C-MIDN: Coupled Multiple Instance Detection Network With Segmentation Guidance for Weakly Supervised Object Detection

Yan Gao[1,2,*], Boxiao Liu[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
{gaoyan,liuboxiao,guonan,yexiaochun,youhaihang,fandr}@ict.ac.cn, wanfang13@mails.ucas.ac.cn

Abstract

Weakly supervised object detection (WSOD) that only needs image-level annotations has obtained much attention recently. By combining convolutional neural network with multiple instance learning method, Multiple Instance Detection Network (MIDN) has become the most popular method to address the WSOD problem and been adopted as the initial model in many works. We argue that MIDN inclines to converge to the most discriminative object parts, which limits the performance of methods based on it. In this paper, we propose a novel Coupled Multiple Instance Detection Network (C-MIDN) to address this problem. Specifically, we use a pair of MIDNs, which work in a complementary manner with proposal removal. The localization information of the MIDNs is further coupled to obtain tighter bounding boxes and localize multiple objects. We also introduce a Segmentation Guided Proposal Removal (SGPR) algorithm to guarantee the MIL constraint after the removal and ensure the robustness of C-MIDN. Through a simple implementation of the C-MIDN with online detector refinement, we obtain 53.6% and 50.3% mAP on the challenging PASCAL VOC 2007 and 2012 benchmarks respectively, which significantly outperform the previous state-of-the-arts.

1. Introduction

Recent development of Convolutional Neural Networks (CNN) [18] has helped object detection to achieve superior performance [13, 23, 22, 20]. However, to train such object detectors requires large scale datasets with accurate bounding box annotations, which cost quite a lot of human labor to get. To address this problem, Weakly Supervised Object Detection (WSOD), which needs only image-level annotations during training, becomes increasingly attractive. Compared with bounding box annotations, image-level annotations are much easier to collect, and can also be massively obtained through the Internet.

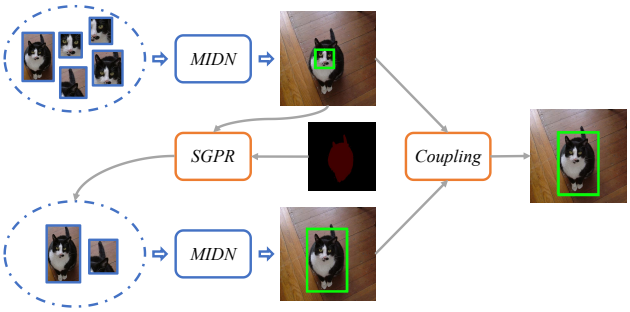

Figure 1. Illustration of the proposed C-MIDN. Green rectangles indicate the top scoring bounding boxes. Two MIDNs work in a complementary way and generate candidates for coupling. Best viewed in color.

To localize objects in cluttered scene without bounding box annotations, a common way is to formulate WSOD as a Multiple Instance Learning (MIL) problem. In recent years, CNN has been introduced into MIL, which is referred as Multiple Instance Detection Network (MIDN), to improve the detection performance. Bilen and Vedaldi [4] propose a concise end-to-end Weakly Supervised Deep Detection Network (WSDDN). WSDDN is effective and convenient to implement, thus many works choose it as a basic MIDN. However, due to the inconsistency between the training objective and supervision, WSDDN tends to localize the most discriminative object parts rather than the entire object.

*Equally-contributed.
†Corresponding author.

Some works propose to use refinement modules combined with WSDDN to solve this problem. Tang *et al*. [31] propose an online detector refinement method to refine the output of WSDNN. Wang *et al*. [37] introduce a collaborative leaning framework which combines WSDDN and Faster RCNN in one end-to-end network to improve the detection performance. However, the performance of these methods is still limited by the performance of basic WSDDN. Once WSDDN converges to parts of objects and fails to generate reasonable initial detection on most training images, these methods have little chance to localize the tight object prediction boxes.

Our motivation is: while single MIDN inclines to converge to the most discriminative parts of objects, we can couple the localization information of MIDNs that work in a complementary manner to alleviate this issue. Based on this idea, we propose a Coupled Multiple Instance Detection Network (C-MIDN). C-MIDN consists of two MIDNs and we use proposal removal to force them to mine different candidate bounding boxes. In particular, we remove the top-scoring proposals of the first MIDN from the input of the second one. If the first detector finds the proposal containing only object parts, such removal can force the second detector to localize the entire object, Fig. 1. Also, the second detector may find new object when there are multiple objects in the image. But the proposal removal must be performed carefully to guarantee that there are still correct object bounding boxes after the removal. Otherwise the removal will destroy the MIL constraint and lead the second detector to go astray. To make the MIDNs more robust, we further introduce a segmentation guided proposal removal algorithm. This is based on the observation that if the detection result cannot cover the segmentation area, the detection result either contains parts of objects or misses some object instances. In both cases, there are tight proposals after removing the result. So we leverage weakly supervised semantic segmentation method to generate the segmentation map, and introduce the segmentation cover rate as a metric to guide the proposal removal operation. Finally, we couple the localization evidence of MIDNs to obtain tighter bounding boxes and localize multiple objects, by applying a priority based suppression algorithm.

Our C-MIDN can be combined with MIDN-based methods. In this paper we implement C-MIDN with popular online detector refinement (ODR) method, and conduct extensive experiments on challenging PASCAL VOC 2007 and 2012 benchmarks. With C-MIDN, we obtain 53.6% and 50.3% mAP on VOC 2007 and VOC 2012 respectively, both significantly outperform the previous state-of-the-arts.

In summary, the contributions of this paper are three folds.

1. We propose a novel coupled multiple instance detection network. By combining a pair of MIDNs with proposal removal and further coupling the results, our method can find complete bounding box and localize multiple instances.

2. We further propose a segmentation guided proposal removal algorithm to make the MIDNs more robust by guaranteeing the MIL constraint after proposal removal.

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

2. Related Work

Traditional Multiple Instance Learning To achieve localization with only image-level annotations provided, most of previous works [17, 7, 28, 5, 14, 3, 2, 24, 27] formulate WSOD as an MIL problem [10]. Under this formulation, an image can be treated as a bag of candidate proposals generated by object proposal methods. Learning procedure alternates between training the detector and selecting positive proposals. Such MIL strategy leads to a non-convex optimization problem, which is sensitive to the initialization and likely to get stuck in local optima. Some works try to find better initialization methods [17, 7, 28, 5, 14]. Jie *et al*. [14] proposed a self-taught approach to harvest high-quality positive object proposals samples. Deselaers *et al*. [7] use objectness score to initialize the object location. Cinbis *et al*. [5] proposed a multi-fold MIL by splitting the training data to multi-fold to escape local optima.

Multiple Instance Detection Network In recent years, many end-to-end frameworks have been proposed to combine MIL and CNN [4, 31, 30, 15, 42, 36, 35, 34, 37, 25, 32, 9]. Bilen and Vedaldi [4] proposed WSDDN, which consists of two parallel data streams to get classification and detection confidence of proposals respectively. A spatial regulariser which forces the features of top scoring region and regions with high overlap to be the same is further added to guarantee the spatial smoothness.

Many works incorporate WSDDN into their framework and improve the detection performance. Tang *et al*. [31] combine WSDDN with several instance classifiers, and propose an online instance classifier refinement method to refine the initial candidates of WSDDN. PCL [30] uses a graph-based center cluster method and average MIL loss based on [31]. Zhang *et al*. [43] propose a Weakly-Supervised to Fully-Supervised Framework(W2F) which use PGA and PGE to mine better pseudo ground truth from MIDN to train a fully-supervised detector. Wang *et al*. [37] introduce a collaborative leaning framework which combine WSDDN detector and Faster RCNN in one end-to-end network, and use feature sharing to improve WSDNN at the same time. These methods have achieved promising results,

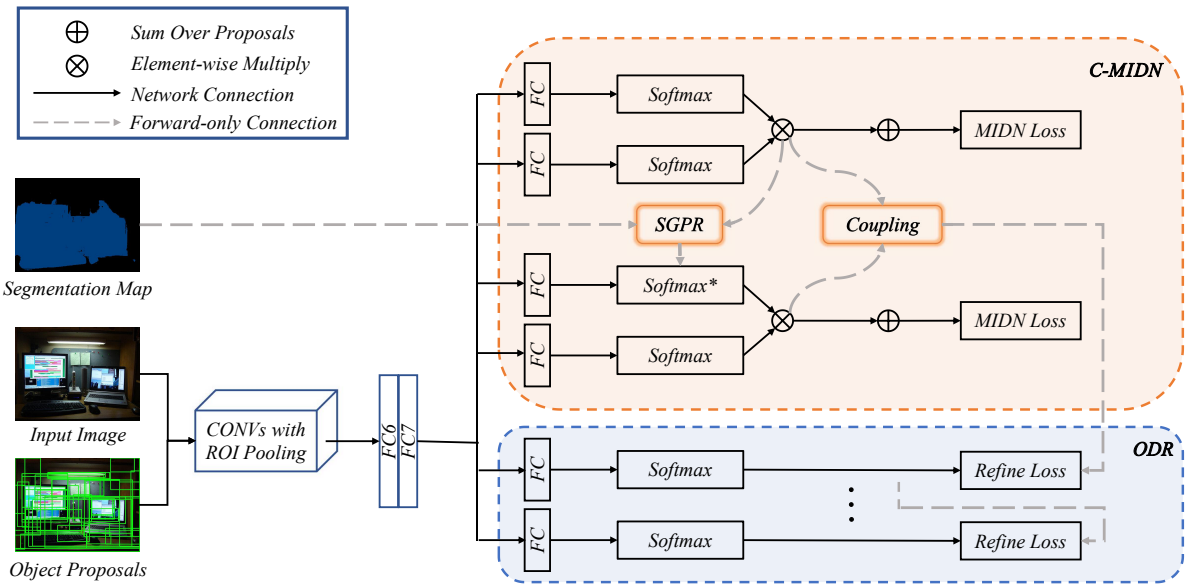

Figure 2. The proposed architecture. A backbone network with ROI-pooling layer is used to get the feature vectors of candidate proposals. Then these feature vectors are fed in two main modules, C-MIDN and ODR. In C-MIDN, two MIDNs work in a complementary way with segmentation guided proposal removal, and the results of two MIDNs are coupled to mine more complete proposals. In ODR, there are several instance classifiers. The supervision of the first stage comes from the coupled result of C-MIDN, and the supervision of other stages comes from their preceding stages. In the second MIDN of C-MIDN, "softmax*" denotes the masked *softmax* layer as in Eq. 4

but their performance is limited by the basic MIDN. We also choose WSDDN as our basic MIDN, but we couple the localization information of two complementary WSDDNs to escape the sub-optimum of detecting object parts.

Some methods propose to leverage Weakly Supervised Semantic Segmentation to improve WSOD [12, 9, 40]. Diba *et al*. [9] use segmentation confidence map to generate better proposals for MIL. Wei *et al*. [40] introduce two segmentation based metrics, purity and completeness, to mine tight boxes. However, WSSS can only provide semantic-level information. When there are several instances near each other in one image, the segmentation map may mix into one big region, as shown in the last row of Fig. 4(c), leading these methods to learn proposals which contain multiple instances. Our method also uses WSSS, but we introduce a new perspective, i.e. to use the coverage of segmentation region to identify whether there remains tight instance bounding boxes have not been found by detector. Then we can inherit the advantage of WSSS and avoid its shortcomings.

Weakly Supervised Semantic Segmentation A significant advance of performance of WSSS has been witnessed in last several years [44, 16, 38, 1, 29, 45, 41]. Class activation map [44] provides a simple and effective way to produce initial segmentation region. Kolesnikov and Lampert [16] introduce three principles and propose an end-to-end network to implement these principles. Wei *et al*. [39] proposed adversarial erasing method to progressively mine object region. Ahn and Kwak [1] propose AffinityNet, which trains a network to predict the affinity between pixels and further employs a random walk algorithm to refine the CAM. Without losing generality, we choose AffinityNet to generate the semantic segmentation map used in our method.

3. Method

In this section, we will first introduce the basic MIDN. Then we describe the proposed Coupled Multiple Instance Network (C-MIDN) in detail. Finally, an implementation of C-MIDN with online detector refinement (ODR) will be presented.

3.1. Multiple Instance Detection Network

By combining CNN and MIL, MIDN provides a simple and efficient pipeline for WSOD. In this paper, we choose WSDDN as our basic MIDN. WSDDN use a weighted-sum pooling strategy to map the proposal scores generated by a latent detector to image-level classification confidence. By optimizing a multi-class cross entropy loss in an end-to-end manner, the latent detector can be trained under only image-level supervision. In particular, for a given image I, the corresponding label is denoted as $Y = \{y_1, y_2, \ldots, y_C\}$, where $y_c = 1$ or 0 indicates the presence or absence of class c in I, and C is the number of classes. We use Selective Search [33] to generate candidate proposals $B = \{b_1, b_2, \ldots, b_{|B|}\}$. The proposals B and image I are fed into a CNN to extract the feature vectors of proposals. As shown in Fig. 3, the network contains two data streams,

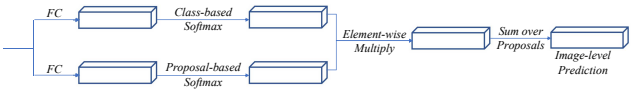

Figure 3. Network structure details of the basic MIDN.

named classification branch and detection branch respectively. Both branches consist of a linear map layer and a *softmax* layer. In the classification branch, the linear map layer maps the feature vectors to a matrix $x^c \in R^{C\times|B|}$, which is then passed through a *softmax* operator defined as $[\sigma_{class}(x^c)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{C} e^{x^c_{kj}}}$. The detection branch also applies a linear map on the feature vector to generate a matrix $x^d \in R^{C\times|B|}$, but the following *softmax* operator is changed to $[\sigma_{det}(x^d)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=0}^{|B|} e^{x^c_{ik}}}$. The final score of each proposal is generated by an element-wise product of the two matrices: $x^R = \sigma_{class}(x^c) \odot \sigma_{det}(x^d)$. Finally, a summation over all proposals is used to obtain the image score $p_c = \sum_{k=1}^{|B|} x^R_{ck}$. The parameters are optimized by a multi-class cross entropy loss L, as in Eq. 1.

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

3.2. Coupled Multiple Instance Detection Network

The basic MIDN inclines to localize the most discriminative object parts, which is undesirable in the detection task. To solve this issue, our C-MIDN contains a pair of MIDNs which work in a complementary way, Fig. 1. The two MIDNs have similar structure, but specific proposals are removed from the input of the second one. In particular, after the forward propagation, the top-scoring proposal of the first detector and adjacent proposals will be removed from the input of the second MIDN. With such removal, the latter detector can avoid being trapped to the same object parts as the first detector, and has more chance to find the entire object or localize new objects. However, if there is only one object in the image and the first MIDN has correctly localized it, Fig. 4(b), such removal will lead to none tight boxes in the remaining proposals and break the MIL constraint, which will confuse the second detector and harm its detection performance. To address this problem, we propose to leverage weakly supervised semantic segmentation to guide the process of proposal removal, named Segmentation Guided Proposal Removal (SGPR). Finally, we couple the localization information of the MIDNs to keep good proposals as many as possible and suppress the bad ones. In the rest of this subsection, we will present the details of the SGPR algorithm and the coupling method.

Segmentation Guided Proposal Removal As justified in [40, 9], semantic segmentation can find more complete object regions. If the segmentation coverage rate of the first

Algorithm 1 SGPR

Input: The final score of the first MIDN x^s; object proposals B; image label Y.
Output: Mask for the second MIDN $M \in \{0,1\}^{C\times|B|}$.

```
Set all M_ck = 1, c ∈ {1,...,C} and k ∈ {1,...,|B|}.
for c = 1 to C do
    if y_c = 1 then
        b_c ← arg max_{b_k∈B} x^s_ck.
        Compute the segmentation coverage rate r_c of b_c.
        if r_c < t_cover then
            for k = 1 to |B| do
                Compute IoU I_k between proposal b_k and b^1_c.
                if I_k > t_remove then
                    M_ck ← 0.
```

MIDN's top scoring box is too small, we speculate that there might be two cases: 1) only one object exists in the image, and the detector only finds part of the object, Fig. 4(a); 2) there are multiple object instances of the same class, and the detector fails to find all of them, Fig. 4(c). In both cases, there are tight instance bounding boxes that have not been found. So we use the segmentation coverage rate as a metric to evaluate whether the removal operation can be performed.

Specifically, we generate the segmentation map offline by weakly supervised semantic segmentation method. Without losing generality, we choose AffinityNet [1], one of the state-of-the-art WSSS methods. Firstly, we check the segmentation coverage rate of the first detector's top proposal. We denote the set of positive pixels in segmentation map for class c as M_c. For every class c that $y_c = 1$, we select the first detector's top-scoring proposal b_c as in Eq. 2, and denote the set of inner pixels of b_c as N_c. Then the coverage rate r_c can be computed according to Eq. 3. If r_c is smaller than a coverage threshold t_{cover}, we perform the proposal removal on class c, otherwise we retain all the proposals.

$$b_c = \arg\max_{b_k\in B} x^s_{ck} \quad (2)$$

$$r_c = \frac{|M_c \bigcap N_c|}{|M_c|} \quad (3)$$

When performing proposal removal, we select the proposals whose IoU with b_c is larger than t_{remove} and remove them from the input of the second detector in a class specific way. In practice, we generate a mask $M \in \{0,1\}^{C\times|B|}$, where $M_{ck} = 0$ indicates that proposal b_k needs to be removed in class c. The *softmax* layer in the detection branch of the second MIDN is modified to achieve proposal removal, as in Eq. 4. Like WSSDN, the score of each proposal in the second detector can be obtained by an element-

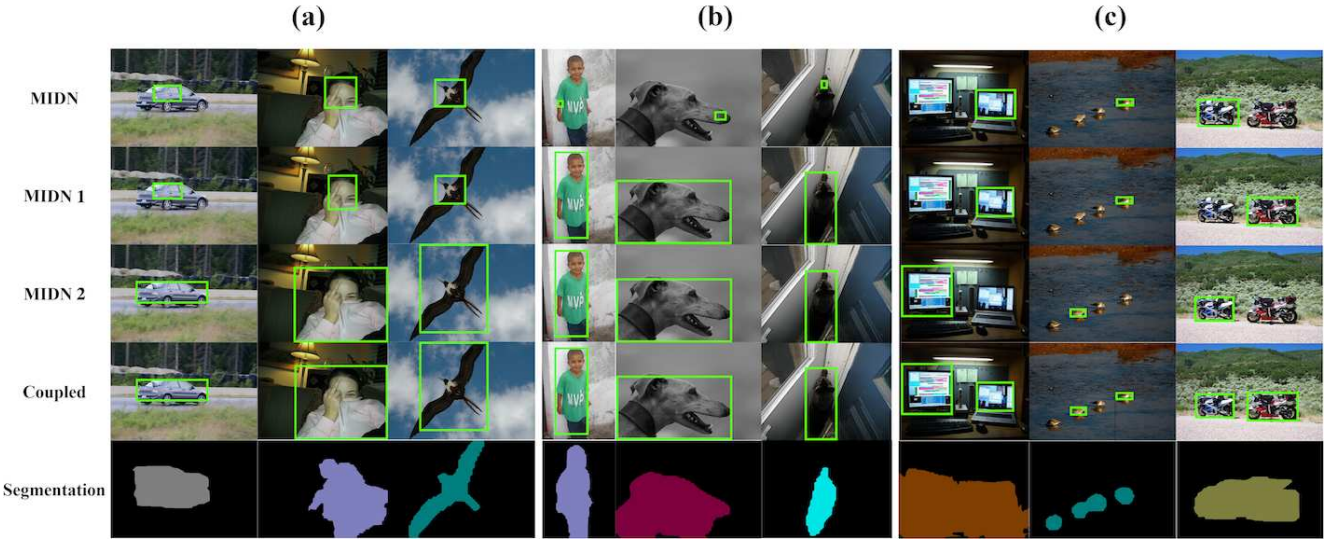

Figure 4. The comparison between different MIDNs and the coupled result. "MIDN" indicates the MIDN in baseline model. "MIDN-1" indicates the first MIDN in proposed C-MIDN, and "MIDN-2" indicates the second. "Coupled" indicates the result after coupling. "Segmentation" indicates the segmentation map generated by WSSS.

wise multiplication. Then the image-level score and loss of $Loss^2_{MIDN}$ can be obtained in the same way as the first detector. The total loss of C-MIDN is the sum of both detector's loss, Eq. 5.

$$\sigma^2_{det}(x^d)_{ij} = \frac{e^{x^c_{ij}} M_{ij}}{\sum_{k=0}^{|B|} e^{x^c_{ik}} M_{ik}} \tag{4}$$

$$Loss_{C-MIDN} = Loss^1_{MIDN} + Loss^2_{MIDN} \tag{5}$$

To make the SGPR algorithm more clear, we summarize the process of SGPR in Algorithm 1.

Candidates Coupling As shown in Fig. 4, the MIDNs in C-MIDN can localize different object regions. To couple the localization evidence of the MIDNs, we choose the top scoring proposals of them as candidate bounding boxes, and then merge the candidates by a priority based suppression method. Specifically, if the IoU of the top proposals is smaller than 0.1, it is highly possible that they belong to different objects, so we keep both of them. Otherwise, they may belong to the same object with good chance, and we keep the top proposal of the second MIDN as it is more likely to find the compete object after some bad proposals have been removed by SPRG.

3.3. Implementation with ODR

In this section, we will describe how to combine C-MIND with the popular Online Detector Refinement (ODR) framework following [31, 30, 32]. As shown in Fig. 2, we add several instance classifiers (ICs) parallel to C-MIDN into the network. The proposal features are extracted from a pretrained VGG [26] model. The coupled result of C-MIDN will be used to generate initial supervision for the first IC of ODR, while the supervision of the k^{th} IC depends on the $\{k-1\}^{th}$ IC's top-scoring proposal.

Formally, we denote the image label vector as $Y = \{y_1, y_2, \ldots, y_C\}$. For each class c that $y_c = 1$, we select the top-scoring proposal of $\{k-1\}^{th}$ IC as the positive seed for the k^{th} IC, and the positive seeds of the first IC come from the coupled result of C-MIDN. Consider a seed s^{ck}, we first compute a set of IoUs $\{I^{ck}_j\}$, where I^{ck}_j is the IoU between the j-th proposal b_j and the seed s^{ck}. Then we denote the set of positive proposals as $B^{ck}_p = \{b_j | I^{ck}_j \geq 0.5\}$ and the set of negative proposals as $B^{ck}_n = \{b_j | 0.1 \leq I^{ck}_j < 0.5\}$. Negative proposals will be labeled to class $\{C+1\}$, which means the background class. Instead of directly labeling the positive proposals to class c, we treat these positive proposals as a bag, and use a averaged MIL pooling method [30]. For seed s^{ck}, the loss of negative proposals is

$$Loss^k_n = -\frac{1}{|B|}\Big(\sum_{b_j \in B^{ck}_n} \log x^k_{(C+1)j}\Big)\Big) \tag{6}$$

and the averaged MIL loss of k^{th} IC is

$$Loss^{ck}_p = -\frac{1}{|B|}\Big(|B^{ck}_p| \log\Big(\frac{\sum_{b_j \in B^{ck}_p} x^k_{cj}}{|B^{ck}_p|}\Big)\Big) \tag{7}$$

Then, the ODR loss of k-th IC $Loss^k_{ODR}$ is a summation of both losses over all seeds and all positive classes. Moreover, we use a weighted loss as suggested in OICR. For more details, please refer to [30, 31].

Finally, We use SGD to train the network end-to-end by combining the loss of C-MIDN and ODR as in Eq. 8.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN [4]	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 [31]	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 [40]	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 [30]	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
MELM [36]	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 [32]	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 [31]	**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
CL [37]	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 [30]	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 [32]	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
W2F [43]	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
Baseline(MIDN+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
C-MIDN	53.3	71.5	49.8	26.1	20.3	70.3	69.9	68.3	**28.7**	65.3	45.1	64.6	58.0	**71.2**	20.0	27.5	54.9	54.9	**69.4**	63.5	52.6
C-MIDN+FRCNN	54.1	**74.5**	**56.9**	26.4	22.2	68.7	68.9	**74.8**	25.2	64.8	46.4	**70.3**	66.3	67.5	21.6	24.4	53.0	59.7	68.7	58.9	**53.6**

Table 1. Detection average precision (%) 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' result as pseudo groundtruth.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
OICR [31]	67.7	61.2	41.5	25.6	22.2	54.6	49.7	25.4	19.9	47.0	18.1	26.0	38.9	67.7	2.0	22.6	41.1	34.3	37.9	55.3	37.9
TS2C [40]	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 [30]	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
OICR+FRCNN [31]	71.4	69.4	55.1	29.8	28.1	55.0	**57.9**	24.4	17.2	59.1	21.8	26.6	57.8	71.3	1.0	23.1	52.7	37.5	33.5	56.6	42.5
CL [37]	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
PCL+FRCNN [30]	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
W2F [43]	**73.0**	69.4	45.8	30.0	28.7	58.8	58.6	56.7	20.5	58.9	10.0	69.5	**67.0**	73.4	7.4	24.6	48.2	**46.8**	50.7	58.0	47.8
Baseline(MIDN+ODR)	68.8	70.4	48.8	**30.4**	29.4	61.2	55.6	45.0	**25.5**	**61.3**	**26.2**	45.4	60.6	**73.9**	7.6	25.0	**54.6**	28.2	58.9	**60.0**	46.8
C-MIDN	72.9	68.9	53.9	25.3	**29.7**	60.9	56.0	78.3	23.0	57.8	25.7	73.0	63.5	73.7	13.1	**28.7**	51.5	35.0	56.1	57.5	50.2
C-MIDN+FRCNN	72.0	70.7	**58.7**	27.2	26.0	59.0	54.3	**82.6**	21.5	55.7	26.0	**78.3**	66.2	72.8	**16.7**	20.4	44.8	37.5	**61.9**	54.3	**50.3**

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

$$Loss = Loss_{C-MIDN} + \sum_{k=1}^{K} Loss_{ODR}^{k} \tag{8}$$

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate our method on the challenging PASCAL VOC 2007, PASCAL VOC 2012 and MS-COCO datasets [11, 19], which are widely used as benchmarks for widely supervised object detection. In all the experiments, only image-level annotations are used for training.

For VOC 2007 and 2012, we use the *trainval* set (5011 images and 11540 images respectively) to train our network, and the *test* set (4952 images and 10991 images respectively) for testing. For evaluation, we use 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.

For MS-COCO, the *train* set (about 80K images) of MS-COCO 2014 is used for training and the *val* set (about 40K images) for testing. For evaluation, we use two metrics mAP@0.5 and mAP@[.5, .95] which are the standard PASCAL criterion and the standard MS-COCO criterion respectively.

4.2. Implementation Details

We use VGG16 as our backbone network, which is pretrained on the ImageNet dataset [6]. Also, we replace the penultimate max-pooling layer and subsequent convolution layers by the dilated convolution layers as recommended in [31]. In SGPR, the coverage threshold t_{cover} is set to 0.3, and the IoU threshold t_{remove} is set to 0.3. The refinement time k is set to 3. The momentum and weight decay are set to 0.9 and 5×10^4 respectively. The mini-batch for training is set to 2, 2, and 4 for VOC 2007, VOC 2012 and MS-COCO respectively. The learning rate is 1×10^{-3} for the first 50K, 100K and 120K iterations, and then decreases to 1×10^{-4} for the following 25K, 50K and 80K iterations for VOC 2007, VOC 2012 and MS-COCO respectively.

We use Seletective Search [33] to generate object proposals for VOC 2007 and 2012 datasets, and use MCG [21] for MS-COCO dataset. The segmentation map of training images are generated offline by AffinityNet, which is trained on the same training images, and we use the original training settings recommended in [1]. For data augmentation, we rescale the shortest side of images to one of these five scales $\{480, 576, 688, 864, 1233\}$ and cap the longest image side to 2000. The scale of a training image is randomly selected and a random horizontal flip is applied. In evaluation, each testing image is augmented with all these five scales and horizontal flip, then the average score of total 10 images is used as the final score. For all the experiments, an NMS of 0.3 is employed to get the final detection

result. Our experiments are implemented based on the PyTorch deep learning framework and run on NVIDIA TITAN X GPUs.

Method	VOC 2007	VOC 2012
WSDDN [4]	58.0	-
OICR [31]	61.2	63.5
WCCN [9]	56.7	-
TS2C [40]	61.0	64.4
PCL [30]	63.0	65.0
MELM [36]	61.4	-
WSRPN [32]	66.9	67.2
OICR+FRCNN [31]	64.3	65.6
CL [37]	64.7	65.2
PCL+FRCNN [30]	66.6	68.0
WSRPN+FRCNN [32]	68.4	69.3
W2F [43]	70.3	69.4
C-MIDN	68.7	71.2
C-MIDN+FRCNN	**71.9**	**73.3**

Table 3. Detection CorLoc (%) on the *trainval* set of VOC 2007 and VOC 2012.

Method	mAP@0.5	mAP@[.5, .95]
PCL [30]	19.4	8.5
PCL+FRCNN [30]	19.6	9.2
C-MIDN	**21.4**	**9.6**

Table 4. Results (mAP@0.5 and mAP@[.5, .95] in %) on the MS-COCO dataset.

4.3. Ablation Studies

We first compare the proposed framework with the baseline model (WSDDN+ODR) to demonstrate the effectiveness of C-MIDN. Additional ablation experiments are presented to illustrate the influence of SGPR and the threshold of IoU in proposal removal, denoted as t_{remove}. Without loss generality, we only perform experiments on VOC 2007.

Influence of C-MIDN framework To show the effectiveness of the proposed C-MIDN, we compare the result of our method with a baseline framework, which replaces the C-MIDN in our framework by a WSDDN and chooses the top proposal of WSDDN as the initial supervision of ODR. From the Table. 1, we can see that our model exceeds the baseline by 3.6 points on mAP, and the increase is about 7%. The CorLoc in Table. 3 shows the same trend as mAP. The performance of almost all classes have been improved. Our model can not only greatly improve the performance on non-rigid classes, such as cat (mAP from 47 to 68.3), dog (mAP from 59.8 to 64.6) and person(mAP from 7 to 20), showing the ability of C-MIDN to avoid being trapped to parts of objects. Meanwhile, our model can also improve the performance on some rigid classes, such as diningtable (mAP from 37.5 to 45.1) and aeroplane (mAP from 44.3 to 53.3). This is because C-MIDN can find more objects and enrich the object patterns by the candidates coupling process.

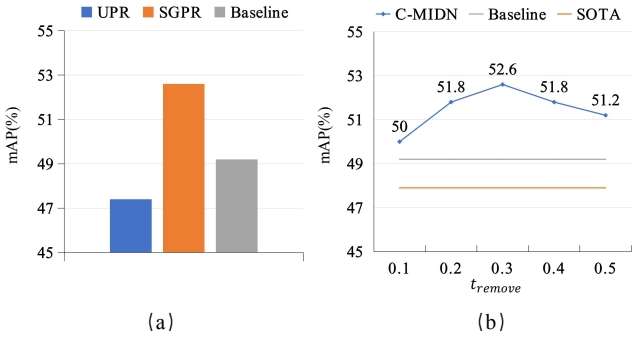

Figure 5. (a) Results of different proposal removal strategies. "UPR" indicates an unconditional proposal removal method. "SGPR" indicates the proposed segmentation guided proposal removal algorithm. "Baseline" indicates the basic framework combining WSDDN with ODR. (b) Comparison of the results for different proposal removal threshold t_{remove}.

Influence of SGPR To validate the effect of SGPR, we conduct an experiment by replacing the SGPR with an Unconditional Proposal Removal method (UPR). To be specific, the same proposal removal as in SGPR will be always performed during the whole training process in UPR. Table 2 shows that with UPR, the performance declines greatly. We think the reason is that the UPR method removes all tight proposals on some images, and breaks the basic assumption of MIL. Thus the second MIDN would be confused and localize background regions falsely, which eventually harms the performance of the entire model.

Influence of t_{remove} We conduct experiments to analyze the influence the removal threshold t_{remove}. As shown in Fig. 5, we can observe that our framework is insensitive to t_{remove}, and all models with different thresholds can outperform the baseline by more than 2.4 in mAP. In particular, performance rises and then decreases as t_{remove} increases continuously, reaching the peak at 0.3. The reason behind this trend may be two folds. When t_{remove} is too small, too many proposals will be removed and there is a high risk of removing all tight proposals, which will broke the MIL constraint and lead the MIDNs go astray. When t_{remove} is too large, only a few proposals will be removed, which may cause that both two detectors are trapped at parts of objects. So in other experiments, we set t_{remove} to 0.3.

4.4. Comparison with State-of-the-Art

In this subsection, we present the result of our C-MIDN 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 52.6 mAP, which outperforms the state-of-the-art method by 9.8%. On VOC 2012, our model obtains 50.2 mAP, and the improvement over the state-of-the-art increases to 15.7%. This increase of improvement is because our model can benefit from better segmentation re-

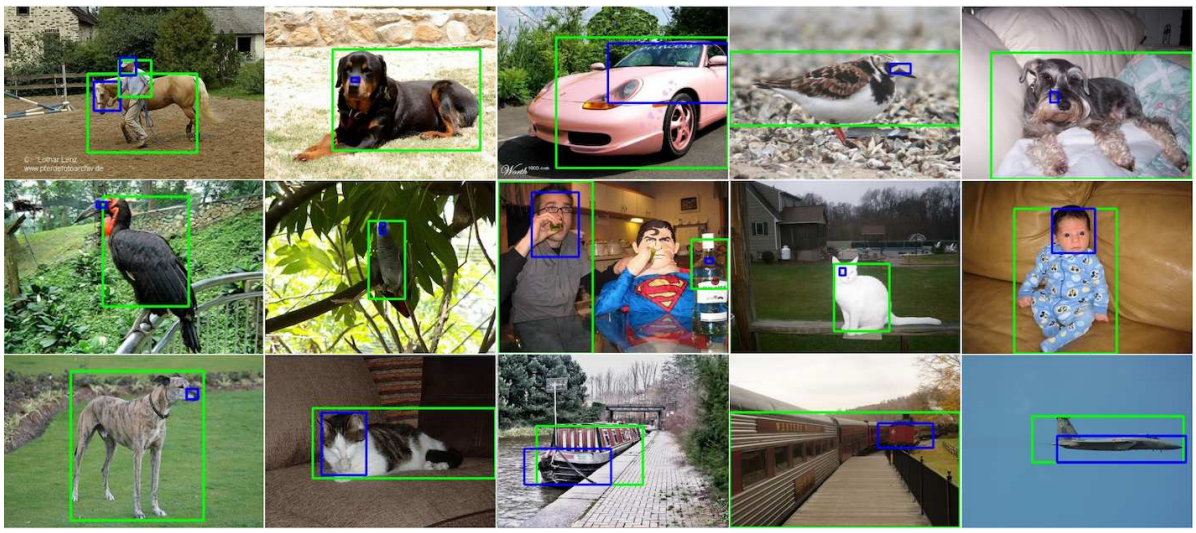

Figure 6. Qualitative results of the baseline model and our framework. Blue rectangle indicates the top-scoring bounding box of the baseline model, and green rectangle indicates ours.

sults trained on larger data sets. As shown in Table. 1, our model achieves the best results on almost all non-rigid classes.

Some 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 [31], we also use the top-scoring proposals produced by C-MIDN as pseudo ground-truth to train a Fast-RCNN. As shown in Table. 1 and Table. 2, the detection performance on VOC 2007 and VOC 2012 of our method are further improved to 53.6 and 50.3 respectively, which are the new state-of-the-arts.

The CorLoc results of C-MIDN on VOC 2007 and VOC 2012 are reported in Table. 3, which also create new state-of-the-arts. To further reveal the robustness of the our method, we conduct experiments on more challenging MS-COCO dataset, and C-MIDN surpasses existing methods on both mAP@0.5 and mAP@[.5, .95] 4.

We illustrate some detection results of our framework in Fig. 6. It can be found that the proposed method can correctively localize the objects while the baseline method is trapped to parts of objects. But the detection result on some classes is still undesirable, and we show some failure cases in Fig 7. The main failures are due to that the second MIDN also finds discriminative part of object instead of the entire object, especially on the class of person.

5. Conclusions

In this paper, we propose a Coupled Multiple Instance Detection Network for WSOD. C-MIDN uses two MIDNs that work in a complementary way by proposal removal. A novel Segmentation Guided Proposal Removal algorithm is further introduced to guarantee the MIL constraint after proposal removal. Finally we couple the output of the MIDNs to get tighter object bounding-boxes and recall more objects. Extensive experiments have been conducted to verify the effectiveness of C-MIDN. Combined with Online Detector Refinement, the proposed framework surpasses all previous methods proposed on WSOD, and creates new state-of-the-arts.

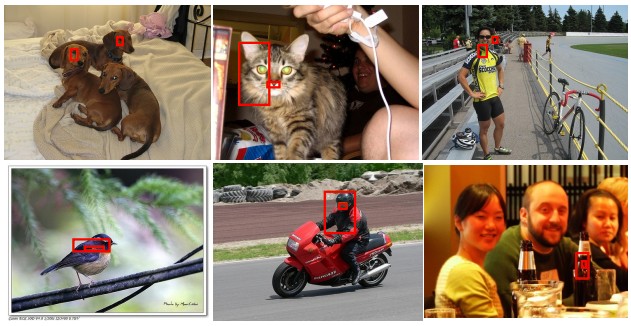

Figure 7. The failure cases in which both MIDNs find different parts of objects. The red rectangles denote the failed detection results of two MIDNs.

Acknowledgments

This work was supported by the National Key Research and Development Program (2018YFB1003501, 2017YFB0202502), the National Natural Science Foundation of China (61732018, 61872335, 61802367), Austrian-Chinese Cooperative R&D Project (FFG and CAS) Grant No. 171111KYSB20170032, the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDA18000000, and the Innovation Project Program of the State Key Laboratory of Computer Architecture (CARCH4505, CARCH4506, CARCH4509). The authors would like to thank Ruiping Wang for helpful discussions.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018.

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

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

[4] 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.

[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] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016.

[17] 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.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] 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.

[21] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016.

[22] 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.

[23] 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.

[24] Mrigank Rochan and Yang Wang. Weakly supervised localization of novel objects using appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4315–4324, 2015.

[25] 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.

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

[27] 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.

[28] Hyun Oh Song, Yong Jae Lee, Stefanie Jegelka, and Trevor Darrell. Weakly-supervised discovery of visual pattern configurations. In *Advances in Neural Information Processing Systems*, pages 1637–1645, 2014.

[29] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1818–1827, 2018.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019.

[35] Fang Wan, Pengxu Wei, Zhenjun Han, Jianbin Jiao, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[36] 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.

[37] 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.

[38] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1354–1362, 2018.

[39] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.

[40] 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.

[41] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018.

[42] 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.

[43] 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.

[44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[45] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018.