Weakly Supervised Region Proposal Network and Object Detection

Peng Tang[1], Xinggang Wang[1], Angtian Wang[1], Yongluan Yan[1], Wenyu Liu[1](✉), Junzhou Huang[2,3], Alan Yuille[4]

[1] School of EIC, Huazhong University of Science and Technology, Wuhan, China
{pengtang,xgwang,angtianwang,yongluanyan,liuwy}@hust.edu.cn
[2] Tencent AI lab, Shenzhen, China
[3] Department of CSE, University of Texas at Arlington, Arlington, USA
jzhuang75@gmail.com
[4] Department of Computer Science, The Johns Hopkins University, Baltimore, USA
alan.l.yuille@gmail.com

Abstract. The Convolutional Neural Network (CNN) based region proposal generation method (*i.e.* region proposal network), trained using bounding box annotations, is an essential component in modern fully supervised object detectors. However, Weakly Supervised Object Detection (WSOD) has not benefited from CNN-based proposal generation due to the absence of bounding box annotations, and is relying on standard proposal generation methods such as selective search. In this paper, we propose a weakly supervised region proposal network which is trained using only image-level annotations. The weakly supervised region proposal network consists of two stages. The first stage evaluates the objectness scores of sliding window boxes by exploiting the low-level information in CNN and the second stage refines the proposals from the first stage using a region-based CNN classifier. Our proposed region proposal network is suitable for WSOD, can be plugged into a WSOD network easily, and can share its convolutional computations with the WSOD network. Experiments on the PASCAL VOC and ImageNet detection datasets show that our method achieves the state-of-the-art performance for WSOD with performance gain of about 3% on average.

Keywords: Object detection, region proposal, weakly supervised learning, convolutional neural network

1 Introduction

Convolutional Neural Networks (CNNs) [22, 24] in conjunction with large scale datasets with detailed bounding box annotations [14, 26, 32] have contributed to a giant leap forward for object detection [15, 16, 30, 37, 43]. However, it is very laborious and expensive to collect bounding box annotations. By contrast, images with only image-level annotations, indicating whether an image belongs to an object class or not, are much easier to acquire (*e.g.*, using keywords to search

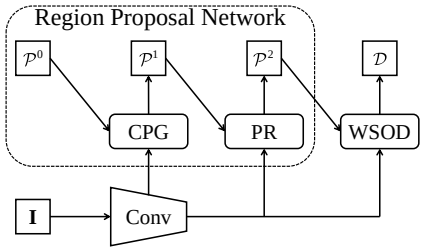

Fig. 1. The overall network architecture. "**I**": input image; "$\mathcal{P}^0$": the initial proposals by sliding window, "$\mathcal{P}^1$": the proposals from the first stage of the network, "$\mathcal{P}^2$": the proposals from the second stage of the network, "$\mathcal{D}$": the detection results, "Conv": convolutional layers, "CPG": coarse proposal generation, "PR": proposal refinement, "WSOD": weakly supervised object detection

on the Internet). Inspired by this fact, in this paper we focus on training object detectors with only image-level supervisions, *i.e.*, Weakly Supervised Object Detection (WSOD).

The most popular pipeline for WSOD has three main steps [4, 5, 9, 12, 20, 21, 25, 34, 38, 39, 42]: region proposal generation (shortened to proposal generation) to generate a set of candidate boxes that may cover objects, proposal feature extraction to extract features from these proposals, and proposal classification to classify each proposal as an object class, or background. Various studies focus on proposing better proposal classification methods [4, 9, 41, 42]. Recently, some methods have trained the last two steps jointly and have achieved great improvements [5, 21, 38, 39].

But most of the previous studies only use standard methods, *e.g.* selective search [40] and Edge Boxes [46], to generate proposals. A previous work [17] has shown that the quality of the proposals has great influence on the performance of fully supervised object detection (*i.e.*, using bounding box annotations for training). In addition, the CNN-based region proposal generation method (*i.e.* region proposal network) [30] is an essential component in the state-of-the-art fully supervised object detectors. These motivate us to improve the proposal generation method, in particular to propose CNN-based methods for WSOD.

In this paper, we focus on proposal generation for WSOD, and propose a novel weakly supervised region proposal network which generates proposals by CNNs trained under weak supervisions. Due to the absence of bounding box annotations, we are unable to train a region proposal network end-to-end as in Faster RCNN [30]. Instead, we decompose the proposal network into two stages, where the first stage is coarse proposal generation which generates proposals $\mathcal{P}^1$ from sliding window boxes $\mathcal{P}^0$ ($|\mathcal{P}^0| > |\mathcal{P}^1|$), and the second stage is proposal refinement which refines proposals $\mathcal{P}^1$ to generate more accurate proposals $\mathcal{P}^2$ ($|\mathcal{P}^1| > |\mathcal{P}^2|$). The proposals $\mathcal{P}^2$ are fed into the WSOD network to produce detection results $\mathcal{D}$. In addition, the proposal network and the WSOD network are integrated into a single three-stage network, see Fig. 1.

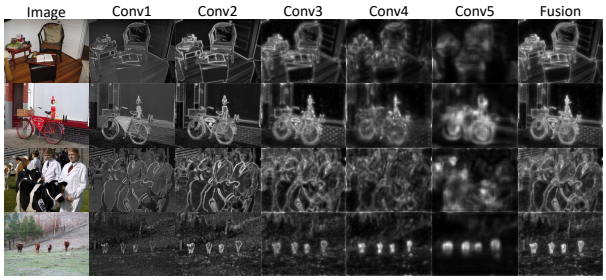

Fig. 2. The responses of different convolutional layers from the VGG16 [36] network trained on the ImageNet [32] dataset using only image-level annotations. Results from left to right are original images, response from the first to the fifth layers, and the fusion of responses from the second layer to the fourth layer

The first stage of our method is motivated by the intuition that CNNs trained for object recognition contain latent object location information. For example, as shown in Fig. 2, the early convolutional layers concentrate on low-level vision features (*e.g.* edges) and the later layers focus on more semantic features (*e.g.* object itself). Because the first and fifth convolutional layers also have high responses on many non-edge regions, we exploit the low-level information only from the second to the fourth convolutional layers to produce edge-like responses, as illustrated in Fig. 2. More specifically, after generating initial proposals $\mathcal{P}^0$ from an exhaustive set of sliding window boxes, these edge-like responses are used to evaluate objectness scores of proposals $\mathcal{P}^0$ (*i.e.* the probability of a proposal being an object), following [46]. Then we obtain some proposals $\mathcal{P}^1$ accordingly.

However, the proposals generated above are still very coarse because the early convolutional layers also fire on background regions. To address this, we refine the proposals $\mathcal{P}^1$ in the second stage. We train a region-based CNN classifier, which is a small WSOD network [38], using $\mathcal{P}^1$, and adapt the network to distinguish whether $\mathcal{P}^1$ are object or background regions instead of to detect objects. The objectness scores of proposals in $\mathcal{P}^1$ are re-evaluated using the classifier. Proposals with high objectness scores are more likely to be objects, which generates the refined proposals $\mathcal{P}^2$. We do not use the region-based CNN classifier on the sliding window boxes directly, because this requires an enormous number of sliding window boxes to ensure high recall and it is hard for a region-based CNN classifier to handle such a large number of boxes efficiently.

The proposals $\mathcal{P}^2$ are used to train the third stage WSOD network to produce detection results $\mathcal{D}$. To make the proposal generation efficient for WSOD, we adapt the alternating training strategy in Faster RCNN [30] to integrate the proposal network and the WSOD network into a single network. More precisely, we alternate the training of the proposal network and the WSOD network, and share the convolutional features between the two networks. After that, the convolutional computations for proposal generation and WSOD are shared, which improves the computational efficiency.

Elaborate experiments are carried out on the challenging PASCAL VOC [14] and ImageNet [32] detection datasets. Our method obtains the state-of-the-art performance on all these datasets, *e.g.*, 50.4% mAP and 68.4% CorLoc on the PASCAL VOC 2007 dataset which surpass previous best performed methods by more than 3%.

In summary, the main contributions of our work are listed as follows.

- We confirm that CNNs contain latent object location information which we exploit to generate proposals for WSOD.
- We propose a two-stage region proposal network for proposal generation in WSOD, where the first stage exploits the low-level information from the early convolutional layers to generate proposals and the second stage is a region-based CNN classifier to refine the proposals from the first stage.
- We adapt the alternating training strategy [30] to share convolutional computations among the proposal network and WSOD network for testing efficiency, and thus the proposal network and WSOD network are integrated into a single network.
- Our method obtains the state-of-the-art performance on the PASCAL VOC and ImageNet detection datasets for WSOD.

2 Related Work

Weakly Supervised Object Detection/Localization. WSOD has attracted a great deal of attention in recent years [4, 5, 9, 12, 20, 21, 34, 38, 39, 41, 42]. Most methods adopt a three step pipeline: proposal generation, proposal feature extraction, and proposal classification. Based on this pipeline, many variants have been introduced to give better proposal classification, *e.g.*, multiple instance learning based approaches [4, 9, 34, 39, 42]. Recently, inspired by the great success of CNNs, many methods train a WSOD network by integrating the last two steps (*i.e.* proposal feature extraction and proposal classification) into a single network [5, 12, 21, 38]. These networks show more promising results than the step-by-step ones. However, most of these methods use off-the-shelf methods [40, 46] for the proposal generation step. Unlike them, we propose a better proposal generation method for WSOD. More specifically, we propose a weakly supervised region proposal network which generates object proposals by CNN trained under weak supervisions, and integrate the proposal network and WSOD network into a single network. This relates to the work by Diba *et al.* [12] who propose a cascaded convolutional network to select some of the most reliable proposals for WSOD. They first generate a set of proposals by Edge Boxes [46], and then choose a few most confident proposals according to class activation map from [44] or segmentation map from [2]. These chosen proposals are used to train multiple instance learning classifiers. Unlike them, we use CNN to generate proposals, and refine proposals using region-based CNN classifiers. In fact, their network can be used as our WSOD network.

Recently, some studies show a similar intuition that CNNs trained under weak supervisions contain object location information and try to localize ob-

jects without proposals [10,18,27,35,44,45]. For example, Oquab *et al.* [27] train a max-pooling based multiple instance learning network to localize objects. But they can only give coarse locations of objects which are independent of object sizes and aspect ratios. The methods in [10, 35, 44, 45] localize objects by first generating object score heatmaps and then placing bounding boxes around the high response regions. However, they mainly test their methods on the ImageNet localization dataset which contains a large portion of iconic-object images (*i.e.*, a single large object located in the center of an image). Considering that natural images (*e.g.* images in PASCAL VOC) contain several different objects located anywhere in the image, the performance of these methods can be limited compared with the proposal-based methods [5,12,21,38]. Zhu *et al.* [45] also suggest a soft proposal method for weakly supervised object localization. They use a graph-based method to generate an objectness map that indicates whether each point on the map belongs to an object or not. However, the method cannot generate "real" proposals, *i.e.*, generate boxes which cover as many as possible objects in images. Our method differs from these methods in that we generate a set of proposals using CNNs which potentially cover objects tightly (*i.e.*, have high Intersection-over-Union with groundtruth object boxes) and use the proposals for WSOD in complex images. In addition, all these methods focus on the later convolutional layers that contain more semantic information, whereas our method exploits the low-level information from the early layers.

Region Proposal Generation. There are many works focusing on region proposal generation [6, 29, 40, 46], where Selective Search (SS) [40] and Edge Boxes (EB) [46] are two most commonly used proposal generation methods for WSOD. The SS generates proposals based on a superpixel merging method. The EB generates proposals by first extracting image edges and then evaluating the objectness scores of sliding window boxes. Our method follows the EB for objectness score evaluation in the first stage. But unlike EB which adopts edge detectors trained on datasets with pixel-level edge annotations [13] to ensure high proposal recall, we exploit the low-level information in CNNs to generate edge-like responses, and use a region-based CNN classifier to refine the proposals. Experimental results show that our method obtains much better WSOD performance.

There are already some CNN-based proposal generation methods [23,28,30]. For example, the Region Proposal Network (RPN) [30] uses bounding box annotations as supervisions to train a proposal network, where the training targets are to classify some sliding window style boxes (*i.e.* anchor boxes) as object or background and regress the box locations to the real object locations. These RPN-like proposals are standard for recent fully supervised object detectors. However, to ensure their high performance, these methods require bounding box annotations [23,31] and even pixel-level annotations [28] to train their networks, which deviates from the requirement of WSOD that only image-level annotations are available during training. Instead, we show that CNNs trained under weak supervisions have the potential to generate very satisfactory proposals.

Others. The works by [3, 33] also show that the different CNN layers contain different level visual information. Unlike our approach, Bertasius *et al.* [3] aim

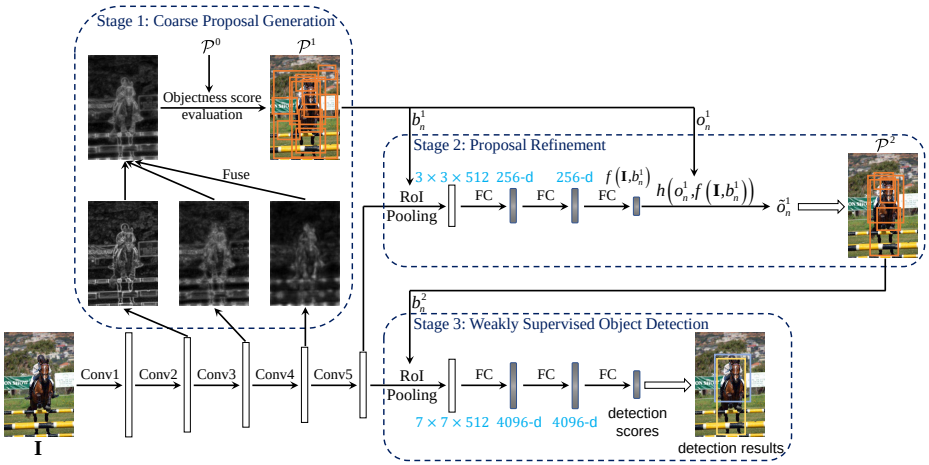

Fig. 3. The detailed architecture of our network. The first stage "Coarse Proposal Generation" produces edge-like responses which can evaluate objectness scores of sliding window boxes $\mathcal{P}^0$ to generate coarse proposals $\mathcal{P}^1$. The second stage "Proposal Refinement" uses a small region-based CNN classifier to re-evaluate the objectness scores of each proposal in $\mathcal{P}^1$ to get refined proposals $\mathcal{P}^2$. The third stage "Weakly Supervised Object Detection" uses a large region-based CNN classifier to classify each proposal in $\mathcal{P}^2$ as different object classes or background, to produce the object detection results. The proposals $\mathcal{P}^t, t \in \{0, 1, 2\}$ consist of boxes $\{b_n^t\}_{n=0}^{N^t}$ and objectness scores $\{o_n^t\}_{n=0}^{N^t}$

to fuse information from different layers for better edge detection which requires pixel-level edge annotations for training. Saleh *et al.* [33] choose more semantic layers (*i.e.* later layers) as foreground priors to guide the training of weakly supervised semantic segmentation, whereas we show that the low-level cues can be used for proposal generation.

3 Method

The architecture of our network is shown in Fig. 1 and Fig. 3. Our architecture consists of three stages during testing, where the first and second stages are the region proposal network for proposal generation and the third stage is a WSOD network for object detection. For an image $\mathbf{I}$, given initial proposals $\mathcal{P}^0$ which are an exhaustive set of sliding window boxes, the coarse proposal generation stage generates some coarse proposals $\mathcal{P}^1$ from $\mathcal{P}^0$, see Section 3.1. The proposal refinement stage refines the proposals $\mathcal{P}^1$ to generate more accurate proposals $\mathcal{P}^2$, see Section 3.2. The WSOD stage classifies the proposals $\mathcal{P}^2$ to produce the detection results, see Section 3.3. The proposals consist of bounding boxes and objectness scores, *i.e.*, $\mathcal{P}^t = \{(b_n^t, o_n^t)\}_{n=1}^{N^t}, t \in \{0, 1, 2\}$, where b_n^t and o_n^t are the box coordinates and the objectness score of the n-th proposal respectively. $o_n^0 = 1, n \in \{1, ..., N^0\}$ because we have no prior knowledge on the locations

of objects so we consider that all initial proposals have equal probability to cover objects. To share the conv parameters among different stages, we use an alternating training strategy, see Section 3.4.

3.1 Coarse Proposal Generation

Given the initial proposals $\mathcal{P}^0 = \{(b_n^0, o_n^0)\}_{n=1}^{N^0}$ of image $\mathbf{I}$ which are an exhaustive set of sliding window boxes with various sizes and aspect ratios, together the conv features of the image, the coarse proposal generation stage evaluates the objectness scores of these proposals coarsely and filters out most of the proposals that correspond to background. This stage needs to be very efficient because the number of initial proposals is usually very large (hundreds of thousands or even millions). Here we exploit the low-level information, more specifically the edge-like information from the CNN for this stage.

Let us start from Fig. 2. This visualizes the responses from different conv layers of the VGG16 network [36] trained on the ImageNet classification dataset (with only image-level annotations). Other networks have similar results and could also be chosen as alternates. Specially, we pass images forward through the network and compute the average value over the channel dimension for each conv layer to obtain five response maps (as there are five conv layers). Then these maps are resized to the original image size and are visualized as the second to the sixth columns in Fig. 2. As we can see, the early layers fire on low-level vision features such as edges. By contrast, the later layers tend to respond to more semantic features such as objects or object parts, and the response maps from these layers are similar to the saliency map. Obviously, these response maps provide useful information to localize objects. Here we propose to make use of the second to the fourth layers to produce edge-like response maps for proposal generation, as shown in Fig. 3.

More specifically, suppose the output feature map from a conv layer is $\mathbf{F} \in \mathbb{R}^{C\times W\times H}$, where C, W, H are the channel number, weight, and height of the feature map respectively. Then the response map $\mathbf{R} \in \mathbb{R}^{W\times H}$ of this layer is obtained by Eq. (1) which computes the average over the channels first and the normalization then, where f_{cwh} and r_{wh} are elements in $\mathbf{F}$ and $\mathbf{R}$ respectively.

$$r_{wh} = \frac{1}{C}\sum_{c=1}^{C} f_{cwh},\ r_{wh} \leftarrow \frac{r_{wh}}{\max\limits_{w',h'} r_{w'h'}}. \tag{1}$$

As we can see in Fig. 2, both the second to the fourth conv layers have high responses on edges and relative low responses on other parts of the image. Hence we fuse the response maps from the second to the fourth conv layers by first resizing them to the original image size and sum them up, see the 7^{th} column in Fig. 2 for examples. Accordingly we obtain the edge-like response map. We do not choose the response maps from the first and the fifth conv layers, because the former has high responses on most of the image regions and the later tends to fire on the whole object instead of the edges.

After obtaining the edge-like response map, we evaluate the objectness scores of the initial proposals $\mathcal{P}^0$ by using the Edge Boxes (EB) [46] to count the number of edges that exist in each initial proposal. More precisely, we follow the strategies in EB to generate $\mathcal{P}^0$, evaluate objectness scores, and perform Non-Maximum Suppression (NMS), so this stage is as efficient as Edge Boxes. Finally, we rank the proposals according to the evaluated objectness scores and choose N^1 ($N^1 < N^0$) proposals with the highest objectness scores. Accordingly we obtain the first stage proposals $\mathcal{P}^1 = \{(b_n^1, o_n^1)\}_{n=1}^{N^1}$.

In fact, the edge-like response map generated here is not the "real" edge in the sense of the edges generated by a fully supervised edge detector [13]. Therefore, directly using EB may not be optimal. We suspect that this stage can be further improved by designing more sophisticated proposal generation methods that consider the characteristics of the edge-like response map. In addition, responses from other layers can also be used as cues to localize objects, such as using saliency based methods [1]. Exploring these variants is left to future works and in this paper we show that our simple method is sufficient to generate satisfactory proposals for the following stages.

No direct loss is required in this stage and any trained network can be chosen.

3.2 Proposal Refinement

Proposals generated by the coarse proposal generation stage are still very noisy because there are also high responses on the background regions of the edge-like response map. To address this, we refine proposals using a region-based CNN classifier to re-evaluate the objectness scores, as shown in Fig. 1 and Fig. 3.

Given the proposals $\mathcal{P}^1 = \{(b_n^1, o_n^1)\}_{n=1}^{N^1}$ from the first stage and the conv features of the image, the task of the proposal refinement stage is to compute the probability that each proposal box b_n^1 covers an object using a region-based CNN classifier $f(\mathbf{I}, b_n^1)$, to re-evaluate the objectness score $\tilde{o}_n^1 = h\left(o_n^1, f(\mathbf{I}, b_n^1)\right)$, and to reject proposals with low scores. To do this, we first extract the conv feature map of b_n^1 and resize it to $512 \times 3 \times 3$ using the RoI pooling method [15]. After that, we pass the conv feature map through two 256-dimension Fully Connected (FC) layers to obtain the object proposal feature vector. Finally, an FC layer and a softmax layer are used to distinguish whether the proposal is object or background (we omit the softmax layer in Fig. 3 for simplification). Accordingly we obtain proposals $\tilde{\mathcal{P}}^1 = \{(b_n^1, \tilde{o}_n^1)\}_{n=1}^{N^1}$ with re-evaluated objectness score $\tilde{o}_n^1$. Here we use a simple multiplication to compute $h(\cdot, \cdot)$ as in Eq. (2).

$$\tilde{o}_n^1 = h\left(o_n^1, f(\mathbf{I}, b_n^1)\right) = o_n^1 \cdot f(\mathbf{I}, b_n^1). \tag{2}$$

There are other possible choices like addition, but we find that multiplication works well in experiments.

To get final proposals we can simply rank the proposals according to the objectness score $\tilde{o}_n^1$ and select some proposals with top objectness scores. But there are many redundant proposals (*i.e.* highly overlapped proposals) in $\tilde{\mathcal{P}}^1$. Therefore, we apply NMS on $\tilde{\mathcal{P}}^1$ and keep N^2 proposals with the highest objectness scores. Accordingly we obtain our refined proposals $\mathcal{P}^2 = \{(b_n^2, o_n^2)\}_{n=1}^{N^2}$.

To train the network using only image-level annotations, we train the state-of-the-art WSOD network given in [38], and adapt the network to compute $f(\mathbf{I}, b_n^1)$ instead of to detect objects. The network in [38] has a multiple instance learning stream which is trained by an image classification loss, and some instance classifier refinement streams which encourage category coherence among spatially adjacent proposals. The loss to train the network in the second stage network has the form of $\mathrm{L}^2(\mathbf{I}, \mathbf{y}, \mathcal{P}^1; \mathbf{\Theta}^2)$, where $\mathbf{y}$ is the image-level annotation and $\mathbf{\Theta}^2$ represents the parameters of the network. Please see [38] for more details. Other WSOD networks [5,12,21] can also be chosen as alternates. Specially, the output of proposal box b_n^1 by [38] is a probability vector $\mathbf{p}_n^1 = [p_{n0}^1, ..., p_{nK}^1]$, where p_{n0}^1 is for background, $p_{nk}^1, k > 0$ is for the k-th object class, and K is the number of object classes. We transfer this probability to the probability that b_n^1 covers an object by $f(\mathbf{I}, b_n^1) = 1 - p_{n0}^1 = \sum_{k=1}^{K} p_{nk}^1$. We use a smaller network than the original network in [38] to ensure the efficiency.

3.3 Weakly Supervised Object Detection

The final stage, *i.e.* WSOD, classifies proposals $\mathcal{P}^2$ into different object classes, or background. This is our ultimate goal. Similar to the previous stage, we use a region-based CNN for classification, see Fig. 3.

Given the proposals $\mathcal{P}^2 = \{(b_n^2, o_n^2)\}_{n=1}^{N^2}$ from the second stage and the conv features of the image, for each proposal box b_n^2, $512 \times 7 \times 7$ feature map and two 4096-dimension FC layers are used to extract the proposal features. Then a $\{K+1\}$-dimension FC layer is used to classify the b_n^2 as one of the K object classes or background. Finally, NMS is used to remove redundant detection boxes and produces object detection results.

Here we also train the WSOD network given in [38] and make some improvements. Then the loss to train the third stage network has the form of $\mathrm{L}^3(\mathbf{I}, \mathbf{y}, \mathcal{P}^2; \mathbf{\Theta}^3)$, where $\mathbf{\Theta}^3$ represents the parameters of the network. Both of the multiple instance detection stream and instance classifier refinement streams in [38] produce proposal classification probabilities. Given a proposal box b_n^2, suppose the proposal classification probability vector from the multiple instance detection stream is $\boldsymbol{\varphi}_n$, then similar to [5], we multiply $\boldsymbol{\varphi}_n$ by the objectness score o_n^2 during the training to exploit the prior object/background knowledge from the objectness score. More improvements are described in the supplementary material. We use the original version network in [38] rather than the smaller version in Section 3.2 for better detection performance.

3.4 The Overall Network Training

If we do not share the parameters of the conv layers among the different stages, then each proposal generation stage and the WSOD stage has its own separate network. Suppose $\mathbb{M}^{\mathrm{pre}}$, $\mathbb{M}^1$, $\mathbb{M}^2$, and $\mathbb{M}$ are the ImageNet pre-trained network, the proposal network for the first stage, the proposal network for the second stage, and the WSOD network for the third stage, respectively, we train the

Algorithm 1 Proposal network training

Input: Training images with image-level annotations; an initial CNN network $\mathbb{M}^{\text{init}}$.
Output: Proposal networks $\mathbb{M}^1, \mathbb{M}^2$; proposals $\mathcal{P}^2$.
1: Generate initial proposals $\mathcal{P}^0$ for each image and initialize $\mathbb{M}^1$ by $\mathbb{M}^{\text{init}}$.
2: Generate proposals $\mathcal{P}^1$ for each image using $\mathcal{P}^0$ and $\mathbb{M}^1$.
3: Train the proposal network $\mathbb{M}^2$ on $\mathbb{M}^{\text{init}}$ using $\mathcal{P}^1$.
4: Generate $\mathcal{P}^2$ for each image using $\mathcal{P}^1$ and $\mathbb{M}^2$.

Algorithm 2 The alternating network training

Input: Training images with image-level annotations; $\mathbb{M}^{\text{pre}}$.
Output: Proposal networks $\mathbb{M}^1, \mathbb{M}^2$; WSOD network $\mathbb{M}$.
1: Train proposal networks $\mathbb{M}^1, \mathbb{M}^2$ on $\mathbb{M}^{\text{pre}}$ and generate proposals $\mathcal{P}^2$ for each image, see Algorithm 1.
2: Train WSOD network $\mathbb{M}'$ on $\mathbb{M}^{\text{pre}}$ using $\mathcal{P}^2$.
3: Re-train proposal networks $\mathbb{M}^1, \mathbb{M}^2$ on $\mathbb{M}'$, fix the parameters of conv layers, and re-generate proposals $\mathcal{P}^2$ for each image, see Algorithm 1.
4: Re-train WSOD network $\mathbb{M}$ on $\mathbb{M}'$ using $\mathcal{P}^2$ and fix the parameters of conv layers.

proposal networks and the WSOD network step-by-step, because in our architecture each network requires outputs generated from its previous network for training. That is, we first initialize $\mathbb{M}^1$ by $\mathbb{M}^{\text{pre}}$ and generate $\mathcal{P}^1$, then use $\mathcal{P}^1$ to train $\mathbb{M}^2$ and generate $\mathcal{P}^2$, and finally use $\mathcal{P}^2$ to train $\mathbb{M}$.

Although we can use different networks for different stages, this would be very time-consuming during testing, because it requires passing image through three different networks. Therefore, w adapt the alternating network training strategy in Faster RCNN [30] in order to share parameters of conv layers among all stages. That is, after training the separate networks $\mathbb{M}^1$, $\mathbb{M}^2$, and $\mathbb{M}$, we re-train proposal networks $\mathbb{M}^1$ and $\mathbb{M}^2$ on $\mathbb{M}$, fixing the parameters of the conv layers. Then we generate proposals to train the WSOD network on $\mathbb{M}$, also fixing the parameters of the conv layers. Accordingly the conv computations of all stages are shared. We summarize this procedure in Algorithm 2. It is obvious that the shared method is more efficient than the unshared method because it computes the conv features only one time rather than three times.

4 Experiments

In this section we will give experiments to analysis different components of our method and compare our method with previous state of the arts.

4.1 Experimental Setups

Datasets and Evaluation Metrics. We choose the challenging PASCAL VOC 2007, 2012 [14], and ImageNet [32] detection datasets for evaluation. We only use image-level annotations for training.

Table 1. Result comparison (AP and mAP in %) for different methods on the PASCAL VOC 2007 `test` set. The upper/lower part are results by single/multiple model. Our method obtains the best mAP. See Section 4.2 for definitions of the Ours-based methods

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN-VGG16 [5]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
WSDDN+context [21]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	**42.0**	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
OICR-VGG16 [38]	**58.0**	62.4	31.1	**19.4**	13.0	65.1	62.2	28.4	**24.8**	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	**64.3**	**62.6**	41.2
Ours-VGG16	57.9	**70.5**	**37.8**	5.7	**21.0**	**66.1**	**69.2**	**59.4**	3.4	**57.1**	**57.3**	35.2	**64.2**	**68.6**	**32.8**	**28.6**	**50.8**	**49.5**	41.1	30.0	**45.3**
WSDDN-Ens. [5]	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
OM+MIL+FRCNN [25]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	**21.8**	57.1	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
WCCN [12]	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
HCP+DSD+OSSH3 [20]	54.2	52.0	35.2	25.9	15.0	59.6	67.9	**58.7**	10.1	**67.4**	27.3	37.8	54.8	67.3	5.1	19.7	52.6	43.5	56.9	**62.5**	43.7
OICR-Ens.+FRCNN [38]	**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
Ours-Ens.	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
Ours-Ens.+FRCNN	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**

There are 9,962 and 22,531 images for 20 object classes in the PASCAL VOC 2007 and 2012 respectively. The datasets are divided into the `train`, `val`, and `test` sets. Following [5, 21, 38], we train our network on the `trainval` set. For evaluation, the Average Precision (AP) and mean of AP (mAP) [14] is used to evaluate our network on the `test` set; the Correct Localization (CorLoc) [11] is used to evaluate the localization accuracy on the `trainval` set.

There are hundreds of thousands of images for 200 object classes in the ImageNet detection dataset which is divided into the `train`, `val`, and `test` sets. Following [16], we divide the `val` set into `val1` and `val2` sets, randomly choose no more than 1000 images per-class from the `train` set ($\texttt{train}_{\texttt{1k}}$ set), combine the $\texttt{train}_{\texttt{1k}}$ and `val1` sets for training, and report mAP on the `val2` set.

Implementation Details. We choose the VGG16 network [36] pre-trained on ImageNet classification dataset [32] as our initial CNN network $\mathbb{M}^{\text{pre}}$ in Section 3.4. The two 256-dimension FC layers in Section 3.2 are initialized by sub-sampling the parameters of the FC parameters in the original VGG16 network, following [8]. Other new added layers are initialized by sampling from a Gaussian distribution with mean 0 and standard deviation 0.01.

During training, we choose Stochastic Gradient Descent and set the batchsize to 2 and 32 for PASCAL VOC and ImageNet respectively. We train each network 50K, 80K, and 20K iterations for the PSACAL VOC 2007, 2012, and ImageNet datasets, respectively, where the learning rates are 0.001 for the first 40K, 60K, and 15K iterations and 0.0001 for the other iterations. We set the momentum and weight decay to 0.9 and 0.0005 respectively.

As stated in Section 3.2 and Section 3.3, we choose the best performed WSOD network by Tang *et al.* [38] for region classification, while other WSOD networks can also be chosen. We use five image scales $\{480, 576, 688, 864, 1024\}$ along with horizontal flipping for data augmentation during training and testing, and train a Fast RCNN (FRCNN) [15] using top-scoring proposals by our method as pseudo groundtruths following [12, 25, 38]. For the FRCNN training, we also use our proposal network through replacing the "WSOD network" in the second line and fourth line of Algorithm 2 by the FRCNN network. Other hyper-parameters are as follows: the number of proposals from the first stage of the network is set to 10K (*i.e.* $N^1 = 10\text{K}$), the number of proposals from the second stage of

Table 2. Result comparison (CorLoc in %) among different methods on the PASCAL VOC 2007 `trainval` set. The upper/lower part are results by single/multiple model. Our method obtains the best mean of CorLoc. See Section 4.2 for definitions of the Ours-based methods

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
WSDDN-VGG16 [5]	65.1	58.8	58.5	33.1	39.8	68.3	60.2	59.6	34.8	64.5	30.5	43.0	56.8	82.4	25.5	41.6	61.5	55.9	65.9	63.7	53.5
WSDDN+context [21]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
OICR-VGG16 [38]	81.7	80.4	48.7	**49.5**	32.8	**81.7**	85.4	40.1	**40.6**	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	**81.4**	60.6
SP-VGG16 [45]	**85.3**	64.2	**67.0**	42.0	16.4	71.0	64.7	**88.7**	20.7	63.8	58.0	**84.1**	**84.7**	80.0	**60.0**	29.4	56.3	**68.1**	**77.4**	30.5	60.6
Ours-VGG16	77.5	**81.2**	55.3	19.7	**44.3**	80.2	**86.6**	69.5	10.1	**87.7**	**68.4**	52.1	84.4	**91.6**	57.4	**63.4**	**77.3**	58.1	57.0	53.8	**63.8**
OM+MIL+FRCNN [25]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
WSDDN-Ens. [5]	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
WCCN [12]	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
HCP+DSD+OSSH3 [20]	72.7	55.3	53.0	27.8	35.2	68.6	81.9	60.7	11.6	71.6	29.7	**54.3**	64.3	88.2	22.2	53.7	72.2	52.6	68.9	75.5	56.1
OICR-Ens.+FRCNN [38]	**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
Ours-Ens.	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
Ours-Ens.+FRCNN	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**

Table 3. Result comparison (mAP and CorLoc in %) for different methods on the PASCAL VOC 2012 dataset. Our method obtains the best mAP and CorLoc

Method	mAP	CorLoc
WSDDN+context [21]	35.3	54.8
WCCN [12]	37.9	-
HCP+DSD+OSSH3 [20]	38.3	58.8
OICR-Ens.+FRCNN [38]	42.5	65.6
Ours-VGG16	40.8	64.9
Ours-VGG16-Ens.	43.4	67.2
Ours-VGG16-Ens.+FRCNN	**45.7**	**69.3**

Table 4. Result comparison (mAP in %) for different methods on the ImageNet detection dataset. Our method obtains the best mAP

Method	Results
Wang *et al.* [41]	6.0
OM+MIL+FRCNN [25]	10.8
WCCN [12]	16.3
Ours-VGG16	**18.5**

the network is set to 2K (*i.e.* $N^2 = 2\text{K}$) which is the same scale as the Selective Search [40], and the NMS thresholds for three stages are set to 0.9, 0.75, and 0.3, respectively. We only report results from the method that shares conv features, because there is no performance difference between the shared and unshared methods.

All of our experiments are carried out on an NIVDIA GTX 1080Ti GPU, using the Caffe [19] deep learning framework.

4.2 Experimental Results

The result comparisons among our method and other methods on the PASCAL VOC datasets are shown in Table 1, Table 2, and Table 3. As we can see, using our proposals (Ours-VGG16 in tables), we obtain much better performance than other methods that use a single model [5, 21, 38], in particular the OICR-VGG16 method [38] which is our WSOD network. Following other methods which combine multiple models through model ensemble or training FRCNN [5, 12, 20, 38], we also do model ensemble for our proposal results and

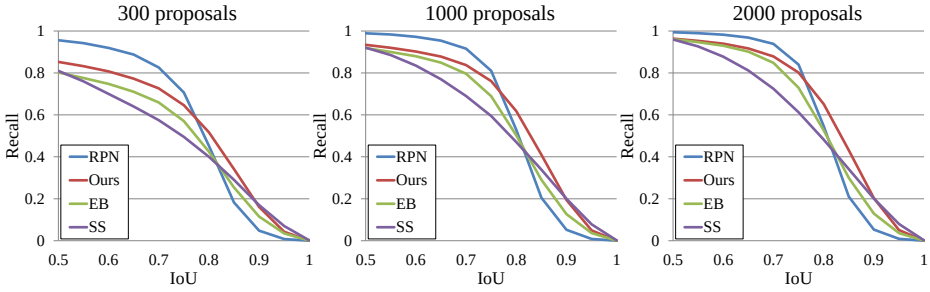

Fig. 4. Recall *vs.* IoU for different proposal methods on the VOC 2007 `test` set. Our method outperforms all methods except the RPN [30] which uses bounding box annotations for training

selective search proposal results (Ours-VGG16-Ens. in tables). As the tables show, performance is improved a lot, which show that our proposals and selective search proposals are complementary to some extent. We also train a FRCNN network using the top-scoring proposals from Ours-VGG16-Ens. as pseudo labels (Ours-VGG16-Ens.+FRCNN in tables). It is clear that the results are boosted further. Importantly, our results outperform the results of the state-of-the-art proposal-free method (*i.e.*, localize objects without proposals) [45], which confirms that the proposal-based method can localize objects better in complex images. Some qualitative results can be found in the supplementary material.

We also report the Ours-VGG16 result on the ImageNet detection dataset in Table 4. Using a single model already outperforms all previous state-of-the-arts [12, 25, 41]. Confidently, our result can be further improved by combining multiple models.

4.3 Ablation Experiments

We conduct some ablation experiments on the PASCAL VOC 2007 dataset to analyze different components of our method, including proposal recall, detection results of different proposal methods, and the influence of the proposal refinement. Also see the supplementary material for more ablation experiments.

Proposal Recall. We first compute the proposal recall at different IoU thresholds with groundtruth boxes. Although the recall to IoU metric is loosely correlated to detection results [7, 17], it can give a reliable result to diagnose whether proposals cover objects of desired categories well [30]. In Fig. 4, we observe that our method obtains higher recall than the Selective Search (SS) and Edge Boxes (EB) methods for IoU<0.9, especially when the number of proposals is small (*e.g.* 300 proposals). This is because our region-based classifier refines proposals. It is not strange that the recall of Region Proposal Network (RPN) [30] is higher than ours, because they train their network using the bounding box annotations. But we do not use the bounding box information because we do weakly supervised learning.

Detection Results of Different Proposal Methods. Here we compare the detection results of different proposal methods, using the same WSOD network [38] (with the improvements in this paper). For fair comparison, we generate about 2K proposals for each method. The results are as follows: 41.6% mAP and 60.7% CorLoc for EB, 42.2% mAP and 60.9% CorLoc for SS, and 46.2% mAP and 65.7% for RPN [30]. Our results (45.3% mAP and 63.8% CorLoc) are much better than the results of EB and SS which were used by most previous WSOD methods. The results demonstrates the effectiveness of our method for WSOD. As before, the RPN obtains the best results because it uses the bounding box annotations for training. These results also show that better proposals can contribute to better WSOD performance.

The Influence of the Proposal Refinement. Finally, we study whether the proposal refinement stage improves the WOSD performance or not. If we only perform the coarse proposal generation stage, we obtain mAP 37.5% and CorLoc 57.3% which are much worse than the results after proposal refinement, and even worse than the EB and SS. This is because the early conv layers also fire on background regions, and the responses of the early conv layers are not "real" edges, thus directly applying EB may not be optimal. The results demonstrates that it is necessary to refine the proposals. It is also possible to perform more proposal generation stages by using more proposal refinement stages. We plan to explore this in the future.

5 Conclusion

In this paper, we focus on the region proposal generation step for weakly supervised object detection and propose a weakly supervised region proposal network which generates proposals by CNN trained under weak supervisions. Our proposal network consists of two stages where the first stage exploits low-level information in CNN and the second stage is a region-based CNN classifier which distinguishes whether proposals are object or background regions. We further adapt the alternating training strategy in Faster RCNN to share convolutional computations among all proposal stages and the weakly supervised object detection network, which contributes to a three-stage network. Experimental results show that our method obtains the state-of-the-art weakly supervised object detection performance with performance gain of about 3% on average. In the future, we will explore better ways to use both low-level and high-level information in CNN for proposal generation.

Acknowledgements. We really appreciate the enormous help from Yan Wang, Wei Shen, Zhishuai Zhang, Yuyin Zhou, and Baoguang Shi during the paper writing and rebuttal. This work was partly supported by NSFC (No.61733007, No.61503145, No.61572207), ONR N00014-15-1-2356, and China Scholarship Council. Xinggang Wang was sponsored by CCF-Tencent Open Research Fund, Hubei Scientific and Technical Innovation Key Project, and the Program for HUST Academic Frontier Youth Team.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. TPAMI **34**(11), 2189–2202 (2012)
2. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: Whats the point: Semantic segmentation with point supervision. In: ECCV. pp. 549–565 (2016)
3. Bertasius, G., Shi, J., Torresani, L.: Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In: CVPR. pp. 4380–4389 (2015)
4. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: CVPR. pp. 1081–1089 (2015)
5. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR. pp. 2846–2854 (2016)
6. Carreira, J., Sminchisescu, C.: CPMC: Automatic object segmentation using constrained parametric min-cuts. TPAMI (7), 1312–1328 (2011)
7. Chavali, N., Agrawal, H., Mahendru, A., Batra, D.: Object-proposal evaluation protocol is 'gameable'. In: CVPR. pp. 835–844 (2016)
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR (2015)
9. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. TPAMI **39**(1), 189–203 (2017)
10. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: NIPS. pp. 6970–6979 (2017)
11. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. IJCV **100**(3), 275–293 (2012)
12. Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: CVPR. pp. 914–922 (2017)
13. Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. TPAMI **37**(8), 1558–1570 (2015)
14. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111**(1), 98–136 (2015)
15. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
16. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. TPAMI **38**(1), 142–158 (2016)
17. Hosang, J., Benenson, R., Dollár, P., Schiele, B.: What makes for effective detection proposals? TPAMI **38**(4), 814–830 (2016)
18. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: CVPR. pp. 7014–7023 (2018)
19. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM MM. pp. 675–678 (2014)
20. Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: CVPR. pp. 1377–1385 (2017)
21. Kantorov, V., Oquab, M., Cho, M., Laptev, I.: Contextlocnet: Context-aware deep network models for weakly supervised localization. In: ECCV. pp. 350–365 (2016)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)

23. Kuo, W., Hariharan, B., Malik, J.: Deepbox: Learning objectness with convolutional networks. In: ICCV. pp. 2479–2487 (2015)
24. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
25. Li, D., Huang, J.B., Li, Y., Wang, S., Yang, M.H.: Weakly supervised object localization with progressive domain adaptation. In: CVPR. pp. 3512–3520 (2016)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)
27. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR. pp. 685–694 (2015)
28. Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P.: Learning to refine object segments. In: ECCV. pp. 75–91 (2016)
29. Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping for image segmentation and object proposal generation. TPAMI **39**(1), 128–140 (2017)
30. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. TPAMI **39**(6), 1137–1149 (2017)
31. Ren, W., Huang, K., Tao, D., Tan, T.: Weakly supervised large scale object localization with multiple instance learning and bag splitting. TPAMI **38**(2), 405–416 (2016)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
33. Saleh, F., Aliakbarian, M.S., Salzmann, M., Petersson, L., Alvarez, J.M., Gould, S.: Incorporating network built-in priors in weakly-supervised semantic segmentation. TPAMI **40**(6), 1382–1396 (2018)
34. Shi, M., Caesar, H., Ferrari, V.: Weakly supervised object localization using things and stuff transfer. In: ICCV. pp. 3381–3390 (2017)
35. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
37. Tang, P., Wang, C., Wang, X., Liu, W., Zeng, W., Wang, J.: Object detection in videos by short and long range object linking. arXiv preprint arXiv:1801.09823 (2018)
38. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: CVPR. pp. 2843–2851 (2017)
39. Tang, P., Wang, X., Huang, Z., Bai, X., Liu, W.: Deep patch learning for weakly supervised object classification and discovery. Pattern Recognition **71**, 446–459 (2017)
40. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV **104**(2), 154–171 (2013)
41. Wang, C., Ren, W., Huang, K., Tan, T.: Weakly supervised object localization with latent category learning. In: ECCV. pp. 431–445 (2014)
42. Wang, X., Zhu, Z., Yao, C., Bai, X.: Relaxed multiple-instance svm with application to object discovery. In: ICCV. pp. 1224–1232 (2015)
43. Zhang, Z., Qiao, S., Xie, C., Shen, W., Wang, B., Yuille, A.L.: Single-shot object detection with enriched semantics. In: CVPR. pp. 5813–5821 (2018)

44. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)
45. Zhu, Y., Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Soft proposal networks for weakly supervised object localization. In: ICCV. pp. 1814–1850 (2017)
46. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV. pp. 391–405 (2014)