Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 3, 2023

Improved Regional Proposal Generation and Proposal Selection Method for Weakly Supervision Object detection

Yujiao Wang, Hua Huo

School of Henan University of Science and Technology, Luoyang 471000, China

Abstract: In recent years, object detection has made great progress with the continuous development of deep neural network. At present, there are many different fully supervised object detection algorithms in the field of computer vision, which are basically saturated, while object detection in a weakly supervised manner is more challenging than strongly supervised object detection. Since nowadays mature object detection algorithms rely heavily on strongly labeled datasets, but strong labeled datasets are very expensive and require huge datasets to support in order to train a better object detection model, weakly supervised object detection has received more and more attention. In this paper, a new module can be embedded in the framework of weakly supervised object detection, three modules are introduced into the weakly supervised object detection framework, which is used to generate high-quality proposals and screen these proposals, and finally selecting more accurate proposal boxes that are beneficial for subsequent training, and demonstrate their effectiveness on the PASCAL VOC2007 and PASCAL VOC2012 datasets, in which this paper achieves a significant improvement over the existing classic weakly supervised object detection algorithms with significant improvements.

Keywords: Weakly supervised, Object detection, Proposals.

1. Introduction

One of the most fundamental tasks under the direction of computer vision, object detection [3,4,8,9,17,18,19,21,27,35], has made remarkable progress with the continuous development of convolutional neural networks [10,13,14] in re-cent years, and its accuracy has reached a very good level. Simply put, object detection is based on image classification by framing objects in the form of enclosing frames, that is, locating and classifying example images.

At present, these object detection algorithms rely heavily on precisely annotated large-scale datasets [6,7,20,23,24], and the acquisition of such instance-level strongly labeled datasets are very labor-intensive and costly. In addition, the strongly supervised object detection algorithms still have some inevitable limitations, such as the possibility of inadvertently introducing labeling noise during the manual labeling of data, which makes it more difficult for the detector to learn a good model. Therefore, researchers have begun to explore weakly supervised object detection that requires only image-level labeled data for training, meaning that the dataset no longer has precise bounding box annotations, but only annotations of image categories. It is because of the very simple and noisy labeling of their datasets that although many methods [1,2,5,25,26,28,29, 37, 38] for weakly supervised object detection have been proposed, its performances are still far from those of strongly supervised object detection.

From the recent work, a number of approaches have been proposed for solving the WSOD problem. When the dataset only has image-level annotations, most of them are formulated as a multi-instance learning problem. Integrating the idea of multi-instance learning into CNN can compensate for the deficiencies of training set labels and improve the detection performance better.

The main problem of weakly supervised object detection lies in the poor localization accuracy due to the lack of precise labels. The wraparound box is overly focused on the part of the feature and the SS [30] and EB [41] algorithms are generally used in the generation of the proposed boxes, which is very time consuming. As shown in Fig. 1, this is the classic problem that weakly supervised object detection will encounter.

Figure 1. Typical WSOD problem, you can see the partial, correct and oversize detection results of an object instance from the first, second and third rows respectively.

Both OICR [2] and PCL [1] are weakly supervised object detection based on multiple instance learning, and since they both use the output of the initial object detector as the true annotation label, their performance is very dependent on the accuracy of the initial object detection results and do not learn the key step of bounding box regression . WSOD2 [11] precisely builds on OICR to obtain the initial object bounding box, based on the localization of each proposed bounding box, we put the bottom-up object evidence to use, which will guide the conversion from image-level to instance-level annotation.

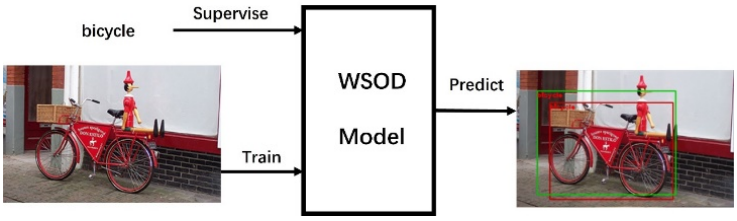

Figure 2. Weakly supervised object detection using image level labels as supervision

The challenge of weakly supervised object detection is that the dataset is weakly labeled, with only image-level labels available, as shown in Fig. 2. But we need to train a good detector with such a dataset, and the result of detection is to get both category information and location information. The limitation of object detection based on multiple example learning is that the most discriminative of all instances can be easily distinguished, while making the network can easily fall into local optima. So how to generate high quality proposals and which method to use to select high quality proposals becomes the key for weakly supervised object detection.

This paper proposes a framework is based on OICR [2] as a baseline and introduce three modules for generating proposals and performing proposal screening. Firstly, we generate high-quality proposals specifically for weakly supervised object detection, and for the proposal generation part we choose to combine the selection search algorithm [30] with an improved version of the gradient-weighted class activation-based mapping [31], and on top of this we add an improved attention module to extract an enhanced feature map from the CNN, and then have ROI pooling to process the generated regions with a combination of bottom-up and A combination of two evidences, bottom-up and top-down, is used to filter the proposals. It is also fed into the basic multi-instance detector and K-level instance optimizer and bounding box regression branch for iterative training as a way to improve its performance.

The contributions of this paper are summarized as follows:

1.In the proposal generation module, this paper uses a combination of Grad CAM++ based class activation graph and selection search algorithm, and incorporate an improved CBAM attention mechanism to achieve better results making it possible to generate high quality candidate frames in the end.

2.In the proposal selection module, in order to better select positive target proposals for weakly supervised target detection tasks, this paper can combine bottom-up target evidence and top-down class confidence scores in a new way to better select the most suitable bounding boxes.

3.This paper adds a bounding box regression branch, and introduces three modules to generate and select proposals respectively, which are unified into a weakly supervised object detection framework for end-to-end training.

2. Related Work

2.1. Weakly supervised object detection

In recent years, weakly supervised object detection has attracted a lot of attention from researchers. The classic framework on WSOD, WSDDN , is to solve the WSOD problem with a multiple-instance learning (MIL) approach, which contributes by using dual streams to perform object localization and classification simultaneously, but since only image-level labeled data can be accessed during the training phase, the most discriminative parts receive more attention during training than the whole object instance, leading to the model suffers from a discriminative region problem, which is improved by the later work.

In order to alleviate the problem of distinguished regions, online instance classifier refinement strategy (OICR) [2] takes WSDDN as the baseline and adds three more instance classifier refinement processes after the baseline, which improves the performance of weakly supervised target detection but also easily falls into local optimum because only the most distinguished instances are selected for refinement. By combining WADDN and OICR, Zhang et al [40] designed a framework from weakly supervised to fully supervised, which is also implemented with MIL. PCL is a further improvement of the above OICR, which proposes to use proposal clusters on top of OICR to divide all proposals into different pouches and then apply classifiers for refinement, i.e., proposal clustering.

Arun et al [32] designed a new phase difference coefficient-based WSOD framework that implements the WSOD task by minimizing the difference between the annotation agnostic prediction distribution and the annotated perceptual conditional distribution. Shen et al [33] proposed a framework called weakly supervised joint detection and segmentation (WS-JDS) by combining these two tasks into a multi-task learning framework. Li et al [34] proposed a segmentation collaboration network that uses segmentation graphs as a priori information to supervise the learning of object detection. Ze Chen et al [36] proposed a spatial likelihood voting module to converge the localization process of proposed frames without any bounding box annotation. Chenhao Lin et al [37] proposed an end-to-end object instance mining weakly supervised object detection framework that introduces a spatial graph and appearance graph based information propagation mechanism to try to mine all object instances in each image during iterative network learning.

2.2. Boundary box regression

Because only image-level labels are available, they only indicate whether the target category has appeared or not. However, in order to train a standard target detector with a regression task, it is necessary to mine instance-level supervised information, e.g., bounding box annotations. Therefore, Yang et al [22] here introduces a MIL branch to obtain pseudo-GT annotation information, and chooses to use a WSDDN-based OICR network for end-to-end training. Bounding box regression is used after refinement using multiple box classifications and only once.

C-WSL [28] also explores bounding box regression for weakly supervised object detection networks as in [22]. And both use bounding box regression in an online fashion, C-WSL uses a box regressor to refine each box classifier after the MIL branch.

Bounding box regression is a key step in object detection

for predicting rectangular boxes to locate targets, so almost all recent fully supervised object detection [3,4,12,13,21,24] used bounding box regression, which can reduce the localization error of prediction boxes. However, since it is weakly supervised learning and the data lacks the labeling information of the bounding box, only a small number of works have introduced the bounding box into the target detection, and some of them consider the bounding box regression as a post-processing module.

2.3. Attention mechanism

The use of attention modules first appeared in natural language and was later introduced into computer vision. Mixed spatial and channel attention mechanisms are widely used in weakly supervised target detection because they can not only focus on important parts of the image but also assign more weight to important channels.

In this paper, the attention module of CBAM [39] is used and improved to make it better embedded in the network. Attention mechanisms are very similar to human ones in that both tend to focus on one part of the information and ignore the others when they see things. The neural network first learns to new features by channel attention, and then learns to the location of key features by serial structure to the spatial attention module, and makes efforts to acquire the features with discriminative nature for images to achieve the effect of adaptive attention of the network.

3. Method

In this section, we will describe in detail the introduced proposal generation module and the proposal selection and attention modules.

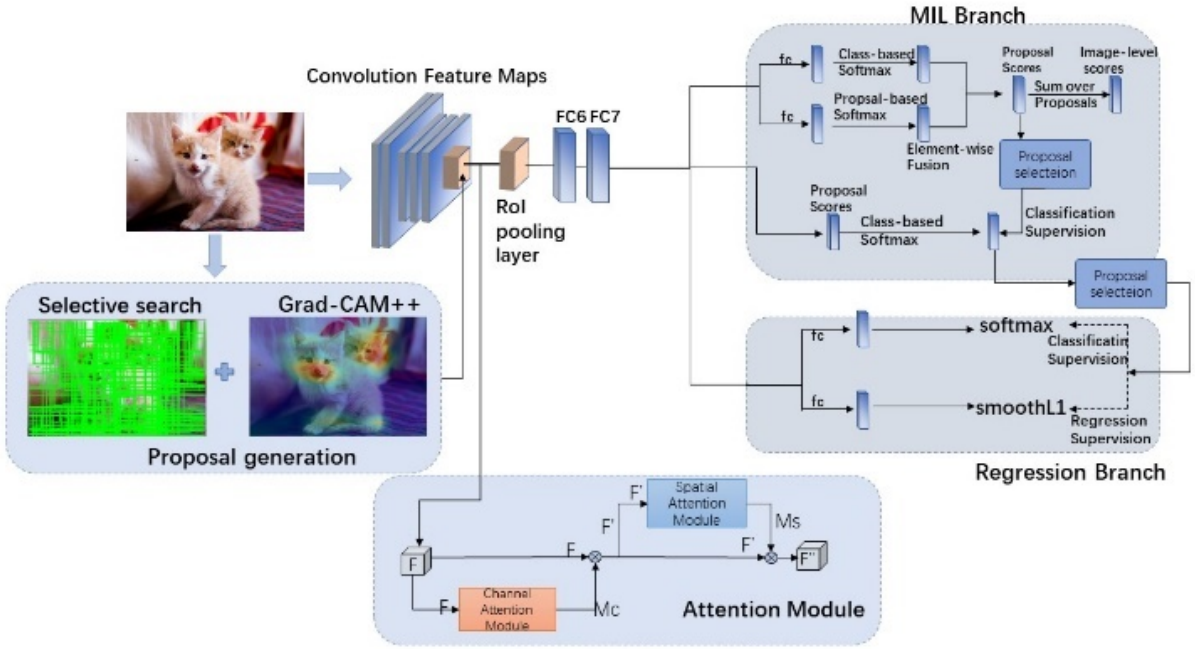

Figure 3. Network structure of our method. Each proposed feature is extracted using a base network with VGG16. Then, the proposed features are passed through two fully connected layers and the generated feature vectors are branched to the basic MIL module and to a new module (reclassification branch). In the basic MIL module, there is one WSDDN branch and three refinement branches. The average classification scores of the three refinement branches are input to the new module to generate supervision.

The overall architecture of the proposed network framework is shown in Fig. 3. This paper puts forward a framework based on OICR, and introduces three modules for generating proposals and filtering them. Firstly, high-quality proposals are generated specifically for weakly supervised target detection. For the proposal generation part, the selection search algorithm is combined with the improved gradient-weighted class activation mapping. On this basis, an attention module is added to extract enhanced feature maps from CNN, and then the enhanced feature maps are sent to the ROI pool layer to process the generated areas. In the proposal selection module, the proposals are screened by combining low-level semantic information with high-level semantic information, and sent to the basic multi-instance detector, the K-level instance optimizer and the bounding box regression branch for iterative training, so as to improve its performance.

The input picture passes through the convolution layer, ReLU activation function and pooling layer of convolutional neural network to generate the feature map of the image, which is used to extract the proposal box later. The selection search algorithm is combined with Grad CAM++ to generate proposals, and an improved CBAM attention module is added to generate an enhanced feature map. The proposals and the enhanced feature map are sent to the ROI pooling layer to generate a 7×7 ROI pooled feature. Finally, the feature vector is processed by the multi-instance learning module and the refined branch instance detector module for subsequent classification and boundary box regression, and the object category and positioning prediction results are output.

During the forward propagation of training, the extracted proposal features are sent to the basic MIL module to generate proposal score matrices. After the proposal selection module, more plausible positive proposals are selected, and subsequently, these proposal score matrices are used for

subsequent training supervision.

3.1. Proposal generation

At first, the VGG16 model is used to train the basic multi-instance classifier with only image-level labels, and the multi-class cross entropy loss function is used in Eq.1:

$$S = -\sum_{i=1}^{C} (y_i \log P_i + (1 - y_i) \log(1 - P_i)) \quad (1)$$

Where c is the total number of image categories, y_i is the label representation of the i th image category, and p_i is the prediction result of the i-th sigmoid classifier, which finally constitutes this loss function. For each image containing category C, a group of feature maps are weighted and combined by using the basic multi-instance classifier to obtain its category-specific activation map, as shown in Eq.2:

$$M_c = \mathrm{Re}\,LU(\sum_k w_k^c A_k) \quad (2)$$

Among them, A_{k} is the k-th convolution feature map, and w_k^c is the importance of the feature map A_{k} of class C in the object, which is calculated as follows in Eq.3:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot \mathrm{re}lu(\frac{\partial y_c}{\partial (A_{\mathrm{k}})_{ij}}) \quad (3)$$

Grad-CAM++ further improves on Grad-CAM, which can better locate the complete object position compared to Grad-CAM, Grad-CAM++ improves the representation when there are multiple targets in the image. It obtains the importance of each pixel in the feature map mainly by adding ReLU and pixel-level weighting to the weights of the feature map output of the corresponding classification to find out more accurate position information. This paper is mainly based on the combination of Grad-CAM++ and SS to generate a large number of object proposals with higher target overlap based on specific categories.

3.2. Attention Module

In order to better generate high quality proposal candidate frames, an attention module is added on top of the previously described proposal generation method, starting with a brief description of the spatial attention structure. First, the proposed feature maps generated from the SS-based algorithm combined with Grad CAM++ , which will be used as input to the attention module, are then augmented by a modified CBAM module.

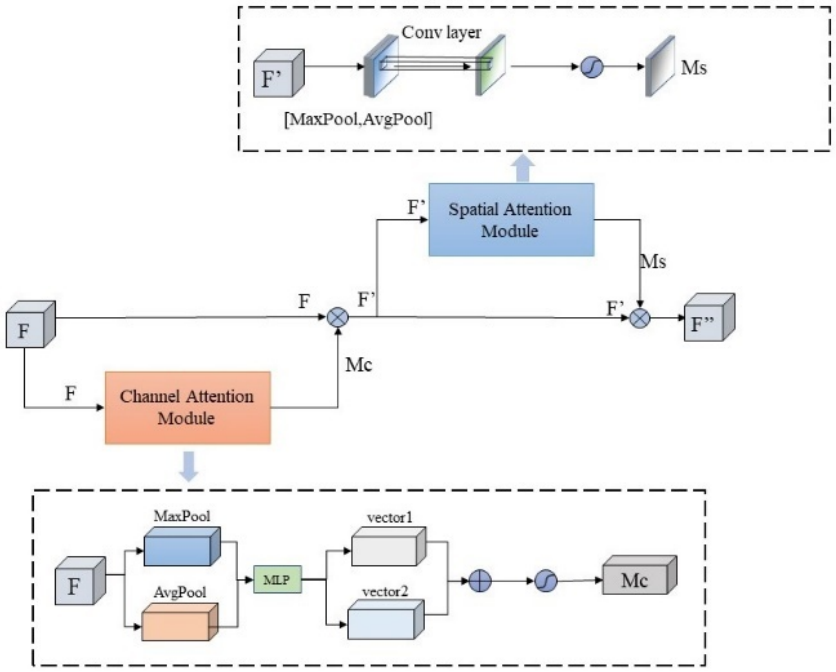

Figure 4. CBAM attention module

As shown in Fig. 4, this is a schematic diagram of the structure of CBAM. First, the size of the feature map F is H×W×C. Then, the global information is extracted through the global average pooling layer and the maximum pooling layer based on width and height to generate a 1×1×C feature map and fed into a two-layer neural network with shared weights, i.e., a Multi-Layer Perceptron (MLP, Multi-Layer Perceptron), which learns through inter-channel dependencies. Dimensionality reduction is achieved between the two neural layers by compression ratio r. The channel attention weighting factor equation is shown in Eq.4:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F_{avg}^C)) + W_1(W_0(F_{\max}^C))) \end{aligned} \quad (4)$$

W_1 and W_0 is the full connection weight of two layers contained in MLP, with hidden layer and ReLU activation function in the middle, σ represents Sigmoid activation function. As shown in Figure 4, spatial attention takes the output characteristic map of channel attention module as the input characteristic map of this module, focusing on the most informative part, which is a supplement to channel attention. Firstly, the maximum pooling and average pooling are carried

out in the channel dimension to fuse the information of different channels in the same position, which is used as the feature information of this position. Then, the position information obtained by the maximum pooling and average pooling is spliced in the channel dimension, and the heat map of spatial importance is obtained through convolution. Finally, the real heat map is generated through Sigmoid activation function, and multiplied by the original input to obtain the calibrated feature map Ms(F), which encodes the position that needs attention or suppression.

$$M_S(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) = \sigma(f^{7\times7}([F^S_{avg}; F^S_{\max}])) \quad (5)$$

As shown in Eq.5, two feature maps are obtained by two pooling operations in the spatial dimension, namely, F^s_{avg} and $F^s_{\max}$. These two feature maps are spliced based on the channel dimension, and then the channel dimension is reduced by using a 7×7 convolution kernel, $f^{7\times7}$ represents a convolution operation with a filter size of 7×7, and the dimension is reduced to a single channel feature map. Finally, the weight of the spatial dimension is generated by learning the dependency relationship between spatial elements through sigmoid.

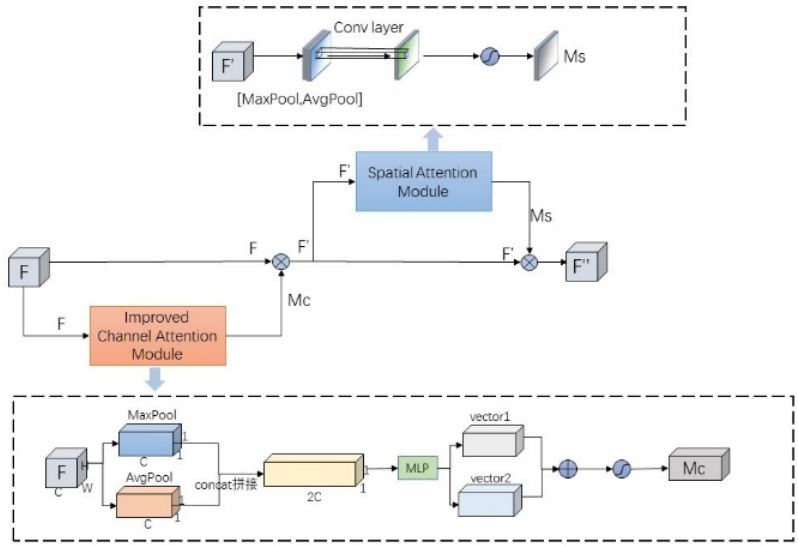

Figure 5. Improved CBAM attention module

As shown in Fig. 5, the channel attention module in CBAM is simply improved in this section, so that the two groups of features that have undergone maximum pooling and average pooling are concat spliced, and then the weights W_0 and W_1 are trained by MLP, and the formula is shown in Eq.6:

$$M_C(F) = \sigma(MLP[MaxPool(F); AvgPool(F)]) = \sigma(W_1(W_0([F^C_{\max}; F^C_{avg}]))) \quad (6)$$

The fused features after splicing are sent to MLP, which is composed of two fully connected layers. The input features X of the first fully connected layer are reduced in dimension to obtain feature Y_0, and the second fully connected layer is upgraded in dimension to obtain output feature Y_1, as shown in Formula Eq.7 and Eq.8:

$$Y_0 = W_0 \times X \quad (7)$$

$$Y_1 = W_1 \times Y_0 \quad (8)$$

It can be seen that the weight parameters of the first fully connected layer in the improved attention module mentioned above have increased, and the model performance has been relatively enhanced. This improved attention module is embedded into the weak supervision network architecture proposed in this chapter to achieve better detection results.

3.3. Basic Multiple Instance Detector

This paper mainly takes OICR as the main framework, and OICR is divided into two parts. The first part is to train the MIDN of the basic case classifier, which is transformed from WSDDN network; The second part is the refinement classifier, and the supervision of the refinement classifier is determined by the output of the previous stage. On this basis, three modules are introduced, namely, proposal generation and proposal selection and attention module. Firstly, Grad-CAM++ is combined with SS algorithm to generate several candidate boxes, and an improved attention module is added to achieve better results.

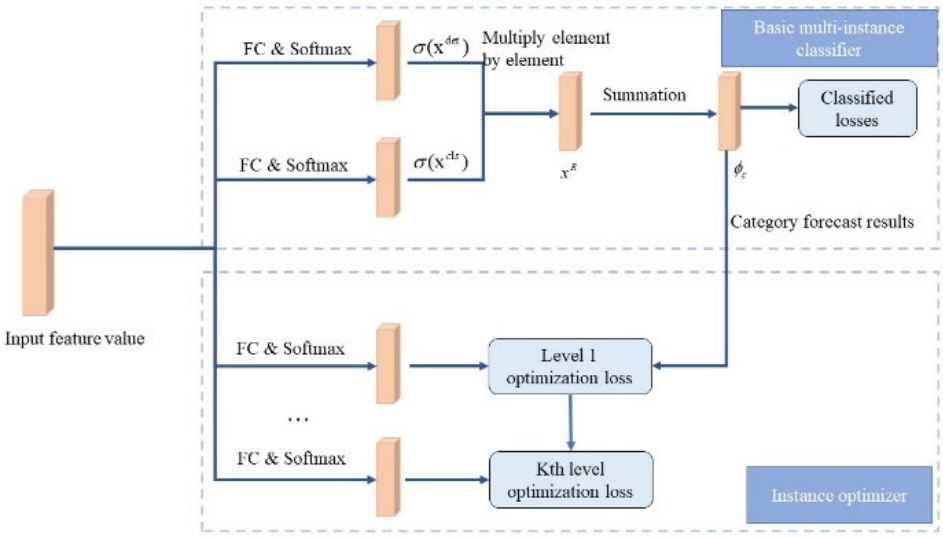

Figure 6. Basic multi-instance learning detector and refinement branch instance detector

Because of weakly supervised learning, only image-level annotations can be used, that is to say, there is only classification information but no location information in the data set. In order to better understand the semantic information inside the image, it is necessary to examine the map to the regional level and analyze the characteristics of each bounding box. Firstly, a basic detector is used to obtain the preliminary detection results, and the basic detector is optimized by transforming the weak supervised object detection problem into a multi-label classification problem following the idea of WSDDN using multi-instance learning. The proposed score obtained from the basic detector can guide the first level of the multi-level case optimizer, and the supervision of the case optimizer is determined by the output of its previous level. Multiple refinements at the first level can gradually detect a larger part of the target, as shown in Fig. 6.

The regional features x are then fed into the two streams by two separate fully connected layers and produce two feature matrices denoted as x^{cls} and $\mathrm{x}^{\det} \in R^{C\times|R|}$, where C denotes the number of categories and $|R|$ denotes the number of proposals. The two softmax functions are applied to x^{cls} and x^{cls} for two different directions, as shown in the following Eq.9 and Eq.10:

$$[\sigma(x^{cls})]_{ij} = \frac{e^{x_{ij}^{c}}}{\sum_{k=1}^{C} e^{x_{kj}^{c}}} \tag{9}$$

$$[\sigma(x^{\det})]_{ij} = \frac{e^{x_{ij}^{\det}}}{\sum_{k=1}^{|R|} e^{x_{ik}^{\det}}} \tag{10}$$

The formula for generating the region fraction by multiplying the elemental aspects is as follows:

$$x^{R} = \sigma(x^{cls}) \mathrm{Z}\ \sigma(x^{d}) \tag{11}$$

Finally, the category C image score can be obtained by summing all the proposed scores with the following equation:

$$\phi_c = \sum_{\mathrm{r}=1}^{|R|} x_{cr}^{R} \tag{12}$$

Given an image label $Y = [\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_C]^T \in R^{C\times1}$, $y_c = 1$ or $y_c = 0$ indicates whether the image has target c. In this training phase, we can perform the multi-label classification task by the standard multi-category cross-entropy loss function like the following equation, and then the instance classifier can be obtained according to the proposed score x^R. In this training phase, the loss function can be formulated as Eq.13:

$$L_{base} = -\sum_{c=1}^{C} \{y_c \log \varphi_c + (1-y_c)\log(1-\varphi_c)\} \tag{13}$$

3.4. Proposal Selection

After all the regional proposal boxes and scores are obtained through the above modules, how to adaptively select high-quality proposals becomes the key. Due to the lack of accurate location labels in the data, it is difficult for the weakly supervised object detector to select the most suitable bounding box from all the proposals of the object. Suggestions that get the highest classification score usually cover the different parts of the object, while many other suggestions that cover a larger part often have lower scores. Inspired by WSOD2, a simple strategy is used to combine low-level semantic information to train the weakly supervised object detector. Low-level semantic information summarizes the boundary characteristics of common objects, which is helpful to make up for the shortage of CNN in boundary discovery.

In OICR method, given an image, which contains the category of the target object, it selects only the candidate box with the highest category score and the candidate box with spatial overlap, and the rest are all negative examples. However, if the image contains multiple target objects of the same category, it is impossible to distinguish the positive and negative examples well, which will lead to the omission of some actual and valuable positive suggestion candidate boxes and the introduction of some inaccurate negative suggestion

boxes. In this paper, a simple but very effective suggestion selection strategy is proposed, which combines low-level semantic information to screen high-quality suggestions.

Firstly, a simple strategy is used to select the proposal candidate box with high score, and then the low-level semantic information is used to screen and adjust the candidate box. Specifically, an image is input, and a group of proposals $R = \{R_1, ..., R_{|R|}\}$ and corresponding proposal scores x^R are generated through the above-mentioned module. Each proposal in category C has an objetness score [42], marked as $O_{bu}(r)$, the selection of each proposal is as follows:

1. If, it means that the image contains at least one object whose target is Class C, then the proposal j_c with the highest score is selected according to the following formula, and it is marked as a pseudo-label class C, $y_{cj_c} = 1$, as shown in the following formula:

$$j_c = \arg\max_r x_{cr}^{\Re} \tag{14}$$

2. If the IOU between the proposal box and is higher than the value we defined (IOU=0.5) and it's $O_{bu}(r)$ has the highest score, this paper marks the proposal box as category c.

3. This paper continues to select the highest scoring proposal boxes in addition to the previously selected ones as described above.

4. Repeat this step until the IOU of a proposal box with the highest score is higher than 0.5.

3.5. Object Detector Refinement

After the base multi-instance learning detector, K classifier branches are iteratively trained, and the refined instance detector section contains K classifier branches from Cls 1 to Cls K. The final Bbox is obtained by box regression after the last classifier branch Cls K. Each classifier outputs a pseudo-label as the supervision of the next classifier, so the whole process Only the initial classifier Cls 0, which is the base multi-instance learning detector, uses the real image labels. The subsequent k classifiers are trained with pseudo-labels, and for the kth classifier, its loss function is as the following Eq.15:

$$L_{ref}^k = -\frac{1}{|R|}\sum_{r\in R}(w_r^k \cdot CE(p_r^k, \hat{p_r^k})) \tag{15}$$

The CE part of the equation is as follows:

$$CE(p_r^k, \hat{p_r^k}) = -\sum_{c=0}^{C} \hat{p_{rc}^k} \log(p_{rc}^k) \tag{16}$$

$\hat{p_{rc}^k}$ and p_{rc}^k are respectively the prediction result and label of proposal r for the c-th category, which is the pseudo-label generated by the previous classifier.

The focus is on this weighting factor, which is obtained from the top-down information and the bottom-up information, as shown in Eq.17:

$$w_r^k = \alpha O_{bu}(r) + (1-\alpha) O_{td}^k(r) \tag{17}$$

The bottom-up object evidence is the similarity score objectness mentioned before, which is the four similarity measures, namely MS(Multi-scale Saliency), CC(Color Constrast), ED(Edge Density) and SS(Superpixels Straddling), which is the classification score calculated according to the classification result obtained by the previous one, namely the k-1st classifier, as shown in Eq.18:

$$O_{td}^k(r) = \sum_{c=0}^{C}(p_{rc}^{k-1} \cdot \hat{p_{rc}^k}) \tag{18}$$

α is a balance factor set by itself to balance the weight of these two information. The intuitive understanding of this loss function is to penalize the classification result of the current classifier for each proposal with the pseudo-label generated by the previous classifier, and the higher the weight of the proposal the stronger the penalty.

3.6. Bounding Box Regression

Since it is weakly supervised learning, there is no strong supervised information in the dataset. In OICR, it relies on the location of the highest scoring region proposal in the multi-instance learning branch, but this label is a coarse label, and this coarse prediction result will definitely give a bad effect to the detector. So this paper adds a regression branch to the previous module.

Although convolutional neural networks can learn features well, they have shortcomings in discovering boundaries, so during training, we explore how to use bottom-up object evidence to guide the target's bounding box for updating.

An objective detector is actually a bounding box sorting function, where an important factor is the objective metric. In weakly supervised target detection, if the classification confidence is considered as an objective score, the shortcoming is that even very good detectors have difficulty in distinguishing complete objects from obviousness object parts or irrelevant backgrounds. In target detection, the most important thing is said to be clear boundaries and centers. Therefore we expect to eventually find a bounding box that completely encloses the complete object, and the above-mentioned (bottom-up object) features with object boundaries can exactly compensate for CNN's deficiency in its aspect.

The position loss function uses L1, L2 or smooth loss functions to regress the four coordinate values. The goal of the regressor is to output a correction for each box for each of the four parameters x,y,w,h:

$$t_r = (t_r^x, t_r^y, t_r^w, t_r^h) \tag{19}$$

A total of K classifier branches containing Cls 1 to Cls K are divided, and the final bbox is obtained by bounding box regression after the last classifier branch Cls K, the formula as shown in Eq.20:

$$L_{box} = \frac{1}{|R_{pos}|}\sum_{r=1}^{|R_{pos}|}(w_r^K \cdot smooth_{L1}(t_r, \hat{t}_r)) \quad (20)$$

The final loss function for all modules combined is:

$$L = L_{base} + \lambda_1 \sum_{k=1}^{K} L_{ref}^k + \lambda_2 L_{box} \quad (21)$$

3.7. Overall Training Framework

Firstly, given an image, a region proposal R is generated by selective search combined with Grad-CAM++, and then the region features are extracted by CNN and ROI pooling layers and two fully connected layers. Then, the region features enter two streams through two full connections, one classification stream and one localization stream, and the region proposal score is obtained by multiplying the corresponding elements according to the formula. Next, the dimensions of region R are aggregated to obtain the image-level classification vector, and the image-level labels are used as supervision to guide the training network training by applying a binary cross-entropy loss function optimization and summarizing its boundary features with bottom-up objects.

4. Experimental Results and Analysis

4.1. Experimental setup

This paper evaluates the method proposed in this paper on three target detection benchmarks: PASCAL VOC2007 and PASCAL VOC2012. After removing these bounding box annotations provided by the data set, only the image and its classification label information are used for training.

Two data sets like PASCAL VOC2007 and PASCAL VOC2012 are the most widely used benchmarks for weak supervised target detection. Performance is measured by the average accuracy (AP) of the maps of all object classes, and CorLoc, a widely used WSOD evaluation, is also reported. Accuracy, Recall and mean average Precision (mAP) can all be used to evaluate the performance of the target detection algorithm. Among them, the mAP and CorLoc obtained in the experiment of this paper all follow the calculation standard stipulated by PASCAL VOC, that is, the IoU between the prediction result frame and the real frame is greater than 0.5.

This paper generates region proposals by combining a selective search algorithm and Grad CAM++, and the proposed features are fed into a modified CBAM attention module. This paper uses the VGG16 network as the base network, and uses the stochastic gradient descent SGD with an initial learning rate set to 0.001, weight decay set to 0.0005 and momentum set to 0.9. On the VOC2007 dataset, the total number of iteration steps is set to 80,000, and the learning rate is reduced to 0.0001 at the 40,000th step. dataset, we double the number of iteration steps and the learning rate decay step to the 80,000th step. This paper follows the multi-scale settings of PCL and OICR in training, specifically, the short edges of the input image are randomly rescaled to a scale of {480,576,588,864,1280}, and the length of the long edges is restricted to no more than 2000.

4.2. Ablation experiments

In order to prove the effectiveness of the three modules, namely, regional suggestion generation (PG), regional suggestion selection (PS) and CBAM module, the improved weakly supervised target detection network is ablated in the test set based on PASCAL VOC2007 data set, and the best detection results are displayed in bold, so we can see the performance of the three module methods introduced in this chapter on weakly supervised target detection in these 20 categories, as shown in Table 1:

Table 1. Performance of different methods and modules on 20 classes of VOC 2007 test data set

Method	MIL	MIL+CBAM	MIL+PG-PS	MIL+CBAM+PG-PS
Aero	56.2	55.2	60.3	66.0
Bicycle	62.1	62.5	61.4	65.2
Bird	39.4	43.0	47.1	58.3
Boat	21.8	22.1	24.5	39.1
Bottle	10.3	12.7	14.1	22.3
Bus	63.6	66.1	67.6	66.7
Car	60.6	62.0	63.0	68.9
Cat	31.8	38.2	68.1	63.4
Chair	24.8	26.3	22.8	31.7
Cow	45.9	48.9	52.9	67.8
Table	35.3	37.7	40.2	43.3
Dog	24.1	26.1	59.7	62.2
Horse	36.7	45.5	62.9	72.0
Mbike	63.3	64.3	61.2	69.5
Person	13.1	12.4	10.1	20.4
Plant	23.1	24.6	22.3	27.8
Sheep	39.4	42.1	45.1	59.1
Sofa	49.1	46.6	50.8	59.8
Train	64.7	65.8	69.2	52.6
TV	60.3	62.3	64.4	64.9

It can be clearly seen from Table 1 that in the weak supervised object detection algorithm model proposed in this chapter, the addition of each sub-module improves the performance of the model to a certain extent. In the MIL baseline, after adding PG and PS modules, the model has been improved obviously, and after adding the improved CBAM

module on this basis, the performance has been improved relatively obviously, which shows that it is useful for improving CBAM.

As shown in Table 2, the first column represents the final effect of the image directly generated by the Selective Search algorithm and sent to the basic multi-instance detector for multi-instance detection; The second column represents the introduction of PG(Proposal Generation) and PS(Proposals Selection) modules on the basis of MIL detector; The third column represents that an improved CBAM module is added after the candidate box is generated to enhance the feature map. The third column represents adding PG-PS module and CBAM module on the basis of MIL detector, which is the improved method proposed in this chapter, and it can be seen that it is obviously improved compared with the previous method.

Table 2. MAP of different method modules in VOC 2007 test data set

MIL+CBAM	MIL+PG-PS	MIL+CBAM+PG-PS	mAP(%)
-	-	-	47.0
✓	-	-	49.2
-	✓	-	51.1
-	-	✓	**54.0**

From Table 2, we can clearly see that the addition of each sub-module in the weak supervised object detection algorithm model we proposed improves the performance of the model to a certain extent. We can find that in the OICR network baseline, the model has been significantly improved after adding PG and PS modules, and on this basis, the performance has also been significantly improved after adding the improved CBAM module, indicating that we have played a role in the improvement of CBAM.

4.3. Comparison with other methods

Table 3 and Table 4 respectively show the detection performance and positioning performance of the weakly supervised target detection model proposed in this paper and other weakly supervised target detection models in 20 categories of VOC 2007 data set. The bold mark is the highest accuracy in this category. We can see from the table that our model has achieved the highest accuracy in 11 categories of aircraft, birds, cars, cats, chairs, cows, dogs, horses, sheep, sofas and TVS, and the highest positioning accuracy in 7 categories of birds, cats, chairs, cows, dogs, horses and sheep, significantly improving the local positioning problem that is very prone to occur in animal categories. That is, only the head of the animal was detected, and the whole animal was ignored.

Table 3. Comparison of detection performance of each category

Method	WSDDN	Context-LocNet	OICR	PCL	C-WSL	WSCDN	WSOD2	Ours
Aero	39.4	57.1	58.0	54.4	62.9	61.2	65.1	**66.0**
Bicycle	50.1	52.0	62.4	**69.0**	64.8	66.6	64.8	65.2
Bird	31.5	31.5	31.1	39.3	39.8	48.3	57.2	**58.3**
Boat	16.3	7.6	19.4	19.2	28.1	26.0	**39.2**	39.1
Bottle	12.6	11.5	13.0	15.7	16.4	15.8	**24.3**	22.3
Bus	64.5	55.0	65.1	62.9	69.5	66.5	**69.8**	66.7
Car	42.8	53.1	62.2	64.4	68.2	65.4	66.2	**68.9**
Cat	42.6	34.1	28.4	30.0	47.0	53.9	61.0	**63.4**
Chair	10.1	1.7	24.8	25.1	27.9	24.7	29.8	**31.7**
Cow	35.7	33.1	44.7	52.5	55.8	61.2	64.6	**67.8**
Table	24.9	**49.2**	30.6	44.4	43.7	46.2	42.5	43.3
Dog	38.2	42.0	25.3	19.6	31.2	53.5	60.1	**62.2**
Horse	34.4	47.3	37.8	39.3	43.8	48.5	71.2	**72.0**
Mbike	55.6	56.6	65.5	67.7	65.0	66.1	**70.7**	69.5
Person	9.4	15.3	15.7	17.8	10.9	12.1	**21.9**	20.4
Plant	14.7	12.8	24.1	22.9	26.1	22.0	**28.1**	27.8
Sheep	30.2	24.8	41.7	46.6	52.7	49.2	58.6	**59.1**
Sofa	40.7	48.9	46.9	57.5	55.3	53.2	59.7	**59.8**
Train	54.7	44.4	64.3	58.6	60.2	**66.2**	52.2	52.6
TV	46.9	47.8	62.6	63.0	66.6	59.4	64.8	**64.9**
mAP	34.8	36.3	41.2	43.5	46.8	48.3	53.6	**54.0**

As shown in Table 4, we can intuitively see the performance of the positioning performance of the method

proposed in this chapter and other methods in different categories of VOC 2007 data sets. Compared with other methods, the positioning performance in eight categories such as birds and cats has been improved to some extent, which proves the effectiveness of the method in this chapter.

As shown in Table 5, it shows the comparison of the accuracy and positioning performance of the weakly supervised target detection model proposed in this chapter with other weakly supervised target detection models on VOC 2012 data sets. It can be seen that the method proposed in this chapter has improved to some extent compared with other methods, which fully proves the effectiveness of this work.

Compared with the WSOD2 algorithm model, the method proposed in this chapter has been improved to some extent from the aspect of improving the quality of proposal box, and the quality of proposal box has been effectively improved from the aspects of proposal generation and proposal selection. The improved attention module has a good impact on the follow-up training and improved the performance of the weakly supervised target detection model. In PCL algorithm, candidate frame clustering is used to solve the multi-instance problem of weakly supervised target detection according to the standard of whether candidate frames overlap or not, but the method in this chapter has made many improvements from the proposal generation part and used a simple and effective proposal selection strategy, which is obviously more advantageous. By comparing the existing methods, the effectiveness of the algorithm in this chapter is fully verified, and its performance improvement is mainly due to the improvement of the candidate box.

Table 4. Localization performance of each model on PASCAL VOC 2007 training verification set

Method	WSDDN	Context-LocNet	OICR	PCL	C-WSL	WSCDN	WSOD2	Ours
Aero	65.1	83.3	81.7	79.6	85.8	85.8	**87.1**	86.8
Bicycle	58.8	68.6	80.4	**85.5**	81.2	80.4	80.0	81.2
Bird	58.5	54.7	48.7	62.2	64.9	73.0	74.8	**74.9**
Boat	33.1	23.4	49.5	47.9	50.5	42.6	**60.1**	55.9
Bottle	39.8	18.3	32.8	**37.0**	32.1	36.6	36.6	36.8
Bus	68.3	73.6	81.7	83.8	**84.3**	79.7	79.2	79.6
Car	60.2	74.1	85.4	83.4	**85.9**	82.8	83.8	84.8
Cat	59.6	54.1	40.1	43.0	54.7	66.0	70.6	**71.3**
Chair	34.8	8.6	40.6	38.3	43.4	34.1	43.5	**44.4**
Cow	64.5	65.1	79.5	80.1	80.1	78.1	88.4	**88.9**
Table	30.5	**47.1**	35.7	50.6	42.2	36.9	46.0	46.5
Dog	43.0	59.5	33.7	30.9	42.6	68.6	**74.7**	**76.3**
Horse	56.8	67.0	60.5	57.8	60.5	72.4	87.4	**88.0**
Mbike	82.4	83.5	88.8	90.8	90.4	91.6	**90.8**	90.5
Person	25.5	35.3	21.8	27.0	13.7	22.2	**44.2**	42.6
Plant	41.6	39.9	57.9	**58.2**	57.5	51.3	52.4	51.1
Sheep	61.5	67.0	76.3	75.3	82.5	79.4	81.4	**82.5**
Sofa	55.9	49.7	59.9	**68.5**	61.8	63.7	61.8	62.7
Train	65.9	63.5	75.3	**75.7**	74.1	74.5	67.7	68.3
TV	63.7	65.2	81.4	78.9	**82.4**	74.6	79.9	80.1
CorLoc	53.5	55.1	60.6	62.7	63.5	64.7	69.5	**69.7**

Table 5. Comparison of various models on PASCAL VOC 2012 data set

method	mAP	CorLoc
OICR	37.9	62.1
PCL	40.6	63.2
WSCDN	43.3	65.2
WSOD2	47.2	71.9
Ours	47.9	72.5

4.4. Experimental Results

Fig. 7 shows some detection results of the algorithm model proposed in this paper on PASCAL VOC2007 data set, where the green box is the real label of the image and the red box is the detection result of the algorithm proposed in this paper. It can be seen that the prediction results of the algorithm model proposed in this paper are basically close to the real tags, but there may be local optimization problems for human detection, in which the detection frame is too small, but this will not happen for other objects recognition, which is an improvement compared with OICR and WSOD2.

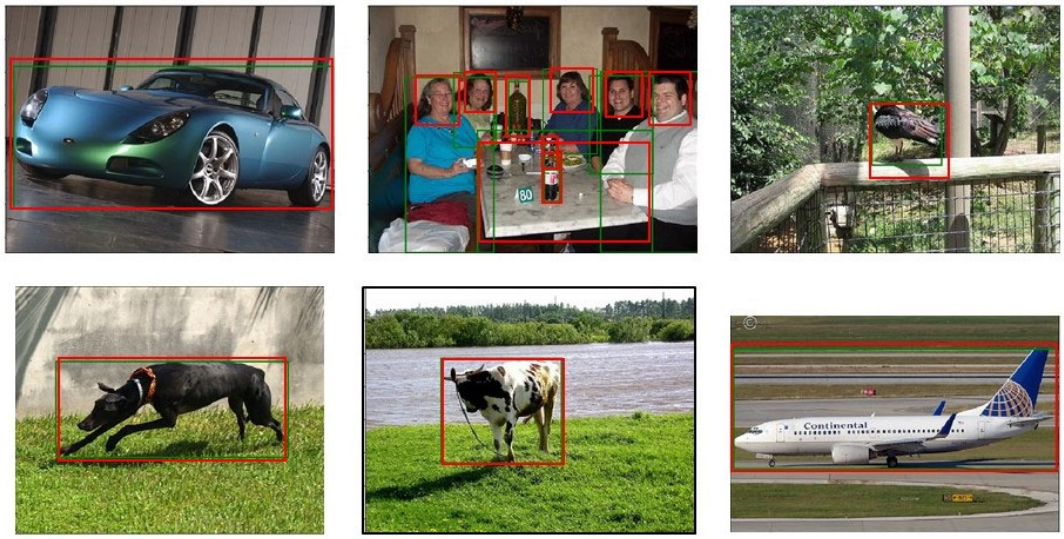

Figure 7. Experimental part of the results show

5. Conclusion

In this paper, firstly, a detection method based on multi-instance learning idea is used to obtain the initial object bounding box, and the weakly supervised object detection problem is understood as a multi-instance learning problem, in which the input image is equivalent to a set of object proposals. In this paper, three modules can be embedded in the framework of weakly supervised target detection, which are used to generate high-quality proposals and filter them. Finally, more accurate proposals that are beneficial to subsequent training are selected, and their effectiveness is demonstrated on PASCAL VOC 2007 and PASCAL VOC 2012 data sets, and the existing weakly supervised target detection algorithms are significantly improved.

References

[1] Tang P, Wang X, Bai S, et al. Pcl: Proposal cluster learning for weakly supervised object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 42(1): 176-191.

[2] Tang P, Wang X, Bai X, et al. Multiple instance detection network with online instance classifier refinement [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2843-2851.

[3] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In NIPS, pages 379–387, 2016.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In ICCV, pages 764–773, 2017.

[5] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In Proc. CVPR, 2016.

[6] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. International journal of computer vision, 111(1):98–136, 2015.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. IJCV, 88(2):303–338, 2010.

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

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[11] Zeng Z, Liu B, Fu J, et al. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8292-8300.

[12] Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L.: Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017).

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

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

[15] Jie, Z., Wei, Y., Jin, X., Feng, J., Liu, W.: Deep self-taught learning for weakly supervised object localization. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2017).

[16] Ren, Z., Yu, Z., Yang, X., Liu, M.Y., Lee, Y.J., Schwing, A.G., Kautz, J.: Instanceaware, context-focused, and memory-efficient weakly supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020).

[17] Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[18] Kaiming He, Georgia Gkioxari, Piotr Doll´ar, and Ross Girshick. Mask r-cnn. In Proc. ICCV, 2017.

[19] 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.

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv:1811.00982, 2018.

[21] 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.

[22] Yang K, Li D, Dou Y. Towards precise end-to-end weakly supervised object detection network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8372-8381.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755, 2014.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3):211–252, 2015.

[25] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In Proc. ECCV, 2016.

[26] Wan F, Wei P, Jiao J, et al. Min-entropy latent model for weakly supervised object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1297-1306.

[27] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In Proc. ECCV, 2018.

[28] Gao M, Li A, Yu R, et al. C-wsl: Count-guided weakly supervised localization[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 152-168.

[29] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[30] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104: 154-171.

[31] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018: 839-847.

[32] Arun A, Jawahar C V, Kumar M P. Dissimilarity coefficient based weakly supervised object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9432-9441.

[33] Shen Y, Ji R, Wang Y, et al. Cyclic guidance for weakly supervised joint detection and segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 697-707.

[34] Li X, Kan M, Shan S, et al. Weakly supervised object detection with segmentation collaboration[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9735-9744.

[35] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proc. CVPR, 2016.

[36] Chen Z, Fu Z, Jiang R, et al. Slv: Spatial likelihood voting for weakly supervised object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12995-13004.

[37] Lin C, Wang S, Xu D, et al. Object instance mining for weakly supervised object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11482-11489.Zhang D, Han J, Zhang Y. Supervision by fusion: towards unsupervised learning of deep salient object detector[A]. IEEE International Conference on Computer Vision[C]. Honolulu, Hawaii, USA: IEEE, 2017. 4048 - 4056.

[38] Wang L, Lu H, Wang Y, et al. Learning to detect salient objects with image-level supervision[A]. IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolu-lu, HI, USA: IEEE, 2017. 136 - 145.

[39] Woo S, Park J, Lee JY, et al. CBAM: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018, 3-19.

[40] Zhang Y, Bai Y, Ding M, et al. W2f: A weakly-supervised to fully-supervised framework for object detection [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 928-936.

[41] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 391-405.

[42] Alexe B, Deselaers T, Ferrari V. What is an object?[C]//2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010: 73-80.