# W2F: A Weakly-Supervised to Fully-Supervised Framework for Object Detection

Yongqiang Zhang[1,2,*] Yancheng Bai[1,3] Mingli Ding[2] Yongqiang Li[2] Bernard Ghanem[1]
[1] Visual Computing Center, King Abdullah University of Science and Technology (KAUST)
[2] School of Electrical Engineering and Automation, Harbin Institute of Technology (HIT)
[3] Institute of Software, Chinese Academy of Sciences (CAS)
{zhangyongqiang, dingml, liyongqiang}@hit.edu.cn {yancheng.bai, bernard.ghanem}@kaust.edu.sa

## Abstract

*Weakly-supervised object detection has attracted much attention lately, since it does not require bounding box annotations for training. Although significant progress has also been made, there is still a large gap in performance between weakly-supervised and fully-supervised object detection. Recently, some works use pseudo ground-truths which are generated by a weakly-supervised detector to train a supervised detector. Such approaches incline to find the most representative parts of objects, and only seek one ground-truth box per class even though many same-class instances exist. To overcome these issues, we propose a weakly-supervised to fully-supervised framework, where a weakly-supervised detector is implemented using multiple instance learning. Then, we propose a pseudo ground-truth excavation (PGE) algorithm to find the pseudo ground-truth of each instance in the image. Moreover, the pseudo ground-truth adaptation (PGA) algorithm is designed to further refine the pseudo ground-truths from PGE. Finally, we use these pseudo ground-truths to train a fully-supervised detector. Extensive experiments on the challenging PASCAL VOC 2007 and 2012 benchmarks strongly demonstrate the effectiveness of our framework. We obtain 52.4% and 47.8% mAP on VOC2007 and VOC2012 respectively, a significant improvement over previous state-of-the-art methods.*

## 1. Introduction

Object detection is a fundamental problem in computer vision, since it is the basic technology of some advanced tasks such as object segmentation, object tracking, action analysis and detection, etc. Recently, many state-of-the-art methods [11, 12, 5, 20] based on deep Convolutional Neural Networks (CNNs) [22, 15] have been proposed and superior performances have been achieved. The key to their successes is the strong learning ability (*i.e.* regression ability) of fully-supervised deep CNN models and the availability of large scale labeled datasets [30, 25], which include tight bounding box annotations. However, collecting such accurate annotations are expensive and time-consuming. Moreover, these annotations usually have bias and errors introduced by the subjectivity of annotators, which could lead the learned models to converge to an undesirable solution.

To address these problems, some weakly-supervised detectors [4, 19, 2] are trained by only utilizing image-level labels (*e.g.* "dog", "cat", etc.) as the supervised information. Building a training dataset with only image-level an-

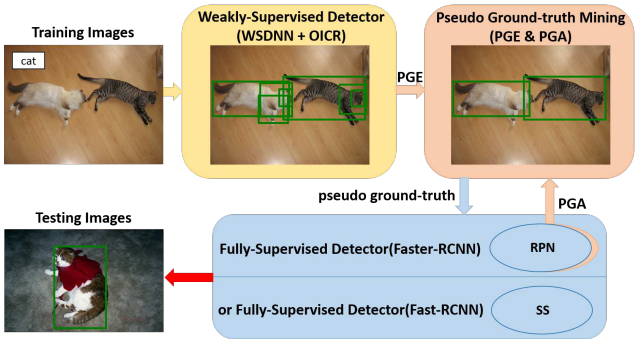


Figure 1. An illustration of our weakly-supervised to fully-supervised framework for object detection (W2F). Given an image collection with only image-level labels, we first combine a weakly supervised deep detection network (WSDDN) with online instance classifier refinement (OICR) to train a weakly-supervised detector, and here both tight bounding boxes and discriminative boxes of object parts are found. Then, we propose a pseudo ground-truth excavation (PGE) algorithm to preserve those tight bounding boxes as pseudo ground-truth, which is in turn used to train a supervised detector whose RPN (region proposal network in Faster-RCNN) makes use of our proposed pseudo ground-truth adaptation (PGA) algorithm to further fine-tune the pseudo ground-truths.

*This work is done when Yongqiang Zhang was a visiting PhD student at KAUST.

notations is much easier than compiling one with accurate bounding-box annotations, since these image-level annotations are easily obtained online in many cases (*e.g*. tags or keywords of an image). However, to the best of our knowledge, the performance of weakly-supervised detectors remains far behind 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 not be adequate due to the lack of location annotations.

**Main idea:** Can we design an architecture that inherits the advantages of both fully-supervised (regression ability) and weakly-supervised detection (inexpensive training annotation) and avoids their shortcomings (*i.e*. expensive annotations and poor detection performance)? To this end, we propose our weakly-supervised to fully-supervised framework for object detection (W2F). Given an image collection with only image-level labels, we first employ Multiple Instance Learning (MIL) to train a weakly-supervised detector, and then we propose a pseudo ground-truth excavation algorithm to seek pseudo ground-truth boxes, which are in turn refined using our pseudo ground-truth adaptation algorithm and used to train a supervised detector. Figure 1 illustrates the pipeline of our weakly-supervised to fully-supervised framework.

In practice, there are two issues in the W2F. (1) How to train an actuate weakly-supervised detector. (2) How to mine the tight pseudo ground-truth (*i.e*. the bounding-box surrounding the whole body of object tightly) for each instance in the image.

As for training a weakly-supervised object detector, most existing methods [13, 31, 23, 16, 29, 35] treat it as a Multiple Instance Learning (MIL) problem, and the result is only discriminative object parts are highlighted instead of the whole object, which is detrimental when a tight pseudo ground truth box is required to span the whole object instance. To alleviate this issue, we follow [32] and combine MIL with online instance classifier refinement (OICR) [32] to implement our weakly-supervised object detector.

In term of mining the tight pseudo ground-truths, a natural way is selecting the highest score proposal from a weakly-supervised detector. However, this procedure has two drawbacks. First, they only seek one ground-truth box per class even though many instances are existing in this category. Second, the most representative parts (like head) of an object rather than the whole body of objects are usually highlighted, as shown in Figure 2(a). To solve these problems, we put forward a pseudo ground-truth mining method which includes two components: pseudo ground-truth excavation (PGE) and pseudo ground-truth adaptation (PGA). In the PGE, we propose a iterative algorithm to retrieve the more accurate pseudo ground truth of each object instances, as shown in Figure 2(b). Moreover, we further design PGA algorithm to refine the pseudo ground truths generated by PGE, as shown in figure 2(c).

To sum up, we make the following three contributions for weakly-supervised object detection in this work: **(1)** We propose a novel 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 boxes, instead of only one box of the discriminative part of an object per class is found. After that, we propose the pseudo ground-truth adaption (PGA) algorithm to further refine pseudo ground-truths. **(3)** Our W2F framework surpasses state-of-the-art weakly supervised detection methods by a large margin on two challenging benchmarks: an absolute mAP improvement of 5.4% on PASCAL VOC 2007 and 5.3% on PASCAL VOC 2012. Interestingly, our method works particularly well in detecting non-rigid objects, such as "cat", "dog" and "person", where the performance gain ranges from 15% to 49%.

## 2. Related Work

**Weakly-supervised detection.** Most existing methods formulate weakly-supervised detection as an MIL problem [1, 31, 23, 16, 29, 18]. These approaches divided training images into positive and negative parts, where each image is considered as a bag of candidate object instances. The main task MIL-based detectors is to learn the discriminative representation of the object instances and then select them from positive images to train a detector. However, positive object instances often focus on the most discriminative parts of an object (*e.g*. the head of a cat, etc.) and not 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 tends to get trapped in local optima.

Some works try to solve these problems via finding better initialization methods. For instance, Jie *et al*. [18] propose a self-taught learning approach to progressively harvest high-quality positive samples. Li *et al*. [23] propose classification adaptation 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 *etc*. [2] present a two-stream CNN weakly supervised deep detection network (WSDDN), which selects the positive samples by multiplying the score of recognition and detection.

In addition, many efforts have been made to improve the optimization strategy. In [18], relative improvement of output CNN scores are used instead of relying on the static absolute CNN score at training iterations. Cinbis *et al*. [4] pro-

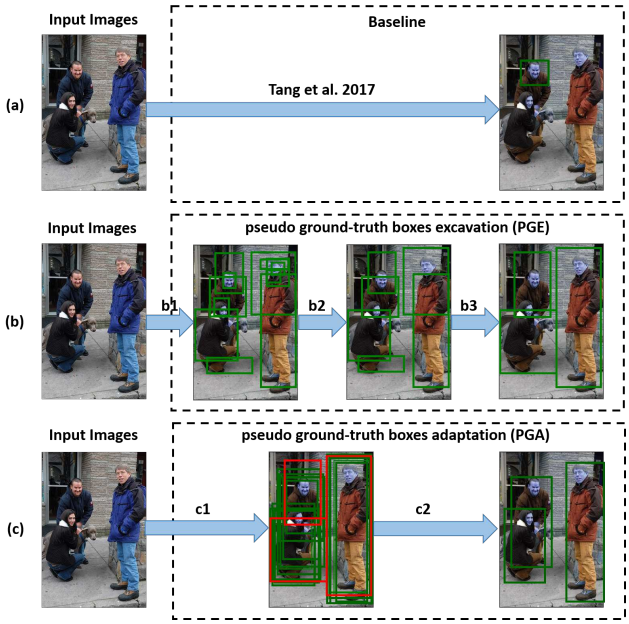


Figure 2. An illustration of the proposed pseudo ground-truth mining method. (a) Only one representative part of an object (*i.e.* head) is found, even though there are three persons in the image [32]. (b) Our method retrieves the pseudo ground-truth boxes of all instances using our pseudo ground-truth excavation (PGE) algorithm, where b1 is the NMS process, b2 is the procedure of removing discriminative boxes and b3 is the procedure of merging boxes. (c) These rough 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.

pose a multi-fold MIL strategy to prevent the detector from being locked into erroneous object locations. Tang *et al.* [32] design an online instance classifier refinement (OICR) algorithm to alleviate the local optimum problem.

In this paper, we consider the initialization and optimization problems simultaneously. We follow the MIL pipeline and combine the two-stream WSDDN [2] and OICR algorithms [32] to implement our basic weakly-supervised detector (*i.e.* the first part of our framework).

**Pseudo ground-truth mining.** Due to the strong regression ability of fully-supervised learning, here we cast the weakly-supervised problem to the supervised one. The key problem is how to mine accurate pseudo ground-truths from predicted boxes of a weakly-supervised detector to train a supervised detector. A framework [21] is proposed that exploits tracked object boxes from videos to serve as pseudo ground-truths to train an object detector. However, an extra video dataset is required and it must share the same categories with the image dataset, making this method not efficient. We would like to emphasize that our framework does not need an extra dataset, and the only training data needed is images with image-level labels (possibly crawled from online sources as in [36, 3, 10, 9], or from a standard object detection dataset [8, 6]).

The most similar approach to our framework is the work of Tang *et al.* [32]. In their method, the highest scoring predicted box from weakly-supervised detector (WSD) is selected as the pseudo ground-truth, thus, leading to some shortcomings. For instance, they only seek one ground-truth box per class in an image even though many same-class instances may exist. Moreover, the most representative parts of objects are usually found instead of the tight object prediction boxes. In contrast, more accurate and tighter pseudo ground-truth boxes can be generated by our PGE and PGA algorithms (Section 3.2 will have a detailed explanation).

**Fully-supervised detection.** With the development of deep learning, many methods have been proposed, such as the Fast RCNN [11], faster RCNN [28] and its other variants [5, 14, 24]. Specifically, Faster RCNN [28] has achieved a balance between detection performance and computational efficiency. And it becomes the *de facto* framework for fully-supervised object detection. Though great improvements have been achieved, 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.

## 3. Proposed Method

In this section, we introduce our framework in details. Figure 1 shows the architecture of the proposed method. We first describe the weakly-supervised detector. Then, pseudo ground-truth mining methods (PGE and PGA) are presented, which greatly improve the quality of the pseudo ground-truths. Finally, we simply summarize our fully-supervised detector.

### 3.1. Weakly-Supervised Detector(WSD)

Given an image $\mathbf{I}$, we denote the image-level labels $\mathbf{y}=[y_1, y_2, \ldots, y_C] \in \mathbb{R}^{n\times 1}$, 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, we employ MIL to implement the weakly-supervised detector, where the instance-level annotations (*i.e.* bounding box and label) are required. However, only image-level (*i.e.* label) annotations are available in the training dataset, and there are many works [1, 34, 23, 2], which can capture the instance-level annotations. Here, we follow [2] to achieve them, in which the WSDNN model branches into two data streams: the classification and detection data streams.

For each input image $\mathbf{I}_i$, object proposals $\mathbf{R} = (r_1, \ldots, r_n)$ are generated by the selective search method [33]. The features of each proposal are extracted by a VGG16 model pre-trained on ImageNet [6], 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 $\phi_{fc8c}$, and the results are a matrix $\mathbf{x}^c \in \mathbb{R}^{C\times|R|}$, where $|R|$ denotes the number of proposals, which then goes through a softmax layer and the output is: $[\sigma_{class}(\mathbf{x}^c)]_{ij}=\frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}}$. The second stream performs instead detection by using a second linear map $\phi_{fc8d}$, and also resulting a matrix $\mathbf{x}^d \in \mathbb{R}^{C\times|R|}$. It then passes through another softmax layer and the output is: $[\sigma_{det}(\mathbf{x}^d)]_{ij}=\frac{e^{x^d_{ij}}}{\sum_{k=1}^{|R|} e^{x^d_{ik}}}$. After that, the score of each proposal is generated by element-wise product $\mathrm{x}^R=\sigma_{class}(\mathbf{x}^c) \odot \sigma_{det}(\mathbf{x}^d)$. Finally, the $c_{th}$ class prediction score at the image-level can be obtained by summation over all proposals: $p_c=\sum_{r=1}^{|R|} x^R_{cr}$. During the training stage, the loss function can be formulated as following:

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

Since WSDNN tends to converge to the discriminative part of an object and the performance is unsatisfactory, we adopt the online instance classifier refinement (OICR) method [32] to refine the WSDNN. Specifically, refining branches are added in the training network, and they are parallel to the two data steams as mentioned above. Different from the classification and detection data streams, the output of the refining branch is a $\{C+1\}$-dimensional score vector $\mathbf{x}^{Rk}_j$ for proposal $j$, where $\{C+1\}$ denotes $C$ different classes and background and $k$ denotes the $k^{th}$ time refinement. The label $y^k_{cr}$ of proposals in the $k^{th}$ branch comes from the $\{k-1\}^{th}$ branch. For more details about how to get the label $y^k_{cr}$, please refer to [32]. Based on the achieved supervision, we train the refining instance classifier by considering the loss function $Loss_r$ in Eq.(2).

$$Loss_r = -\frac{1}{|R|}\sum_{r=1}^{|R|}\sum_{c=1}^{C+1} w^k_r y^k_{cr}\log(x^{Rk}_{cr}) \quad (2)$$

where $w^k_r$ is the loss weight of each refinement step.

Finally, we train the weakly-supervised detector end-to-end by combining the loss functions of WSDNN ($Loss_w$) and OICR ($Loss_r$), as in Eq.(3).

$$Loss = Loss_w + \sum_{k=1}^{K} Loss^k_r \quad (3)$$

where $K$ represents the total number of refinement times.

### 3.2. Pseudo Ground-truth Mining

After training the weakly-supervised detector, we find that the weakly-supervised detector can indeed find tight boxes of an object, and it is just that the scores of these tight boxes are lower than the discriminative ones. While these tight boxes with lower score boxes are discarded during selecting the pseudo ground-truth by previous weakly-supervised detectors, as shown in Figure 2 (a). In this paper, We exploit them through our proposed pseudo ground-truth mining algorithm, which comprises pseudo ground-truth excavation (PGE) followed by pseudo ground-truth adaptation (PGA).

**Pseudo ground-truth excavation (PGE).** Let $G$=$f(P)$ denotes the set of pseudo ground-truth boxes, where $P$ is the prediction boxes generated by the WSD and $f$ is the excavation function. PGE includes three components: **(i)** The first component selects the candidate pseudo ground-truth boxes, in which NMS operates on all the predictions $P$ and only boxes whose score larger than a pre-defined threshold $T_{score}$ are maintained. In doing so, key discriminative boxes with a high score as well as tight boxes with a low score are retained, as shown in the second image of Figure 2(b). **(ii)** Since the key discriminative boxes are usually completely surrounded by the tight boxes, we delete these discriminative boxes in this step. We propose an iterative algorithm to remove all these smaller boxes. Specifically, we first choose the biggest prediction box generated by the weakly supervised detector, delete all the smaller discriminative boxes that are completely surrounded by this biggest box, and save the biggest box. Then, we choose the second biggest prediction box and do the same process, and so on. This step prevents the tiny discriminative part of an object from being chosen as a ground-truth. **(iii)** In some times, some object instances may not have a tight box. For this case, the result of step ii is some bigger discriminative detection boxes are reserved as shown in the third image of Figure 2(b). The detection performance is not satisfactory while using these bigger discriminative boxes as the pseudo ground-truths to train a fully-supervised detector. To further improve the performance, we leverage those bigger discriminative boxes of each object parts to generate a tight box. The procedure is that we choose the biggest discriminative boxes from step ii and merge all the discriminative boxes whose intersection-over-union (IoU) is larger than a threshold $T_{fusion}$ with this biggest discriminative boxes (*i.e.* choosing the minimum left-top coordinate and the maximum right-bottom coordinate), and save the merged box. Then, we choose the second biggest one among the rest of the discriminative boxes and do the same process, and so on. These three steps define our PGE algorithm, which is detailed in Algorithm 1 and visualized in Figure 2(b).

**Pseudo ground-truth adaptation (PGA).** After obtaining the pseudo ground-truth boxes from PGE, we seek to improve them by taking advantage of a region proposal network (RPN) as used in [28]. Since only image-level labels are available during training, the pseudo ground-truths se-

**Algorithm 1** Pseudo Ground-truth Excavation (PGE)

**Input:** $P, T_{nms}, T_{score}, T_{fusion}$
  **while** $i < n$, $n$ is the number training data **do**
    **for** $j$ in $C$, $C$ is the list of training data class **do**
      $keep = nms(P_i, T_{nms})$
      $G_{nms} = P_i[keep, :]$
      $score\_index = G_{nms}[:, -1] > T_{score}$
      $G_{nms} = G_{nms}[score\_index, :]$
      $G_{del} = h(G_{nms})$, where $h$ is the function of step(ii)
      $iou = IoU(G_{del}, max(G_{del}))$
      **if** $iou > T_{fusion}$ **then**
        $G_{fusion} = f(G_{del})$, where $f$ is the function of step(iii)
        $G_{ij} = G_{fusion}$
      **else**
        $G_{ij} = G_{del}$
      **end if**
    **end for**
  **end while**
**Output:** Pseudo ground-truth boxes $G$

lected by PGE may 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. Our motivation is that the proposals generated by RPN usually have a closer outline than those retrieved pseudo ground-truth bounding boxes. In particular, we train an RPN using the pseudo ground-truth boxes from PGE. For each pseudo grounding-truth box, we choose all the proposals $P_{ro}$ generated by RPN, whose IoU with this pseudo grounding-truth box are larger than a pre-defined threshold $T_{iou}$, and then average the pixel coordinates of these proposals as the final pseudo ground-truth, as shown in Figure 2(c). The procedure is detailed in Algorithm 2.

In Figure 3, we illustrate examples of the pseudo ground-truth mining by our method and the baseline. (*i.e.* selecting the top proposal with the highest predicted score).

### 3.3. Fully-Supervised Detector(FSD)

After generating the refined ground-truths, weakly-supervised detection can be cast as a supervised problem, where the advantages (*i.e.* regression ability) of fully-supervised learning are employed to further improve overall detection performance. In this paper, we choose Fast-RCNN and Faster-RCNN based on VGG16 as our fully-supervised detector. In Faster-RCNN, we train a region proposal network based on the pseudo grounding-truths from PGE, in which we further fine-tune the pseudo ground-truths using the PGA algorithm, and then train it by using these higher-quality pseudo ground-truths. We would like to note that our fully-supervised detector is not specific and any off-the-shelf detectors can be used here, such as YOLO [27], SSD [26], R-FCN [5], etc..

**Algorithm 2** Pseudo Ground-truth Adaptation (PGA)

**Input:** $G$ from PGE algorithm, $T_{iou}, P_{ro}$
  **while** $i < n$ , $n$ is the number training data **do**
    **for** $j$ in $C$, $C$ is the list of training data class **do**
      $iou = IoU(G_{ij}, P_{ro_i})$
      $keep = iou > T_{iou}$
      $G_{ada_{ij}} = mean(P_{ro_i}[keep, :])$
      ${G^*}_{ij} = G_{ada_{ij}}$
    **end for**
  **end while**
**Output:** Final pseudo ground-truth boxes $G^*$

## 4. Experiments

In this section, we experimentally validate our W2F framework and analyze each of its components for weakly-supervised object detection.

### 4.1. Datasets and Evaluation Metrics

**Datasets.** We evaluate our framework on two challenging and widely used benchmarks in weakly-supervised object detection: PASCAL VOC 2007 and 2012, which have 9,963 and 22,531 images from 20 object categories, respectively. For VOC 2007, we use the $trainval$ set for training and use the $test$ set for testing. For VOC 2012, we choose the $trainval$ set to train our network and evaluate on the $test$ set. We stress that we only use image-level labels during training (no bounding-box annotations are used).

**Evaluation metrics.** We follow the standard metrics for weakly-supervised object detection. We use mean average precision (mAP) [8] as the evaluation metric for evaluating our model on the testing set. Also, the correct location (CorLoc) [7] metric is used to evaluate the localization accuracy of our model on the training set. Both metrics comply with the PASCAL criterion, where a positive detection has an IoU>0.5 with the ground-truth.

### 4.2. Implementation Details

Our framework utilizes VGG16 as the backbone network, which is pre-trained on the ImageNet dataset [30]. In the weakly-supervised detector, we refine the instance classifier three times (*i.e.* $K$=3). During training, the total number of iterations is 70K, and the learning rate is 0.001 for the first 40K iterations and then divided by 10 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 $T_{nms}$ for NMS is set to 0.3, while $T_{score}$ and $T_{fusion}$ are set to 0.2 and 0.4 respectively. In the PGA, the IoU threshold $T_{iou}$ is set to 0.5. For the fully-supervised detector (*i.e.* Fast-RCNN and Faster-RCNN) training, all the hyper-parameters are the same as [11, 28]. NMS with 30% IoU threshold is used to calculate mAP and CorLoc.

For data augmentation, we fix the original aspect ratio of

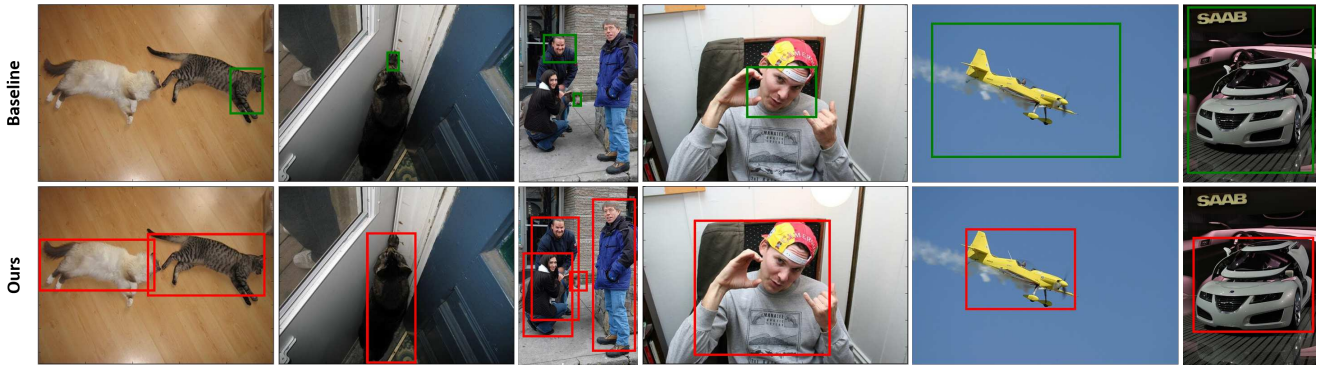


Figure 3. Some examples of pseudo ground-truth boxes generated by different weakly supervised detection methods. The top row shows the results of baseline [32] (*i.e.* selecting the top proposal with the highest predicted score as the pseudo ground-truth). The bottom row shows some pseudo ground-truth boxes mined by our method (*i.e.* PGE and PGA).

images and resize the shortest side to one of these five scales {480, 576, 688, 864, 1,200} for both training and testing, and ensure the longest side is not larger than 2,000 simultaneously. Furthermore, we randomly flip images in the horizontal direction during training. In all our experiments, we run the publicly available deep learning framework Caffe [17] on an NVIDIA GTX TITAN X GPU.

### 4.3. Ablation Studies

We first conduct an ablation experiment to prove the effectiveness of our W2F framework. And then, to validate the contribution of each component including PGE and PGA, we also perform ablation studies by cumulatively adding each of them to the baseline (WSD+FSD), which selecting the highest score of predicted boxes from WSD as the pseudo ground-truths to train an FSD.

**Influence of the W2F framework.** From Table 1 (specifically the $1^{st}$ and $2^{nd}$ rows of the bottom part), we see that our baseline (WSD+FSD1) improves mAP by 4.2% compared to 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. We attribute this to the effect of the pseudo ground-truth and the regression ability of fully-supervised learning. Table 2 shows that the Corloc metric undergoes a similar trend as mAP, where WSD+FSD1 boosts the performance from 61.4% to 65.0%, which further confirms the effectiveness of our framework.

**Influence of the PGE.** To validate the effect of PGE, we conduct an ablation experiment between WSD+FSD1 and WSD+PGE+FSD1. From Table 1 (the $2^{nd}$ and $3^{rd}$ row of the bottom part), we observe that PGE brings about 6% improvement in mAP. Interestingly, our PGE algorithm is more effective for 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.), the reason is that the baseline WSD+FSD1 chooses the topmost scoring detected boxes from the WSD as the ground-truths and only one pseudo ground-truth is found per class even though multiple instances of this class are existing. However, PGE retrieves the pseudo ground-truth box for each instance, and more accurate and tighter pseudo ground-truth boxes are mined than the baseline (*i.e.* WSD + FSD1). In Table 2, we show that Corloc has a similar trend as mAP, whereby PGE brings about 4.4% improvement. Again, we see that Corloc of all non-rigid objects experience a huge boost, and the performance of each class can been found in Table 1 of supplementary material. Figure 4 illustrates the improvement in mAP of each category by the PGE algorithm.

**Influence of the PGA.** We also validate the contribution of the PGA algorithm in the RPN of Faster-RCNN (*i.e.* WSD+PGE+PGA+FSD2). From Table 1 (the $3^{rd}$ and $4^{th}$ row of the bottom part) and Figure 4, PGA further improves the mAP from 51.7% to 52.4%, because proposals generated by RPN are usually closer to the outline of the object than pseudo ground-truths mined by PGE, especially for those pseudo ground-truth boxes including excessive background. Similarly, PGA improves the performance from 69.4% to 70.3% in Corloc as shown in Table 2.

### 4.4. Comparison with State-of-the-Art

We compare the proposed method to other state-of-the-art methods for weakly-supervised object detection, including MIL-based methods [4, 1, 34, 19, 2, 23, 18] and pseudo ground-truth based methods [32, 21].

Table 1 shows mAP performance on the VOC 2007 $test$ set. Our method achieves the highest performance (52.4%), outperforming the state-of-the-art MIL-based method [18] and the state-of-the-art pseudo ground-truth based method [32] by 10.7% and 5.4% respectively. Compared to MIL-based methods, our performance boost mainly comes from two contributions: (1) The combination of the WSDNN [2] and OICR [32] to train a WSD, in which the refinement network guide the weakly-supervised detector to learn the

| Method | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cinbis *et al*. 2017 [4] | 38.1 | 47.6 | 28.2 | 13.9 | 13.2 | 45.2 | 48.0 | 19.3 | 17.1 | 27.7 | 17.3 | 19.0 | 30.1 | 45.4 | 13.5 | 17.0 | 28.8 | 24.8 | 38.2 | 15.0 | 27.4 |
| Bilen *et al*. 2015 [1] | 46.2 | 46.9 | 24.1 | 16.4 | 12.2 | 42.2 | 47.1 | 35.2 | 7.8 | 28.3 | 12.7 | 21.5 | 30.1 | 42.4 | 7.8 | 20.0 | 26.8 | 20.8 | 35.8 | 29.6 | 27.7 |
| Wang *et al*. 2014 [34] | 48.9 | 42.3 | 26.1 | 11.3 | 11.9 | 41.3 | 40.9 | 34.7 | 10.8 | 34.7 | 18.8 | 34.4 | 35.4 | 52.7 | 19.1 | 17.4 | 35.9 | 33.3 | 34.8 | 46.5 | 31.6 |
| Kantorov *et al*. 2016 [19] | 57.1 | 52.0 | 31.5 | 7.6 | 11.5 | 55.0 | 53.1 | 34.1 | 1.7 | 33.1 | 49.2 | 42.0 | 47.3 | 56.6 | 15.3 | 12.8 | 24.8 | 48.9 | 44.4 | 47.8 | 36.3 |
| Bilen *et al*. 2016† [2] | 46.4 | 58.3 | 35.5 | 25.9 | 14.0 | 66.7 | 53.0 | 39.2 | 8.9 | 41.8 | 26.6 | 38.6 | 44.7 | 59.0 | 10.8 | 17.3 | 40.7 | 49.6 | 56.9 | 50.8 | 39.3 |
| Li *et al*. 2016 [23] | 54.5 | 47.4 | 41.3 | 20.8 | 17.7 | 51.9 | 63.5 | 46.1 | 21.8 | 57.1 | 22.1 | 34.4 | 50.5 | 61.8 | 16.2 | **29.9** | 40.7 | 15.9 | 55.3 | 40.2 | 39.5 |
| Tang *et al*. 2017(OICR) [32] | 58.0 | 62.4 | 31.1 | 19.4 | 13.0 | 65.1 | 62.2 | 28.4 | **24.8** | 44.7 | 30.6 | 25.3 | 37.8 | 65.5 | 15.7 | 24.1 | 41.7 | 46.9 | **64.3** | 62.6 | 41.2 |
| Jie *et al*. 2017 [18] | 52.2 | 47.1 | 35.0 | 26.7 | 15.4 | 61.3 | 66.0 | 54.3 | 3.0 | 53.6 | 24.7 | 43.6 | 48.4 | 65.8 | 6.6 | 18.8 | 51.9 | 43.6 | 53.6 | 62.4 | 41.7 |
| Krishna *et al*. 2016 [21] | 53.9 | - | 37.7 | 13.7 | - | - | 56.6 | 51.3 | - | 24.0 | - | 38.5 | 47.9 | 47.0 | - | - | - | - | 48.4 | - | 41.9 |
| Tang *et al*. 2017† [32] | **65.5** | 67.2 | 47.2 | 21.6 | **22.1** | 68.0 | 68.5 | 35.9 | 5.7 | **63.1** | **49.5** | 30.3 | **64.7** | 66.1 | 13.0 | 25.6 | 50.0 | 57.1 | 60.2 | 59.0 | 47.0 |
| WSD | 61.4 | 65.6 | 35.3 | 27.7 | 10.1 | 67.0 | 60.9 | 27.3 | 24.7 | 41.4 | 35.0 | 21.6 | 37.6 | 64.1 | 12.6 | 23.8 | 40.0 | 50.9 | 62.6 | 62.7 | 41.6 |
| WSD+FSD1 | 60.9 | 68.7 | 47.1 | 31.7 | 14.2 | 71.2 | 68.9 | 24.5 | 23.5 | 57.6 | 43.6 | 20.9 | 47.9 | 66.0 | 11.3 | 22.3 | **56.4** | 57.7 | 61.1 | 60.1 | 45.8 |
| WSD+PGE+FSD1 | 64.0 | 67.4 | 49.9 | **32.8** | 15.0 | 71.8 | **69.2** | 70.6 | 24.2 | 55.2 | 49.2 | 64.9 | 54.3 | 65.3 | 24.3 | 23.0 | 49.6 | **60.1** | 60.0 | **62.8** | 51.7 |
| WSD+PGE+PGA+FSD2 | 63.5 | **70.1** | **50.5** | 31.9 | 14.4 | **72.0** | 67.8 | **73.7** | 23.3 | 53.4 | 49.4 | **65.9** | 57.2 | **67.2** | **27.6** | 23.8 | 51.8 | 58.7 | 64.0 | 62.3 | **52.4** |

Table 1. Average precision(AP) (%) of our method and other state-of-the-art methods on the PASCAL VOC 2007 *test* set. The † denotes the results of combining multiple models, others are the results of using single model. FSD1 means Fast-RCNN, and FSD2 represents Faster-RCNN. 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 (*i.e*. using pseudo ground-truths to train a fully-supervised detector).

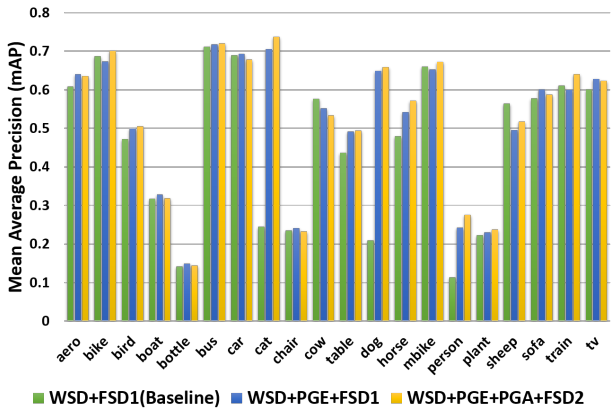


Figure 4. The mAP of each class in different ablation versions of our framework on VOC 2007 *test* set.

| Method | CorLoc(%) |
|---|---|
| Cinbis *et al*. 2017 [4] | 47.3 |
| Bilen *et al*. [1] | 43.7 |
| Wang *et al*. 2014 [34] | 48.5 |
| Kantorov *et al*. 2016 [19] | 55.1 |
| Bilen *et al*. 2016† [1] | 39.3 |
| Li *et al*. 2016 [23] | 52.4 |
| Tang *et al*. 2017(OICR) [32] | 60.6 |
| ie *et al*. 2017 [18] | 56.1 |
| Krishna *et al*. 2016 [21] | 64.3 |
| Tang *et al*. 2017† [32] | 64.3 |
| WSD | 61.4 |
| WSD+FSD1 | 65.0 |
| WSD+PGE+FSD1 | 69.4 |
| WSD+PGE+PGA+FSD2 | **70.3** |

Table 2. Correct localization (CorLoc)(%) of our method and other state-of-the-art methods on the PASCAL VOC 2007 *trainval* set. †, FSD1 and FSD2 have the same meanings as Table1.

more accurate detection bounding-boxes. (2) The regression ability of the fully-supervised detector trained by the pseudo ground-truths. Compared to the pseudo ground-truth based methods, the reason for improvement is that our pseudo ground-truth mining algorithm can retrieve more accurate and tighter pseudo ground-truths. We would like to note that it is unfair to compare our performance with the methods [32] and [21], because the result of [32] (47.0%) is obtained by combining multiple different models. However, our model is trained using a single model. For fair comparison, we compare our method with the baseline, which uses the single model, and the mAP improves by 6.6%. As for the method in [21], their reported mAP is averaged across only ten categories that do not include some difficult categories such as "bottle", "person", etc. On the other hand, our method includes all 20 categories. In this case, the performance of their method is 41.9%, which is lower than our result by 10.5%. Under such unfair conditions, our method is still able to outperform previous state-of-the-art by a large margin, which confirms the effectiveness of our framework.

Table 2 shows the Corloc performance on the VOC 2007 *trainval* set. Our method achieves 70.3% of average Corloc, outperforming all the state-of-the-art methods. The performance of each class are presented in Table1 in the supplementary material. We can observe that all the classes have a better performance than other methods. All the previous state-of-the-art methods [2, 32] encounter a dilemma that the detector inclines to highlight the discriminative parts of an object instead of the whole object leading to poor performance. To the best of our knowledge, our proposed framework is the first to avoid and address these pitfalls.

Table 4 shows our performance in terms of mAP and Corloc on the PASCAL VOC 2012 *test* and *trainval* sets, respectively. Using our framework and the pseudo ground-truth mining algorithm, we achieve state-of-the-art performance. The proposed approach outperforms the second highest performance by 5.3% and 3.8% in mAP and Corloc respectively. For the performance of each class, please refer to Table 2 and Table 3 in the supplementary material.

### 4.5. Run-time of inference

The baseline methods [2, 32] adopt multi-scale testing without horizontal flip. We follow this same setting for fair comparison. In Table 3, we report the inference run-time

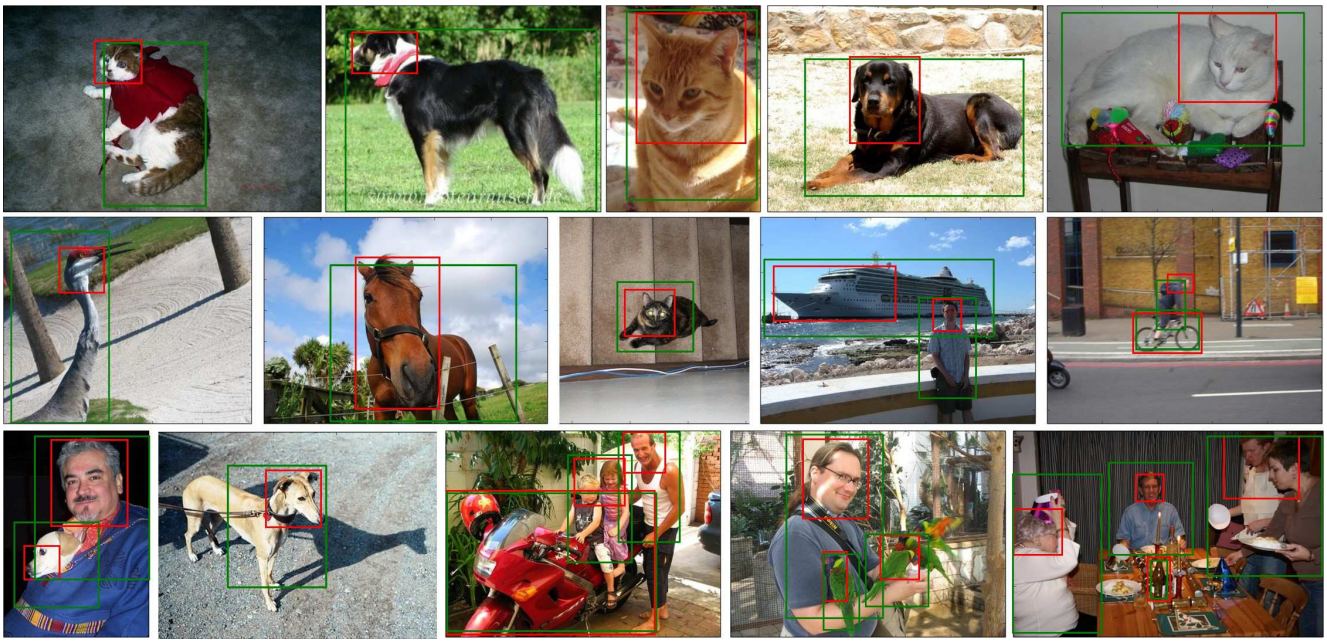

Figure 5. Qualitative detection results of our method (WSD+PGE+PGA+FSD2) and the baseline (WSD+FSD). Green bounding boxes indicate objects detected by our method, while red ones correspond to those detected by the baseline.

(Fast-RCNN run on Pascal TITAN X) and the detection performance under different settings on PASCAL VOC 2007. Table 3 shows that the run-time is similar to the baseline methods, as well as, the fully-supervised Fast-RCNN.

| Training scales | Testing scales | mAP | Run-time (s/img) |
|---|---|---|---|
| multi (480∼1200) | multi (480∼1200) | 52.4 | 0.80 |
| | single (600) | 50.0 | 0.14 |
| single (600) | multi (480∼1200) | 51.6 | 0.80 |
| | single (600) | 49.0 | 0.14 |

Table 3. The performance and the run-time of inference under different settings on PASCAL VOC 2007.

### 4.6. Qualitative Results

In Figure 5, we illustrate some detection results generated by our framework as compared to those generated from the baseline methods. It can been found that the bounding boxes detected by our method surround the object tightly, while the baseline only highlights the most discriminative object parts. This is due to the high quality pseudo ground-truths mined by our PGE and PGA algorithms. Moreover, we visualize some failure results as the last three images in the last row. In these cases, a single retrieved bounding box includes not only one object instance, but it contains multiple adjacent similar instances. So, there is still room for improvement.

## 5. Conclusions

In this paper, we present a novel weakly-supervised to fully-supervised framework (W2F) for object detection.

| Method | mAP(%) | CorLoc(%) |
|---|---|---|
| Kantorov *et al.* 2016 [19] | 35.3 | 54.8 |
| Tang *et al.* 2017(OICR) [32] | 37.9 | 62.1 |
| Jie *et al.* 2017 [18] | 38.3 | 58.8 |
| Tang *et al.* 2017† [32] | 42.5 | 65.6 |
| WSD | 39.6 | 63.0 |
| WSD+FSD1 | 42.4 | 65.5 |
| WSD+PGE+FSD1 | 47.3 | 69.0 |
| WSD+PGE+PGA+FSD2 | **47.8** | **69.4** |

Table 4. Performance of our method and other state-of-the-art methods on the PASCAL VOC 2012. †, FSD1 and FSD2 have the same meanings as Table 1.

Different from previous work, our framework combines the advantages of fully-supervised and weakly-supervised learning. We first use WSDNN and OICR to train a weakly-supervised detector (WSD) end-to-end. And then by the virtue of pseudo ground-truth excavation (PGE) and pseudo ground-truth adaption (PGA), our approach finds high quality pseudo ground-truths from the WSD. Finally, those pseudo ground-truths are fed into a fully-supervised detector to produce the final detection results. Extensive experiments on PASCAL VOC 2007 and 2012 demonstrate the substantial improvements (5.4% and 5.3% in mAP respectively) of our method compared with previous state-of-the-art weakly-supervised detectors.

## Acknowledgments

This work was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research and by Natural Science Foundation of China, Grant No. 61603372.

## References


[1] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, pages 1081–1089, 2015. 2, 3, 6, 7

[2] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2016. 1, 2, 3, 6, 7

[3] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *ICCV*, pages 1431–1439, 2015. 3

[4] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *TPAMI*, 39(1):189–203, 2017. 1, 2, 6, 7

[5] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016. 1, 3, 5

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 3

[7] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012. 5

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 3, 5

[9] C. Gan, C. Sun, L. Duan, and B. Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*, pages 849–866. Springer, 2016. 3

[10] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *CVPR*, pages 923–932, 2016. 3

[11] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 1, 3, 5

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 1

[13] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, pages 2409–2416, 2014. 2

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988. IEEE, 2017. 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[16] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko. Detector discovery in the wild: Joint multiple instance and representation learning. In *CVPR*, pages 2883–2891, 2015. 2

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 6

[18] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. *arXiv preprint arXiv:1704.05188*, 2017. 2, 6, 7, 8

[19] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, pages 350–365. Springer, 2016. 1, 6, 7, 8

[20] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen. Ron: Reverse connection with objectness prior networks for object detection. In *CVPR*, volume 1, page 2, 2017. 1

[21] K. Kumar Singh, F. Xiao, and Y. Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *CVPR*, pages 3548–3556, 2016. 3, 6, 7

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998. 1

[23] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, pages 3512–3520, 2016. 2, 3, 6, 7

[24] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 5

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 5

[28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 3, 4, 5

[29] M. Rochan and Y. Wang. Weakly supervised localization of novel objects using appearance transfer. In *CVPR*, pages 4315–4324, 2015. 2

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1, 5

[31] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, pages 1637–1645, 2014. 2

[32] P. Tang, X. Wang, X. Bai, and W. Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017. 2, 3, 4, 6, 7, 8

[33] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 3

[34] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, pages 431–445. Springer, 2014. 3, 6, 7

[35] Y. Zhang, M. Ding, W. Fu, and Y. Li. Reading recognition of pointer meter based on pattern recognition and dynamic three-points on a line. In *ICMV*, pages 103410K–103410K. International Society for Optics and Photonics, 2017. 2

[36] B. Zhuang, L. Liu, Y. Li, C. Shen, and I. Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In *CVPR*, 2017. 3