object detection. We assume $\alpha \in (0, 1)$ portion of training images have bounding-box-level annotations. We aim at closing the performance gap between fully-supervised object detection and weakly-supervised object detection with only a

few extra bounding box level annotations. We propose WASS R-CNN based on a recent proposed weakly-supervised object detector, OICR[8], and the famous fully-supervised object detector, Fast R-CNN[1]. OICR and Fast R-CNN have a common base network, which contains a deep CNN (e.g., the VGG-Net[18]) and a spatial pyramid pooling (SPP) layer[19] to generate feature vectors for each object proposal[20]. The OICR detection network contains a classification subnet and several detection subnets. A WASS R-CNN is trained in a stage-wise fashion (see Fig.1). Specifically, training a WASS R-CNN contains the following three stages.

1) training a weakly-supervised OICR detection network using image-level annotations;

2) only finetuning the k-th detection subnet using the partial fully-supervised training images;

3) training a Fast R-CNN detector using the noisy bounding boxes produced by the detector in the previous step as ground truth on all training images.

In the following parts, we will give the details of the three stages.

3.1 Weakly-Supervised Learning: The OICR Detector

Since image-level annotations are easy to be obtained and recent studies show that weakly-supervised detectors have obtained impressive object detection results[7,8], we train a weakly-supervised object detection network as the first component in our WASS R-CNN. The weakly-supervised network serves as a ground base of the proposed detection system. In this paper, we adopt the most recent weakly-supervised detection network, the OICR detector[8], which obtained the state-of-the-art performance. OICR contains a base CNN for object proposal feature extraction, a subnet with an image classification loss function, and several subnets with object detection loss functions. We name them base network, classification subnet, and detection subnets, respectively. The classification subnet follows the design in [7], which is formulated as a multi-instance learning problem. In this classification subnet, the object proposals (instances) are classified into object classes and background to achieve object detection by combining with non-maximum suppression. The merit of OICR is a novel detection subnet cascade. The first detection subnet is supervised by the classification subnet, and the k-th detection subnet is supervised by the $(k-1)$-th detection subnet. The overall structure of OICR is shown in Fig.1. The network cascades in OICR are trained online in an end-to-end fashion.

Given a training image I of OICR, it has image-level label $\boldsymbol{y}$. We first generate a set of object pro-

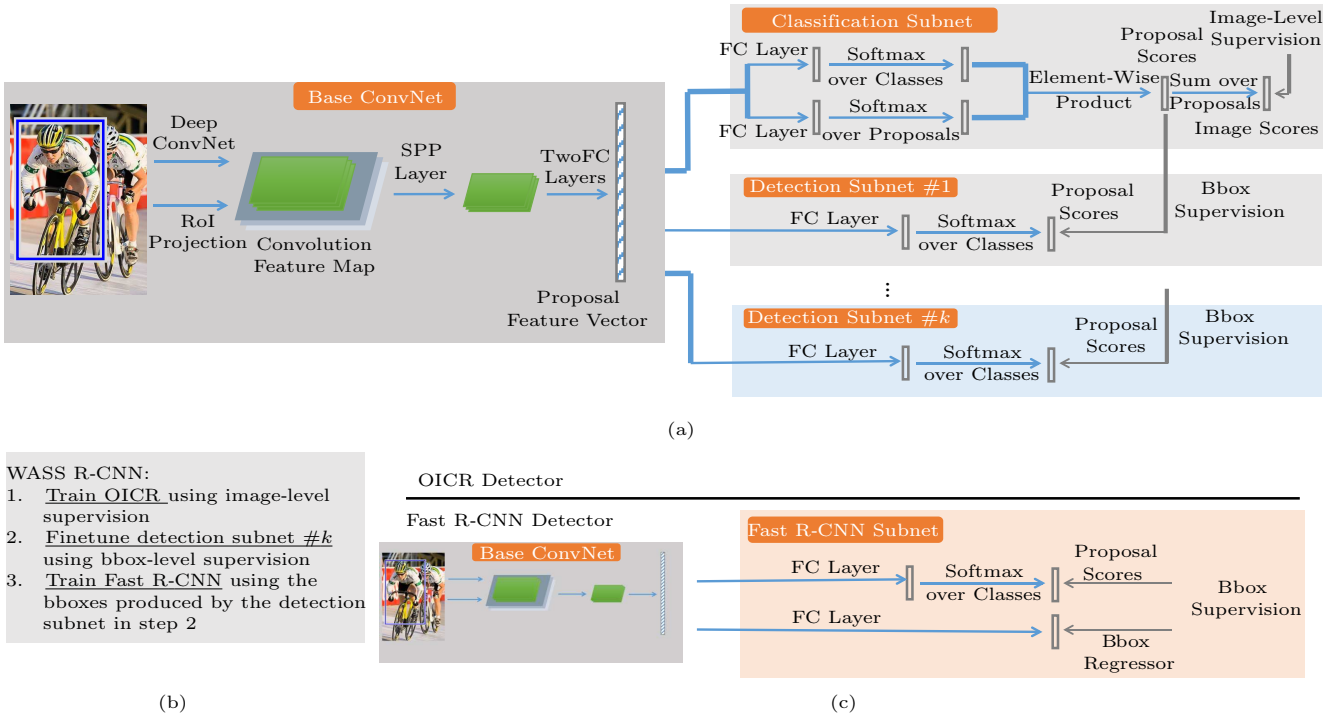


Fig.1. Network cascades of WASS R-CNN together with the stage-wise training strategy. Bbox: bounding box. TwoFC: two fully-connected.

posals $R = \{R_1, R_2, ..., R_n\}$ from I using the selective search method[20]. There are around 2 000 proposals per image. Here R_i only contains two coordinates $R_i = \{R_1^{x1}, R_1^{y1}, R_1^{x2}, R_1^{y2}\}$. Then the image is put into a series of convolutional layers to produce a feature map. Next, the feature map and the proposals are fed into an ROI pooling layer[19] and two fully-connected layers to obtain a fixed-size feature vector for each of the proposals, which is called the proposal feature vector. The proposal feature vectors in I are fed into the classification subnet and the detection subnets. Without the ground-truth bounding boxes, the proposal with the highest score may cover the most discriminative part of the object rather than the whole object. Proposals with high overlap may cover the whole object, or at least contain a larger portion of the object; therefore the cascaded detection subnets are proposed to refine the classification results. For details of the detection subnets and the effectiveness, please refer to the paper [8] and its source code①.

3.1.1 Classification Subnet

In the classification subnet, it contains two streams to produce two matrices $\boldsymbol{X}^c, \boldsymbol{X}^d \in \mathbb{R}^{C\times|R|}$, in which C denotes the number of object categories and $|R|$ denotes the number of proposals in the image. $\boldsymbol{X}^c$ stores the class score for each proposal and is passed through a softmax operator $[\sigma(\boldsymbol{X}^c)]_{ij} = \frac{e^{\boldsymbol{X}^c_{ij}}}{\sum_{k=1}^{C} e^{\boldsymbol{X}^c_{kj}}}$ to represent which class the proposal belongs to. $\boldsymbol{X}^d$ stores the proposal score for each class and is passed through a softmax operator $[\sigma(\boldsymbol{X}^d)]_{ij} = \frac{e^{\boldsymbol{X}^d_{ij}}}{\sum_{k=1}^{|R|} e^{\boldsymbol{X}^d_{ik}}}$. This softmax operator chooses the proposal that has the highest contribution to correctly classify the image. The final score of each proposal is obtained by the element-wise product $\boldsymbol{X}^R = \sigma(\boldsymbol{X}^c) \odot \sigma(\boldsymbol{X}^d)$. Then the classification scores of I over all the classes, which is a C-dimension vector, can be obtained by summing over all proposals: $\phi_c = \sum_{r=1}^{|R|} \boldsymbol{X}^R_{cr}$. Since we have the image-level label $\boldsymbol{y}$, the classification subnet can be supervised by the image-level label $\boldsymbol{y} = (y_1, y_2, ..., y_C)^{\mathrm{T}} \in \mathbb{R}^{C\times 1}$ with a standard multi-class cross entropy loss shown in (1) as follows:

$$L_{\mathrm{cls}} = -\sum_{c=1}^{C}\{y_c \log \phi_c + (1-y_c)\log(1-\phi_c)\}. \quad (1)$$

For more details of the classification subnet, please refer to the original WSDDN paper[7].

3.1.2 Detection Subnet

Training the detection subnet is essentially different from training the classification network, as we need the label of each object proposal. Specifically, the loss function is different. The score vector from the k-th detection subnet for proposal j is $\boldsymbol{x}_j^{Rk} \in \mathbb{R}^{(C+1)\times 1}$ which is a $(C+1)$-dimensional vector. Meanwhile, for the detection subsets, each proposal has its own supervision. That is, in the k-th detection subset, the supervision of proposal j is a label vector $\boldsymbol{y}_j^k = (y_{1j}^k, y_{2j}^k, ..., y_{(C+1)j}^k)^{\mathrm{T}} \in \mathbb{R}^{(C+1)\times 1}$, where $y_{cj}^k \in \{0,1\}$ indicates whether proposal j belongs to class c. But for weakly-supervised object detection, such proposal labels are unaccessible as there are no bounding box-level annotations. To deal with this problem, if an image has object class c and the proposal j has the highest score for class c from the $(k-1)$-th subnet, the proposals which have high spatial overlaps with j are classified into class c (i.e., $y_{cj}^k = 1$ and $y_{c'j}^k = 0, c' \neq c$), and the other proposals are labeled as background (i.e., $y_{0j}^k = 1$ and $y_{c'j}^k = 0, c' \neq 0$). Then the K-th classifier takes these pseudo labels as supervision to refine itself with a well-designed weighted loss, as shown in (2), and produce pseudo supervision for the $(k+1)$-th subnet at the same time, and so on, where r denotes the index of the top ranking proposal, x_{cr}^{Rk} is the score of the top ranking proposal, y_{cr}^k is the label of the top ranking proposal, and w_r^k is a weight parameter. For more details, please refer to [8]. Therefore the loss function for detection network is,

$$L_{\mathrm{det}}^k = -\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c=1}^{C+1} w_r^k y_{cr}^k \log x_{cr}^{Rk}. \quad (2)$$

After that, the loss of the whole OICR network is defined as (3).

$$L_{\mathrm{oicr}} = L_{\mathrm{cls}} + \sum_{k=1}^{K} L_{\mathrm{det}}^k. \quad (3)$$

3.2 Semi-Supervised Learning: Finetuning the Detection Subnet

Though some promising results are obtained by weakly-supervised object detectors, the performance is still limited compared with the fully-supervised methods. For example, the state-of-the-art weakly-supervised detection result is 47% mAP on PASCAL VOC 2007[8], whereas the fully-supervised one, i.e., Fast R-CNN, obtains 66.9% mAP[1]. To improve the results

①https://github.com/ppengtang/oicr, Oct. 2019.

of weakly-supervised ones, here we put forward a semi-supervised method, which can use only partial fully-labeled data to obtain competitive performance, which can liberate the requirement of fully-labeled data to some extent.

In a given image classification task with limited training examples, an effective and efficient way to obtain good performance is to finetune a deep CNN from its pre-trained model, which adapts the pre-trained model to the target domain.

Similar to this idea, here we use these partial fully labeled data to finetune the last detection subnet of OICR. Since there is a notable drawback in the classification subnet when evaluating the contributions of each proposal for judging whether the image contains an object of this class, a softmax operator is used, that is, for each class only one proposal that tightest contains the object of this class is picked out. This does not make sense for the situation that multiple objects of the same class appear in an image at the same time, unfortunately, which is quite often in the natural scene. When this happens, the performance of the classification subnet is degraded; therefore the performance of the whole network is limited. While the detection subnets do not have that kind of softmax operators, which avoids this limitation to some degree. And after the pre-training process, the last detection subnet performs pretty well (on PASCAL VOC 2007, mAP obtained by the last detection subnet is 38.5% and by the whole network is 39.0%).

Therefore, we only pick the last detection subnet, the strongest detector in OICR, and finetune it using the partial fully-supervised training images. In this case, as each proposal only has one single label, we choose the popular softmax loss as in (4). In the experiments, we can observe that the performance of the K-th detection subnet is significantly boosted only using 10% random selected fully-supervised training images.

$$L_{\text{det}}^{k} = -\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c=1}^{C+1} y_{cr}^{k}\log x_{cr}^{Rk}. \tag{4}$$

Different from the traditional semi-supervised learning setting which has a part of fully-supervised training data and the rest unlabeled training data, in this paper, we are in a weakly- and semi-supervised supervised (WASS) setting. In this setting, semi-supervised learning may take advantage of transfer learning. In the deep CNN framework, transfer learning is mainly implemented based on finetuning[21].

3.3 Refining Detection Using Fast R-CNN

The merit of semi-supervised learning is propagating the labeling information of the fully-supervised training data to the rest of the training data without full supervision. In WASS R-CNN, after finetuning the detection subnet, we can get better bounding box labels for the weakly labeled training images.

In the third stage, we test the finetuned detection subnet on all training images and apply NMS to obtain detection results. The detection results are used as the ground truth for images with only image-level labels to train a Fast R-CNN detector, i.e., using the top-scoring proposals as ground truths for weakly-labeled images. This is a noisy training for Fast R-CNN. The purpose of this stage is to enjoy the bounding box regression benefit in Fast R-CNN since in the previous two stages, the pseudo ground-truth is too rough to conduct bounding box regression.

Reasons for Using OICR and Fast R-CNN. Although OICR is not most recently proposed, it is simple and efficient. Almost all recent weakly-supervised object detection methods are based on OICR. Fast R-CNN is also not new enough compared with some other object detection methods, such as SSD[2], YOLO[3] and Faster R-CNN[22]. However, Fast R-CNN can take advantage of the classic algorithm to produce reliable proposals based on low-level cues in images, such as selective search. When the ratio of supervised images is small, i.e., a few annotated bounding boxes, learning to generate proposals in the way of SSD[2], YOLO[3] and Faster R-CNN[22] is not possible.

4 Experiments

In this section, we will present the results and detailed analysis of the proposed WASS R-CNN.

4.1 Datasets and Evaluation Measures

We evaluate our method on PASCAL VOC 2007 dataset[23], which is a very challenging and commonly used benchmark for object detection in images. The dataset contains 2 501 training images, 2 510 validation images and 4 952 test images in 20 object classes. For each image, bounding box annotations are given in the task of object detection. In the experiments, for all object detection methods, the train and val (validation) sets are used for training and the test set is used for evaluation. Thus, it has 5 011 images for training and 4 092 images for testing. The evaluation metric

is the mean of average precision (mAP). The standard PASCAL criterion mAP@0.5, i.e., IoU > 0.5 between ground-truth bounding boxes and predicted boxes, is used for evaluation.

4.2 Implementation Details

As described in the method section, WASS R-CNN contains three stages: the OICR detector training stage, the finetuning stage, the refining stage. Even without the third refining stage, the previous two stages can accomplish the WASS object detection task. Thus, for convenience, we name the three stages as OICR, WASS R-CNN I and WASS R-CNN II, respectively. WASS R-CNN II is much stronger than WASS R-CNN I. We use two different base networks, which are VGG_M and VGG_16 pre-trained on the ImageNet dataset[24].

For training OICR using weakly-supervised training images, we follow its original setting. Please refer to [8] and its source code②. For the training of WASS R-CNN I, we randomly select 10% of the images in the trainval (train + validation) set; their bounding box annotations are used for finetuning the last detection subnet. To reduce the impact of randomness and make our results more persuasive, we repeat the experiments three times, record the class-specific average precision for each time, and calculate the mean and standard deviation of the three groups of results. For data augmentation in both training and testing, we use five different scales $\{480, 576, 688, 864, 1\,200\}$, each of which is used as the minimum of width or height for resizing images while keeping the largest image-side less than 2 000[19]. Meanwhile, random horizontal flips are used in data augmentation. During the finetuning stage, we use SGD to optimize the network. The mini-batch size is set to 2. For the first 40 000 iterations, the learning rate is set to 0.000 01; for the next 30 000 iterations, it decades to 0.000 001. The momentum and the weight decay are set to 0.9 and 0.000 5 respectively.

In the testing stage, we still use the same multiple scales and average the scores of a proposal on all scales. Then, non-maximum suppression (NMS) (with 30% IoU threshold) is applied to obtain the final detection results. In the refining stage, we train a Fast R-CNN[1] based on the detection results of WASS R-CNN I. For each object class, we choose the highest scored proposal as ground-truth in this stage; for taking full advantage of the fully labeled training images, if the bounding box annotations of the image are provided, we replace the noisy labels (outputted by WASS R-CNN I) with the well-annotated labels. During training, we still use the multiple scales $\{480, 576, 688, 864, 1\,200\}$, as well as the random horizontal flipping. The bounding box regression in Fast R-CNN is applied. The learning rate for the first 30 000 iterations is 0.001, and for the next 10 000 iterations is 0.000 1. The momentum and the weight decay are set to 0.9 and 0.000 5 respectively.

The WASS R-CNN system is implemented based on the Caffe[25] deep learning framework using the Python interface. All our experiments are carried out on a Linux server with an NVIDIA GTX Titan X GPU.

4.3 Results

The main results of WASS R-CNN are given in Table 1. In the table, we compare the proposed WASS R-CNN I and WASS R-CNN II to some weakly-supervised object detectors, including WSSDN[7] and OICR[8] and the fully-supervised Fast R-CNN detector[1]. When using VGG_M as the base network, the mAPs over 20 classes are 37.9%, 44.9%, 54.3% and 61.3% for OICR, WASS R-CNN I, WASS R-CNN II and Fast R-CNN, respectively. When using VGG_16 as the base network, the mAPs are 41.2%, 50.1%, 59.2% and 69.0% for OICR, WASS R-CNN I, WASS R-CNN II and Fast R-CNN, respectively.

From the results, we can observe that: 1) by using 10% fully-supervised training images, WASS R-CNN I has 7.0% mAP and 8.9% mAP improvement over OICR using VGG_M and VGG_16 as base network respectively; 2) WASS R-CNNs have a great improvement in the Person class, which is a very difficult yet important class for weakly-supervised detectors; after using the semi-supervised training data, the performance of WASS R-CNN has been significantly improved; 3) the WASS R-CNN II reaches 88.6% of fully supervised Fast R-CNN's mAP when using VGG_M as the base network and 85.5% of the fully supervised Fast R-CNN's mAP when using VGG_16 as the base network. The results are very impressive, since they are getting closer and closer to the fully-supervised Fast R-CNN detectors.

Robustness to the Random Selection of Fully-Supervised Images. Since 10% fully-supervised training images are randomly selected from the trainval set, it is necessary to check the influence of the random selection. In Table 2, we give the standard deviation of the APs when running WASS R-CNN for three times.

②https://github.com/ppengtang/oicr, Oct. 2019.

Table 1. Average Precision (%) for Different Methods on VOC 2007 Test Set

Method	Aero	Bike	Bird	Boat	Bott	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Perso	Plant	Sheep	Sofa	Train	TV	mAP
WSDDN VGG_M[7]	43.6	50.4	32.2	26.0	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	37.8	46.9	53.4	47.9	34.9
WSDDN VGG_16[7]	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
OICR VGG_M[8]	53.1	57.1	32.4	12.3	15.8	58.2	56.7	39.6	0.9	44.8	39.9	31.0	54.0	62.4	4.5	20.6	39.2	38.1	48.9	48.6	37.9
OICR VGG_16[8]	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
WCCN VGG_16[26]	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
Jie VGG_16[27]	54.2	52.0	35.2	25.9	15.0	59.6	67.9	58.7	10.1	67.4	27.3	37.8	54.8	67.3	5.1	19.7	52.6	43.5	56.9	62.5	43.7
DPM[28]	32.8	56.8	2.5	16.8	28.5	39.7	51.6	21.3	17.9	18.5	25.9	8.8	49.2	41.2	36.8	14.6	16.2	24.4	39.2	39.1	30.2
Fast R-CNN VGG_M[1]	70.9	70.9	62.5	46.7	28.0	70.9	72.7	77.4	33.7	66.6	61.6	70.3	74.8	69.8	62.2	30.1	59.6	62.1	70.0	65.4	61.3
Fast R-CNN VGG_16[1]	77.6	78.6	71.0	61.3	39.6	78.6	78.2	83.5	43.7	74.4	67.7	82.0	81.4	75.5	67.8	32.2	68.0	69.1	78.6	70.2	69.0
Fast R-CNN(511) VGG_M[1]	41.8	51.5	29.5	22.9	8.7	45.3	59.8	46.1	17.1	43.3	36.0	43.8	51.3	53.0	44.3	7.6	38.0	26.6	43.5	49.2	38.0
Fast R-CNN(511) VGG_16[1]	44.0	60.9	44.1	26.6	20.0	50.5	67.8	60.7	22.0	48.6	30.9	54.7	57.0	57.5	56.0	15.1	45.2	43.9	52.0	55.1	45.6
WASS R-CNN I VGG_M	53.0	58.9	34.8	26.9	18.7	57.9	61.8	54.8	18.1	58.1	40.3	46.9	55.1	58.3	42.3	20.8	49.1	36.1	52.4	54.1	44.9
WASS R-CNN I VGG_16	58.4	64.2	41.4	28.7	17.1	63.6	67.5	67.3	28.0	52.0	47.1	57.5	54.7	65.1	42.5	24.3	45.0	50.4	65.9	62.0	50.1
WASS R-CNN II VGG_M	60.0	67.1	51.8	38.5	23.6	65.7	66.7	71.6	20.3	65.2	56.3	64.1	68.5	63.9	50.6	24.1	54.9	53.5	62.8	57.0	54.3
WASS R-CNN II VGG_16	66.1	71.9	59.7	41.9	21.7	71.9	70.7	77.2	31.4	68.2	64	73.7	68.1	67.8	50.7	26.4	58.1	60.6	72.0	61.5	59.2

Note: The first part, from the 2nd row to the 7th row shows the results of two effective weakly-supervised object detectors (WSSDN and OICR). The second part, from the 8th row to the 12th row, shows the results of the fully-supervised Fast R-CNN detector, and among them Fast R-CNN(511) means that we train a Fast R-CNN detector with 511 fully annotation images in the WASS setting. The third part, from the 13th to the 16th row, shows the results of WASS R-CNN I & II using 10% fully-supervised training images. Mbike: motobike.

Table 2. Standard Deviation of APs (%) When Running WASS R-CNNs for Three Times

Method	Aero	Bike	Bird	Boat	Bott	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Perso	Plant	Sheep	Sofa	Train	TV	mStd
WASS R-CNN I VGG_M	1.6	1.4	1.0	1.1	0.5	0.9	0.6	0.4	0.5	1.5	1.4	1.7	1.0	1.5	0.4	0.2	2.3	1.5	2.0	1.7	0.2
WASS R-CNN I VGG_16	1.7	1.5	0.8	1.6	1.5	0.7	0.4	2.0	1.0	4.1	2.6	1.0	2.5	2.5	0.6	1.3	2.2	3.6	2.2	0.8	0.3
WASS R-CNN II VGG_M	1.7	0.9	0.6	2.7	0.9	1.3	0.2	1.7	2.0	1.1	2.7	1.8	2.3	1.7	0.2	1.4	0.7	1.7	2.4	2.7	0.3
WASS R-CNN II VGG_16	1.7	1.8	1.5	1.2	0.2	2.1	0.6	0.5	0.2	0.8	2.0	0.4	0.9	1.6	0.4	1.7	1.3	1.3	0.3	1.6	0.2

In each of time, 10% semi-supervised training images are randomly selected. The mean standard deviations over 20 classes are very small, close to 0.2% or 0.3%, which shows WASS R-CNNs are robust to the random selection of the fully-supervised training images.

Different Proportion of Fully-Supervised Images. To study the effect of different numbers of fully-supervised images used in WASS R-CNN, we conduct experiments by changing the ratio of images that contain bounding box annotations from 10% to 90% with the step size of 10%. The results are summarized in Fig.2. For each ratio, we randomly sample fully-supervised training images for three times, run WASS R-CNN I and WASS R-CNN II for three times, and take the mean of mAP. The results show that when using 50% of fully-supervised training images, the mAP of WASS R-CNN II is getting very close to Fast R-CNN and 10% is the turning point of the performance of WASS R-CNN II.

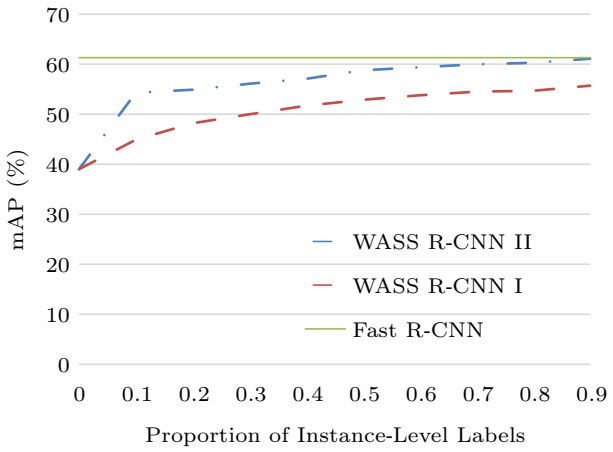


Fig.2. Results (mAP in %) of WASS R-CNN I and WASS R-CNN II with VGG_M trained using different portions of fully-supervised information. The dash line shows the mAP of Fast R-CNN which is treated as the upper bound of WASS R-CNN.

Error Analysis. Knowing the shortages of the current WASS R-CNN is important for consistently improving the WASS R-CNN detectors. To examine the disparity between WASS R-CNN II and Fast R-CNN, we use the methodology and tools of Hoiem *et al.*[29] to analyze the error types for false positives, including localization error (correct class, $0.1 < \text{IoU} < 0.5$), similar class error (misclassified to the similar classes, $\text{IoU} > 0.1$), other class error (misclassified to the other classes, $\text{IoU} > 0.1$) and background error ($\text{IoU} < 0.1$). For more details about this tool, please refer to [29]. Fig.3 shows the breakdown of each error type averaged across all 20 classes. The results show that the localization error is the most obvious shortage of WASS R-CNN II. This message makes sense since we have only 10% bounding box annotations. In the future study of WASS R-CNN, more effort should be made in improving the localization accuracy.

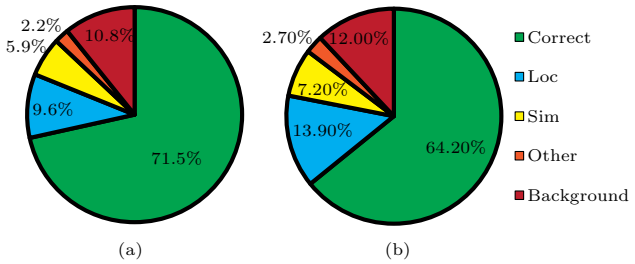


Fig.3. Error analysis: (a) Fast R-CNN VGG_16. vs (b) WASS R-CNN II VGG_16. Loc: localization; Sim: similar.

5 Discussion

Due to the importance of the problem of object detection using little supervision, besides image-level supervision and the studied weakly- and semi-supervision, there are other kinds of little supervision methods that have been studied. [8] and [27] are very effective image-level supervised methods. [8] is an on-line refinement method to help the detector find the proposal that covers the whole object, and [27] has a novel online supportive sample harvesting (OSSH) strategy to improve the quality of positive samples such that the ability of the detector can be substantially enhanced. [31] adds the extra class-specific saliency maps to help select highly confident object proposals. However, [17] uses a different kind of supervision, i.e., click. In [17], human annotators give the center of the object and another point that shows the scale of object. With the two-click annotations, the distance of the two points and the statistics of the distance with the scale of an object, they can get a rough scale of the object to improve the quality of positive samples. In [16], the authors used human verification to select the positive samples, while this method is only suitable for images that contain one single object to detect. The results are shown in Table 3. By comparing these results, we can see that selecting the positive samples of high quality is of great importance for object detection. Among them, with the rough information of the object scale, the performance of the two-click supervision method can get close to human verification. In the second stage of our WASS R-CNN, we use some fully-annotated images to fine-tune the last detection subnet, which solves the problem that OICR cannot handle the situation where more than one object from the same category appears in the same image; therefore the performance of our method outperforms that in [8] and [16].

Table 3. Object Detection Results of Some Object Detection Methods Using Little Supervision

Supervision	Method & Description	mAP(%)
Image level	OICR[8], MIL	47.0
Image level	OSSH[27], self-taught learning	43.7
Image level + saliency map	SGWSOD[30], saliency-guided	43.9
Two-click supervision	MIL with click supervision[17]	57.5
Human verification	Yes/no verification[16]	58.0
WASS	WASS R-CNN II	59.2

6 Conclusions

In this paper, we studied the problem of learning a Fast R-CNN detector in a new WASS setting and obtained significant improvement over traditional weakly-supervised object detection methods. There are many effective studies on weakly-supervised object detection, while the gap of performance between them and fully supervised object detection is huge. The WASS setting is a trade-off between performance and annotation cost, thereby it is important to achieve the goal of learning effective object detectors using limited human annotations. The experimental results can support this conclusion. The proposed WASS R-CNN method takes the advantages of weakly-supervised learning, semi-supervised learning via finetuning and refining detection by training Fast R-CNN. It has achieved a very impressive detection accuracy on the challenging PASCAL VOC 2007 dataset. In the future, we will focus on developing more effective WASS detectors by modeling some prior information in low-level image understanding, such as image saliency[32−34] and contour detection[35,36].

Acknowledgements We thank the anonymous reviewers for the helpful comments.

References

[1] Girshick R. Fast R-CNN. In *Proc. the 2015 IEEE International Conference on Computer Vision*, December 2015, pp.1440-1448.

[2] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.6517-6525.

[3] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg A. SSD: Single shot multibox detector. In *Proc. the 14th European Conference on Computer Vision*, October 2016, pp.21-37.

[4] Dai J, Li Y, He K, Sun J. R-FCN: Object detection via region-based fully convolutional networks. In *Proc. the 2016 Annual Conference on Neural Information Processing Systems*, December 2016, pp.379-387.

[5] Lin T Y, Dollr P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. arXiv:1612.03144, 2016. https://arxiv.org/pdf/1612.03144.pdf, August 2019.

[6] Chu W, Cai D. Deep feature based contextual model for object detection. *Neurocomputing*, 2018, 275: 1035-1042.

[7] Bilen H, Vedaldi A. Weakly supervised deep detection networks. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.2846-2854.

[8] Tang P, Wang X, Bai X, Liu W. Multiple instance detection network with online instance classifier refinement. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.3059-3067.

[9] Dietterich T, Lathrop R, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence,* 1997, 89(1/2): 31-71.

[10] Viola P, Platt J, Zhang C. Multiple instance boosting for object detection. In *Proc. the 2005 Annual Conference on Neural Information Processing Systems*, December 2005, pp.1417-1424.

[11] Zhu J Y, Wu J, Xu Y, Chang E, Tu Z. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(4): 862-875.

[12] Wu J, Yu Y, Huang C, Yu K. Deep multiple instance learning for image classification and auto-annotation. In *Proc. the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp.3460-3469.

[13] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(9): 1901-1907.

[14] Wang X, Zhu Z, Yao C, Bai X. Relaxed multiple-instance SVM with application to object discovery. In *Proc. the 2015 IEEE International Conference on Computer Vision*, December 2015, pp.1224-1232.

[15] Cinbis R G, Verbeek J, Schmid C. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Parttern Analysis and Machine Intelligence*, 2017, 39(1): 189-203.

[16] Papadopoulos D P, Uijlings J R, Keller F, Ferrari V. We don't need no bounding-boxes: Training object class detectors using only human verification. In *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.854-863.

[17] Papadopoulos D P, Uijlings J R, Keller F, Ferrari V. Training object class detectors with click supervision. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.180-189.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In *Proc. the 3rd International Conference on Learning Representations*, May 2015, Article No. 4.

[19] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Proc. the 13th European Conference on Computer Vision*, September 2014, pp.346-361.

[20] Uijlings J R, van de Sande K E, Gevers T, Smeulders A W. Selective search for object recognition. *International Journal of Computer Vision*, 2013, 104(2): 154-171.

[21] Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In *Proc. the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp.1717-1724.

[22] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. the 2015 Annual Conference on Neural Information Processing Systems*, December 2015, pp.91-99.

[23] Everingham M, Zisserman A, Williams C K, Van Gool L, Allan M, Bishop C M, Chapelle O, Dalal N, Deselaers T, Dorkó G. The pascal visual object classes challenge 2007 (voc2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, Sept. 2019.

[24] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In *Proc. the 26th Annual Conference on Neural Information Processing Systems*, December 2012, pp.1106-1114.

[25] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In *Proc. the 2014 ACM International Conference on Multimedia*, November 2014, pp.675-678.

[26] Diba A, Sharma V, Pazandeh A, Pirsiavash H, van Gool L. Weakly supervised cascaded convolutional networks. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.5131-5139.

[27] Jie Z, Wei Y, Jin X, Feng J, Liu W. Deep self-taught learning for weakly supervised object localization. In *Proc. the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.4294-4302.

[28] Felzenszwalb P F, Girshick R, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(9): 1627-1645.

[29] Hoiem D, Chodpathumwan Y, Dai Q. Diagnosing error in object detectors. In *Proc. the 12th European Conference on Computer Vision*, October 2012, pp.340-353.

[30] Bai X, Lai S. Saliency guided end-to-end learning for weakly supervised object detection. In *Proc. the 26th International Joint Conference on Artificial Intelligence*, August 2017, pp.2053-2059.

[31] Han L, Li X, Dong Y. Convolutional edge constraint-based U-Net for salient object detection. *IEEE Access*, 2019, 7: 48890-48900.

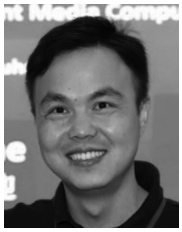
[32] Liu Y, Cheng M M, Hu X, Bian J W, Zhang L, Bai X, Tang J. Richer convolutional features for edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(8): 1939-1946.

[33] Li X, Liu K, Dong Y. Superpixel-based foreground extraction with fast adaptive trimaps. *IEEE Transactions on Cybernetics*, 2018, 48(9): 2609-2619.

[34] Xie S, Tu Z. Holistically-nested edge detection. *International Journal of Computer Vision*, 2017, 125(1-3): 3-18.

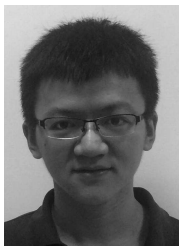
[35] Shen W, Wang X, Wang Y, Bai X, Zhang Z. DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proc. the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp.3982-3991.

[36] Li X, Liu K, Dong Y, Tao D. Patch alignment manifold matting. *IEEE Trans. Neural Network and Learning System*, 2018, 29(7): 3214-3226.



Xing-Gang Wang received his B.S. and Ph.D. degrees in electronics and information engineering from Huazhong University of Science and Technology (HUST), Wuhan, in 2009 and 2014, respectively. He is currently an associate professor with the School of Electronic Information and Communications, HUST, Wuhan. His research interests include computer vision and machine learning.

Jia-Si Wang received her B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, in 2019. Her research field is computer vision.



Peng Tang received his B.S. and Ph.D. degrees in electronics and information engineering from Huazhong University of Science and Technology (HUST), Wuhan, in 2014 and 2019 respectively. He is going to be a researcher in Salesforce Research, San Francisco. His research interests include image classification and object detection in images/videos.

Wen-Yu Liu received his B.S. degree in computer science from Tsinghua University, Beijing, in 1986, and his M.S. and Ph.D. degrees, both in electronics and information engineering, from Huazhong University of Science and Technology (HUST), Wuhan, in 1991 and 2001, respectively. He is now a professor and the chair of the Artificial Intelligence Institute in the School of Electronic Information and Communications, HUST, Wuhan. His current research areas include computer vision, multimedia, and machine learning.

JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

Volume 34, Number 6, November 2019

JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

《计算机科学技术学报》

Volume 34 Number 6 2019 (Bimonthly, Started in 1986)

Indexed in: **SCIE, Ei, INSPEC, JST, AJ, MR, CA, DBLP**

Edited by:

THE EDITORIAL BOARD OF JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

Guo-Jie Li, Editor-in-Chief, P.O. Box 2704, Beijing 100190, P.R. China

Managing Editor: Feng-Di Shu E-mail: jcst@ict.ac.cn http://jcst.ict.ac.cn Tel.: 86-10-62610746

Sponsored by: Institute of Computing Technology, CAS & China Computer Federation

Supervised by: Chinese Academy of Sciences

Undertaken by: Institute of Computing Technology, CAS

Published by: SCIENCE PRESS, BEIJING, CHINA

Printed by: Beijing Kexin Printing House

Distributed by:

China: All Local Post Offices

Other Countries: ORDER DEPT., SPRINGER, P.O. BOX 322, AH DORDRECHT, THE NETHERLAND

Available Online: www.springerlink.com

ISSN 1000-9000

国内统一连续出版物号: CN 11-2296/TP 国内邮发代号: 2-578 RMB ¥150.00