International Core Journal of Engineering Volume 7 Issue 6, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

Multi-scale Feature Learning based Keypoints Mapping for Object Detection

Zhiyong Zhang[1,a], Yongsheng Dong[1], Boshi Zheng[1] and Hong Liu[1]

[1]The School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China.

[a]zhangzhiyong18@163.com

Abstract

We propose a simple, flexible, and common framework for object detection. Our method can effectively detect small objects in an image while generating friendly bounding boxes. This method is called "Multi-scale feature learning based keypoints mapping for object detection (MFLKM)". MFLKM has four main contributions: 1. Using the method of key-point range mapping, balancing positive and negative samples and improving the performance of the model; 2. Expanding the loss of center point offset, excavating difficult samples, improving the model optimization ability; 3. Width and high scale mapping, speeding up model convergence; 4. MFLKM performing better than other representative algorithms on multiple data sets (PAS VOCALC and Microsoft COCO). In addition, MFLKM is easy to train and the inference phase requires no additional steps to achieve real-time detection speeds (46 frames per second).

Keywords

Object Detection; Multi-scale; Keypoints Mapping; Convolutional Neural Network.

1. Introduction

The object detector based on convolutional neural network (CNN) [1-4] has achieved the latest results in a variety of challenging benchmark tests. A common component of state-of-the-art methods is anchor boxes, which are of various sizes and aspect ratios and can be used as detection candidates. Anchor boxes are widely used in one-stage detector [5, 6] and this method can obtain highly competitive results with two-stage detector, which is very effective. The one-stage detector designs dense anchor boxes on the image, and generates the final prediction frame by scoring the anchor boxes and refining its coordinates.

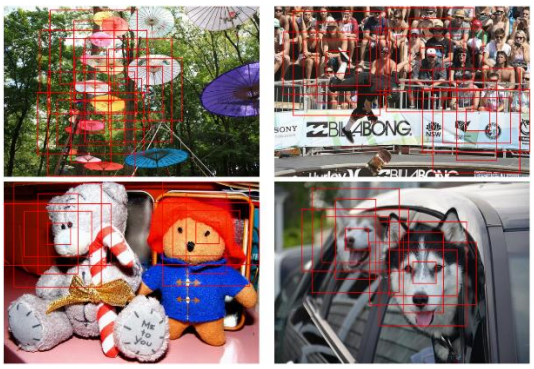

Figure 1. A large number of anchor boxes are designed manually for one-stage detector, resulting in the imbalance of positive and negative samples.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

Figure 1 shows anchor boxes, it can be seen from Figure 1 that there are two disadvantages in using anchor boxes: first, this method usually needs to design a large number of anchor boxes with different aspect ratios. For example, the number of anchor boxes in deconvolutional single shot detector (DSSD) [7] is more than 40K, while the data in RetinaNet [8] is more than 100K. This is because a large number of anchor boxes are required to ensure full overlap with most ground truth. Therefore, a small part of the anchor boxes will overlap with the ground truth, which will cause a huge imbalance between the positive and negative samples and slow down the training speed. Secondly, the use of anchor boxes introduces many super parameters and design options. Such choices are largely made through temporary heuristics, which may become more complex when combined with multi-scale architectures.

In order to overcome the disadvantages of anchor boxes, we propose a multi-scale feature learning method based keypoints mapping for object detection (MFLKM). Our method represents each object by the central key point and aspect ratio, which bypasses the need of anchor box and achieves the latest level of object detection accuracy. We use a single convolution network to predict the center point heat map, center point offset and aspect ratio of the different object. Finally, the accurate bounding box is generated by fusing multi-scale features. This method greatly simplifies the output of the network and eliminates the need to design the anchor box. MFLKM explores the central region of each object and has the ability to perceive the visual pattern of the central region of each object, so that it can correctly classify each bounding box. We notice that even if the center point is offset, the bounding box can still have a high intersection over union (IoU) with the ground truth. Therefore, we define all the points within a certain range of the central key point as positive samples according to the size of the object. This method can effectively detect objects of different scales, balance positive and negative samples, mine difficult samples, and improve the ability of model optimization.

In order to detect the key points and bounding box better, we propose three strategies to better locate the center and fine tune the bounding box. The first strategy is called range mapping, which is used to predict central keys. The center key can perceive more recognizable visual patterns in the object, thus it is easier to obtain higher classification confidence. We achieve this goal by large object large range mapping, small object small range mapping. The second strategy is called strong optimization of center point offset, which can expand the loss, excavate difficult samples and improve the detection accuracy of small objects. We can achieve this by finding the square root of the offset on the feature map. The third strategy is called the aspect ratio mapping strategy, which can improve the ability of feature extraction and accelerate the convergence of the model. We can achieve this goal by normalizing the width and height of different objects on the feature map. The experiment shows that these three strategies are more robust to the noise at the feature level, which is helpful to the detection of small objects, and improves the accuracy and recall rate.

The main contributions of this paper are as follows:

(1). We propose the strategy of large objects large range mapping, small objects small range mapping. The strategy can perceive more recognizable visual patterns in the object, thus it is easier to obtain higher classification confidence.

(2). We propose the strategy of strong optimization of center point offset. The strategy can expand the loss, excavate difficult samples and improve the detection accuracy of small objects.

(3). We propose the strategy of the aspect ratio mapping. The strategy can improve the ability of feature extraction and accelerate the convergence of the model.

(4). Experimental results reveal that our proposed MFLKM has satisfactory object detection performance when compared with nine representative object detection methods. MFLKM gains state-of-the-art detection accuracy in multiple categories while maintaining real-time detection on PASCAL VOC and Microsoft COCO datasets.

The rest of this paper is organized as follows. Sec. 2 introduces the recent object detection methods and analyzes their advantages and disadvantages. Sec. 3 introduce the proposed three strategies and

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

the optimize details. Sec. 4 describe the experimental results, compare the differences between the proposed method and the current method. And conclusion is made in Sec. 5.

2. Related Work

Object detection includes locating and classifying objects. Object detection methods can be roughly divided into two main types: the traditional methods based on manually designed features [9-11] and the methods based on deep convolution neural network. There are two methods based on deep convolution neural network: two-stage methods [2-4] and one-stage methods [5-7].

2.1 Traditional Methods

In order to recognize different objects, traditional object detection algorithms need to extract visual features that can provide semantic and robust representation. Histogram of oriented gradient (HOG) [10] feature is a kind of feature descriptor used for object detection in computer vision and image processing. It constructs features by calculating and counting the histogram of the gradient direction of the local region of the image. In addition, the deformable part model (DPM) [11] is a flexible model, which adopts the strategy of "divide and rule", and combines the object parts with deformation cost to deal with severe deformation. But the traditional methods of object detection has the following disadvantages [12-14]: (1). Each stage is carried out separately, it is difficult to optimize; (2). The feature extracted manually is not robust and generalization is weak, it can not effectively transfer learning.

2.2 Two-stage Methods

The two-stage object detection task is divided into two stages: extracting region of interest (RoI), then classifying and regressing RoIs. Region-cnn (R-CNN) [2] uses selective search [15] method to extract RoIs, and uses CNN-based classifier to classify RoIs independently. Fast R-CNN [3] uses a single CNN to extract RoIs to improve R-CNN. Faster R-CNN [4] generates ROIs by introducing regional proposal network (RPN). The anchor frame used in RPN is widely used in later object detection tasks. On the basis of Faster R-CNN, Mask R-CNN [16] adds a branch to predict the mask, which can detect the object and predict its mask at the same time. Cascade R-CNN [17] trains the detector by increasing the IoU threshold in order to solve the problems of over fitting and low data quality. Other object detection algorithms put forward meaningful work for different problems. For example, [18] focuses on context relations, [19] focuses on multi-scale objects, [20] focuses on architecture design.

2.3 One-stage Methods

The one-stage method does not extract RoIs, but directly classifies and regresses prior anchor boxes. YOLO [5] transforms the object detection into the problem of returning the bounding box and the category probability, and realizes the end-to-end training. SSD [6] uses more intensive prior anchor frames to regress and classify multi-scale objects. DSSD [7] enters the deconvolution module and combines the feature maps of different depths to detect the objects. RetinaNet [8] proposes to focal loss, mine difficult samples and solve the problem of sample imbalance. RefineDet [21] adjusts the position and size of the prior anchor boxes twice, which combines the advantages of the two methods. CornerNet [22] uses the object's vertex to detect the objects, and proposes a novel corner pool. CenterNet [23] is another method of using key points to detect objects. Different from CornerNet, it uses top left, bottom right and center points.

Feature pyramids are widely used in one-stage and two-stage object detectors, which can effectively detect multi-scale objects [24-30]. The layers of the network are deep enough to extract high-level semantic information [31-36]. The full convolution network can reduce the parameters without reducing the accuracy [37-40]. [41-43] improves the accuracy of the model by fusing different levels of features, removing noise and adding constraints. In addition, a large number of experiments show that the residual network can still maintain good performance in extremely deep network [44-49]. But the existing methods based on anchor boxes and key points lack the ability to highlight the object

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

boundary clearly, and can only achieve partial context reliability. Therefore, we will introduce our approach in the next section.

3. Our Proposed Method

In this chapter, we propose a multi-scale feature learning based keypoints mapping for object detection (MFLKM). In this method, a single convolutional neural network is used to map the object boundary box to the central key point and aspect ratio. By regressing the object detection to the key point, it is unnecessary to design a set of anchor boxes which are usually used for existing one-stage detectors. In addition to our novel network structure and loss function, we use an adaptive method to map all the points within a certain range of the object center to positive sample points, which can effectively balance the positive and negative samples and accelerate the convergence speed of the model.

3.1 Architecture

In MFLKM, we take the object detection as the central key point and aspect ratio. In order to better detect the center key point and aspect ratio, we propose a key point mapping strategy. In this strategy, the points near the geometric center of the object are weighted according to the distance, because when the center point has a small offset, the prediction box still has a high IoU with ground truth. This can balance the positive and negative samples and accelerate the convergence of the model. In addition, we map the width and height of the object to the proportion of the image, which is similar to the normalization processing, which can speed up the training speed and reduce the training deviation caused by different scale objects. However, if the key point mapping is used directly, the center point of the small object will be deviated too much, resulting in low recall rate and inaccurate bounding box. Therefore, we introduce the center point offset loss to optimize.

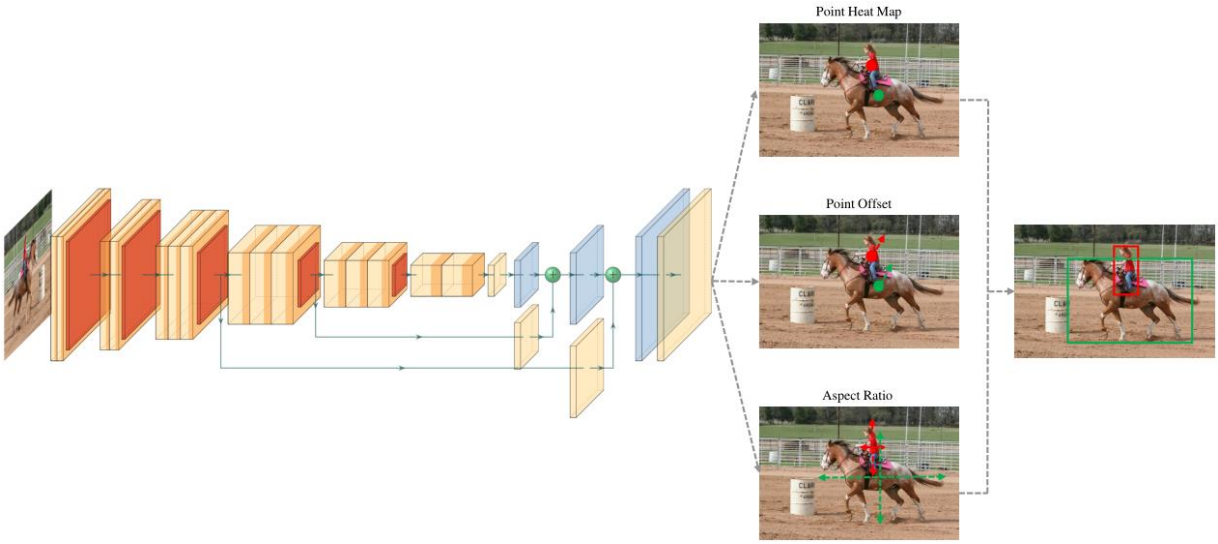

Figure 2. The structure of MFLKM. The backbone network outputs object center point map, center point offset and aspect ratio.

The whole network architecture is shown in Figure 2. It can be seen from Figure 1 that we use a central key point and aspect ratio to represent each object. Specifically, we map the central key to a heat map and predict the offset of the central key point. Then, we use the predicted value of aspect ratio to generate the predicted bounding box. In order to remove the redundant prediction box, we take the following operations: first, select the top k key points with high weight according to the model output, then map the k key points back to the original image according to the weight, and finally check whether the object center area of the original image contains these key points, if any, keep the boundary box, otherwise delete it.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

MFLKM uses residual network [50] as the backbone network, and uses a structure similar to hourglass in [51]. The hourglass network is followed by three prediction modules, one for the center point, one for the offset of the center point, and the other for the aspect ratio. The center point module has C channels, where C is the number of categories and the size is H*W. There are no background channels, and each channel is a binary mask. The size of the center point offset module is H*W, and the number of channels is 2, which is used to output the center point offset. The module size of aspect ratio is H*W, and the number of channels is 2, which is used to output the ratio of width and height of the object.

For the center prediction module, there is only one positive sample and the others are negative samples. In the training process, we do not punish these negative samples, but reduce the punishment of negative samples within the radius of positive samples. Because less offset can still generate a bounding box with higher IoU than ground truth. For the center point offset prediction module, the image center point will use rounding operation in the down sampling. When we remap some positions from the heat map to the input image, the center point of the image will deviate, which may seriously affect the IOU of small size bounding box. To solve this problem, we need to reduce the prediction error. For the aspect ratio prediction module, we use the ratio of the object aspect ratio to the image to optimize the loss. In this way, the width and height of the object can be normalized to (0, 1), which can effectively reduce the error caused by the size of the object. At the same time, it can also speed up the training speed of the whole model.

3.1.1 Hourglass Module

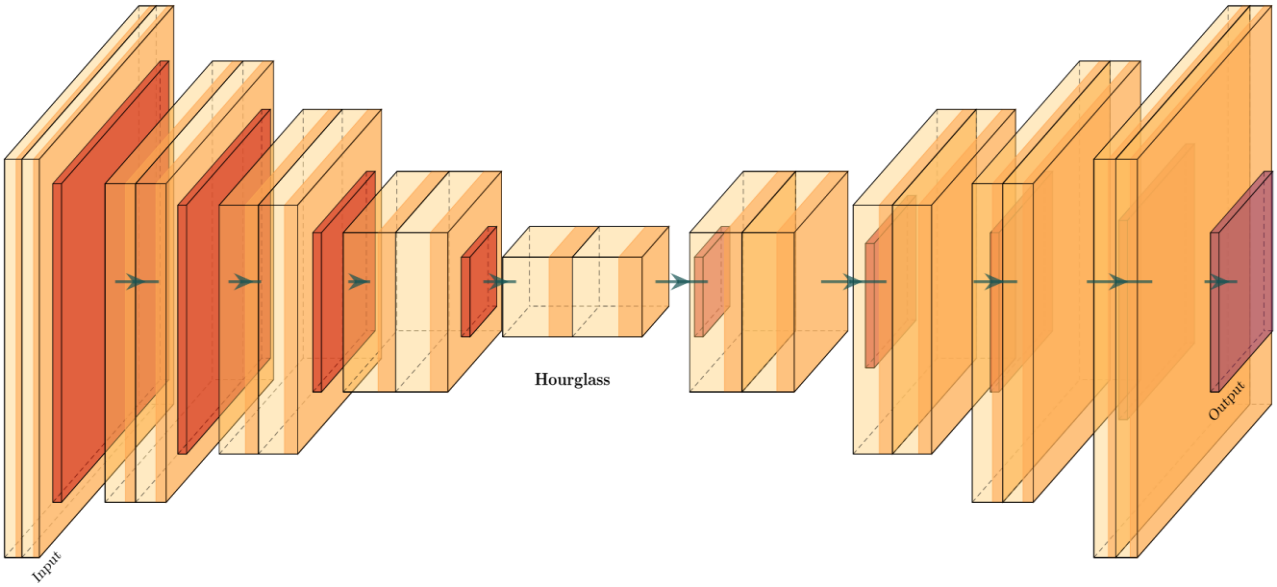

Figure 3. Hourglass module. It is a full convolution neural network composed of one or more hourglass modules.

MFLKM uses hourglass structure [51]. As shown in Figure 3. It can be seen from Figure 3 that it is a full convolution neural network composed of one or more hourglass modules. The hourglass module first downsampling the input image through a series of convolution and pooling layers. Then, the feature is resize to the original resolution through a series of upsampling and convolution layers. Since the details are lost in the pooling layer, skip connection are used to bring the details back to the upsampled feature map. Hourglass module can capture global and local features at the same time. When multiple hourglass modules are stacked in the model, the hourglass module can reprocess features to capture higher-level information. The hourglass features also make these modules ideal for detection. In addition, we use multiple 1*1 convolution kernels in the network, which can effectively reduce the number of parameters.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

3.2 Training

We use tensorflow framework [52] to implement our method. In the training, we use mosaic data enhancement technology used in YOLO [5] to expand the scale of the data set, ensure the quality of the data set, and improve the robustness and generalization of the model feature extraction. We use a smaller convolution kernel, which will increase the training time to a certain extent, but can improve the ability of small object feature extraction.

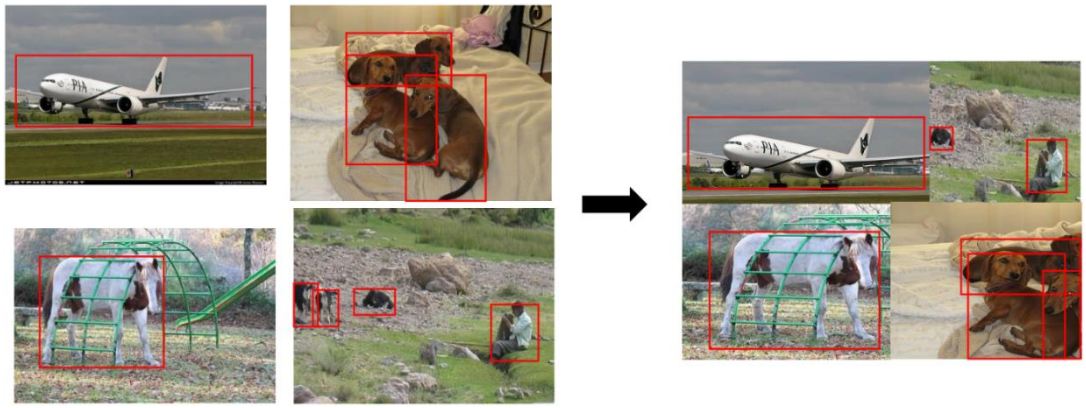

Figure 4. Mosaic data enhancement.

Figure 4 shows the mosaic data enhancement technique. It can be seen from Figure 4 that mosaic data enhancement is used to twist, clip, zoom, flip and adjust the color gamut of four images, and then the processed four images and frames are spliced. This greatly enriches the background information in object detection, and makes the model have more data to train. This can effectively avoid the problem of over fitting.

3.2.1 Key Point Mapping

First, we map the central key point to the heat map. In the hear map, the closer to the central key point, the greater the weight. The attenuation formula of weight is as follows:

$$Y_{x,y,j} = \exp\left(-\frac{(x-\hat{x})^2+(y-\hat{y})^2}{2\sigma^2}\right) \quad (1)$$

where, x is the abscissa of the central key point on the heat map, y is the ordinate of the central key point on the heat map, $\hat{x}$ and $\hat{y}$ are the abscissa and ordinate near the central key point, σ is the super parameter, we set it to 1/4. Through formula (1), we map the central key point of the object to the heat map $Y \in [0,1]^{\frac{w}{R}*\frac{h}{R}*C}$, where R is the multiple of down sampling.

Figure 5. Heat map. The larger scale object has a larger central area, while the smaller scale object has a smaller central area.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

Figure 5 shows the heat map with the original image mapped as the central key point. It can be seen from Figure 5 that the larger scale object has a larger central area, while the smaller scale object has a smaller central area. Under the same deviation value, a larger object can still have a higher IoU with ground truth, which is acceptable. However, due to the inaccurate prediction box and low confidence, the small-scale object has a low recall rate, which further affects the accuracy.

Then, we use the following formula to optimize our central key point:

$$Loss_{point} = \frac{1}{N}\sum_{x,y,c}\begin{cases}(1-\hat{Y}_{x,y,c})^{\alpha}\log(\hat{Y}_{x,y,c}) & \text{if } Y_{x,y,c}=1\\ (1-Y_{x,y,c})^{\beta}(\hat{Y}_{x,y,c})^{\alpha}\log(1-\hat{Y}_{x,y,c}) & \text{otherwise,}\end{cases} \quad (2)$$

where N is the number of key points in the image center, α and β are super parameters, which are set to 2 and 4 respectively.

Because of convolution calculation, there will be deviation in the process of image center key point down sampling, especially for small-size objects, so we also design the center key point offset function for optimization, The formula is as follows:

$$Loss_{offset} = \frac{1}{N}\sum\sqrt{\left|\frac{x}{R}-\hat{x}\right|}+\sqrt{\left|\frac{y}{R}-\hat{y}\right|} \quad (3)$$

Here, we use the square root, which can improve the error, mining difficult samples, and make the model achieve better optimization effect.

3.2.2 Multi-scale Learning

Small objects detection is always one of the challenging problems in object detection. The main reasons are as follows: 1. Small scale objects are usually fuzzy and can not extract effective information; 2. Convolution operation causes further loss of boundary information of small objects, resulting in inaccurate prediction box; 3. Small scale objects are usually clustered, and nearby small objects will also block each other, resulting in the failure to extract useful features. In order to improve the accuracy of small object detection, we propose a multi-scale feature learning method. We normalize the width and height of the objects.

The loss function of width and height ratio is as follows:

$$Loss_{size} = \frac{1}{N}\sum_{k}^{N}|\hat{p}_k - t_k| \quad (4)$$

where, N is the number of key points in the image center, $\hat{p}_k$ is the proportion of predicted width and height, and t_k is the proportion of width and height of the detected object to the image.

The total loss function is the weighted sum of three part losses:

$$\text{Loss} = Loss_{point} + \gamma Loss_{offset} + \lambda Loss_{size} \quad (5)$$

where, γ and λ are super parameters, which are 0.9 and 1, respectively.

3.2.3 Evaluation Index

We evaluate our approach using different evaluation criteria in PASCAL VOC [53] and Microsoft COCO [54] datasets. For Microsoft COCO data set, we use the more rigorous evaluation method proposed by Microsoft coco to calculate recall rate and accuracy rate of objects at different scales. Here, the object less than 32^2 pixels is defined as a small object, the object between 32^2 and 96^2 pixels is defined as a medium object, and the object larger than 96^2 pixels is defined as a large object.

Microsoft COCO data set adopts more refined evaluation criteria, using four major items: average precision (AP), AP across scales, average recall (AR), AR across scales, a total of 12 sub items: AP (refers to the average value of 10 calculation results after every 0.05 change of IoU from 0.5 to 0.95), $AP^{IoU=0.5}$ (the measured AP value when IoU = 0.5), $AP^{IoU=0.75}$ (the AP value measured when IOU = 0.75), AP^{small} (the AP value measured when the size is less than 32^2), AP^{medium} (the AP value measured when the size is more than 32^2 and less than 96^2), AP^{large} (the AP value measured when the size is more than 96^2), $AR^{max=1}$ (the maximum of the detection results when the number of detection is 1 in a picture), $AR^{max=10}$ (in a picture, the maximum retrieval of the detection results when the given

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

number of detection is 10), $AR^{max=100}$ (in a picture, the maximum retrieval of the detection results when the given number of detection is 100), AR^{small} (the AR value measured for the object whose size is less than 32^2), AR^{medium} (the AR value measured for the object whose size is more than 32^2 but less than 96^2), and AR^{large} (AR values measured for object larger than 96^2).

4. Experimental Result and Analysis

The proposed method is analyzed for ablation and compared with other representative methods in average accuracy and vision, including Faster R-CNN [4], Mask R-CNN [16], Cascade R-CNN [17], YOLO [5], SSD [6], DSSD [7], RetiaNet [8], CornerNet [22], CenterNet [23]. We evaluate the AP for different types of objects on PASCAL VOC data set. We evaluate and average AP and AR at different scale and thresholds on Microsoft COCO data set.

4.1 Ablation Analysis

In the width and height prediction module, we use the normalization method to get the ratio of the width and height of multiple groups of detected objects relative to the whole image. In order to evaluate the relationship between different down sampling times and detection speed and accuracy, we selected 2, 4, 8 and 16 times down sampling for ablation experiments, and the results are shown in Table 1.

Table 1. Ablation Analysis

Times	Size	TrainData	TestData	AP	FPS
16	512	COCO	COCO	39.4	46
8	512	COCO	COCO	41.1	41
4	512	COCO	COCO	42.9	39
2	512	COCO	COCO	43.0	18

Table 1 shows the relationship between different down sampling times and detection speed and accuracy. It can be seen from Table 1 that the average accuracy of the model increases with the decrease of the lower sampling multiple. When the current sampling multiple is 4, the lower sampling multiple is reduced, and the effect of improving the accuracy of the model is not obvious. The lower sampling multiple is reduced from 4 times to 2 times, and the accuracy is only improved by less than one percentage point. However, the detection speed has shown a significant regression, from 39 frames per second to 18 frames per second. If the input size of the image increases, it will not meet the real-time detection. Therefore, we choose the lower sampling multiple of 4 as the final down sampling multiple of the model. This allows for a balance between average accuracy and frame rate.

4.2 Average Accuracy

Table 2. PASCAL VOC 2012 test detection result. Our method has the highest detection accuracy in many categories.

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster	70.1	84.3	78.0	**74.6**	54.5	48.0	**77.1**	75.4	87.8	44.5	77.0	55.7	86.9	81.4	80.0	79.3	40.6	72.5	61.0	81.1	61.3
Mask	68.8	82.7	79.4	71.5	52.0	53.1	73.8	69.5	85.1	46.9	74.5	52.7	85.4	81.1	79.7	72.7	39.5	72.1	59.8	77.0	**68.3**
Cascade	70.1	**84.9**	79.4	73.6	53.9	50.2	77.3	75.5	87.9	44.9	76.3	54.4	**87.3**	81.8	81.4	78.1	40.5	**73.0**	59.8	79.7	62.1
YOLO	57.8	76.8	67.4	56.7	38.2	23.2	68.5	56.0	81.6	35.9	60.3	48.2	76.9	71.8	71.1	63.2	28.1	53.1	53.8	74.1	51.2
SSD	66.2	71.2	75.2	62.3	57.0	36.1	73.3	77.9	77.1	45.3	66.7	**68.1**	74.1	80.0	75.9	72.3	40.1	62.2	67.0	75.9	66.2
DSSD	67.8	83.0	78.1	70.1	57.1	40.1	75.1	66.1	85.6	**50.1**	72.1	55.9	85.7	77.9	79.1	76.5	40.2	67.5	63.2	73.2	61.8
CornerNet	69.7	83.9	75.1	73.1	56.1	43.1	71.1	73.4	81.5	49.1	**81.0**	57.3	82.1	81.3	81.1	74.1	42.5	70.7	**69.3**	81.1	66.5
CenterNet	70.7	83.2	77.6	73.5	55.7	**54.2**	78.3	80.1	86.5	48.7	74.1	52	84.1	81.5	76.5	**82.1**	**47.8**	72.5	59.5	79.3	65.8
RetinaNet	69.9	83.0	79.4	71.6	51.9	51.1	73.2	72.1	85.6	48.3	73.4	57.8	86.1	80.0	80.7	70.4	46.6	69.6	67.8	75.9	74.1
Ours	**71.2**	***84.5***	78.7	***73.7***	55.5	***53.7***	***78.6***	**80.1**	**88.9**	***49.5***	75.1	54.1	86.4	**83.1**	**82.1**	75.2	40.6	70.3	67.5	**81.2**	65.4

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

The average accuracy of the proposed method is compared with that of the existing representative methods. Table 2 and Table 3 show the average accuracy of the ten methods. It can be seen from Table 2 that our method has the highest detection accuracy for cars, cats, horses, motorcycles and trains, and the accuracy is more than 80%. Especially for cats, the detection accuracy is 88.9%. These object categories with the highest detection are all variable in shape, with multiple scales, and some of them are occluded with each other. This further proves the effectiveness of our method. In addition, the detection accuracy of our method for aircraft, bird, cup, bus and chair is only better than the best detection results, and the difference is within one percentage point. These differences are mainly related to the two-stage detector. This is because the proposed region is extracted by the two-stage detector.

The accuracy of Faster R-CNN [4], Mask R-CNN [16] and Cascade R-CNN [17] is similar, which is closely related to their use of convolution network to extract the proposal region. The low accuracy of the one-stage detectors, such as YOLO [5], SSD [6] and DSSD [7], it is due to the pursuit of too fast detection speed. RetinaNet [8] has the same accuracy as the two-stage detector, which is due to its focal loss and difficult sample mining, which solves the problem of positive and negative sample imbalance to a certain extent. CornerNet and CenterNet do not need to design the anchor, which is widely used in the one-stage detector, and the combination of multiple modules, which enhances the performance of the model and achieves high detection accuracy. Our method belongs to one-stage detector, but it has more performance than two-stage detector in accuracy, which is closely related to our central key point mapping and multi-scale feature learning.

Table 3. Microsoft COCO test detection result. Our method has the highest detection accuracy for different scale objects.

	AP	AP^{50}	AP^{75}	AP^{s}	AP^{m}	AP^{l}	AR^{1}	AR^{10}	AR^{100}	AR^{s}	AR^{m}	AR^{l}
Faster	33.3	55.4	35.4	12.2	36.3	50.5	31.3	49.0	51.8	27.2	56.4	71.2
Mask	39.7	61.5	43.2	21.0	41.4	50.9	-	-	-	-	-	-
Cascade	42.7	60.7	45.2	22.5	43.9	55.1	-	-	-	-	-	-
YOLO	22.5	42.8	18.2	4.8	24.1	36.6	20.2	29.6	31.6	8.0	34.8	53.0
SSD	30.7	49.4	31.4	10.2	33.1	47.9	27.3	40.5	43.1	17.4	47.6	64.5
DSSD	32.4	52.4	33.5	11.7	33.7	50.8	28.4	42.6	45.1	20.6	48.0	64.8
CornetNet	37.6	54.3	44.1	21.0	43.2	52.5	33.9	55.4	59.7	38.0	61.9	75.8
CenterNet	41.3	58.6	43.1	21.8	43.0	52.5	34.6	54.6	59.1	37.0	62.1	75.9
RetinaNet	39.1	57.2	42.2	21.5	42.7	48.2	-	-	-	-	-	-
Ours	**42.9**	**61.7**	**45.7**	**23.3**	**44.2**	**55.3**	**35.4**	**56.0**	**60.2**	**38.1**	**62.6**	**77.1**

Table 3 shows Microsoft COCO test detection result. It can be seen from Table 3 that our method has the highest detection accuracy in all indicators. For small objects, our detection accuracy reaches 23.3%, which is 0.8% higher than other best methods. Because we use multi-scale feature learning method to improve the weight of center point offset and aspect ratio optimization. At the same time, our method has the highest recall rates for all scale objects, which are 38.1%, 62.6% and 77.1% respectively. This means that our method has the lowest classification error rate. This further proves that our method improves the detection accuracy of small objects to a certain extent, and proves that our method is feasible. In addition, our method has the highest recall rate when the maximum number

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

of recalls is 1, 10 and 100, which also means the highest classification accuracy. Under the commonly used threshold of IoU = 0.5 and IoU = 0.75, the accuracy of our method is 61.7% and 45.7% respectively. Compared with other methods, the accuracy of our method is the slowest when the threshold is raised. This proves that the features extracted from our model have higher robustness and generalization.

We also notice that the detection accuracy of one-stage detectors YOLO, SSD and DSSD for small objects is 4.8%, 10.2% and 11.7% respectively, which obviously can not meet the requirements. This is also a common deficiency of one-stage detectors. The detection accuracy of RetinaNet for small object detection reaches 21.5%, which is due to the focal loss, the lack of difficult samples and the improvement of the loss of small object detection, which is very effective. Faster R-CNN, Mask R-CNN and Cascade R-CNN, two-stage objext detection algorithms, generally have higher detection accuracy than one-stage object detection algorithms because of the extraction of proposal regions. Neither CornerNet nor CenterNet uses anchor, which combines multiple modules and achieves the detection accuracy higher than that of the secondary detector.

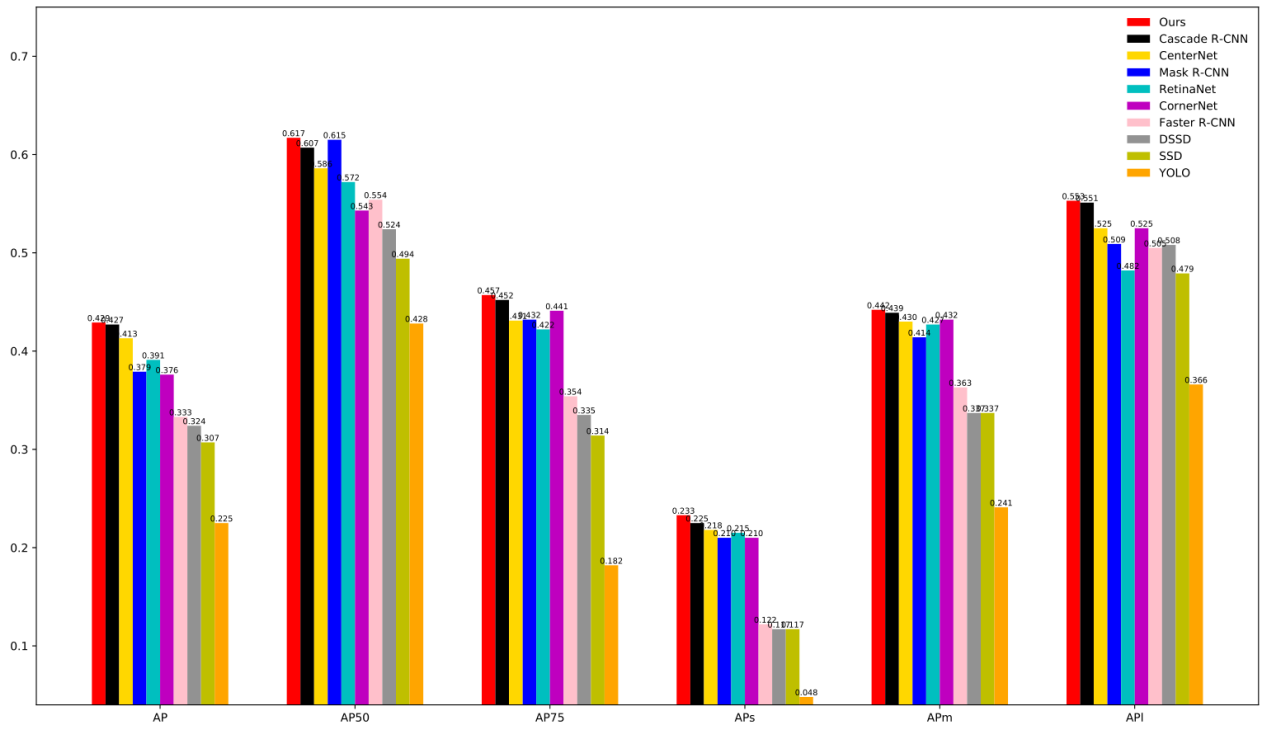

Figure 6. The histogram of different scale object detection accuracy by ten methods.

Figure 6 shows the histogram of different scale object detection accuracy by ten methods. It can be seen from Figure 6 that with the increase of object scale, the detection accuracy of all methods are gradually improved, which is in line with the common sense. Because, with the continuous down sampling of the network, small objects will lose more boundary information, resulting in low classification confidence. This is also the problem to be solved by all means. Moreover, with the increase of IoU threshold, the detection accuracy of all methods is declining, which is equivalent to evaluating the model in a higher standard. This requires a smaller offset to achieve this. In addition, our method has the highest detection accuracy in all indicators, especially for small scale objects. This also further verifies that our method is very effective in improving the accuracy of small object detection.

4.3 Qualitative test Results

Figure 7 shows some qualitative results of our method on the Microsoft COCO validation data set. It can be seen from Figure 7 that our proposed method not only has high classification accuracy, but

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

also produces consistent boundaries. It is worth mentioning that the method of multi-scale feature learning and central key point mapping is adopted. Our method makes the confidence of prediction box higher and the boundary more accurate. Thus, the recall rate is improved and a more friendly visual effect is obtained. This makes our results closer to the real situation, so it is better than other methods.

In addition, for small objects and mutual occlusion objects. We use a priori width and height to improve the matching degree of the object bounding box, which is very effective. When the object has only partial features, our method can still recognize the object more accurately. This is because the central key point can collect more abundant feature information, which is also in line with common sense. The central key point mines the visual pattern of the object to be identified at the minimum cost, which can improve the recall and precision.

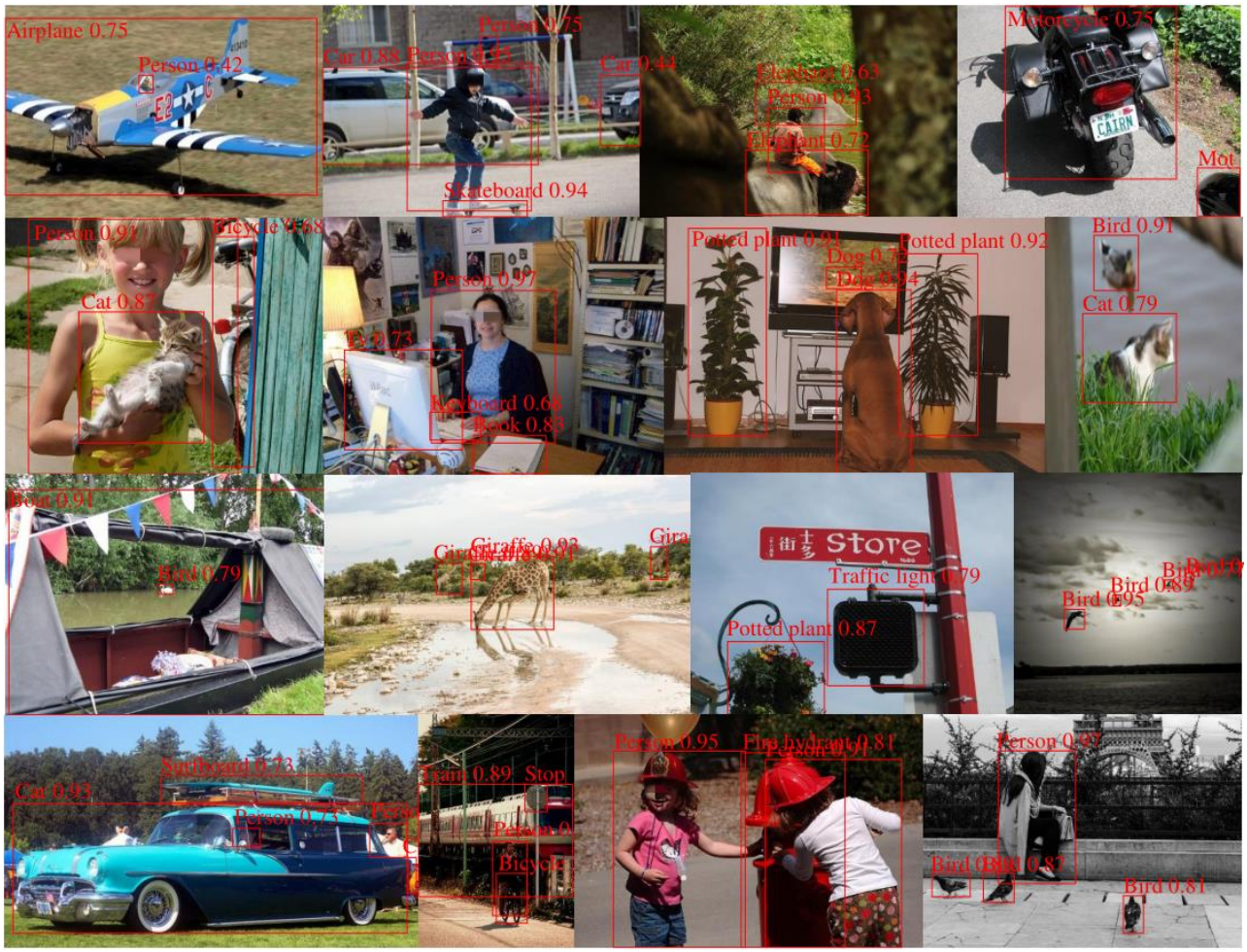

Figure 7. Qualitative detect results of Microsoft COCO.

5. Conlusion

In this paper, we propose a new one-stage object detector. Our detector uses the central key point mapping method to extract features. A new loss function is designed to expand the weight of center point offset and aspect ratio. By learning multi-scale features, our detector achieves the same detection accuracy as the two-stage detector, and has real-time detection speed. The main contributions of our method are as follows: (1). A new method of center key point mapping is proposed to realize end-to-end training and prediction without designing prior anchor boxes; (2). Increasing the weight of center point offset, mining difficult samples, enhancing the optimization performance of the model, and improving the detection accuracy of small objects; (3). By normalizing the width and height of the object, we improve the detection accuracy of different scales. (4).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

Compared with other secondary and primary detectors, our method has the highest detection accuracy. Experimental results show that the features extracted by our algorithm are more robust and generalization, and our detector is more advanced.

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton. ImageNet classification with deep convolutional neural networks, Neural Information Processing Systems, (2012), p. 1106-1114.

[2] R. Girshick, J. Donahue, T. Darrell and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, IEEE International Conference on Computer Vision and Pattern Recognition, (2014), p. 580-587.

[3] R. Girshick. Fast r-cnn, IEEE International Conference on Computer Vision, (2015), p. 1440-1448.

[4] S. Ren, K. He, R. Girshick and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, (2017), No. 6, p. 1137-1149.

[5] J. Redmon, A. Farhadi. You only look once: Unified, real-time object detection, IEEE International Conference on Computer Vision and Pattern Recognition, (2016), p. 779-788.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu and A. C. Berg. Ssd: Single shot multibox detector, European Conference on Computer Vision, (2016), p. 21-37.

[7] C. Fu, W. Liu, A. Ranga, et al. Dssd: Deconvolutional single shot detector, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[8] T. Lin, P. Goyal, R. Girshick, et al. Focal loss for dense object detection, IEEE International Conference on Computer Vision, (2017), p. 2980-2988.

[9] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks, International Conference on Learning Representations, 2014.

[10] N. Dalal, B. Triggs. Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, (2005), p. 886-893.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model, IEEE Conference on Computer Vision and Pattern Recognition, (2008), pp. 1–8.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features, International Conference on Computer Vision and Pattern Recognition, Vol. 1, (2001), p. I-I.

[13] P. Viola and M. J. Jones. Robust real-time face detection, International journal of computer vision, vol. 57, (2004), no. 2, p. 137–154.

[14] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection, International Conference on Computer Vision, (1998), p. 555-562.

[15] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W.Smeulders. Segmentation as selective search for object recognition, IEEE International Conference on Computer Vision, (2011), p. 1879–1886.

[16] K. M He, G. Gkioxari, P. Dollár and R. Girshick. Mask r-cnn, IEEE International Conference on Computer Vision, (2017), p. 2980-2988.

[17] Z. Cai, N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection, IEEE International Conference on Computer Vision and Pattern Recognition, (2018), p. 6154-6162.

[18] S. Bell, C. Lawrence, K. Bale and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, IEEE Conference on Computer Vision and Pattern Recognition, (2016), p. 2874-2883.

[19] Z. Cai, Q. Fan, R. S. Feris and N. Vascaoncelos. A unified multi-scale deep convolutional neural network for fast object detection, European Conference on Computer Vision, (2016), p. 354-370.

[20] H. Lee, S. Eum and H. Kwon. Me r-cnn: Multi-expert r-cnn for object detection, IEEE Transactions on Image Processing, Vol. 29, (2020), p. 1030-1044.

[21] S. Zhang, L. Wen, X. Bian, et al. Single-shot refinement neural network for object detection, IEEE Conference on Computer Vision and Pattern Recognition, (2018), p. 4203-4212.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

[22] H. Law, J. Deng. Cornernet: Detecting objects as paired keypoints, European Conference on Computer Vision, (2018), pp. 734-750.

[23] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian. Centernet: Keypoint triplets for object detection, International Conference on Computer Vision, (2019), p. 6569-6578.

[24] P. Dollár, R. Appel, S. Belongie and P. Perona. Fast feature pyramids for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, (2014), No. 8, p. 1532-1545.

[25] F. Sun, T. Kong, W. Huang, C. Tan, B. Fang and H. Liu. Feature pyramid reconfiguration with consistent loss for object detection, IEEE Transactions on Image Processing, Vol. 28, (2019), No. 10, p. 5041-5051.

[26] F. Sun, T. Kong, W. Huang, C. Tan, B. Fang and H. Liu. Feature pyramid reconfiguration with consistent loss for object detection, IEEE Transactions on Image Processing, Vol. 28, (2019), No. 10, pp. 5041-5051.

[27] X. Wang, Z. Hou, W. Yu, Z. Jin, Y. Zha and X. Qin. Online scale adaptive visual tracking based on multilayer convolutional features, IEEE Transactions on Cybernetics, Vol. 49, (2019), No. 1, p. 146-158.

[28] C. Huang, J. Chen, Y. Pan, H. Lai, J. Yin and Q. Huang. Clothing landmark detection using deep networks with prior of key point associations, IEEE Transactions on Cybernetics, Vol. 49, (2019), No. 10, pp. 3744-3754.

[29] X. Li, Y. Zhang, Q. Cui, X. Yi and Y. Zhang. Tooth-marked tongue recognition using multiple instance learning and cnn features, IEEE Transactions on Cybernetics, Vol. 49, (2019), No. 2, p. 380-387.

[30] W. Wu, Y. Yin, X. Wang and D. Xu. Face detection with different scales based on faster r-cnn, IEEE Transactions on Cybernetics, Vol. 49, (2019), No. 11, pp. 4017-4028.

[31] K. M. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition, European Conference on Computer Vision, (2014), pp. 346–361.

[32] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie. Feature pyramid networks for object detection, IEEE Conference on Computer Vision and Pattern Recognition, (2017), p. 936-944.

[33] Z. Li, C. Peng, G. Yu, et al. Detnet: A backbone network for object detection, European Conference on Computer Vision, 2018.

[34] Q. Zhao, T. Sheng, Y. Wang, et al. M2det: A single-shot object detector based on multi-level feature pyramid network, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, (2019), p. 9259-9266.

[35] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition, Computer Science, 2014.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition, (2014), p. 1-9.

[37] L. Wang, L. Wang, H. Lu, P. Zhang and X. Ruan. Salient object detection with recurrent fully convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, (2019), No. 7, p. 1734-1746.

[38] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao and H. Lu. Attention couplenet: Fully convolutional attention coupling network for object detection, IEEE Transactions on Image Processing, Vol.28,(2019), No. 1, 113-126.

[39] G. Cheng, J. Han, P. Zhou and D. Xu. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection, IEEE Transactions on Image Processing, Vol. 28, (2019), No. 1, p. 265-278.

[40] H. Li, G. Li and Y. Yu. Rosa: Robust salient object detection against adversarial attacks, IEEE Transactions on Cybernetics, (2019), doi: 10.1109/TCYB.2019.2914099.

[41] X. L. Li, D. W. Song and Y. S. Dong. Hierarchical feature fusion network for salient object detection, IEEE Transactions on Image Processing, Vol. 29, (2020), 9165-9175.

[42] G. C. Liu, L. L. Li, L C. Jiao, Y. S. Dong and X. L. Li. Stacked fisher autoencoder for SAR change detection, Pattern Recognition, Vol. 96, (2019), No. 106971, p. 1-12.

[43] L. Han, X. L. Li and Y. S. Dong. Convolutional edge constraint baded U-Net for salient object detection, IEEE Access, Vol. 7, (2019), p. 48890-48900

ISSN: 2414-1895 DOI: 10.6919/ICJE.202106_7(6).0052

[44] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu and P. H. S. Torr. Deeply supervised salient object detection with short connections, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, (2019), No. 4, p. 815-828.

[45] M. Boroumand, M. Chen and J. Fridrich. Deep residual network for steganalysis of digital images, IEEE Transactions on Information Forensics and Security, Vol. 14, (2019), No. 5, p. 1181-1193.

[46] O. Costilla-Reyes, R. Vera-Rodriguez, P. Scully and K. B. Ozanyan. Analysis of spatio-temporal representations for robust footstep recognition with deep residual neural networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, (2019), No. 2, pp. 285-296.

[47] Z. Wu, C. Shen and A. Hengel. Wider or deeper: Revisiting the resnet model for visual recognition, Pattern Recognition, Vol. 90, (2019), p. 119-133.

[48] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza and F. Pla. Deep pyramidal residual networks for spectral–spatial hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing, Vol. 57, (2019), No. 2, p. 740-754.

[49] L. Mou and X. X. Zhu, Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network, IEEE Transactions on Geoscience and Remote Sensing, Vol. 56, (2018), No. 11, p. 6699-6711.

[50] K. M. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition, (2016), p. 770-778.

[51] A. Newell, K. Yang and J. Deng. Stacked hourglass networks for human pose estimation, European Conference on Computer Vision, (2016), p. 483-499.

[52] Abadi M, Barham P, Chen J, et al. Tensorflow: A system for large-scale machine learning, Symposium on Operating Systems Design and Implementation, (2016), p. 265-283.

[53] M. Everingham, L. Van Gool, C. Williams, et al. The pascal visual object classes (voc) challenge, International Journal of Computer Vision, Vol. 88, (2010), No. 2, p. 303-338.

[54] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar and C. Zitnick. Microsoft coco: Common objects in context, European Conference on Computer Vision, (2014), p. 740-755.