# Utilizing the Instability in Weakly Supervised Object Detection

Boxiao Liu[1,2,*], Yan Gao[1,2,*], Nan Guo[1,2], Xiaochun Ye[1,2], Fang Wan[2],
Haihang You[1,2], and Dongrui Fan[1,2,†]

[1]State Key Laboratory of Computer Architecture, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China

{liuboxiao,gaoyan,guonan,yexiaochun,youhaihang,fandr}@ict.ac.cn, wanfang13@mails.ucas.ac.cn

## Abstract

*Weakly supervised object detection (WSOD) focuses on training object detector with only image-level annotations, and is challenging due to the gap between the supervision and the objective. Most of existing approaches model WSOD as a multiple instance learning (MIL) problem. However, we observe that the result of MIL based detector is unstable, i.e., the most confident bounding boxes change significantly when using different initializations. We quantitatively demonstrate the instability by introducing a metric to measure it, and empirically analyze the reason of instability. Although the instability seems harmful for detection task, we argue that it can be utilized to improve the performance by fusing the results of differently initialized detectors. To implement this idea, we propose an end-to-end framework with multiple detection branches, and introduce a simple fusion strategy. We further propose an orthogonal initialization method to increase the difference between detection branches. By utilizing the instability, we achieve 52.6% and 48.0% mAP on the challenging PASCAL VOC 2007 and 2012 datasets, which are both the new state-of-the-arts.*

## 1. Introduction

Weakly supervised object detection (WSOD) has attracted intensive attention recently [24, 1, 5, 2, 26, 33, 31, 29, 12]. Unlike fully supervised object detection, WSOD aims at training detectors with only image-level annotations, which cost much less human labor than bounding boxes annotations.

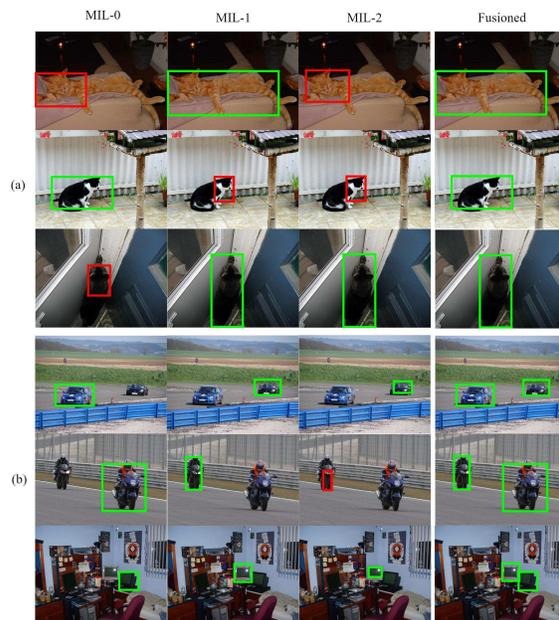A popular solution for WSOD is to formulate it as a



Figure 1. The instability of MIL-based detector. Each column corresponds to one MIL-based detector, and each row corresponds to one image. Green rectangles indicate the positive bounding boxes and red rectangles indicate the negative ones. (a) Images that contain only one object. (b) Images that contain multiple objects. Best viewed in color.

multiple instance learning (MIL) problem. Training images are treated as labeled bags, which consist of multiple candidate bounding boxes. The learning procedure alternates between selecting the most confident proposals and using them to train a detector [5, 24, 16]. Recently, many works combine convolutional neural networks (CNN) with MIL and get promising results [2, 26, 25, 27, 33, 30, 31, 9, 29, 32, 14, 12]. Bilen *et al*. [2] propose a concise end-to-end Weakly Supervised Deep Detection Net-

---

*Equally-contributed.
†Corresponding author.

work (WSDDN), using two parallel branches to get classification and detection information. Based on WSDDN, some works propose to leverage regularization [2], online refinement [26, 25, 27] and curriculum learning [32] to further improve the performance. Weakly supervised semantic segmentation is also introduced in WSOD to provide objectness information [9, 31, 12]. However, there is still a remarkable performance gap compared with fully supervised detectors [13, 20, 19, 18].

We observe that the result of MIL-based detector is unstable. Specifically, detectors with different initializations may localize different regions on the same images. For example, the MIL-0 in the first column of Fig. 1 (a) can correctly localize the cat in the image of the second row, but converge to the head of cat in the first and the third rows. However, MIL-1 in the second column of Fig. 1 (a) succeeds to localize the cat in the first and third rows but fails in the second row. Also, if there are multiple objects in the image, detectors with different initial parameters may localize different one, Fig. 1(b).

The instability of MIL-based detector seems notorious as it limits the performance and leads to a high variance of the result, but we propose to utilize it on the contrary. Our motivation is that, by fusing the results of detectors with different initializations, we can keep the good candidate proposal and suppress the bad ones, so as to improve the detection performance. To implement this idea, we introduce a novel end-to-end framework to utilize the instability. Specifically, we design a multi-branch network structure consisting of a classification branch and multiple detection branches, inspired by WSDDN. The results of detection branches are coupled with the classification branch and the coupled results are further processed by a fusion strategy, called Surrounded Candidate Suppression (SCS). We further refine the fused result by training instance classifiers, following the popular practice [26, 25, 27, 31, 33]. Moreover, in order to further increase the difference between different detection results, we propose a novel category orthogonal parameter initialization method which makes the initialization parameters of the same category in different detection branches orthogonal.

To show the effectiveness of utilizing the instability, we conduct extensive experiments on the challenging PASCAL VOC 2007 and 2012 benchmarks. With the proposed framework, we obtain 52.6% and 48.0% mAP on VOC 2007 and VOC 2012 respectively, which are the new state-of-the-arts.

In summary, the contributions of this paper are listed as follows:

1. We analyze the instability of MIL based detector, providing quantitative evidence and empirical explanation. Based on the analysis, we propose to utilize the instability to improve detection performance by fusing the results of differently initialized detectors.

2. As training multiple detectors with different initialization is time consuming, we propose a simple but effective end-to-end framework and an online fusion strategy to utilize the instability. An orthogonal parameter initialization method is further proposed to increase the difference between detection branches.

3. The proposed framework significantly outperforms previous methods, and creates new state-of-the-arts both on PASCAL VOC2007 and VOC2012 datasets.

## 2. Related Work

The majority of existing methods formulate WSOD as an MIL [10] problem. Under this formulation, a training image is seen as a bag of candidate proposals [5]. Given the labels of bags, the objective of MIL is to train a classifier to correctly separate positive proposals from negative ones.

However, the loss function of MIL is non-convex, and the optimization of MIL is sensitive to initialization [7, 5, 1, 24]. In order to solve this issue, some works introduce better initialization methods. Deselaers *et al.* [7] propose to initialize object locations based on the objectness score. Cinbis *et al.* [5] propose to split the training data into multi folds to escape local optima. Beyond generating better initialization, some works propose to smooth the optimization of MIL to alleviate the non-convexity problem. Bilen *et al.* [1] introduce a smoothed version of MIL that softly labels object instances. Song *et al.* [24] propose to use Nesterov's smoothing technique in latent SVM model. The proposed method is also related to the non-convexity of MIL, but we propose to utilize the instability, which is partly caused by the non-convexity.

In recent years, many works combine deep convolutional neural networks with MIL and achieve promising results [2, 26, 25, 27, 33, 30, 31, 9, 29, 32, 14, 12]. The pretrained CNN models provide generic visual feature representations and reliable initialization for MIL. Bilen *et al.* [2] proposed a two-stream weakly supervised deep detection network (WSDDN), which can be trained with image-level labels in an end-to-end manner. Tang *et al.* [26] add into WSDDN several instance classifiers, and propose an online instance classifier refinement method. Wan *et al.* [29] focus on reducing the randomness of localization during the optimization of network. Also, some works propose to leverage weakly supervised semantic segmentation to improve detection performance. Diba *et al.* [9] generate candidate proposals based on the segmentation result and perform MIL among them. Wei *et al.* [31] introduce two metrics to select reliable candidate proposals by using segmentation result.

Moreover, some researchers propose to combine fully supervised detectors with MIL based ones [30, 33, 22].
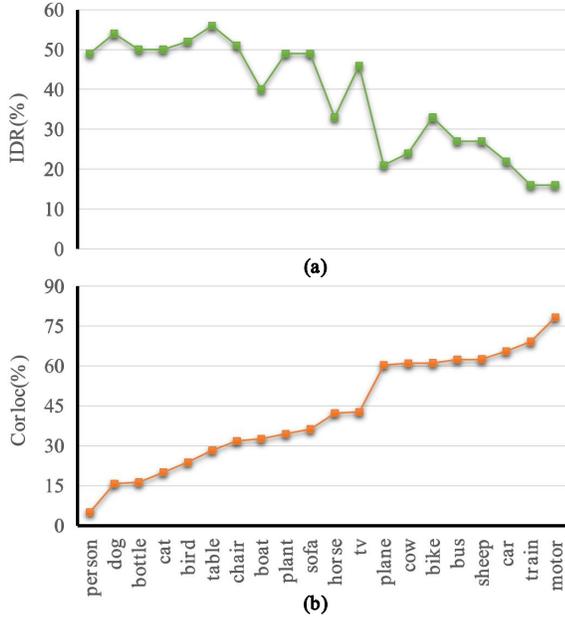
Figure 2. (a) Averaged CorLoc of 10 randomly initialized WS-DDNs on each class. (b) Averaged IDR of each class, obtained by randomly sampling two WSDDNs 10 times.

Wang *et al*. [30] introduce a collaborative learning framework and a consistence-based loss to combine WSDDN and Faster-RCNN. Zhang *et al*. [33] propose to mine better pseudo ground truths from the result of MIL based detector to train a fully supervised detector. Shen *et al*. [22] use a generative adversarial learning framework to build the connection between fully and weakly supervised detectors.

The proposed framework consists of multiple detection branches and is related to [4, 3, 17] that also propose to use multi-branch network to improve detection result. Cheng *et al*. [4] propose to add a decoupled classification refinement module to suppress hard false positive results. Li *et al*. [17] use multiple branches with different dilated convolution layers. Our method is different to these methods because the branches in our method have the same structure. Also, the branches in the proposed framework have different initializations in order to utilize the instability, which is not involved in these methods.

## 3. Analysis of Instability

We observe that the result of MIL based detector shows significant instability, *i.e.*, detectors trained with different initializations often localize different regions in the same image. To quantitatively analyze the instability of MIL based detectors, we first introduce a metric, Inconsistent Detection Rate (IDR), representing the inconsistency between the results of two detectors. In an image, if the IoU of the top scoring bounding boxes of two detectors is less than 0.5, we say the results are inconsistent. IDR indicates
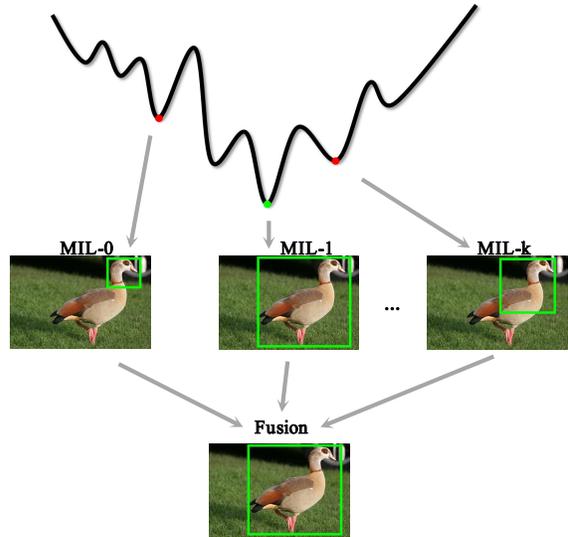


Figure 3. The non-convexity of MIL. Differently initialized detectors may be trapped in different local minimum, leading to different detection results. By fusing the results of different detectors, we may get more accurate localization of object.

the rate of images where the results of two detectors are inconsistent. Formally, the IDR on class $c$ is defined as

$$IDR^c = \frac{|\{I_k^c, \text{ where } IoU(b_{1,k}^c, b_{2,k}^c) < 0.5\}|}{|\{I_k^c\}|}, \quad (1)$$

where $I_k^c$ denotes the $k$th training image with positive label on class $c$, $b_{1,k}^c$ and $b_{2,k}^c$ denote the top scoring bounding boxes of two detectors on the $k$th image. The mean IDR over all classes is defined as

$$mIDR = \frac{\sum_{c=1}^{C} IDR^c}{C}, \quad (2)$$

where $C$ denotes the number of classes.

We choose WSDDN, a popular MIL based detector, as a representative to show the instability problem. We train WSDDN for 10 times with different initial parameters. Then we randomly sample two detectors to compute IDR and mIDR for 10 times and show their averaged values. As shown in Fig. 2 (a), the mIDR reaches 38.3% and the IDRs on some classes are even greater than 50%, which means the detection result in about a half of images changes significantly if we change initial parameters of a detector. Also, the instability is more serious on classes with poor localization performance. From Fig. 2 (a) we can find that the class-specific IDR shows a negative relation with CorLoc.

We think the reasons of this instability are two folds. Firstly, MIL is inherently non-convex and may have many local optimum. Secondly, the proposals in the same image have strong spatial relationship, which is related to the detection task. Specifically, negative proposals that only contain part of object always appear together with the positive
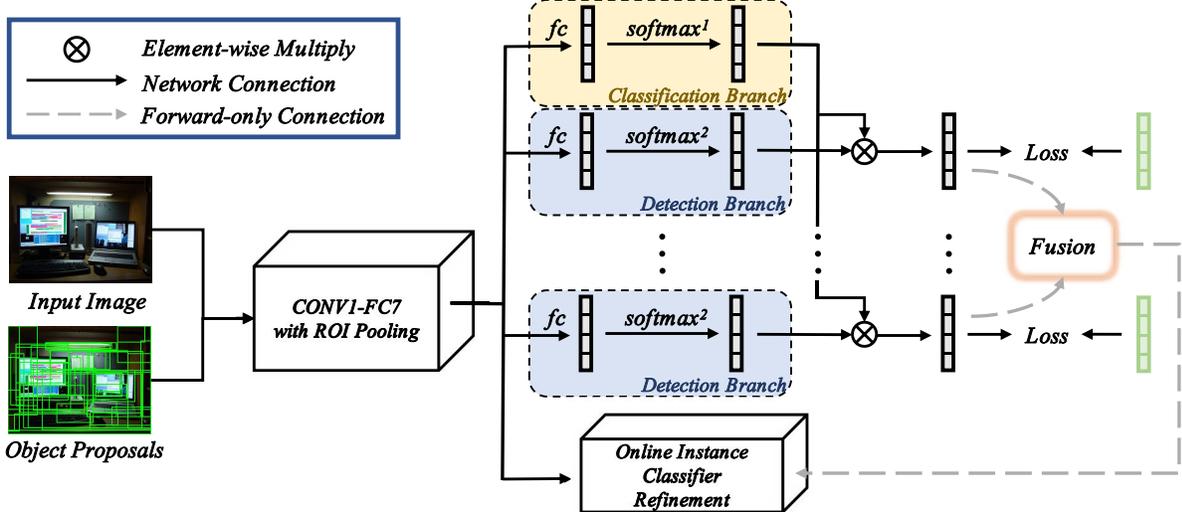
Figure 4. The network architecture of our method. A CNN backbone with ROI pooling and two fully connected layers is used to get the feature vectors of proposals. Then the feature vectors are fed into a classification branch, $K$ detection branches and several instance classifiers. Softmax[1] indicates softmax operation over the classes, and Softmax[2] indicates softmax operation over the proposals.

proposals in the positive bags. However, such negative proposals never appear in a negative bag. Both of such negative proposals and positive proposals may correspond to local optimums of MIL, Fig. 3. For example, an image containing a bird always has a positive proposal that contains the entire bird and some negative proposals that only contain the head of the bird, but images that contain such negative proposals are never labeled as negative. Without instance-level annotations, this spatial correlation between proposals usually makes the detector confused, leading to strong instability.

It seems that the instability is notorious as it limits the detection performance and results in unreliable localization. However, we argue that the instability can be utilized to improve detection performance on the contrary. Considering different top scoring proposals generated by randomly initialized detectors, some of them may be tight bounding boxes while others may be not. If we can fuse these proposals, keep the good ones and discard the bad ones, the detection performance will be improved, Fig. 3.

## 4. Method

To utilize the instability of MIL based detector, we propose to fuse the results of differently initialized detectors. A natural way is to train the detector, such as WSDDN, for several times, and then fuse the results after all the training processes. However, this procedure is time-consuming. So we further propose a novel end-to-end network and an online fusion strategy to utilize the instability of MIL based detector. The overall network architecture is shown in Fig. 4. The main part of the network is a multi-branch structure consisting of a classification branch and multiple detection branches, which are initialized with different pa-

rameters. The results of all branches will be further fused and refined online. Moreover, in order to enlarge the difference between detection branches, we introduce an orthogonal initialization strategy for detection branches. We will elaborate the details of the proposed network, the online fusion strategy and the orthogonal initialization method in the rest of this section.

### 4.1. Multi-branch Network Structure

To utilize the instability of MIL, we need to generate different sets of localization information and fuse them to get better detection results. Inspired by WSDDN, we design a multi-branch network structure to implement this idea. Formally, given a training image $I$ and the corresponding candidate proposals $\mathcal{B}$ generated by selective search method [28], we feed $I$ and $\mathcal{B}$ into a CNN backbone with ROI pooling, which is pretrained on ImageNet dataset. We use the output of FC7 as the features of candidate proposals. Then the network branch into a classification branch and $K$ detection branches. Each branch consists of a fully connected layer and a softmax layer. In the classification branch, the fully connected layer maps the features of proposals into a matrix $x^c \in \mathbb{R}^{C \times |\mathcal{B}|}$, where $C$ is the number of categories. Then a softmax operation along the class-axis is performed to get the classification scores of proposals

$$[\sigma_{cls}(x^c)]_{ij} = \frac{e^{x_{ij}^c}}{\sum_{n=1}^{C} e^{x_{nj}^c}}. \tag{3}$$

In the $k$th detection branch, the matrix after fully connected layer $x^{d,k} \in \mathbb{R}^{C \times |\mathcal{B}|}$ is passed through a softmax operator

along proposal-axis, defined as:

$$[\sigma_{det}(x^{d,k})]_{ij} = \frac{e^{x_{ij}^{d,k}}}{\sum_{n=1}^{|\mathcal{B}|} e^{x_{in}^{d,k}}}. \tag{4}$$

The detection scores of each detection branch will be further coupled with the classification scores of the same classification branch to get the final scores by an element-wise product, defined as:

$$\sigma^k = \sigma_{cls}(x^c) \odot \sigma_{det}(x^{d,k}). \tag{5}$$

For each detection branch, we can compute the image classification score $p_c$ on class $c$ with a summation over all proposals, defined as:

$$p_c^k = \sum_{n=1}^{|\mathcal{B}|} \sigma_{c,n}^k. \tag{6}$$

With the image label $\mathcal{Y} = \{y_1, y_2, \ldots, y_C\}$, the loss for $k$th detection branch $L_k$ and the total loss $L$ are defined as:

$$L^k = -\sum_{c=1}^{C} \{y_c \log p_c^k + (1 - y_c) \log(1 - p_c^k)\} \tag{7}$$

$$L = \sum_{k=1}^{K} L^k \tag{8}$$

Instead of fusing the results of different detectors after the training procedure, we propose a simple online fusion strategy. After forward propagation in each training step, we can get the final scores of detection branches, and then fuse the results of different detection branches based on these scores. The fused results will be further refined by training instance classifiers. Here we follow [26, 25, 27] to design the refinement module, which refines the fused results in a cascaded manner. For more details, please refer to [26, 25, 27].

## 4.2. Online Fusion Strategy

After getting the final scores of each detection branches, we introduce a simple online fusion strategy, called surrounded candidates suppression (SCS), to fuse the results of $K$ detection branches in every training step. The proposed strategy is based on the observation that WSDDN tends to localize the discriminative object parts or the whole object, rather than selecting bounding boxes that contain multiple objects or too many background regions. Thus, if a top scoring bounding box of a detector is surrounded by that of another detector, it is very likely that the surrounded bounding box only contains part of object and should be discarded. Formally, we first get the set of top scoring proposals of all detection branches $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$. Then we remove from $\mathcal{P}$ all proposals that are surrounded by another

---

**Algorithm 1** Surrounded Candidates Suppression (SCS)

**Input:** The final score of $K$ detection branches $\{\sigma^1, \sigma^2, \ldots, \sigma^K\}$; object proposals $\mathcal{B}$; image label $\mathcal{Y}$.
**Output:** Fused detection result $B_{fused}$.
1: Set $B_{fused} \leftarrow \{\}$.
2: **for** $c = 1$ **to** $C$ **do**
3:     **if** $y_c = 1$ **then**
4:         Set $B_{top} \leftarrow \{\}$.
5:         **for** $k = 1$ **to** $K$ **do**
6:             $b_c^k \leftarrow \arg\max_{b_n \in B} \sigma_{c,n}^k$.
7:             $B_{top} \leftarrow B_{top} \bigcup \{b_c^k\}$.
8:         **end for**
9:         **for** $b_k^c \in B_{top}$ **do**
10:            **if** $\exists b \in B_{top}$ s.t. $b_k^c$ is surrounded by $b$ **then**
11:                $B_{top} \leftarrow B_{top} \setminus \{b_k^c\}$.
12:            **end if**
13:        **end for**
14:        $B_{fused}^c \leftarrow \text{NMS}(B_{top})$.
15:        $B_{fused} \leftarrow B_{fused} \bigcup B_{fused}^c$.
16:    **end if**
17: **end for**

---

proposal in $\mathcal{P}$. Finally, a standard NMS with threshold 0.1 is performed among the remained proposals to get the fused result. With SCS, we can discard the bad candidate proposals and keep the good ones as many as possible. To make it clear, we summarize the process of SCS in Alg. 1.

## 4.3. Orthogonal Initialization

The detection branches in the proposed network is supposed to be different with each other, so the fused results can be better than the results of original branches. The difference of detection branches comes from the randomness of initialization, which is not reliable. We argue that the proposed method can benefit from more significant difference between the initial parameters of detection branches. So we propose an orthogonal initialization method, making sure that the parameters of the fully connected layers of different detection branches are orthogonal with each other on every class.

Similar initialization methods have been proposed in [21] to avoid the vanishment of gradient in recurrent neural networks. We follow the implementation in [21] to design our orthogonal initialization method. For the $k$th detection branch, the parameters of the fully connected layer is denoted as a matrix $m^k \in \mathbb{R}^{l \times C}$, where $l$ denotes the length of the feature vector of a proposal. For each class $c$, we construct get an orthogonal matrix $q^c \in \mathbb{R}^{l \times K}$ by performing QR factorization on a random matrix. Then we assign the value of the $k$th column in $q$ to the $c$th column of $m^k$.

| Method | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WSDDN [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 |
| OICR [26] | 58.5 | 63.0 | 35.1 | 16.9 | 17.4 | 63.2 | 60.8 | 34.4 | 8.2 | 49.7 | 41.0 | 31.3 | 51.9 | 64.8 | 13.6 | 23.1 | 41.6 | 48.4 | 58.9 | 58.7 | 42.0 |
| WCCN [9] | 49.5 | 60.6 | 38.6 | 29.2 | 16.2 | 70.8 | 56.9 | 42.5 | 10.9 | 44.1 | 29.9 | 42.2 | 47.9 | 64.1 | 13.8 | 23.5 | 45.9 | 54.1 | 60.8 | 54.5 | 42.8 |
| TS2C [31] | 59.3 | 57.5 | 43.7 | 27.3 | 13.5 | 63.9 | 61.7 | 59.9 | 24.1 | 46.9 | 36.7 | 45.6 | 39.9 | 62.6 | 10.3 | 23.6 | 41.7 | 52.4 | 58.7 | 56.6 | 44.3 |
| PCL [25] | 57.1 | 67.1 | 40.9 | 16.9 | 18.8 | 65.1 | 63.7 | 45.3 | 17.0 | 56.7 | 48.9 | 33.2 | 54.4 | 68:3 | 16:8 | 25.7 | 45.8 | 52.2 | 59.1 | 62.0 | 45.8 |
| MLEM [29] | 55.6 | 66.9 | 34.2 | 29.1 | 16.4 | 68.8 | 68.1 | 43.0 | 25.0 | 65.6 | 45.3 | 53.2 | 49.6 | 68.6 | 2.0 | 25.4 | 52.5 | 56.8 | 62.1 | 57.1 | 47.3 |
| WSRPN [27] | 60.3 | 66.2 | 45.0 | 19.6 | 26.6 | 68.1 | 68.4 | 49.4 | 8.0 | 56.9 | 55.0 | 33.6 | 62.5 | 68.2 | 20.6 | 29.0 | 49.0 | 54.1 | 58.8 | 58.4 | 47.9 |
| OICR+FRCNN [26] | 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 |
| ZLDN [32] | 55.4 | 68.5 | 50.1 | 16.8 | 20.8 | 62.7 | 66.8 | 56.5 | 2.1 | 57.8 | 47.5 | 40.1 | 69.7 | 68.2 | 21.6 | 27.2 | 53.4 | 56.1 | 52.5 | 58.2 | 47.6 |
| CL [30] | 61.2 | 66.6 | 48.3 | 26.0 | 15.8 | 66.5 | 65.4 | 53.9 | 24.7 | 61.2 | 46.2 | 53.5 | 48.5 | 66.1 | 12.1 | 22.0 | 49.2 | 53.2 | 66.2 | 59.4 | 48.3 |
| PCL+FRCNN [25] | 63.2 | 69.9 | 47.9 | 22.6 | 27.3 | 71.0 | 69.1 | 49.6 | 12.0 | 60.1 | 51.5 | 37.3 | 63.3 | 63.9 | 15.8 | 23.6 | 48.8 | 55.3 | 61.2 | 62.1 | 48.8 |
| WSRPN+FRCNN [27] | 63.0 | 69.7 | 40.8 | 11.6 | 27.7 | 70.5 | 74.1 | 58.5 | 10.0 | 66.7 | 60.6 | 34.7 | 75.7 | 70.3 | 25.7 | 26.5 | 55.4 | 56.4 | 55.5 | 54.9 | 50.4 |
| Baseline(WSDDN+ODR) | 44.3 | 71.0 | 45.6 | 24.2 | 15.4 | 70.0 | 69.5 | 47.0 | 21.8 | 65.9 | 37.5 | 59.8 | 52.7 | 70.4 | 7.2 | 26.4 | 59.8 | 60.5 | 67.5 | 64.4 | 49.0 |
| Ours | 63.4 | 70.5 | 45.1 | 28.3 | 18.4 | 69.8 | 65.8 | 69.6 | 27.2 | 62.6 | 44.0 | 59.6 | 56.2 | 71.4 | 11.9 | 26.2 | 56.6 | 59.6 | 69.2 | 65.4 | 52.0 |
| Ours+FRCNN | 62.7 | 69.1 | 43.6 | 31.1 | 20.8 | 69.8 | 68.1 | 72.7 | 23.1 | 65.2 | 46.5 | 64.0 | 67.2 | 66.5 | 10.7 | 23.8 | 55.0 | 62.4 | 69.6 | 60.31 | 52.6 |

Table 1. Detection average precision (AP %) on the PASCAL VOC 2007 test set. The upper part shows the results of weakly supervised detectors, and the second part shows the results of fully supervised detector trained by using the output of weakly supervised detectors as pseudo ground truth.

| Method | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WSDDN+context[15] | 64.0 | 54.9 | 36.4 | 8.1 | 12.6 | 53.1 | 40.5 | 28.4 | 6.6 | 35.3 | 34.4 | 49.1 | 42.6 | 62.4 | 19.8 | 15.2 | 27.0 | 33.1 | 33.0 | 50.0 | 35.3 |
| WCCN [9] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 37.9 |
| OICR [26] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 37.9 |
| TS2C [31] | 67.4 | 57.0 | 37.7 | 23.7 | 15.2 | 57.0 | 49.1 | 64.8 | 15.1 | 39.4 | 19.3 | 48.4 | 44.5 | 67.2 | 2.1 | 23.3 | 35.1 | 40.2 | 46.6 | 45.8 | 40.0 |
| PCL [25] | 63.4 | 64.2 | 44.2 | 25.6 | 26.4 | 54.5 | 55.1 | 30.5 | 11.6 | 51.0 | 15.8 | 39.4 | 55.9 | 70.7 | 8.2 | 26.3 | 46.9 | 41.3 | 44.1 | 57.7 | 41.6 |
| MLEM [29] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 42.4 |
| OICR+FRCNN [26] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 42.5 |
| ZLDN [32] | 54.3 | 63.7 | 43.1 | 16.9 | 21.5 | 57.8 | 60.4 | 50.9 | 1.2 | 51.5 | 44.4 | 36.6 | 63.6 | 59.3 | 12.8 | 25.6 | 47.8 | 47.2 | 48.9 | 50.6 | 42.9 |
| CL [30] | 70.5 | 67.8 | 49.6 | 20.8 | 22.1 | 61.4 | 51.7 | 34.7 | 20.3 | 50.3 | 19.0 | 43.5 | 49.3 | 70.8 | 10.2 | 20.8 | 48.1 | 41.0 | 56.5 | 56.7 | 43.3 |
| WSRPN [27] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 43.4 |
| PCL+FRCNN [25] | 69.0 | 71.3 | 56.1 | 30.3 | 27.3 | 55.2 | 57.6 | 30.1 | 8.6 | 56.6 | 18.4 | 43.9 | 64.6 | 71.8 | 7.5 | 23.0 | 46.0 | 44.1 | 42.6 | 58.8 | 44.2 |
| TS2C+FRCNN [31] | 73.9 | 64.2 | 45.7 | 30.7 | 16.4 | 62.0 | 56.7 | 62.4 | 16.1 | 52.2 | 20.0 | 39.5 | 54.0 | 72.1 | 2.7 | 25.9 | 46.6 | 44.7 | 47.9 | 54.4 | 44.4 |
| WSRPN+FRCNN [27] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 45.7 |
| Ours | 72.7 | 68.8 | 51.6 | 29.4 | 29.1 | 60.3 | 58.0 | 59.0 | 22.6 | 61.9 | 22.4 | 52.3 | 59.8 | 74.0 | 7.2 | 28.1 | 53.4 | 33.5 | 54.5 | 60.7 | 48.0 |

Table 2. Detection average precision (AP %) on the PASCAL VOC 2012 test set.

# 5. Experiments

## 5.1. Datasets and Evaluation Metrics

We evaluate our method on two widely used datasets, PASCAL VOC 2007 and 2012 [11]. For each dataset, we use the *trainval* set for training, and the *test* set for testing. Only image-level labels are used to train the network.

For evaluation, we choose two kinds of measurements: 1) Average Precision (AP) and the mean of AP (mAP) on the *test* set, following the standard PASCAL VOC protocol; 2) CorLoc [8] on the *trainval* set to evaluate the localization accuracy. Based on the PASCAL criterion, a bounding box is considered to be positive if it has an $IoU \geq 0.5$ with the ground-truth for both metrics.

## 5.2. Implementation Details

We built our model on a VGG16 [23] network pretrained on ImageNet [6]. We remove the last fully connected layer, and replace the last max-pooling layer with an ROI pooling layer. The mini-batch for training is set to 2. The momentum and weight decay is set to 0.9 and $5 \times 10^{-4}$ respectively. The learning rate is $5 \times 10^{-4}$ for the first 10 epochs and then decrease to $5 \times 10^{-5}$ for the following 5 epochs.

The image proposals are generated by selective search [28]. For data augmentation, we use five image scales $\{480, 576, 688, 864, 1200\}$, with horizontal flips for
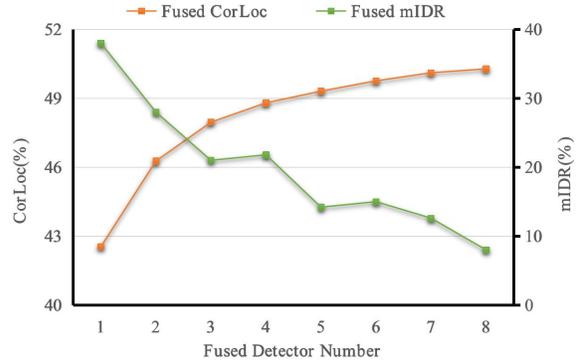


Figure 5. Averaged CorLoc and mIDR of fused results with different fused detector numbers $K$, obtained by randomly sampling $K$ WSDDNs 10 times.

both training and testing. In each training step, we randomly choose a scale to resize the image and then the image is randomly flipped. For all experiments, an NMS of 0.3 is employed to get final detection result. The average score of 10 augmented images is used as the final proposal scores. Our experiments are implemented based on PyTorch deep learning framework, and are conducted on NVIDIA GTX TitanX GPU.

## 5.3. The Effectiveness of Utilizing the Instability

To demonstrate the effectiveness of utilizing the instability, we choose WSDDN as the basic detector, and fuse

| Method | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | mbike | person | plant | sheep | sofa | train | tv | CorLoc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WSDDN [2] | 68.9 | 68.7 | 65.2 | 42.5 | 40.6 | 72.6 | 75.2 | 53.7 | 29.7 | 68.1 | 33.5 | 45.6 | 65.9 | 86.1 | 27.5 | 44.9 | 76.0 | 62.4 | 66.3 | 66.8 | 58.0 |
| OICR [26] | 85.4 | 78.0 | 61.6 | 40.4 | 38.2 | 82.2 | 84.2 | 46.5 | 15.2 | 80.1 | 45.2 | 41.9 | 73.8 | 89.6 | 18.9 | 56.0 | 74.2 | 62.1 | 73.0 | 77.4 | 61.2 |
| WCCN [9] | 83.9 | 72.8 | 64.5 | 44.1 | 40.1 | 65.7 | 82.5 | 58.9 | 33.7 | 72.5 | 25.6 | 53.7 | 67.4 | 77.4 | 26.8 | 49.1 | 68.1 | 27.9 | 64.5 | 55.7 | 56.7 |
| TS2C [31] | 84.2 | 74.1 | 61.3 | 52.1 | 32.1 | 76.7 | 82.9 | 66.6 | 42.3 | 70.6 | 39.5 | 57.0 | 61.2 | 88.4 | 9.3 | 54.6 | 72.2 | 60.0 | 65.0 | 70.3 | 61.0 |
| PCL [25] | 81.7 | 82.4 | 63.4 | 41.0 | 42.4 | 79.7 | 84.2 | 54.9 | 23.4 | 78.8 | 54.4 | 46.0 | 75.9 | 89.6 | 22.8 | 51.3 | 72.2 | 66.1 | 74.9 | 76.0 | 63.0 |
| MLEM [29] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 61.4 |
| WSRPN [27] | 81.2 | 81.2 | 60.7 | 36.7 | 52.3 | 80.7 | **89.0** | 65.1 | 20.5 | 86.3 | 61.6 | 49.5 | 86.4 | 92.4 | 41.4 | **62.6** | 79.4 | 62.4 | 73.0 | 75.6 | 66.9 |
| OICR+FRCNN [26] | 85.8 | 82.7 | 62.8 | 45.2 | 43.5 | **84.8** | 87.0 | 46.8 | 15.7 | 82.2 | 51.0 | 45.6 | 83.7 | 91.2 | 22.2 | 59.7 | 75.3 | 65.1 | 76.8 | 78.1 | 64.3 |
| ZLDN [32] | 74.0 | 77.8 | 65.2 | 37.0 | 46.7 | 75.8 | 83.7 | 58.8 | 17.5 | 73.1 | 49.0 | 51.3 | 76.7 | 87.4 | 30.6 | 47.8 | 75.0 | 62.5 | 64.8 | 68.8 | 61.2 |
| CL [30] | 85.8 | 80.4 | **73.0** | 42.6 | 36.6 | 79.7 | 82.8 | 66.0 | 34.1 | 78.1 | 36.9 | 68.6 | 72.4 | 91.6 | 22.2 | 51.3 | 79.4 | 63.7 | 74.5 | 74.6 | 64.7 |
| PCL+FRCNN [25] | 83.8 | 85.1 | 65.5 | 43.1 | 50.8 | 83.2 | 85.3 | 59.3 | 28.5 | 82.2 | 57.4 | 50.7 | 85.0 | 92.0 | 27.9 | 54.2 | 72.2 | 65.9 | 77.6 | **82.1** | 66.6 |
| WSRPN+FRCNN [27] | 83.8 | 82.7 | 60.7 | 35.1 | **53.8** | 82.7 | 88.6 | 67.4 | 22.0 | 86.3 | **68.8** | 50.9 | **90.8** | **93.6** | **44.0** | 61.2 | 82.5 | 65.9 | 71.1 | 76.7 | 68.4 |
| Baseline | 64.2 | 83.5 | 63.1 | 45.2 | 38.5 | 82.2 | 86.7 | 57.6 | 35.5 | 83.6 | 41.8 | 69.5 | 69.0 | 90.4 | 20.1 | 56.8 | 83.5 | **66.9** | 78.3 | 79.6 | 64.8 |
| Ours | 84.2 | 84.7 | 59.5 | 52.7 | 37.8 | 81.2 | 83.3 | 72.4 | 41.6 | 84.9 | 43.7 | 69.5 | 75.9 | 90.8 | 18.1 | 54.9 | 81.4 | 60.8 | 79.1 | 80.6 | 66.9 |
| Ours+FRCNN | **86.7** | **85.9** | 63.4 | **55.3** | 42.0 | **84.8** | 85.2 | **78.2** | **47.2** | **88.4** | 49.0 | **73.3** | 84.0 | 92.8 | 20.5 | 56.8 | **84.5** | 62.9 | **82.1** | 78.1 | **70.0** |

Table 3. CorLoc on the *trainval* set of VOC 2007.

| Method | VOC 2012 |
|---|---|
| WSDDN+context[15] | 54.8 |
| ZLDN [32] | 61.5 |
| OICR [26] | 63.5 |
| TS2C [31] | 64.4 |
| PCL [25] | 65.0 |
| CL [30] | 65.2 |
| OICR+FRCNN [26] | 65.6 |
| WSRPN [27] | 67.2 |
| PCL+FRCNN [25] | 68.0 |
| WSRPN+FRCNN [27] | **69.3** |
| Ours | 67.4 |

Table 4. CorLoc on the *trainval* set of VOC 2012.



Figure 6. Results of different settings of branch number $K$ and initialization strategies. "Orthogonal" indicates the orthogonal initialization method. "Gaussian" indicates the Gaussian initialization method.

the results of WSDDNs that are initialized with different parameters. We use the same fusion strategy introduced in Section 4.2, while we only keep the proposal with largest final score after NMS for the convenience of calculating CorLoc. As shown in Fig. 5, the CorLoc of the fused result increases monotonically as the number of fused detectors increases, and the instability of fused result decreases. Even by fusing two WSDDNs, the CorLoc increases from 42.54% to 46.27%, showing the effectiveness of utilizing the instability.

## 5.4. Ablation Studies

We first compared the proposed method with the baseline model, which combines WSDDN and the refinement module. Then we discuss the influence of detection branch number and the class-specific orthogonal initialization strategy. Without loss generality, we only conduct the ablation experiments on VOC 2007.

**Comparison with Baseline**  To show the effectiveness of the proposed framework, we compare the result of our method with a baseline framework containing a original WSDDN. As shown in Table. 1, our method improves the mAP from 49.0% to 52.0%. The performance on almost all classes has been improved, such as aeroplane (mAP from 44.3% to 63.4%), cat (mAP from 47.0% to 69.6%) and chair

(mAP from 21.8% to 27.2%). Our model can mine more complete object bounding boxes by the fusion of multiple detectors while WSDDN can only find object parts. Also, the fusion strategy has the capacity of generating multiple candidate proposals with high confidence for the refinement module, further improving the detection performance.

**Influence of Detection Branch Number**  In Fig. 6, we illustrate the result of ablation study on different numbers of detection branches. Even adding one more detection branch can significantly boost the performance (mAP from 49.0% to 51.6%), which confirms the effectiveness of our method. With 3 detection branches, the performance achieve the peak. The detection performance decreases slightly as the number of branches further increases. We think the reason may be that the online fusion strategy introduce the risk of localizing too big bounding boxes and this risk outweighs the gain of adding more branches when the number of detection branches is greater than 3. Thus, we set the detection branch number $K$ to 3 in other experiments.

**Influence of Orthogonal Initialization**  To validate the effectiveness of orthogonal initialization, we compare the proposed initialization method with a popular Gaussian ini-

Figure 7. Some detection results of the proposed method. The green rectangles denote the correct detections (IoU $\geq 0.5$), and the red rectangles denote the failed ones.

tialization method, which samples values from Gaussian distribution to initialize the parameters. As shown in Fig. 6, orthogonal initialization method significantly improves the detection performance. The effectiveness of orthogonal initialization further confirms the analysis in Section 3.

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

In this subsection, we present the result of our framework compared with other state-of-the-art methods. Table. 1 shows the result on VOC 2007 dataset, and Table. 2 shows the result on VOC 2012 dataset. On VOC 2007, our model obtains 52.0% mAP. On VOC 2012, our model obtains 48.0% mAP. Our method outperforms previous methods with a large margin on both datasets.

Many works propose to train a fully supervised detector by using the result of MIL based detector as pseudo ground-truth, and show significant improvement of performance. Following Tang *et al*. [26], we also use the top-scoring proposals produced by the proposed framework as pseudo ground-truth to train a Fast-RCNN [13]. As shown in Table. 1, the detection performance on VOC 2007 is further improved to 52.6% in mAP, which is the new state-of-the-art. The CorLoc results of ours on VOC 2007 and VOC 2012 are reported in Table. 3 and Table. 4, which also show the same trend.

We illustrate some detection results of our framework in Fig. 7. Although our model creates the new state-of-the-art, the detection results on some classes, such as person, chair and bottle, are still undesirable. The main failure for person is that the proposed method only finds different parts of person, such as face and hand. Although the inconsistency between the multiply detection branches is large, they all con-

verge to object parts. As for indoor objects such as chairs and bottles, the co-occurrence of objects and backgrounds, or of different objects, is more common and makes it difficult to separate objects from contexts or from each other.

## 6. Conclusions

In this paper, we analyze the instability of MIL-based detector and introduce a metric IDR to measure the instability. Although the instability seems harmful, we propose to utilize it to get more accurate localization result. We propose an end-to-end network architecture and introduce an online fusion strategy to reduce computation cost. Also, a novel orthogonal initialization method is introduced to increase the difference between detection branches. Combined with refinement module, the proposed framework surpasses all previous methods and creates new state-of-the-art.

# References

[1] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly supervised object detection with posterior regularization. In *British Machine Vision Conference*, volume 3, 2014.

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

[3] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Decoupled classification refinement: Hard false positive suppression for object detection. *arXiv preprint arXiv:1810.04002*, 2018.

[4] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018.

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

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer, 2010.

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

[9] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 914–922, 2017.

[10] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

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

[12] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1277–1286, 2018.

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

[14] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017.

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

[16] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[17] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. *arXiv preprint arXiv:1901.01892*, 2019.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

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

[21] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[22] Yunhan Shen, Rongrong Ji, Shengchuan Zhang, Wangmeng Zuo, and Yan Wang. Generative adversarial learning towards fast weakly supervised detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5764–5773, 2018.

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

[24] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.

[25] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Loddon Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[26] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017.

[27] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018.

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

[29] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1306, 2018.

[30] Jiajie Wang, Jiangchao Yao, Ya Zhang, and Rui Zhang. Collaborative learning for weakly supervised object detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 971–977. AAAI Press, 2018.

[31] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 434–450, 2018.

[32] Xiaopeng Zhang, Jiashi Feng, Hongkai Xiong, and Qi Tian. Zigzag learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4262–4270, 2018.

[33] Yongqiang Zhang, Yancheng Bai, Mingli Ding, Yongqiang Li, and Bernard Ghanem. W2f: A weakly-supervised to fully-supervised framework for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 928–936, 2018.