www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 11 Issue 01January 2022, Page **No.25493-25507**
ISSN: 2319-7242 DOI: 10.18535/ijecs/v11i02.4650

Strong Representation Learning for Weakly Supervised Object Detection

Song Yu, Li Min*, Du Weidong, He Yujie, Gou Yao, Wu Zhaoqing, Lv yilong

Xi 'an High-tech Research Institute

Corresponding Author's Email:proflimin@163.com

Abstract:
To solve the problem that the feature maps generated by feature extraction network of traditional weakly supervised learning object detection algorithm is not strong in feature, and the mapping relationship between feature space and classification results is not strong, which restricts the performance of object detection, a weakly supervised object detection algorithm based on strong representation learning is proposed in this paper. Due to enhance therepresentation ability of feature maps, the algorithm weighted the channels of feature maps according to the importance of each channel, to strengthen the weight of crucial feature maps and ignore the significance of secondary feature maps. Meanwhile, a Gaussian Mixture distribution model with better classification performance was used to design the object instance classifier to enhance further the representation of the mapping between feature space and classification results, and a large-margin Gaussian Mixture (L-GM) loss was designed to increase the distance between sample categories and improve the generalization of the classifier. For verifying the effectiveness and advancement of the proposed algorithm, the performance of the proposed algorithm is compared with six classical weakly supervised target detection algorithms on VOC datasets. Experiments show that the weakly supervised target detection algorithm based on strong representation learning has outperformed other classical algorithms in average accuracy (AP) and correct location (CorLoc), with increases of 1.1%~14.6% and 2.8%~19.4%, respectively.

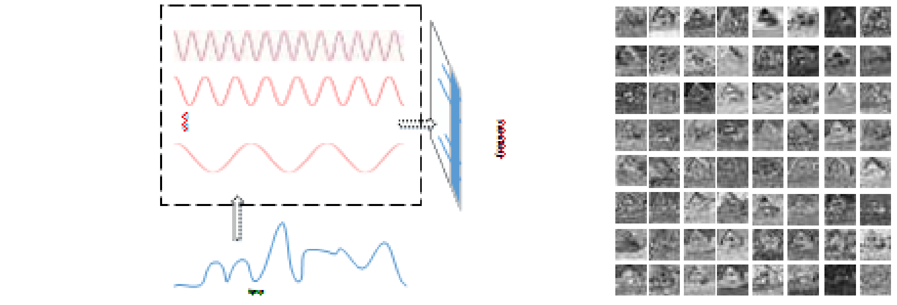
Key words: Object detection; Weakly supervised learning; Strong representation learning; Gaussian mixture distribution

Introduction

Object detection is an essential task in computer vision. With the continuous development of the deep neural network, object detection algorithms based on this method have become a research hotspot, generally divided into one-stage and two-stage. The representative algorithms are YOLO series[1~3] and R-CNN[4~6] series. Deep learning dramatically improves the detection effect and detection performance of the algorithm. Nevertheless, the efficiency of these algorithms has to rely on a large number of accurately labeled data sets. However, there is only classification information for the target (image-level tags) and no location tags in many cases. Therefore, researchers pay more and more attention to weakly supervised object detection.

However, the feature extraction network of most current algorithms adopts the traditional neural network structure, and the feature maps generated by each convolutional layer are not representational enough to efficiently utilize the feature maps of different channels, thus affecting the target detection results. As shown in Figure 1 (a), each channel in the feature maps is similar to the frequency domain information after the time-frequency transformation of the signal. A continuous signal in the time domain (analogous to the original image) can be represented by several salient signals in the frequency domain. In contrast, other secondary frequency domain signals are ignored. Figure 1 (b) shows that the 64 channels generated by the feature graph output from the convolution layer make different contributions to subsequent detection tasks.

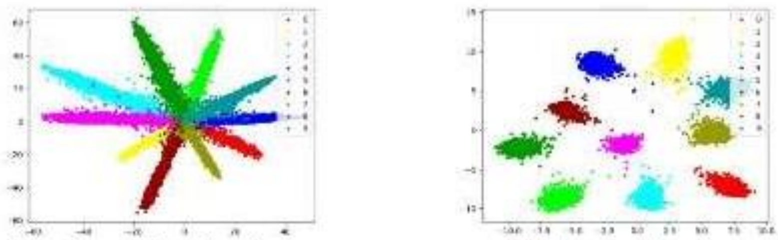
Different weights are assigned to each channel of the feature maps to indicate the importance of information. The channel attention mechanism can strengthen the weight of key features and ignore the secondary features, thus enhancing the representativeness of feature maps. Therefore, a feature extraction network based on strong representation learning is introduced in this paper, which makes the network learn to selectively emphasize the critical information features and suppress the secondary feature maps.



[a] The signal time and frequency domain (b)The visualization of feature maps

Figure 1 Relationship diagram of each channel in feature maps

For the classification algorithm of instance classifier in the weakly supervised network, the discriminant function of traditional neural network primarily uses the Softmax function to calculate the category score. Finally, it discriminates the object category according to the score. Taking image classification as an example, a linear transformation is performed onthe extracted depth features, and the discriminant scores of each category corresponding to the input samples are calculated. However, the classification score of the Softmax function cannot accurately represent the probability distribution of the feature space of training samples. This will have some influence on the object detection algorithm based on weakly-supervised learning. Figure 2 is the comparison diagram of the Softmax and Gaussian mixture distribution classification results of MNIST data sets by the two classification calculation methods. Figure 2 (a) is the classification result of using Softmax as the loss function, and Figure 2 (b) is the classification result of using Gaussian mixture distribution as the loss function. It can be seen that the representation of classification results of Gaussian mixture distribution is obviously better than that of Softmax. In this paper, the classification algorithm is proposed to take advantage of sample Gaussian mixture distribution characteristics and use L-GM large interval Gaussian mixture [7] distribution to calculate the loss function.



(a) The classification results using Softmax (b) The classification results using GM

Figure 2 Comparison of classification results between Softmax and GM

Combined with the above methods, this paper proposes a new target detection algorithm based on weakly-supervised learning. The network structure of the algorithm is generally divided into two parts. The first part is the strong representational feature extraction network, and the second part is the target detection network. The backbone of the feature extraction network uses the improved VGG16 network, that is, the last convolutional module in the network, to complete the distribution of attention to the feature maps extracted from the network. In order to improve the utilization efficiency of feature maps, the weights of crucial feature maps were strengthened, and the weights of secondary feature maps were ignored. This method can improve feature extraction performance at a lower cost and provide more characteristic feature maps for detection networks. The MIL (multi-instance learning) part of the target detection network usually completes the two tasks of instance classification and location. For instance, classification algorithm, Gaussian Mixture model is used to calculate l-GM loss function. The method assumes that the feature value of the samples conforms to the Gaussian distribution and calculates the probability distribution density of the sample category to complete the category prediction of the samples. Through many comparison experiments, it can be seen that AP and

CorLoc of the proposed method are both higher than those of the benchmark.

Related work

Weakly supervised learning object detection

Literature [8] solves the problem of weakly supervised target positioning through classification and detection adaptation. Use the selected object proposal to fine-tune all layers to produce a fully adaptive detection network. Literature [9] uses the local space and semantic patterns captured in the convolution layer of the classification network to propose an image multi-object detection and location method based on beam search. The location strategy in literature [8] is relatively complex, and the candidate proposal and fine-tune steps need to be completed. The method in literature [9] is improved compared with that in literature [8]. Although end-to-end training can be completed, the overall detection effect is not outstanding. Literature [10] creates a new approach for weakly supervised target detection algorithm. Many subsequent classical algorithms [11~15] are improved based on this algorithm. Literature [10] proposes a weakly supervised end-to-end deep detection algorithm, which operates at the level of image region and performs region selection and classification. This algorithm has become a mainstream algorithm for weakly supervised target detection. Researchers have designed many variants based on this algorithm and made a breakthrough in detection accuracy. Literature [11] designed a network with multiple instance learning and bounding box regression branches. A guide attention module based on classification loss was introduced to extract implicit location information in features effectively. Literature [12] integrates the MIL and instance classifier optimization process into a single deep network and conducts end-to-end training on the network with only image-level supervision. Literature [13] proposed a spatial likelihood voting (SLV) module to converge the locating process of the proposal. All area suggestion boxes in a given image act as voters, voting for possibilities in each category. After fine-tuning in regions with large likelihood values, the voting results are regularized into boundary boxes for final classification and locating. Literature [14] combined selective search with gradient-weighted Class Activation Mapping technology, and the generated proposal could better cover the whole object. In terms of proposal selection, as many positive proposals with confidence as possible should be selected. The weight of loss of hard negative proposals should be improved to make the training more effective. Literature [15] adopts the strategy of generating proposal clusters to learn refined instance classifiers through an iterative process. In convolutional neural networks, multiple lines are used to implement refinement of iterative instance classifiers. The first one is the MIL network, and the rest are instance classifiers supervised by the previous network.

The main idea of the algorithms above is to use the traditional neural network for feature extraction and then use the

classifier for example classification to achieve the purpose of final classification and location. Especially after the algorithm proposed in the literature [10], most weakly supervised target detection adopts its idea for improvement. Many algorithms are improved in training strategies, box selection techniques and locating methods. However, weakly supervised target detection performance is rarely enhanced by considering features and a strong representation of sample space.

Strong representation of sample feature

In the computer vision task, sample features are more representational after the attention mechanism is introduced. Attention mechanism has been widely used in various fields of artificial intelligence. According to the domain of attention, it can be divided into spatial domain [16-17], channel domain [18], layer domain [19], time domain [20], and mixed domain [21-22]. Literature [16] puts forward a Spatial Transformer module, which can transform spatial domain information in pictures to extract critical information. Literature [17] ultimately gets rid of convolution operation based on attention mechanism. These models are superior in quality and require significantly less training time. Literature [18] proposes a new architectural unit, called the "squeeze-Congestion" (SE) block, which adaptively recalibrates the channel feature response by explicitly modeling the interdependencies between the channels. These blocks can be stacked together to form a SENet architecture. SE blocks bring significant performance improvements to CNN at a slight additional computing cost. Literature [19] introduces a Feature Pyramid Attention module to perform spatial pyramid attention structure on high-level output and combining global pooling with learning a better feature representation. A

Global Attention Upsample module on each decoder layer to provide global context as guidance of low-level features to select category localization details. Literature [20] is suitable for data with temporal characteristics. In the field of computer vision, the attention mechanism is regarded as the sampling of a region (sequence) point on an image, which is the point that needs attention. A Convolutional Block Attention Module (CBAM) is proposed in the literature [21], which calculates Attention sequentially according to channel and space dimensions. Literature [22] proposed the dual attention network (DANet) to integrate local features and their global relevance adaptively. Two attention modules are attached to the expanded FCN to build semantic dependency models on spatial and channel dimensions, respectively. Finally, the output of the two attention modules is added to improve the feature representation further, to obtain more accurate segmentation results.

The attention mechanism has been a popular feature weighting method in recent years. Key features can be assigned more weight while less critical features are ignored to improve feature utilization efficiency and enhance feature characterization. Especially after Google put forward the Transformer method, the attention mechanism is mushrooming by researchers. Furthermore, the channel attention mechanism is the best of the one in these attention mechanisms. The attention mechanism is simple to implement, enhances the relationship between feature map channels, and has a stronger representation of samples. All these advantages are applicable to a weakly supervised target detection network of instance classification.

Methods

The overall architecture of the algorithm in this paper is shown in Figure 3. It consists of two parts—a strong representational feature extraction network and an object detection network. Feature extraction network adopts channel weighted VGG16 network. The object detection network is divided into four branches. The 0th branch is the primary MIL network, which completes the initial target instance classification and positioning. The remaining branches are involved in the instance classification calculation of this branch with the classification results of the (k-1)th as false labels.

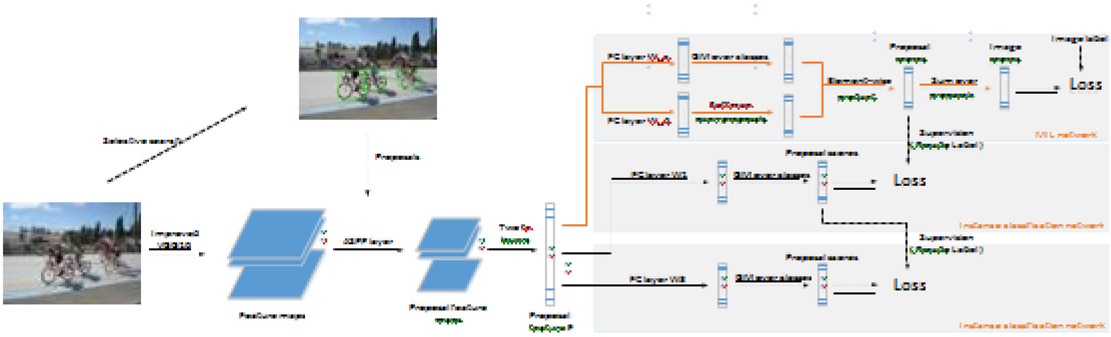


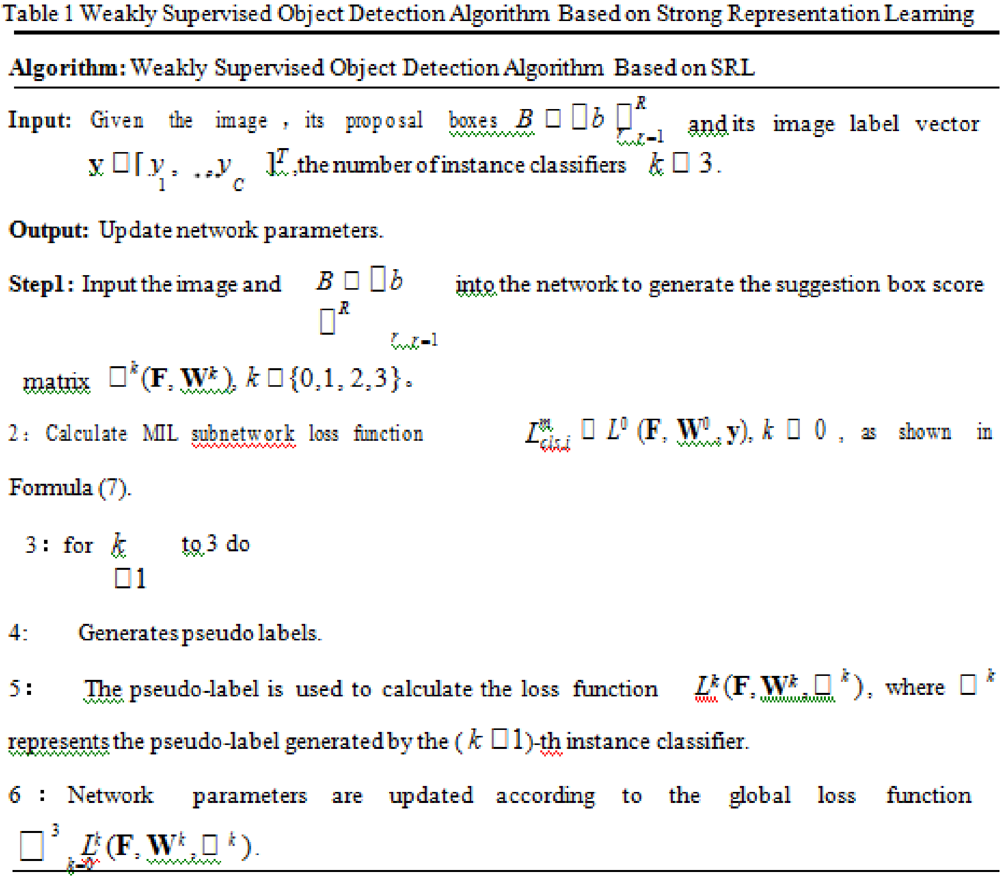
Figure. 3 Algorithm structure diagram of this paper

The algorithm flow is described in detail below. As shown in Figure 3, given an image, instance proposals boxes B □ □ b □$_{r}^{R}$ $_{r=1}$ are generated from Selective Search[23] or EdgeBox[24], and b_r represents the r-th proposal box. During

The forward process of training, images and these proposals are sent to the improved VGG16 network and the subsequent ASPP layer [25], and feature maps of fixed size are generated for each proposal. Channel attention mechanism is used to improve channel utilization (performance) in feature extraction networks. The proposals feature maps are then fed into two fully connected (FC) layers to generate proposal feature vectors $\mathbf{F}$. The object detection network is divided into four branches. The 0th branch is the basic MIL network, and the

subsequent three branches are instance classifiers. In MIL network, weakly supervised object detection is accomplished through object category scoring and object location scoring. The specific operation is as follows: The proposal conv feature map extracted by the feature extraction network generates proposal features through two fully connected layers. After entering the 0th branch, the Proposal features pass through two parallel full-connection layers to generate two C □ R matrices, where C represents the number of classes and R represents the number of proposals. The two matrices are used for category scoring and location scoring respectively. For category scoring, this algorithm uses Gaussian mixture distribution model (GM) instead of traditional Softmax scoring. The category score of the sample is determined by the probability density of the Gaussian mixture distribution of the sample. In location scoring, softmax scoring in the traditional algorithm is still used to obtain a matrix of position scores. The corresponding elements of the two matrices are multiplied to generate a Proposal scores matrix, which can not only be used as the pseudo-label of the next branch, but also calculate the sum of each row of the matrix as image scores, and finally calculate the loss function of this branch with image label. Feature maps of proposals are still used as input of subsequent branches. Classified scores are obtained by softmax function, and scores generated by the previous branch are used as pseudo-labels of thistest for supervised learning. The overall algorithm of this paper (SRL) is shown in the following table:

Table 1 Weakly Supervised Object Detection Algorithm Based on Strong Representation Learning



Algorithm: Weakly Supervised Object Detection Algorithm Based on SRL

Input: Given the image , its proposal boxes B □ □b □$^{R}_{r,r=1}$ and its image label vector $\mathbf{y}$ □ [y_1, . . y_C]^T ,the number of instance classifiers k □ 3 .

Output: Update network parameters.

Step1: Input the image and B □ □b □$^{R}_{r,r=1}$ into the network to generate the suggestion box score matrix □^{k}($\mathbf{F}$, $\mathbf{W}^k$), k □ {0,1, 2, 3} 。

2：Calculate MIL subnetwork loss function $L^{m}_{cls,i}$ □ L^0 ($\mathbf{F}$, $\mathbf{W}^0$, $\mathbf{y}$), k □ 0 , as shown in Formula (7).

3：for k □1 to 3 do

4: Generates pseudo labels.

5： The pseudo-label is used to calculate the loss function L^k($\mathbf{F}$, $\mathbf{W}^k$, □^{k}) , where □^{k} represents the pseudo-label generated by the (k □1)-th instance classifier.

6 ： Network parameters are updated according to the global loss function □$^{3}_{k=0}$ L^k($\mathbf{F}$, $\mathbf{W}^k$, □^{k}) .

Strong representational design of feature extraction network

The convolutional neural network is used in traditional feature extraction networks to extract feature images. In this way, although the weight of critical features and other secondary features in the feature maps is learned by using the convolution kernel, the importance of each feature maps is not different. The representational ability of feature maps can not better reflect the object.

Channel attention mechanism can obtain the difference of importance of each feature map through specific methods, put more computing resources of the neural network into more critical tasks, and use the task results to reverse guide the weight update of the feature map. In recent years, the attention mechanism has developed rapidly and has been widely used in computer vision. Especially after The Transformer model proposed by the Google team, the self-attention mechanism is more and more favored by many researchers. However, the Transformer model structure is more complex.

In a paper for CVPR (Computer Vision and Pattern Recognition) in 2018, se-Net (Squeeze-and-Congestion Network) is mentioned, and there are many variations. These models focus on the relationship between feature maps (channels) and are simple in structure and easy to be used in other convolutional networks. In this paper, the attention mechanism of the model is used to improve the VGG16 network to extract more valuable feature maps.

The feature extraction network structure of the proposed algorithm is shown in Figure 4. In this algorithm, only the structure before the fifth convolution module of VGG16 network is used, and the SE-Net structure is introduced after the last convolution module. For any given image, after passing through the first five convolution modules of VGG16 network, the feature map with size $N \square N \square C$ is generated. The feature maps firstly aggregates the feature (global pooling) on its spatial dimension $N \square N$ by Squeeze operation $F_{sq}(\square)$ to generate $1 \square 1 \square C$ feature maps -- channel descriptor. The aggregation is followed by the excitation operation. The excitation operation takes the form of several combinations of excitation functions, taking the channel descriptors as input and generating a set of modulation weights for each channel. These weights are applied to generate feature maps with weighted values. These outputs can be fed directly into subsequent layers of the network.

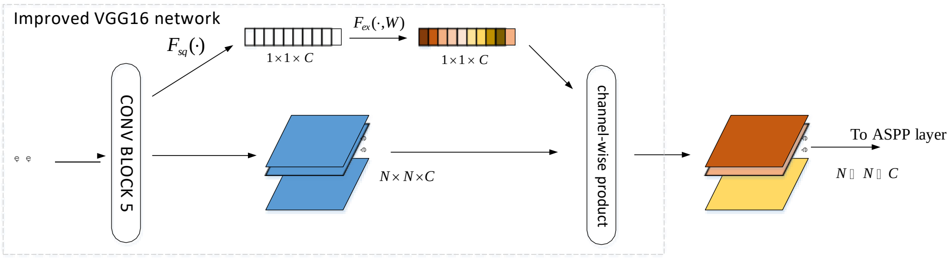


Figure. 4 Network structure diagram of feature extraction based on SRL

Object detection network design

Input the above feature maps into different branches, as shown in figure 3. The first is the MIL network, and the remaining three are interconnected instance classifiers. The MIL network needs to determine target categories and generate target location coordinates. In the classifier, the cross-entropy loss function is improved. Firstly, it is assumed that the features follow the Gaussian mixture distribution, and the conditional probability and prior probability of features are used to calculate the posterior probability of features.

The MIL branch contains two sub-branches. $\mathbf{X}^{cls}(\mathbf{F},\mathbf{W}^{cls}) \square\square^{C\times R}$ and $\mathbf{X}^{dec}(\mathbf{F},\mathbf{W}^{dec}) \square\square^{C\times R}$ are the prediction matrices of the MIL sub-branch. $\mathbf{X}^{cls}$ is the matrix used for classification prediction, and $\mathbf{X}^{dec}$ is the matrix used for positioning prediction. $\mathbf{W}^{cls}$ and $\mathbf{W}^{dec}$ represent the parameters of the two sub-Branch full convolution layers respectively, and C represents the number of target categories. Classification sub-branch predicted classification score matrix by Gaussian mixture distribution model: $[\boldsymbol{\sigma}(\mathbf{X}^{cls})]_{cr} = p(c\,|\,x_{cr}^{cls}) / \square_{c\square\square 1}^{C} p(c'\,|\,x_{c\square r}^{cls})$, where $p(c\,|\,x^{cls}_{cr})$ represents the probability density of category c in the $r\square$ th proposal, and x_{cr}^{cls} is the element of row c and column r in the $\mathbf{X}^{cls}$ matrix. The positioning sub-branch obtains the positioning score prediction matrix: $[\boldsymbol{\sigma}(\mathbf{X}^{dec})]_{cr} \square e^{x_{cr}^{dec}} / \square_{r'=1}^{R} e^{x_{cr'}^{dec}}$ through softmax layer. ($\mathbf{W}^{cls}$, $\mathbf{W}^{dec}$) is represented by $\mathbf{W}^{0}$, and proposal score is $\boldsymbol{\varphi}^{0}(\mathbf{F},\mathbf{W}^{0}) = \boldsymbol{\sigma}(\mathbf{X}^{cls}) \square \boldsymbol{\sigma}(\mathbf{X}^{dec})$. Category c image score is the sum of all proposal scores in this category: $[□(\mathbf{F},\mathbf{W}^0)] = \sum_{r □ 1}^{R} [\boldsymbol{\varphi}^0(\mathbf{F},\mathbf{W}^0)]_{cr}^{c}$, whose value range is (0,1). The given classification label of the image is $\mathbf{y} □ [y_1, \cdots, y_C]^T$. In the following three instance classifiers, each output is $\boldsymbol{\varphi}^k（\mathbf{F},\mathbf{W}^k）$, $\boldsymbol{\varphi}^k\ \mathbf{F},\mathbf{W}^k）=[\boldsymbol{\sigma}(\mathbf{X}^{cls})]_{cr}^{k} = p_k(c \mid x_{cr}^{cls}) / □_{c□ 1}^{C}\ p_k(c' \mid x_{c□ r}^{cls})$, and the highest φ_{cr}^{k-1} corresponding to the label category of $\boldsymbol{\varphi}^{k □ 1}(\ \mathbf{F},\mathbf{W}^{k□1})$ is the pseudo label. (When k=0, supervision of the first refined classifier depends on $\boldsymbol{\varphi}^0（\mathbf{F},\mathbf{W}^0）$ generated by MIL Branch) Literature [7] believes that the mapping between category score using Softmax and probability distribution of sample feature space is not clear and accurate. For example, consider a C categories classification task using Softmax losses. The vector of column r of the feature matrix is represented by x, and its posterior probability $j □ (1, C)$ belonging to the certain class can be represented by the Softmax function (normalized exponential function) of Affinity Score $f(x)$ expressed in Formula (1). In general, the linear transformation function of feature vector x is shown in Formula (2). In practical application, all linear functions of C categories are combined into a linear transformation layer with w_c and b_c as trainable parameters. The higher the value of Affinity score $f_c(x)$ is, the higher the posterior probability of x belonging to class c is.

$$p(j \mid x) □ \frac{e^{f_j(x)}}{\sum_{c□ 1}^{C} e^{f_c(x)}} \qquad （1）$$

$$f_c(x) □ w_c^T x □ b_c, c □ [1, C] \qquad （2）$$

However, $f_c(x)$ cannot be directly used to evaluate the likelihood of x's training feature distribution, because the distribution of training features has not been explicitly stated. Different from softmax loss, we assume that the depth feature x extracted in the training set follows the Gaussian mixture distribution as shown in Formula (3), where $□_c$ and $□_c$ are the mean and covariance of class c in the feature space and are the parameters to be learned by the network. $p(c)$ is the prior probability of class c.

$$p(x) □ \sum_{c□1}^{C} □\ N(x; □_c, □_c)\ p(c) \qquad （3）$$

Under this assumption, the feature vector x_i is the vector of the column r of the classification matrix. The conditional probability distribution of category $z_i □ [1, C]$ is shown in Formula (4). Therefore, the corresponding posterior probability distribution can be expressed as formula (5). Where $□(□)$ represents the probability distribution function of Gaussian mixture distribution.

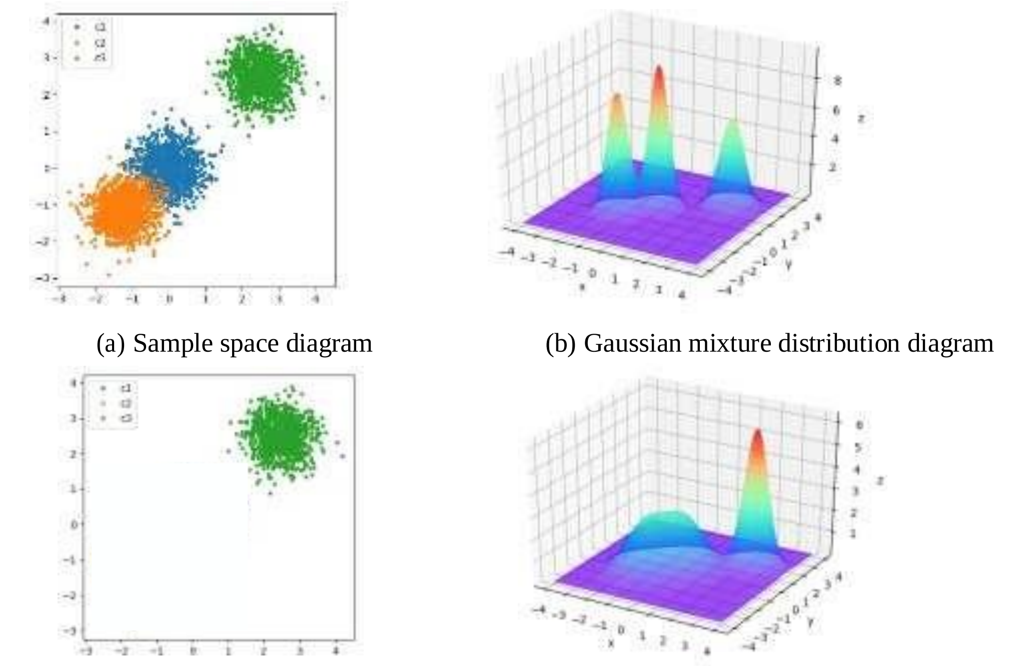
$$p(x_i \mid z_i) □ □ (x_i; □_{z_i},$$

$\square_z$)

（4）

$$p(z_i | x_i) \sqcap \frac{\square(x_i; \square_{z_i}, \square_{z_i})p(z_i}{\sum_{c\square1}^{C}\square(x_i; \square_c, \square_c)p(c)} \quad (5)$$

As shown in Figure. 5, to prove the strong representation of Gaussian Mixture distribution, a schematic diagram of Gaussian mixture distribution is shown below. Figure 5 (a) represents sample space. Suppose the sample space has three categories of samples, namely C1, C2, and C3. Each sample can be represented by a two-dimensional feature vector (x, y). The center point of the feature of the sample of category 1 is at the coordinate (0, 0), the center point of the feature of the sample of category 2 is at the coordinate (-1.2, -1.2), and the center point of the feature of the sample of category 3 is at the coordinate (2.5, 1.5). Gaussian Mixture distribution probability density function is used to obtain the function image, as shown in Figure. 5 (b). The closer the sample is to the mean, the higher the corresponding category probability density value is in the image. It can be seen that sample space and sample category space have strong characterization in Gaussian mixture distribution. Figure. 5 (c) shows the distribution of eigenvalues of a specific type of sample. The distribution of a single type of sample follows the Gaussian distribution. Figure. 5 (d) is the probability density diagram of its mixed Gaussian distribution, indicating the possibility that the sample belongs to the corresponding category. It can be seen that the probability density corresponding to the c3 category is the largest among the three peaks, so it is classified as the C3 category.



(a) Sample space diagram

(b) Gaussian mixture distribution diagram

(c) Single category sample space

(d) Single category Gaussian mixture distribution

Figure 5. sample space and Gaussian mixture model classification results

Therefore, the classification loss L_{cls} can be calculated as the cross entropy between the posterior probability distribution and the class label, as shown in Equation (6). Where, when $z_i \square c$, index function $\square(\ ) \square 1$; Otherwise, 0 is returned.

$$L_{cls} \square \square \frac{1}{N} \square_{i\square1}^{N} \square_{c\square1}^{C} \square(z_i \square c)\log p(c | x_i)$$

$$\square \square \frac{1}{N} \square_{i\square1}^{N} \log \frac{\square(x_i; \square_{z_i}, \square_{z_i})p(z_i)}{\sum_{c\square1}^{C} \square} \quad (6)$$

$(x_i; □_c, □_c)\, p(c)$

In order to classify samples more accurately, distance m is added to L_{cls}, which makes the distance between classes

$□( )$.

increase continuously and makes it easier to distinguish edge samples. The final L-GM loss is shown in Formula □

Where A is equal to 1, otherwise equal to 0.

$$L^{m}_{cls,i} □ □\log \frac{p(z_i)\left|\Sigma_{z_i}\right|^{-1/2} e^{-d_{z_i}-m}}{□_c\, p(c)\left|□_c\right|^{□1/2} e^{□d_c □□(c□z_i)m}} \qquad (7)$$

$$d_c □ (x_i □ □_c)^T □_c^{□1}(x_i □ □_i)/2 \qquad (8)$$

Loss function

The calculation of the overall loss function in this algorithm is shown in Formula (9) :

$$L^k(F,W^k,H^k) □ \begin{cases} □L^m_{cls,i} & \text{if } k=0 \\ □□\_(□\frac{1}{R}\sum_{n□i}^{N} s_n^k M_n^k \log\frac{\sum_{r\ s.t\ b_r□B_n^k}\varphi^k_{y_n r}}{M_n^k}) □ \sum_{r□C^k_{N^k□1}} □_r^k \log □^k_{(C\ 1)r} & \text{if } k □[1,3] \end{cases} \qquad (9)$$

Where, k is the number of branches in the network, s_n^k, M_n^k and φ_{cr}^k respectively represent the cluster confidence score of the n-th target cluster, the number of the proposal in the n-th cluster, and the predicted score of the Proposal in the r-th cluster. In particular, for A, when B (i.e., in the MIL subnetwork), the classification score is represented by the probability distribution density function of Gaussian Mixture distribution, and the traditional the Softmax function is still used in the classifier. This is because the pseudo-label is used in the classifier to complete supervised learning detection. The pseudo-label is not the real distribution of the training sample space, so the traditional the Softmax function is more suitable.

Experiment and evaluation

Operating Environment

Inspired by PCL algorithm, the experiment in this paper is based on TORCH deep learning framework and implemented in Python. All of our experiments ran on NVIDIA RTX and Intel(R) Xeon(R) Silver 4210R CUP (2.40GHz).

PASCAL VOC[26] 2007 dataset contains 9962 images and 20 types of objects. The dataset is divided into *train*, *val* and *test*. VOC2007 training set (5011 images from 2007) was selected to train the network. For the test, mAP and CorLoc were used to evaluate the model. Taking PASCAL VOC datasets as an example, the AP of each class is determined by the Precision and Recall of the class, as shown in Formula (10) and (11). □

$$\text{Precision} □ \frac{TP}{TP □} \qquad (10)$$

$$\text{Recall} □ \frac{FPTP}{TP □ FN} \qquad (11)$$

Where, A represents the number of correctly detected samples, B represents the number of incorrectly detected negative samples as positive samples, and C represents the number of incorrectly detected positive samples as negative samples. AP is shown in Formula (12) :

$$AP □ □_0^1 P(r)dr \qquad (12)$$

Where, r represents Recall, $P(r)$ represents the value of Precision corresponding to r, and AP refers to the integration of Precision on Recall within the interval of (0,1). mAP is the average of AP for each category.

CorLoc [27] is shown in Formula (13). IoU>0.5 between groundtruth boundary box and prediction box in the experimentof this paper.

$$\text{CorLoc} = \frac{TP}{TP + FP} \quad (13)$$

TP and FP have nothing to do with sample category, but only with location.

The improved VGG16 network designed in this paper is added with SENET structure after the last convolution block of the network. In the MIL network part, the GM distribution model is used as the classifier, where the mean and varianceof each category sample feature represented by the Gaussian Mixture model are the parameters to be learned. The category is ultimately determined by the probability density calculated from the mean and variance.

Experimental Results

In order to verify the effectiveness and advancement of the proposed algorithm, it was evaluated in the VOC2007 test set. Due to a large number of targets in the test set, some typical objects are selected to show the test results.

As shown in Figure 6, several representative images were selected in the figure, including four categories, and four samples were selected for each category. The selected samples include small goals, multiple goals, dense goals, and largegoals. The detection effect of the model on targets of different scales can be evaluated.

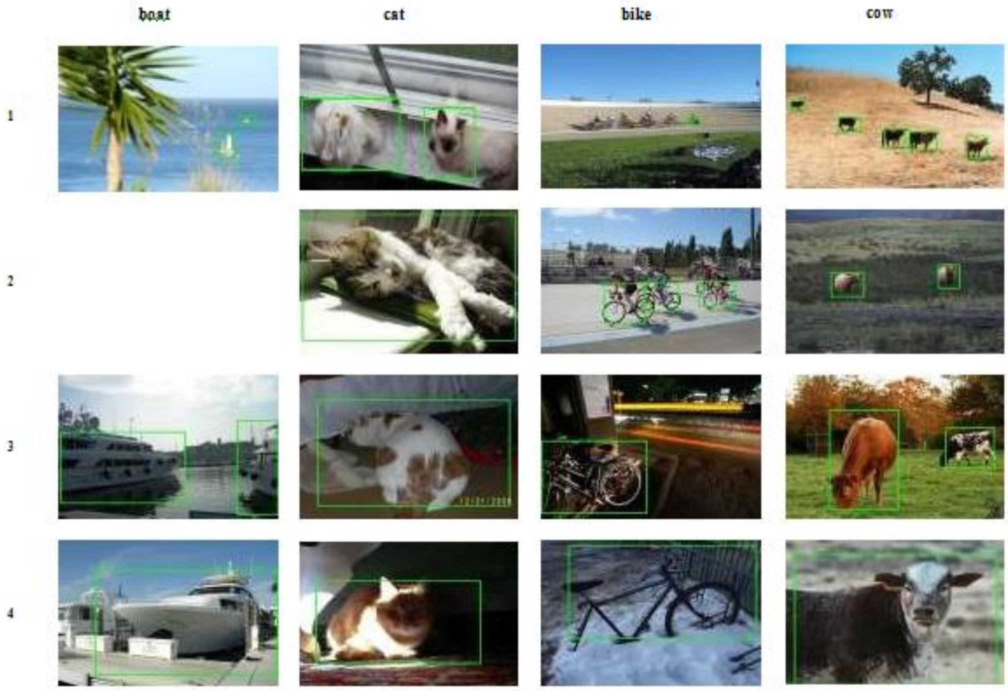


Figure 6 Detection results of the algorithm in this paper

Comparative analysis of experimental results

In this paper, VOC2007 test set is used to compare the more classical algorithms including PCL with the algorithm in this paper [7]. The detection results are shown in Table 2. It can be seen that the detection accuracy (AP) of the algorithm in this paper is better than other algorithms for more than half of the target categories, including *aero*, *bike*, *bird*, *boat*, *car*, *cat*, *chair*, *cow*, *dog*, *mbike*, *plant* and *sheep*. And mAP is also better than other algorithms. Among the 20 categories, *bike* has the highest detection accuracy of 78.0%, and cat has the highest detection accuracy of 14.6%.

Table 2 Detection AP of VOC2007 Test Dataset

Method Classes	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
WSDDN-VGG F	42.9	56.0	32.0	17.6	10.2	61.8	50.2	29.0	3.8	36.2	18.5	31.1	45.8	54.5	10.2	15.4	36.3	45.2	50.1	43.8	34.5
WSDDN-VGG M	43.6	50.4	32.2	26.0	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	37.8	46.9	53.4	47.9	34.9
WSDDN-VGG16	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
PCL-OB-G-Ens.	57.1	67.1	40.9	16.9	18.8	65.1	63.7	45.3	17.0	56.7	48.9	33.2	54.4	68.3	16.8	25.7	45.8	52.2	59.1	62.0	45.8
PCL-OB-G-Ens.+FRCNN	63.2	69.9	47.9	22.6	27.3	71.0	69.1	49.6	12.0	60.1	51.5	37.3	63.3	63.9	15.8	23.6	48.8	55.3	61.2	62.1	48.8
Ours	**67.3**	**78.0**	**55.6**	**40.1**	27.3	68.4	**72.7**	**64.2**	**21.8**	**68.9**	49.8	**47.0**	56.0	**71.5**	14.9	**26.8**	**53.9**	40.0	55.9	62.4	**52.3**

Table 3 is an assessment of CorLoc of the algorithm proposed in this paper on VOC2007 Trainval dataset. The CorLoc of more than half of the categories exceeded typical algorithms, including *aero*, *bike*, *bird*, *boat*, *bottle*, *car*, *cat*, *chair*, *cow*, *mbike*, *plant* and *sheep*. *mbike*'s CorLoc was the highest at 94.4%, while the *boat* category saw the biggest improvement at 19.4 percent. The average CorLoc is 70.3%, which is better than other algorithms.

Table 3 CorLoc of VOC2007 Trainval Dataset

Method Classes	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
WSDDN-VGG F	68.5	67.5	56.7	34.3	32.8	69.9	75.0	45.7	17.1	68.1	30.5	40.6	67.2	82.9	28.8	43.7	71.9	62.0	62.8	58.2	54.2
WSDDN-VGG M	65.1	63.4	59.7	45.9	38.5	69.4	77.0	50.7	30.1	68.8	34.0	37.3	61.0	82.9	25.1	42.9	79.2	59.4	68.2	64.1	56.1
WSDDN-VGG16	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
PCL-OB-G-Ens.	81.7	82.4	63.4	41.0	42.4	79.7	84.2	54.9	23.4	78.8	54.4	46.0	75.9	89.6	22.8	51.3	72.2	66.1	74.9	76.0	63.0
PCL-OB-G-Ens.+FRCNN	83.8	85.1	65.5	43.1	50.8	83.2	85.3	59.3	28.5	82.2	57.4	50.7	85.0	92.0	27.9	54.2	72.2	65.9	77.6	82.1	66.6
Ours	**86.6**	**87.5**	**75.9**	**65.3**	**50.9**	83.2	**87.8**	**74.4**	**50.8**	**87.6**	53.3	**63.6**	75.5	**94.4**	21.2	**57.7**	**89.6**	47.2	72.3	78.7	**70.3**

Figure 7 and Figure 8 show the detection effect of PCL algorithm and the algorithm in this paper. It reflects the detection performance of the two algorithms. This paper still selects four categories of VOC test set, *cat*, *cow*, *boat* and *bike*. In order to be fair, the target boundary box is selected with confidence higher than 40%.

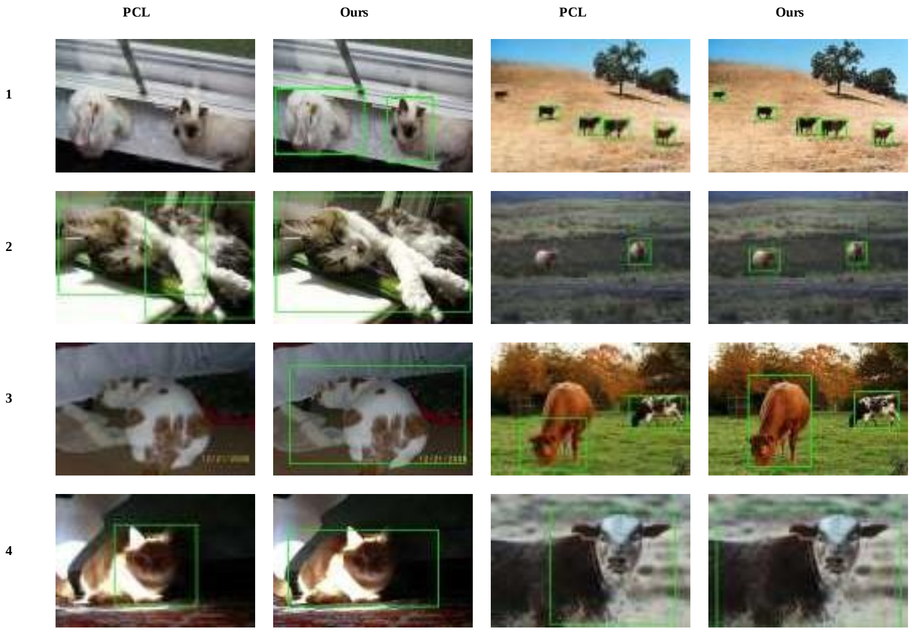


Figure. 7 Comparison between the proposed method and PCL algorithm

Figure 7 shows two algorithms detecting *cat* and *cow* category objects respectively. The first and third columns are the detection of PCL algorithm, and the second and fourth columns are the detection of this algorithm. For *cat* category, we select four images with different object states. When there are multiple objects, the proposed algorithm can correctly distinguish two objects for detection. When a large object appears, the PCL algorithm generates two boundary boxes with

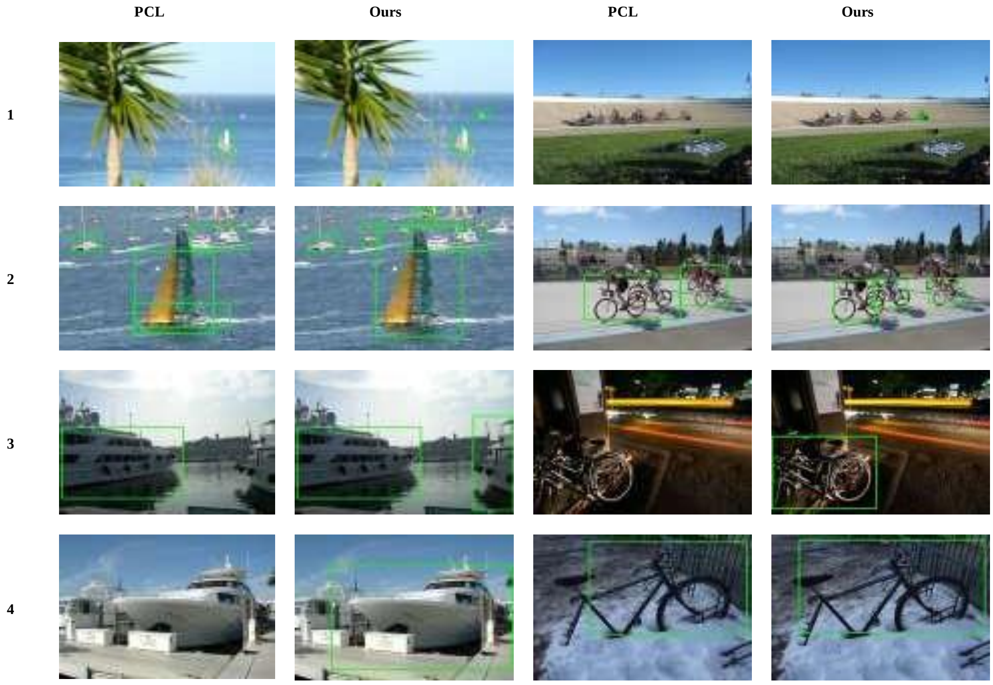


Figure. 8 Comparison between the proposed method and PCL algorithm

For *boat* category, we select several images of different states of the object. When there are small objects, the algorithm can correctly detect small objects. When there are objects with different scales in an image (the second line), the detection performance of PCL algorithm is slightly inferior to that of the algorithm in this paper. It can be seen that the method presented in this paper is more sensitive to small objects and detects more and more accurately. When there is only a part of the object or a small part of the object in the image, the algorithm in this paper can accurately generate boundary boxes, locate and classify them correctly. When a large object appears or the Angle of the object appears changes significantly (line 4), the algorithm in this paper can still generate correct detection results. For *bike* category, mostly presented as dense objects, or dense small objects, PCL algorithm will appear in the missing detection, or misdetection and others. However, the algorithm in this paper can distinguish more different objects. When the light intensity is insufficient, PCL fails to detect the object, but the proposed method can still detect the object correctly (line 3). A special case is given in the figure. In the case of incomplete object (line 4), the algorithm in this paper can locate the object more accurately.

Through the above tables and figures, the performance of the detection algorithm in this paper is systematically and intuitively demonstrated. It can be seen that the index values of 12 categories in VOC2007 dataset of the proposed algorithm, whether mAP or CorLoc indexes, exceed those of other algorithms. The biggest increases were for *boat* and *cat* categories. For the detection of *bike* and *mbike*, two indicators have reached the highest, respectively, *bike* category AP 76%, *mbike* category CorLoc 92.4%. By observing the characteristics of the above four categories, the probability of small objects and multiple objects is relatively high. In this algorithm, a feature extraction network based on strong representation learning is proposed to improve the utilization of feature maps. The extracted features have stronger scale invariance and illumination invariance, which is more conducive to classification and localization. In the classification calculation of objects, this paper proposes to adopt a more representational Gaussian mixture model (GM) and discard the softmax function of the traditional classification algorithm, which makes the model more conducive to accurate classification and improve the performance of object detection.

Conclusion

In this paper, we propose a new weakly supervised object detection algorithm. In this algorithm, a feature extraction network based on strong representation learning is proposed to output more representational feature maps for subsequent detection networks, and the traditional Softmax cross-entropy loss is abandoned in the classifier of detection networks. However, the probability density of a feature belonging to a specific category is calculated by GM. The parameter mean and variance of the Gaussian Mixture model need to be learned. After the above improvements, the performance of weakly supervised target detection has been significantly improved. We believe that the strong representation learning algorithm proposed in this paper can be applied to other fields of computer vision. In the later research, we will focus on this field and more innovative weakly supervised object detection algorithms.

Acknowledgement: This research is supported by National Natural Science Foundation of China, grand number62006240.

Reference

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, et al. You Only Look Once: Unified, Real-Time Object Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 27-30 June, pp.2016:779-788. New York: IEEE

[2] Joseph Redmon, Ali Farhadi. YOLO9000: Better, Faster, Stronger. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Honolulu, HI, USA 21-26 July,2017,.pp.6517-6525. New York: IEEE

[3] Ross Girshick. Jeff Donahue, Trevor Darrell, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA. 23-28 June, 2014, pp.580-587. New York: IEEE,

[4] Ross Girshick. Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV), Santiago, Chile,

[5] December 7-13, 2015, pp.1440-1448. New York: IEEE

[6] Shaoqing Ren, Kaiming He, Ross Girshick, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137-1149.

[7] Weitao Wan; Yuanyi Zhong; Tianpeng Li; Jiansheng Chen. Rethinking Feature Distribution for Loss Functions in Image Classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18-23 June 2018. Salt Lake City, UT, USA.

[8] Dong Li; Jia-Bin Huang; Yali Li; Shengjin Wang; Ming-Hsuan Yang. Weakly Supervised Object Localization with Progressive Domain Adaptation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 27-30 June 2016, Las Vegas, NV, USA

[9] AJ Bency, H Kwon, H Lee, S Karthikeyan, BS Manjunath, Weakly Supervised Localization using Deep Feature Maps. European Conference on Computer Vision. 17 September 2016.

[10] Hakan Bilen; Andrea Vedaldi. Weakly Supervised Deep Detection Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 27-30 June 2016, Las Vegas, NV, USA

[11] Ke Yang; Dongsheng Li; Yong Dou. Towards Precise End-to-end Weakly Supervised Object Detection Network. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 27 Oct.-2 Nov. 2019. Seoul, Korea (South).

[12] Peng Tang; Xinggang Wang; Xiang Bai; Wenyu Liu. Multiple Instance Detection Network with Online Instance Classifier Refinement. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 21-26 July 2017. Honolulu, HI, USA.

[13] Ze Chen, Zhihang Fu, Rongxin Jiang, Yaowu Chen, Xian-sheng Hua. SLV: Spatial Likelihood Voting for Weakly Supervised Object Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2020.

[14] Gong Cheng; Junyu Yang; Decheng Gao; Lei Guo; Junwei Han. High-Quality Proposals for Weakly Supervised Object Detection. IEEE Transactions on Image Processing. 2020, 29: 5794 - 5804

[15] Peng Tang; Xinggang Wang; Song Bai; Wei Shen; Xiang Bai; Wenyu Liu. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020, 42(1): 176 191.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu . Spatial Transformer Networks. Advances in Neural Information Processing Systems. December 7-12, 2015. Montreal, Canada.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. December 2017 Pages 6000–6010

[18] Jie Hu; Li Shen; Gang Sun. Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision andPattern Recognition. 18-23 June 2018. Salt Lake City, UT, USA

[19] Li H , Xiong P , An J. Pyramid Attention Network for Semantic Segmentation. https://arxiv.org/abs/1805.10180v3. 2018.

[20] Ketkar N., Moolayil J. (2021) Recurrent Neural Networks. In: Deep Learning with Python. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-5364-9_7.

[21] Sanghyun Woo, Jongchan Park, Joon-Young Lee, CBAM: Convolutional Block Attention Module. European Conference on Computer Vision. 06 October 2018

[22] J Fu, J Liu, H Tian, Y Li, Y Bao, Z Fang, H Lu. Dual Attention Network for Scene Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 15-20 June 2019. Long Beach, CA, USA

[23] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, Selective search for object recognition. International Journal of Computer Vision, vol. 104, no. 2, pp. 154–171, 2013.

[24] C. L. Zitnick and P. Doll´ar, Edge boxes: Locating object proposals from edges. European Conference on ComputerVision, 2014, pp. 391–405.

[25] K. He, X. Zhang, S. Ren, and J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, 2015

[26] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, vol. 111, no. 1, pp. 98–136, 2015.

[27] T. Deselaers, B. Alexe, and V. Ferrari, Weakly supervised localization and learning with generic knowledge. International Journal of Computer Vision, vol. 100, no. 3, pp. 275–293, 2012.