Weakly-supervised object detection via mining pseudo ground truth bounding-boxes

Yongqiang Zhang\textsuperscript{a,b}, Yaicheng Bai\textsuperscript{b}, Mingli Ding\textsuperscript{a}, Yongqiang Li\textsuperscript{a,}\* Bernard Ghanem\textsuperscript{b}

\textsuperscript{a}School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China
\textsuperscript{b}The Visual Computing Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

A R T I C L E   I N F O

Article history:
Received 23 January 2018
Revised 7 June 2018
Accepted 1 July 2018
Available online 5 July 2018

Keywords:
Weakly-supervised learning
Object detection
Pseudo ground truth
Iterative learning
Deep learning

A B S T R A C T

Recently, weakly-supervised object detection has attracted much attention, since it does not require expensive bounding-box annotations while training the network. Although significant progress has also been made, there is still a large gap on the performance between weakly-supervised and fully-supervised object detection. To mitigate this gap, some works try to use the pseudo ground truths generated by a weakly-supervised detector to train a supervised detector. However, such approaches incline to find the most representative parts instead of the whole body of an object, and only seek one ground truth bounding-box per class even though many same-class instances exist in an image. To address these issues, we propose a weakly-supervised to fully-supervised framework (W2F), where a weakly-supervised detector is implemented using multiple instance learning. And then, we propose a pseudo ground-truth excavation (PGE) algorithm to find the accurate pseudo ground truth bounding-box for each instance. Moreover, the pseudo ground-truth adaptation (PGA) algorithm is designed to further refine those pseudo ground truths mined by PGE algorithm. Finally, the mined pseudo ground truths are used as supervision to train a fully-supervised detector. Additionally, we also propose an iterative ground-truth learning (IGL) approach, which enhances the quality of the pseudo ground truths by using the predictions of the fully-supervised detector iteratively. Extensive experiments on the challenging PASCAL VOC 2007 and 2012 benchmarks strongly demonstrate the effectiveness of our method. We obtain 53.1\% and 49.4\% mAP on VOC2007 and VOC2012 respectively, which is a significant improvement over previous state-of-the-art methods.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

In computer vision, object detection is a fundamental and important task, since it is usually a key technology towards some advanced tasks such as object segmentation, object tracking, action analysis and detection, etc. Object detection has been widely studied over the past few decades, and many state-of-the-art methods [1–4] based on deep Convolutional Neural Networks (CNNs) [5,6] have been proposed and impressive performance has been achieved. The key to their superior performance is the strong learning ability (i.e. regression ability) of fully-supervised deep CNN models and the availability of large scale labeled datasets [7,8], which include the tight bounding-box annotations. However, it is expensive and time-consuming to employ professional annotators for collecting such accurate annotations. Furthermore, these annotations usually have some bias and errors caused by the subjectivity of annotators, which could lead the learned models to converge to an undesirable solution.

In order to address the above problems, some works [9–11] try to train a detector, called weakly-supervised detector, by only utilizing image-level labels (e.g. “dog”, “cat”, etc.) as the supervised information. Their motivation is that it is much simpler to build a training dataset with only image-level annotations than to compile one with accurate bounding-box annotations. More importantly, these image-level annotations are easily obtained online under many circumstances (e.g. tags or keywords of an image). However, we would like to stress that the performance of weakly-supervised detectors remains far behind the fully-supervised detectors. Given a complex image that contains multiple instances of objects especially in the presence of partial occlusion, using only image-level annotations for object detection may be insufficient due to the lack of location annotations.

Toward the contradiction between the superior performance and the expensive annotations, we would like to ask: can we de-
Fig. 1. The pipeline of our weakly-supervised to fully-supervised framework (W2F) for weakly-supervised object detection. Given an image collection with only image-level labels, we first train a weakly-supervised detector by using a weakly-supervised deep detection network (WSDNN) with online instance classifier refinement (OICR) algorithm, where both tight bounding-boxes and discriminative bounding-boxes of the objects are found. Then, we propose a pseudo ground-truth excavation (PGE) algorithm to mine the pseudo ground truths from the predicted bounding-boxes of the weakly-supervised detector. Finally, those mined pseudo ground truths are employed to train a fully-supervised detector, in which the RPN (region proposal network in Faster-RCNN) is in turn used to further fine-tune the pseudo ground truths by our pseudo ground-truth adaptation (PGA) algorithm. We would like to note that either Faster-RCNN with RPN or Fast-RCNN with SS can be used as a fully-supervised detector in our implementation.

As for training a weakly-supervised object detector, most existing methods [12–16] treat it as a Multiple Instance Learning (MIL) problem. But the performance of these methods is unsatisfactory and the detection results are that only discriminative object parts are highlighted instead of the whole object, which fails to fulfill the standard evaluation criteria of object detection (i.e. IoU > 0.5 between predicted bounding-boxes and the ground truths). Moreover, these discriminative bounding-boxes are detrimental when mining the pseudo ground truths, since a tight pseudo ground truth bounding-box is required to span the whole object instance. To alleviate this issue, we combine MIL with online instance classifier refinement (OICR) [17] to implement our weakly-supervised object detector in this paper.

In terms of mining the tight pseudo ground truths, a natural way is selecting predicted bounding-box with the highest score from the weakly-supervised detector. However, this procedure has two main drawbacks: (1) They only seek one ground truth bounding-box per class even though many instances of the same category exist in an image. (2) The most representative parts (like head of person) rather than the whole body of an object are usually highlighted, as shown in Fig. 2(a). To overcome these difficulties, we put forward a pseudo ground truth mining method to process the predicted bounding-boxes of weakly-supervised detector and generate more accurate pseudo ground truth bounding-boxes. The proposed pseudo ground truth mining method includes two components: pseudo ground-truth excavation (PGE) algorithm and pseudo ground-truth adaptation (PGA) algorithm. In the PGE, we can retrieve the more accurate pseudo ground truth for each object instance, as shown in Fig. 2(b). Furthermore, we design the PGA algorithm to further refine the pseudo ground truths generated by PGE, as shown in Fig. 2(c).

The pseudo ground truth bounding-boxes mined by the PGE & PGA may miss some object-related regions (i.e. bounding-boxes include only parts of an object) or may introduce some noise (i.e. bounding-boxes include too much background) due to the inaccurate predicted locations from the weakly-supervised detector. To utilize these ignored object-related regions and alleviate those noises, we further propose a complementary iterative ground-truth learning (IGL) approach to cooperate with PGE & PGA to mine more complete object bounding-boxes and learn better pseudo ground truths. Specifically, we first train a fully-supervised detector with the pseudo ground truths mined by the PGE & PGA algorithms. Then, the predictions of the fully-supervised detector are in turn fed into the PGE & PGA algorithms to generate the pseudo ground truths to train the fully-supervised detector again, which is iterated until the optimal pseudo ground truths are found. Those predictions from the fully-supervised detector own more accurate location information, thus the object-related regions can be mined and noise can be reduced effectively.

In summary, the main contributions of our work for weakly-supervised object detection are listed as follows: (1) We propose a novel W2F framework for weakly-supervised object detection that combines the weakly-supervised detector and the fully-supervised detector by our pseudo ground truth mining algorithm. This framework inherits the advantages of both fully-supervised and weakly-supervised learning, while avoiding their shortcomings. (2) Our pseudo ground-truth excavation (PGE) algorithm can mine more accurate and tighter pseudo ground truth bounding-boxes instead of the discriminative part of an object. Meanwhile, our method can mine the pseudo ground truth bounding-box for each instance instead of only one pseudo ground truth bounding-box per class. After that, the pseudo ground-truth adaptation (PGA) algorithm is proposed to further refine pseudo ground truths. (3) We put forward an iterative ground-truth learning (IGL) approach, which cooperates with PGE & PGA to enhance the quality of pseudo ground truths by utilizing the predictions of the fully-supervised detector iteratively. (4) The performance of our method surpasses state-of-the-art weakly-supervised detection methods by a large margin on two challenging benchmarks: an absolute mAP improvement of 6.1% on the PASCAL VOC 2007 and 6.9% on the PASCAL VOC 2012 respectively. Interestingly, our method works particularly well in detecting non-rigid objects, such as “cat”, “dog” and “person”, etc., where the performance gain ranges from 14% to 40% on the VOC 2007 test set.

The rest of the paper is organized as follows. We review the related work in Section 2. In Section 3, the detailed architecture of our W2F and the IGL is described. In Section 4, we present some
experiments on two object detection benchmarks (i.e. PASCAL VOC 2007 and 2012). In Section 5, some discussions of the proposed method are presented. The conclusions and further work are provided in Section 6.

2. Related work

2.1. Weakly-supervised detection

Recently, weakly-supervised detection draws much attention in the computer vision community. Most existing methods formulate weakly-supervised detection as an MIL problem \cite{13,16,18,22} and these approaches divide training images into positive and negative parts, where each image is considered as a bag of candidate object instances. Positive images are assumed to contain at least one object instance of a certain class, and the negative images do not include object instances from this class. The main task of MIL-based detectors is to learn the discriminative representation of the object instances and then select them from the bag of positive images to train a detector. Nevertheless, the selected object instances often highlight the most discriminative parts of an object (e.g. the head of a cat, etc.) instead of the whole object, which leads to inferior performance of weakly-supervised detectors. Moreover, this underlying MIL optimization is non-convex, it is sensitive to positive instance initialization, and inclines to get trapped in local optima.

Some works \cite{11,14,19} try to find better initialization solutions to the above problems, and achieve the gratifying results. For instance, Jie et. al \cite{19} propose a self-taught learning approach to acquire tight positive samples by making the detector learn the reliable object-level features, and then re-train itself based on the selected positive samples. The result is that the detector progressively improves the detection ability. Li et. al \cite{14} address the initialization problem by progressive domain adaptation with two main steps: classification adaptation and detection adaptation. The classification adaptation is used to fine-tune the network, so that it can collect class specific object proposals, and detection adaptation is used to optimize the representations for the target domain by the confident object candidates. Bilen et. \cite{11} present a two-stream CNN weakly supervised deep detection network (WSDDN), which selects the positive samples by multiplying the score of recognition and detection.

Besides, many efforts \cite{9,17,19} have been made to improve the optimization strategy for solving the non-convex problem. In \cite{19}, relative improvement of output CNN scores instead of the static absolute CNN score is used at training iterations. These relative improvement scores can effectively filter the suspicious samples whose high predicted scores are from undesired overfitting. Cinbis et. al \cite{9} propose a multi-fold MIL strategy to prevent the training from being locked into erroneous object locations. Tang et. al \cite{17} design an online instance classifier refinement (OICR) algorithm to alleviate the local optimum problem.

In this paper, both of the initialization and optimization problems are taken into consideration at the same time. We follow the MIL pipeline and combine the two-stream WSDDN \cite{11} and OICR algorithms \cite{17} to implement our basic weakly-supervised detector (i.e. the first part of our framework).

Fig. 2. Illustration of the baseline and our pseudo ground truth mining methods. (a) An illustration of the baseline \cite{17} for mining pseudo ground truths, in which only one representative part of an object (i.e. head) is found, even though there are three persons in the image. (b) Our PGE algorithm can retrieve the accurate pseudo ground truth bounding-boxes for all instances, where b1 is the NMS process, b2 is the procedure of removing discriminative bounding-boxes and b3 is the procedure of merging bounding-boxes. (c) The mined rough bounding-boxes are further fine-tuned by our pseudo ground-truth adaptation (PGA) algorithm, where c1 is the process of training RPN and c2 is the procedure of calculating final pseudo ground truths. Best seen on the computer, in color and zoomed in.
2.2. Pseudo ground truth mining

Since fully-supervised learning has a strong regression ability, some people try to cast the weakly-supervised problem to the fully-supervised one for improving the weakly-supervised detection performance. Here, the key problem is how to mine accurate pseudo ground truths from predicted bounding-boxes of a weakly-supervised detector to train a supervised detector. Krishna et al. [23] propose a framework that exploits tracked object bounding-boxes from videos to serve as pseudo ground truths to train a fully-supervised object detector. However, an extra video dataset sharing the same categories with the training image dataset is required, making this method inefficient. We would like to stress that our framework does not need an extra dataset, and we only need the training images with image-level labels, which can be possibly crawled from online sources as in [24–27], or from a standard object detection dataset [28,29].

Perhaps the work of Tang et. al [17] is the most similar approach to ours. Tang et al. [17] first trains a weakly-supervised detector (WSD) based on the MIL, and then the highest scoring predicted bounding-box from WSD is selected as the pseudo ground truth to train a fully-supervised detector (FSD). However, their method cannot provide suitable ground truths for training a FSD which requires tight spatial coverage of the whole object instance, and the selected pseudo ground truths have some shortcomings. For instance, they only seek one ground truth bounding-box per class in an image even though many same-class instances exist. Additionally, the most representative parts of objects are usually found instead of the bounding-box that surrounds the whole object body tightly. In contrast, our method (i.e. PGE and PGA algorithms) can mine more accurate and tighter pseudo ground truth bounding-boxes, Section 3.2 will offer a detailed explanation.

2.3. Fully-supervised detection

Many approaches [1,3,30–32] have been proposed for supervised object detection with the development of the deep learning, including the Fast RCNN [1], Faster RCNN [30] and its other variants [3,31,32]. To be specific, Faster RCNN [30] has achieved a balance between detection performance and computational efficiency, and it becomes the de facto framework for fully-supervised object detection. Though great progress has been obtained, these fully-supervised methods require instance-level bounding-box annotations, which are expensive and time-consuming. In this paper, we focus on weakly-supervised object detection, and we generate the pseudo ground truths for training a fully-supervised detector, which can be any general off-the-shelf detectors, such as Fast-RCNN, Faster-RCNN, SSD, etc. .

2.4. Iterative learning methods

Recently, some iterative learning methods (e.g. expectation maximization (EM), curriculum learning [34], self-paced learning [35], etc.) are widely used in the weakly-supervised tasks [9,36–41]. For example, [36] adopts the expectation maximization (EM) algorithm to dynamically predict semantic foreground and background pixels by using an alternative training procedure. Huang et al. win the first place in the CVPR 2017 webvision image [42] classification competition based on the curriculum learning. Wei et al. [38] proposes an adversarial erasing way to progressively mine the most discriminative object regions by using several classification networks, and their task is to solve the weakly-supervised semantic segmentation problems. In their work, a single discriminative object region is highlighted firstly, and then the proposed approach drives the classification network to sequentially find new and complete object regions by erasing the current mined regions in an adversarial way. Finally, these mined regions eventually constitute a dense and tight object region (i.e. pseudo segmentation mask) for learning semantic segmentation. Wei et al. [37] proposes a simple to complex framework for weakly-supervised semantic segmentation based on self-paced learning, in which an initial segmentation network called initial-DCNN is trained with simple images by using saliency maps [43] for supervision. Then, a better network called Enhanced-DCNN is learned with the supervision from the predicted segmentation masks of simple images based on the initial-DCNN. Finally, more pixel-level segmentation masks of complex images, which are inferred by using Enhanced-DCNN, are used as the pseudo ground truths to train a deep network for semantic segmentation.

Motivated by the iterative learning method for mining the pixel-level segmentation masks from the image-level annotations in literature [37], we propose an iterative ground-truth learning (IGL) approach to mine the high-quality pseudo ground truth bounding-boxes from image-level annotations. Different from Wei et al. [37], their method is utilized for finding the pseudo segmentation masks for the task of semantic segmentation, while our IGL method is used to mine the pseudo ground truth bounding-boxes for weakly-supervised object detection. Moreover, Wei et al. [37] needs to separate the training data into the simple and complex subsets in the iterative procedure. However, we do not need this difficult procedure and we use the same training dataset in the iterations of finding high-quality pseudo ground truths.

3. Proposed method

In this section, the proposed method (i.e. W2F and IGL) for weakly-supervised object detection will be detailed. Fig. 1 and 4 show the architecture of the W2F and IGL respectively. We first describe the weakly-supervised detector trained by combining WSDDDN [31] and OICR [17]. Then, pseudo ground truth mining methods (PGE and PGA) are presented, which can mine accurate pseudo ground truth. In addition, we simply summarize our fully-supervised detector. Finally, we explain the proposed iterative ground-truth learning (IGL) approach in detail, where the quality of the pseudo ground truths is improved again.

3.1. Weakly-supervised detector (WSD)

Since only the image-level annotations are available in the task of weakly-supervised object detection and the instance-level supervisions are unavailable, it is necessary to achieve instance-level supervision for training a fully-supervised object detector. The general solution for this problem is that the top-scoring proposal generated by instance classifiers and its adjacent proposals can be labelled to its image label as supervision. So we first introduce our weakly-supervised detector (WSD) to generate the basic instance classifier. There are many possible choices [10,11,14,17,18,44,45] to achieve this. For effectiveness and implementation convenience, we choose the method of Tang which proposes an online instance classifier refinement (OICR) strategy and combines with a two stream weakly-supervised detector for better performance.

Actually, the two stream strategy is a well-known method to solve the problem of weakly-supervised object detection, and it was first proposed by Bilen and Vedaldi [11] in which a weighted MIL pooling strategy combines with two stream detection method gave out the surprising performance. However, the predicted bounding-boxes of the traditional two stream WSD tend to converge to the discriminative part of an object, and the performance is unsatisfactory. To address this problem, the OICR algorithm integrates the basic two stream network and the multi-stage instance-level classifier into a single network and achieves the state-of-art performance, and this is reason why WSD in Section 3.1 gives the
best results and this is also the motivation and the reason why we choose their method as our WSD. Notice that our framework is independent of the special WSD, so any WSD that could be embedded into our framework. The experiments in the Section 5.3 prove that our method is robustness to a weaker WSD. Moreover, we replace the Tang's method [17] with Bilen's method [11], and our approach can still get a comparable performance.

As the pipeline shown in Fig. 1, we first train a weakly-supervised detector to generate the candidate positive object samples for the subsequent steps. Assuming that the training dataset has a total of N images which only have image-level labels, we denote the image-level labels of each image I as \( y = [y_1, y_2, \ldots, y_C] \in \mathbb{R}^{N \times C} \), where C denotes different object classes, and \( y_c = 1 \) or \( y_c = 0 \) indicates the image with or without class c. In this paper, like most of the weakly-supervised object detection methods, we employ MIL to implement our weakly-supervised detector, where the instance-level annotations (i.e. bounding-boxes and labels) are required. However, only image-level annotations (i.e. labels) are available in the training dataset, and many works [11, 14, 18, 44] can capture positive object samples, which are used as the bounding-box annotations in the MIL-based method. Here, we follow Bilen and Vedaldi [11] to achieve them, in which the WSD network branches into two data streams: the classification stream and detection stream, and the positive samples are selected by multiplying the score of recognition and the score of detection.

More formally, given an input image I, about 2000 object proposals \( R = \{r_1, \ldots, r_n\} \) are generated by the selective search method [46]. The feature of each proposal is extracted by a VGG16 model pre-trained on ImageNet [29], and the last fully convolution layer fc7 is followed by two streams as described above. The first stream performs classification by mapping each proposal feature to a C-dimensional score vector. This is achieved by evaluating a linear map \( \phi_{fb} \), and the result is a matrix \( X^f \in \mathbb{R}^{C \times |R|} \), where \(|R|\) denotes the number of proposals. And then, the matrix \( X^f \) goes through a softmax layer and the output is the classification score:

\[
\sigma_{class}(X^f) = \frac{e^{x_{ij}^f}}{\sum_{k} e^{x_{ij}^k}}
\]

The second stream performs detection by mapping each proposal feature to another C-dimensional score vector, which is implemented by using a second linear map \( \phi_{fb} \), and the result is also a matrix \( X^d \in \mathbb{R}^{C \times |R|} \). It then passes through another softmax layer and the output is the detection score:

\[
\sigma_{det}(X^d) = \frac{e^{x_{ij}^d}}{\sum_{k} e^{x_{ij}^k}}
\]

After that, the score of each proposal is generated by element-wise product \( X^f \odot \sigma_{class}(X^f) \odot \sigma_{det}(X^d) \). Finally, the \( c_{ib} \) class prediction score at the image-level can be obtained by summing up all proposals: \( p_{c} = \sum_{i} X_{ij}^f \).

During the training stage, the loss function of the above described WSD can be formulated as Eq. (1):

\[
Loss_w = -\sum_{c=1}^{C} \{y_c \log p_c + (1 - y_c) \log (1 - p_c)\}
\]

However, the predicted bounding-boxes of WSD tend to converge to the discriminative part of an object, and the performance is unsatisfactory. To address this problem, we adopt the online instance classifier refinement (OICR) method [17] to refine the WSD. Specifically, refining branches are added in the training network, and they are parallel to the two data streams (i.e. classification stream and detection stream) in the WSD. Different from the classification and detection data streams, the output of the refining branch is a \( (C + 1) \)-dimensional score vector \( X_{ip}^k \) for proposal j, where \( (C + 1) \) denotes C different classes and background, and k denotes the kth time refinement. The label \( y_{ip}^k \) of proposals in the kth branch comes from the \( (k - 1) \)-th branch, which means the selected positive samples (i.e. the top-scoring proposal) in \( (k - 1) \)th branch are used as supervision to train refinement network in the \( k \)th branch. For more details about how to get the label \( y_{ip}^k \), please refer to Tang et al. [17].

Based on the supervision achieved by the above method, we train the refining instance classifier by the loss function \( Loss_{r} \) in Eq. (2).

\[
Loss_{r} = -\frac{1}{|R|} \sum_{i=1}^{n} \sum_{j=1}^{C} w_{ij} \log (x_{ij}^{rk})
\]

where \( w_{ij} \) is the loss weight of each refinement step.

The above described WSD and OICR are used to solve the initialization and optimization problem in the MIL-based weakly-supervised object detection method, respectively. In this paper, we consider both of the problems simultaneously and train the WSD end-to-end by combining the loss functions of WSD (\( Loss_w \)) and OICR (\( Loss_r \)) as in Eq. (3).

\[
Loss = Loss_{w} + \sum_{k=1}^{K} Loss_{r}^{k}
\]

where \( K \) represents the total number of refinement times.

### 3.2. Pseudo ground-truth mining

In the last subsection, we have finished the first step of our W2E. Here we expound how to mine the pseudo ground truths. Upon observing the predicted bounding-boxes of the WSD, we find that some predicted bounding-boxes are surrounding the objects tightly, and some predicted bounding-boxes only highlight the most discriminative parts of the objects. Usually, the scores of these tight bounding-boxes are lower than the discriminative ones. In practice, these tight bounding-boxes with lower score bounding-boxes are discarded during selecting the pseudo ground truth by the previous weakly-supervised detection methods, as shown in Fig. 2(a). In this paper, we make fully use of both of them through our proposed pseudo ground truth mining algorithm, which comprises pseudo ground-truth excavation (PGE) followed by pseudo ground-truth adaptation (PGA).

#### 3.2.1. Pseudo ground-truth excavation (PGE)

After training the WSD, the predicted bounding-boxes of WSD can be used as the candidate proposals to mine the pseudo ground truth bounding-boxes. Let \( G = f(P) \) denotes the set of pseudo ground truth bounding-boxes, where P is the prediction bounding-boxes from the WSD and f is the function of pseudo ground-truth excavation. PGE algorithm mainly includes the following three steps: (i) The first step is selecting the candidate pseudo ground truth bounding-boxes, in which NMS (i.e. Non Maximum Suppression) operates on all the predictions P and only the bounding-box whose score is larger than a pre-defined threshold \( T_{score} \) will be maintained. In doing so, key discriminative bounding-boxes with a high score as well as tight bounding-boxes with a low score are retained, as shown in the second image of Fig. 2(b). (ii) Since the key discriminative bounding-boxes are usually completely surrounded by the tight bounding-boxes, we remove these discriminative bounding-boxes in this step. Specifically, we first choose the biggest prediction bounding-box generated by the WSD, delete all the discriminative bounding-boxes that are completely surrounded by this biggest bounding-box, and save the biggest bounding-boxes. Then, we choose the second biggest prediction bounding-box and do the same process. The above procedure is continually done on the next biggest one, and so on. This step prevents the tiny discriminative part of an object from being chosen as a ground truth bounding-box. (iii) Due to the weakness of the WSD, some object instances may not have a tight prediction bounding-box. In
this case, the result of step ii is that some bigger discriminative detection bounding-boxes are reserved as shown in the third image of Fig. 2(b). If we use these bigger discriminative bounding-boxes as the pseudo ground truths to train a fully-supervised detector, the detection performance will not be satisfactory. To address this problem, we leverage those bigger discriminative bounding-boxes of each object part to generate a tight bounding-box. The procedure is that we choose the biggest discriminative bounding-box from step ii, merge all the bounding-boxes whose intersection-over-union (IoU) with the biggest discriminative bounding-boxes is larger than a threshold $T_{\text{fusion}}$, and save the merged bounding-box. Then, we choose another biggest one among the rest of the discriminative bounding-boxes and repeat step iii, and so on. Finally, all the bounding-boxes generated by step iii are treated as the pseudo ground truths. Above three steps define our PGE algorithm, which is detailed in Algorithm 1 and visualized in Fig. 2(b).

**Algorithm 1** Pseudo ground-truth excavation (PGE).

**Input:** $P$, $T_{\text{nms}}$, $T_{\text{score}}$, $T_{\text{fusion}}$.

while $i < n$, $i = 0$, $n$ is the number training data do

for $j$ in $C$, $C$ is the list of training data class do

keep = nms($P$, $T_{\text{nms}}$)

$G_{\text{nms}} = Y[\text{keep} : ]$

score_index = $G_{\text{nms}}[\cdot, -1] > T_{\text{score}}$

$G_{\text{nms}} = G_{\text{nms}}[\text{score_index} : ]$

$G_{\text{det}} = h(G_{\text{nms}})$, where $h$ is the function of step(ii)

iou = IoU($G_{\text{det}}$, max($G_{\text{det}}$))

if iou > $T_{\text{fusion}}$ then

$G_{\text{fusion}} = f(G_{\text{det}})$, where $f$ is the function of step(iii)

end if

$G = G_{\text{fusion}}$

end for

$i = i + 1$

end while

**Output:** Pseudo ground truth bounding-boxes $G$

### 3.2.2. Pseudo ground-truth adaptation (PGA)

Since only image-level labels are available during the training stage, the pseudo ground truths selected by PGE may be inaccurate or contain too much context compared with the instance-level annotations labeled by humans. To address this issue, we propose a pseudo ground-truth adaptation (PGA) algorithm, which takes advantage of a region proposal network (RPN) to refine the mined pseudo ground truths by PGE algorithm. Our motivation is that the proposals generated by RPN usually have a closer outline than those retrieved pseudo ground truth bounding-boxes in the PGE algorithm. Specifically, we train an RPN using the pseudo ground truth bounding-boxes from the PGE algorithm. For each pseudo ground truth bounding-box, we choose all the proposals $P_{\text{nms}}$ from the RPN whose IoU with this pseudo ground truth bounding-box is larger than a pre-defined threshold $T_{\text{nms}}$, and then average the pixel coordinates of these selected proposals as the final pseudo ground truths, as shown in Fig. 2(c). The procedure of PGA is detailed in Algorithm 2.

<table>
<thead>
<tr>
<th>Algorithm 2</th>
<th>Pseudo ground-truth adaptation (PGA).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> $G$ from PGE algorithm, $T_{\text{iou}}, P_{\text{nms}}$</td>
<td></td>
</tr>
<tr>
<td>while $i &lt; n$, $i = 0$, $n$ is the number training data do</td>
<td></td>
</tr>
<tr>
<td>for $j$ in $C$, $C$ is the list of training data class do</td>
<td></td>
</tr>
<tr>
<td>iou = IoU($G, P_{\text{nms}}$)</td>
<td></td>
</tr>
<tr>
<td>if iou &gt; $T_{\text{iou}}$ then</td>
<td></td>
</tr>
<tr>
<td>$C_{\text{det}} = \text{mean}(P_{\text{nms}})$</td>
<td></td>
</tr>
<tr>
<td>end if</td>
<td></td>
</tr>
<tr>
<td>$G = C_{\text{det}}$</td>
<td></td>
</tr>
<tr>
<td>end for</td>
<td></td>
</tr>
<tr>
<td>i = i + 1</td>
<td></td>
</tr>
<tr>
<td>end while</td>
<td></td>
</tr>
</tbody>
</table>

**Output:** Final pseudo ground truth bounding-boxes $G$

Learning are employed to further improve the detection performance. In this paper, we choose Fast-RCNN and Faster-RCNN based on VGG16 as our fully-supervised detectors. Particularly, in the Faster-RCNN, we first train a region proposal network by using the pseudo ground truths generated by PGE algorithm, in which the PGA algorithm is proposed to utilize the proposals of the RPN to further fine-tune the pseudo ground truths, and then re-train the network by using these refined pseudo ground truths. Here, we would like to note that our fully-supervised detector is not specific and any off-the-shelf detectors can be used as our fully-supervised detector, such as YOLO [47], SSD [48], R-FCN [3], etc.

### 3.4. Iterative ground-truth learning (IGL)

In the above subsection, we have shown a weakly-supervised to fully-supervised framework for weakly-supervised object detection. The proposed PGE & PGA algorithms provide the pseudo ground truth bounding-boxes for each training image that can be used for training the fully-supervised detection networks. However, some object-related regions may still be lost or background-related regions may be introduced because of the inaccurate predicted bounding-boxes from the weakly-supervised detector. Moreover, those incorrect predictions are used as the ground truths to train the detection network, which may lead to the prediction bounding-boxes drift. To exploit these object-related regions unlabeled by W2F and alleviate the noise brought from the background-related regions, we propose an iterative ground-truth learning method to enhance the quality of the pseudo ground truths.

As shown in Fig. 4, the IGL exploits the FSD model of the W2F to predict the location information for each instance in the training dataset, and these predicted bounding-boxes are further fed to the PGE & PGA algorithms to generate more accurate pseudo ground truths, which are used to train the FSD again. Then, the FSD model in the first iteration of IGL is employed to predict the bounding-boxes for each instance in the training dataset in the second IGL iteration, and the predicted results are processed by PGE & PGA to train the FSD again. The above procedure is iterated until the optimal pseudo ground truths are found. We note that the architecture of the FSD network and the parameters of PGE & PGA are the same as the configuration in the W2F. With IGL, the pseudo ground truth bounding-boxes are getting more accurate, i.e., the missed regions of an object can be mined and the noise from those irrelevant regions can be alleviated effectively. Thus, the detection performance will improve dramatically.

Formally, we denote the image-level label for a training RoI as $u$ and denote the corresponding predicted class label as $p = (p_0, p_1, \ldots, p_c)$, where background category is included. During each iteration, we denote the predicted bounding-box as $t^n = (t^n_1, t^n_2, t^n_3, t^n_4)$, in which $n$ represents the number of iteration, and
Fig. 3. Some examples of pseudo ground truth bounding-boxes generated by the baseline method [17] and our proposed method (i.e. PGA & PGA), respectively. The green bounding-boxes in the top row show the results of baseline (i.e. selecting the proposal with the highest predicted score as the pseudo ground truth). The red bounding-boxes in the bottom row show some pseudo ground truths mined by our pseudo ground truth mining method (i.e. PGE and PGA algorithms). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. The architecture of iterative ground-truth learning (IGL). First, the weakly-supervised to fully-supervised framework (W2F) is conducted. Then, the FSD model in the W2F is used to predict the bounding-boxes of training images in the first iteration of IGL, which plays the same role as the WSD in the W2F, and the predicted bounding-boxes are further fed to the PGE & PGA to generate more accurate pseudo ground truths to train the FSD again. In the IGL, the FSD model in the last iteration is used to generate the supervision for training the FSD in the next iteration.

The predicted bounding-boxes in the $n-1$ iteration are fed to the PGE & PGA algorithm for producing the ground truth bounding-box regression target. We use the multi-task loss as in [1] to train the fully-supervised detection network progressively, which can be formulated as Eq. (4):

$$L(p, u, t_m^u, t_m^{u-1}) = L_{cls}(p, u) + \lambda [u > 1] L_{loc}(t_m^u, g(t_m^{u-1}))$$

(4)
data has the best performance, to produce the final detection results.

4. Experiments

In this section, we make clear the implementation details of the proposed method. And then, we experimentally validate the effectiveness of our method (i.e. W2F framework and IGL approach), and the ablation studies are conducted to analyze the influence of its each component.

4.1. Datasets and evaluation metrics

Datasets. Two challenging and widely used benchmarks in weakly-supervised object detection, i.e. PASCAL VOC 2007 and 2012, are chosen to evaluate the proposed method. The PASCAL VOC datasets include 20 object categories, which have 9963 and 22,531 images in the VOC 2007 and VOC 2012, respectively. For VOC 2007, we use the trainval set for training and use the test set for testing. For VOC 2012, the trainval set is selected to train our network and the test set is used to test the final model. We emphasize that only image-level labels are applied while training, meaning that no bounding-box annotations are used.

Evaluation metrics. As to the standard metrics for weakly-supervised object detection, we adopt the mean average precision (mAP) [28] as the evaluation metric to evaluate our model on the testing set and the correct location (CorLoc) [49] metric to evaluate the localization accuracy of our model on the training set respectively. Here, both metrics comply with the PASCAL criterion, in which the bounding-box of a positive detection has an IoU > 0.5 with the ground truth annotation.

4.2. Implementation details

The backbone network of the proposed method is VGG16, which is pre-trained on the ImageNet dataset [7]. In the weakly-supervised detector (WSD), we follow Tang et al. [17] and refine the instance classifier three times (i.e. K=3). While training the WSD, the total number of iterations is 70K, and the learning rate is 0.001 in the first 40K iterations and 0.0001 in the last 30K iterations. The mini-batch size is 2, and the momentum and weight decay are 0.9 and 0.0005, respectively. In the PGE, the threshold \( T_{\text{th}} \) for NMS is set to 0.3, while \( T_{\text{score}} \) and \( T_{\text{fusion}} \) are set to 0.2 and 0.4 respectively. In the PGA, the IoU threshold \( T_{\text{iou}} \) is set to 0.5. For the fully-supervised detector (i.e. Fast-RCNN and Faster-RCNN) training, all the hyper-parameters are identical to Girshick[1] and He et al. [30]. Finally, NMS with 0.3 IoU threshold is used to calculate the mAP and CorLoc.

For data augmentation, we fix the original aspect ratio of images and resize the shortest side to one of these five scales [480, 576, 688, 864, 1200] for both training and testing, and meanwhile we ensure that the longest side is shorter than 2000. Moreover, we randomly flip images horizontally in the training set. In all of our experiments, we utilize the publicly available deep learning framework Caffe [50] to implement our method, and run it on an NVIDIA GTX TITAN X GPU.

4.3. Ablation analysis

To testify the effectiveness of our W2F framework, we first conduct an ablation experiment by removing the FSD. And then, for validating the contribution of each component in the W2F including PGE and PGA, we perform ablation studies by cumulatively adding each of them to the baseline (WSD+FSD), which selects the highest score of predicted bounding-boxes from WSD as the pseudo ground truths to train an FSD. Finally, we also conduct the experiment to prove the effectiveness of our IGL approach.

The influence of W2F framework. We validate the influence of our W2F framework (i.e. train an FSD by the mined pseudo ground truth bounding-boxes from the WSD) by conducting the experiments between the WSD and the WSD + FSD. Table 1 (the 1st and 2nd rows of the bottom part in particular) indicates that our baseline (WSD+FSD1) improves mAP by 4.2% in comparison with the performance of WSD. Almost all of the categories including rigid objects (e.g. “car”, “train”, “tv”, etc.) and non-rigid objects (e.g. “cat”, “dog”, “person”, etc.) have a better performance, which can be attributed to the effect of the mined pseudo ground truth and the regression ability of fully-supervised learning. Also, we evaluate the localization accuracy (CorLoc) of the WSD + FSD and the WSD on the training set respectively. From Table 2, it is easily seen that the CorLoc metric has a similar trend as mAP, where WSD+FSD1 boosts the performance from 61.4% to 65.0%, which further confirms the effectiveness of our W2F framework.

The influence of PGE. We conduct an ablation experiment by adding the PGE algorithm to the baseline (WSD+FSD1), i.e. WSD+PGE+FSD1, in order to validate the effect of PGE. From Table 1 (the 2nd and 3rd row of the bottom part), it can be easily seen that PGE brings about 6% improvement in mAP. What interests us is that our PGE algorithm is more effective for detecting the non-rigid objects (e.g. 24.5% vs. 73.7% mAP for “cat”, 21.6% vs. 65.9% mAP for “dog”, 12.6% vs. 27.6% mAP for “person”, etc.). For one thing, the baseline WSD+FSD1 only finds one pseudo ground truth bounding-box per class even though multiple instances of this class exist in an image. However, our PGE algorithm can mine more accurate and tighter pseudo ground truth bounding-boxes than the baseline (i.e. WSD + FSD1), and retrieve the pseudo ground truth bounding-box for each instance in the image. We also test the localization accuracy (CorLoc) of the WSD + FSD1 and the WSD + PGE + FSD1 on the training set. In Table 2, we can see that CorLoc has a similar trend as mAP, whereby PGE brings about 4.4% improvement. Again, we see that the CorLoc of all 20 categories experiences a huge boost, especially for non-rigid objects, which further validates the effectiveness of our PGE algorithm. The improvement in mAP of each category by the PGE algorithm is shown in Fig. 5.

The influence of PGA. We also validate the contribution of the PGA algorithm by adding the PGA to our W2F framework, where the mined pseudo ground truths from the PGE are further refined by the proposals in the RPN of Faster-RCNN. From Table 1 (the 3rd and 4th row of the bottom part) and Fig. 5, we can see that the PGA improves the mAP from 51.7% to 52.4%, which is because proposals generated by RPN are usually closer to the outline of the object than pseudo ground truths mined by PGE algorithm, especially for those pseudo ground truth bounding-boxes including excessive background. Similarly, we evaluate the localization accuracy (CorLoc) of the WSD + PGE + FSD1 and the WSD + PGE + PGE + FSD2 on the training set respectively. From Table 2, it can be easily seen that the PGA improves the CorLoc performance from 69.4% to 70.3%. This ablation study proves the contribution of the PGA algorithm in retrieving more accurate pseudo ground truths for training an FSD.

The influence of thresholds in the PGE and PGA algorithm. As for the threshold \( T_{\text{th}} \) in NMS operation in the step of PGE algorithm, we follow the default setting in some previous works (e.g. OICR, Fast-RCNN, Faster-RCNN) and set to 0.3 in all our experiments.

In term of other thresholds (i.e. \( T_{\text{score}} \) and \( T_{\text{fusion}} \)) in PGE algorithm, we do some ablation analysis to find out the optimal so-
<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinbis et al. 2017 [9]</td>
<td>38.1</td>
</tr>
<tr>
<td>Bilen et al. 2015 [18]</td>
<td>46.2</td>
</tr>
<tr>
<td>Wang et al. 2014 [44]</td>
<td>48.9</td>
</tr>
<tr>
<td>Kantorov et al. 2016 [10]</td>
<td>57.1</td>
</tr>
<tr>
<td>Li et al. 2014 [16]</td>
<td>54.5</td>
</tr>
<tr>
<td>Tang et al. 2017 [10]</td>
<td>58.0</td>
</tr>
<tr>
<td>Jie et al. 2015 [47]</td>
<td>52.2</td>
</tr>
<tr>
<td>Krishna et al. 2016 [23]</td>
<td>53.9</td>
</tr>
<tr>
<td>Tang et al. 2017 [17]</td>
<td>65.5</td>
</tr>
<tr>
<td>FSD1</td>
<td>61.4</td>
</tr>
<tr>
<td>FSD1+PFGA-FSD2</td>
<td>63.5</td>
</tr>
<tr>
<td>FSD1+PFGA-FSD2+HGL</td>
<td>64.3</td>
</tr>
</tbody>
</table>

Table 1: Correct localization (CorLoc) (%) of our method and other state-of-the-art methods on the PASCAL VOC 2007 test set. Here, †FSD1 and †FSD2 have the same meanings as Table 1.
Table 3
The AP performance in different IGL iterations on the VOC 2007 test set, where n denotes the number of iteration and n=0 represents the performance without the IGL approach.

<table>
<thead>
<tr>
<th>Iterations</th>
<th>mAP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter 0</td>
<td>63.5</td>
</tr>
<tr>
<td>Inter 1</td>
<td>64.3</td>
</tr>
<tr>
<td>Inter 2</td>
<td>64.3</td>
</tr>
<tr>
<td>Inter 3</td>
<td>65.4</td>
</tr>
</tbody>
</table>

Table 4
Average precision(AP) (%) of our method and other state-of-the-art methods on the PASCAL VOC 2012 test set. Here, †FSD1 and FSD2 have the same meanings as Table 1. The weakly-supervised detectors in the top part are based on MIL learning, and the methods in the middle part are similar to our framework based on the pseudo ground truths (i.e. using the mined pseudo ground truths to train a fully-supervised detector).

<table>
<thead>
<tr>
<th>Method</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kantorov et. al 2016 [10]</td>
<td>64.0</td>
<td>54.9</td>
<td>36.4</td>
<td>8.1</td>
<td>12.6</td>
<td>53.1</td>
<td>40.5</td>
<td>28.4</td>
<td>6.6</td>
<td>35.3</td>
<td>34.4</td>
<td>49.1</td>
<td>42.6</td>
<td>62.4</td>
<td>19.8</td>
<td>15.2</td>
<td>27.0</td>
<td>33.1</td>
<td>33.0</td>
<td>50.0</td>
<td>35.3</td>
</tr>
<tr>
<td>Tang et. al 2017(OICR) [17]</td>
<td>67.7</td>
<td>61.2</td>
<td>41.5</td>
<td>25.6</td>
<td>22.2</td>
<td>54.6</td>
<td>49.7</td>
<td>25.4</td>
<td>19.9</td>
<td>47.0</td>
<td>18.1</td>
<td>26.0</td>
<td>38.9</td>
<td>67.7</td>
<td>2.0</td>
<td>22.6</td>
<td>41.1</td>
<td>34.3</td>
<td>37.9</td>
<td>55.3</td>
<td>37.9</td>
</tr>
<tr>
<td>Tang et. al 2017 [45]</td>
<td>60.8</td>
<td>54.2</td>
<td>34.1</td>
<td>14.9</td>
<td>13.1</td>
<td>54.3</td>
<td>53.4</td>
<td>58.6</td>
<td>3.7</td>
<td>53.1</td>
<td>8.3</td>
<td>43.4</td>
<td>49.8</td>
<td>69.2</td>
<td>4.1</td>
<td>17.5</td>
<td>43.8</td>
<td>25.6</td>
<td>55.0</td>
<td>50.1</td>
<td>38.3</td>
</tr>
<tr>
<td>WSD†</td>
<td>71.4</td>
<td>69.4</td>
<td>55.1</td>
<td>29.8</td>
<td>28.1</td>
<td>55.0</td>
<td>57.9</td>
<td>24.4</td>
<td>17.2</td>
<td>59.1</td>
<td>21.8</td>
<td>24.4</td>
<td>57.8</td>
<td>6.3</td>
<td>1.3</td>
<td>23.1</td>
<td>52.7</td>
<td>37.5</td>
<td>33.5</td>
<td>56.6</td>
<td>42.5</td>
</tr>
<tr>
<td>WSD+FSD1‡</td>
<td>72.3</td>
<td>70.3</td>
<td>51.8</td>
<td>32.4</td>
<td>27.5</td>
<td>58.6</td>
<td>58.7</td>
<td>17.6</td>
<td>13.3</td>
<td>58.1</td>
<td>14.0</td>
<td>29.5</td>
<td>62.2</td>
<td>74.3</td>
<td>1.2</td>
<td>21.6</td>
<td>47.6</td>
<td>45.9</td>
<td>32.6</td>
<td>58.1</td>
<td>42.4</td>
</tr>
<tr>
<td>WSD+FSD1§</td>
<td>71.5</td>
<td>71.0</td>
<td>46.6</td>
<td>27.6</td>
<td>26.6</td>
<td>58.1</td>
<td>59.1</td>
<td>62.1</td>
<td>19.4</td>
<td>59.0</td>
<td>8.9</td>
<td>71.4</td>
<td>64.1</td>
<td>74.2</td>
<td>6.7</td>
<td>23.6</td>
<td>47.4</td>
<td>45.2</td>
<td>44.9</td>
<td>57.5</td>
<td>47.3</td>
</tr>
<tr>
<td>WSD+FSD2†</td>
<td>73.0</td>
<td>69.4</td>
<td>45.8</td>
<td>30.0</td>
<td>28.7</td>
<td>58.8</td>
<td>58.6</td>
<td>56.7</td>
<td>20.5</td>
<td>58.9</td>
<td>10.0</td>
<td>69.5</td>
<td>67.0</td>
<td>73.4</td>
<td>7.4</td>
<td>24.6</td>
<td>48.2</td>
<td>46.8</td>
<td>50.7</td>
<td>58.0</td>
<td>47.6</td>
</tr>
<tr>
<td>WSD+FSD2+IGL§</td>
<td><strong>74.0</strong></td>
<td>69.9</td>
<td>46.9</td>
<td>29.7</td>
<td>27.8</td>
<td><strong>59.4</strong></td>
<td><strong>58.4</strong></td>
<td><strong>76.2</strong></td>
<td><strong>21.8</strong></td>
<td><strong>56.3</strong></td>
<td>10.8</td>
<td><strong>76.2</strong></td>
<td>64.5</td>
<td><strong>74.4</strong></td>
<td><strong>14.1</strong></td>
<td>23.6</td>
<td><strong>47.7</strong></td>
<td><strong>48.5</strong></td>
<td>54.5</td>
<td>54.0</td>
<td><strong>49.4</strong></td>
</tr>
</tbody>
</table>
pseudo ground truth based methods, the reason for the improvement of our performance is that our pseudo ground truth mining algorithm can retrieve more accurate and tighter pseudo ground truths. One point we have to mention is that it is unfair to make a comparison between the performance of our method and the methods of Tang et al. [17] and Zhuang et al. [23], because the result of Tang et al. [17] (47.0%) is obtained by combining various different models while our results are achieved from a single model. For fair comparison, we compare our method with our baseline, which uses a single model, and the mAP performance increases by 7.3%. As for the method in [23], their reported mAP is averaged across only 10 categories that do not include some difficult categories, such as "bottle", "person", etc., while our method includes all 20 categories. In this circumstance, the performance of their method is 41.9%, 11.2% lower than our result. Under such unfair conditions, the proposed method is still able to outperform previous state-of-the-art methods by a large margin, which confirms the effectiveness of our method.

The CorLoc performance on the VOC 2007 trainval set is shown in Table 2, where our method achieves the highest average CorLoc performance (i.e. 71.2%) when comparing with the other state-of-the-art methods. Upon close inspecting the CorLoc performance of each class, we can observe that all the classes have a better performance than other methods, and the reason is the same as the mAP performance promotion. We would like to note that all the previous state-of-the-art methods [11,17] encounter a dilemma, i.e. their detector inclines to highlight the most discriminative parts of an object instead of the whole body of an object, and this is the reason why their detection performance is poor. To the best of our knowledge, our proposed method is the first work to avoid and address these pitfalls.

### Tables 4 and 5

Tables 4 and 5 show our performance in terms of mAP and CorLoc on the PASCAL VOC 2012 test and trainval sets VOC 2012 test, respectively. Since the bounding-box annotations of the VOC 2012 test set are not available, we submit the intermediate files to the PASCAL VOC official evaluation server to achieve the final detection performance and the links of the test results are shown in the caption of Table 4. We achieve state-of-the-art performance in both mAP and CorLoc metrics (i.e. 49.4% mAP and 71.1% CorLoc on the VOC 2012 trainval and test set respectively) by adopting our framework with the pseudo ground truth mining algorithm (i.e. PGE & PGA) and the IGL approach. The proposed approach outperforms the second highest performance by 6.8% and 5.5% in mAP and CorLoc respectively. For fair comparisons, compared with our baseline (i.e. selecting the highest score detected bounding-boxes from WSD as the pseudo ground truths to train a FSD), the mAP and the CorLoc performance increase by 7.0% and 5.6% respectively.

### 4.5 Qualitative results

Some detection results generated by our method as compared to those generated from the baseline method are shown in Fig. 8, where the green bounding-boxes indicate the objects detected by our method and the red ones correspond to those detected by the baseline method. From those detection results, we find that the bounding-boxes detected by our method tightly surround the object, while the baseline method only highlights the most discriminative object parts. This is because of the high quality pseudo ground truths mined by our PGE & PGA algorithms and the IGL approach. Moreover, we also visualize some failure results in the last three images. In these failure cases, a single detected bounding-box not only includes one object instance, but also contains multiple adjacent similar instances. These results indicate that more progress is needed to further improve the weakly-supervised object detection performance.

### 5. Discussion

#### 5.1 Why does the performance of non-rigid object have a huge improvement?

As mentioned in Section 3.2, all the 20 categories are processed by our PGE and PGA algorithm to find the accurate pseudo ground truths, and then we use the generated pseudo ground truths to train a fully-supervised detector. From the Fig. 5, Tables 1 and 4, we can see that the performance of non-rigid objects, like “cat”, “cow”, “dog”, “horse”, “person”, etc., has a huge improvement in mAP ranging from 14% to 40% when comparing with the baseline detector on the VOC 2007 test set. However, the improvement of rigid objects, like “aeroplane”, “boat”, “bottle”, “train”, “tvmonitor”, etc., is limited. Upon close inspecting the mined pseudo ground truths, we find a common phenomenon that the image usually includes more than one non-rigid object, and some parts of body
can represent their characteristics, i.e. we can identify the category of these non-rigid objects through the discriminative part of their body (like the head of person). Under such circumstances, the baseline methods only find one pseudo ground truth even though multiple object instances exist in an image. Moreover, the pseudo ground truth found by baseline method is a part of the object instead of the whole body of an object as shown in Fig. 2(a). In contrast, our method can retrieve the pseudo ground truth of each instance in the image and the pseudo ground truth bounding-boxes are surrounding the objects tightly, as shown in Fig. 2(c). However, there is usually only one rigid object in an image. At the same time, as we all know, those rigid objects do not have the most discriminative parts to highlight their characteristics, and we have to see their whole body if we want to identify the category of these rigid objects. In this case, since no discriminative bounding-boxes exist (or the score of the discriminative bounding-box is very low) in these object categories, the baseline method can find the rough pseudo ground truth bounding-boxes as shown in the last two columns of Fig. 3. Based on the rough pseudo ground truths, the performance of the baseline detector is comparable with our method. The above analysis explains the reason why the performance on non-rigid objects boosts a lot more than rigid objects.

5.2. What is the result without PGE & PGA in the IGL?

For verifying the effectiveness of the PGA & PGE algorithms in the IGL iterations, we conduct an ablation experiment, that is, we remove the PGE & PGA algorithms in the IGL iterations and adopt the baseline method like in [17] to seek the pseudo ground truth bounding-boxes. From Fig. 9, we can see that the mAP performance dropped dramatically (i.e. from 52.48% to 49.22%) on the VOC 2007 test set, which proves that the PGE & PGA algorithms play an important role in the IGL iterations. But the mAP performance rises slightly as the number of iteration increases, which further confirms the effectiveness of our IGL approach on the other hand.
5.3. How is the robustness of the W2F + IGL to the WSD?

To validate the robustness of the proposed method, we use a weaker WSD to replace the state-of-the-art WSD in the W2F framework and test the mAP performance on the VOC 2007 test set. More concretely, we employ the model in the 40,000 iterations as a weaker WSD, whose mAP performance is 39.32%, decreasing by more than 2% compared with the state-of-the-art WSD (i.e. the model in the 70,000 iterations). As Fig. 10 shows, our W2F framework can produce a comparable performance (52.37% vs. 52.42%) in the first iteration even though we replace the current super- rior WSD with a weaker WSD. Moreover, as the red line shows in Fig. 10, the mAP further increases as the iteration progresses, it has a similar tendency with the normal configuration (WSD (70,000 it- erations) + W2F + IGL), which verifies the robustness of our approach.

Also, we replace Tang’s method [17] with Bilen’s method [11], whose mAP performance is 39.3%, to mine pseudo ground truth bounding-boxes and then use W2F and IGL framework to train a fully-supervised detector. We finally get a similar performance (i.e. 52.46% vs. 53.13%) when compared with the performance of pseudo ground truth bounding-boxes mined by Tang’s method, which further verifies the robustness of our method.

6. Conclusions and future work

In this paper, we first present a novel weakly-supervised to fully-supervised framework (W2F) for weakly-supervised object detection. Different from previous work in this task, our framework successfully combines the advantages of fully-supervised and weakly-supervised learning, while avoiding their weakness. Firstly, we employ WSDNN and OICR to train a weakly-supervised detector (WSD) end-to-end. And then, by the virtue of the proposed pseudo ground-truth excavation (PGE) algorithm and pseudo ground-truth adaptation (PGA) algorithm, our method can find accurate pseudo ground truths for each instance in an image. Finally, those pseudo ground truths are utilized to train a fully-supervised detector for producing the final detection results. Moreover, the IGL approach is proposed to enhance the quality of the mined pseudo ground truths again and to further improve the detection performance. Extensive experiments on PASCAL VOC 2007 and 2012 demonstrate the substantial improvements (6.1% and 6.9% in mAP respectively) of our method when comparing with previous state-of-the-art weakly-supervised detectors.

In the future, we plan to use our method to detect the objects in the missing bounding-box annotations scenarios, which is very common in the images of practical applications [51–53], even in the existing annotated benchmarks.

Acknowledgments

The authors would like to thank Peng Tang for his valuable discussions. This work was done when Yongqiang Zhang worked at KAUST as a visiting Ph.D. student. Yongqiang Zhang is supported by funding from Harbin Institute of Technology (HIT) and King Abdullah University of Science and Technology (KAUST). Mingli Ding, and Yongqiang Li are supported by funding from Harbin Institute of Technology (HIT). Bernard Chanem and Yancheng Bai are supported by funding from King Abdullah University of Science and Technology (KAUST).

References

Y. Zhang et al. / Pattern Recognition 84 (2018) 68–81


Yongqiang Zhang received the MS degree in instrument science and technology from Harbin Institute of Technology, Harbin, China, in 2013. He is a PhD student from the Harbin Institute of Technology, and he is currently working at King Abdullah University of Science and Technology (KAUST) as a visiting student. His research areas are computer vision, pattern recognition, and deep learning. His research interests mainly include face detection, object detection, activity detection, image and video understanding in the real-world.

Yancheng Bai is an associate professor in Pattern Recognition and Intelligent System at Institute of Software, Chinese Academy of Sciences. He received the PhD degree from Institute of Automation, Chinese Academy of Sciences in 2014. His research interests include computer vision, pattern recognition, artificial intelligence, and machine learning.

Mingli Ding received the BS, MS and PhD degrees in instrument science and technology from Harbin Institute of Technology, Harbin, China, in 1996, 1997 and 2001, respectively. Currently, he is a professor at Harbin Institute of Technology. Prof. Ding’s research interests are intelligence tests and information processing, automation test technology, and computer vision. He has published over 40 papers in peer-reviewed journals and conferences.

Yongqiang Li received the BS, MS and PhD degrees in instrument science and technology from Harbin Institute of Technology, Harbin, China, in 2007, 2009 and 2014, respectively. He is currently an assistant professor at Harbin Institute of Technology. He worked as a visiting student at Rensselaer Polytechnic Institute, Troy, USA, from September 2010 to September 2012. His areas of research include computer vision, pattern recognition, and human computer interaction.

Bernard Chanem is currently an assistant professor in the CEMSE division and a member of the Visual Computing Center at KAUST. Before that, he was a senior research scientist at the University of Illinois Urbana-Champaign (UIUC) in Singapore, where he still holds an adjunct position. He heads projects that develop algorithms in computer vision, machine learning, and optimization geared towards real-world applications, including semantic video analysis in sports and automated surveillance, content-based image retrieval, large-scale activity recognition, and 2D/3D scene understanding. He received his Bachelor’s degree in Computer and Communications Engineering from the American University of Beirut (AUB) in 2005 and his MS/PhD in Electrical and Computer Engineering from UIUC in 2010. His work has received several awards and honors, including the Henderson Graduate Award from UIUC; two consecutive NSF fellowship awards from UIUC, a Best Paper Award (CVPRW 2013), a two-year KAUST Seed Fund, and a Google Faculty Research Award in 2015. He has co-authored more than 50 peer reviewed conference and journal papers in his field, as well as 4 patents. He is also a co-founder of AutoScount Inc. that provides automated solutions for sports video analytics.