INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.60 Volume 15, 2021

An Improved Object Detection Method using Feature Map Refinement and Anchor Optimization

Yuxia Wang [1, 2], Wenzhu Yang [1, 2, *], Tongtong Yuan [1, 2], Qian Li [1, 2]
[1] School of Cyber Security and Computer, Hebei University, Baoding 071002, China
[2] Hebei Machine Vision Engineering Research Center, Baoding 071002, China

Received: December 14, 2020. Revised: May 20, 2021. Accepted: June 4, 2021. Published: June 8, 2021.

Abstract—Lower detection accuracy and insufficient detection ability for small objects are the main problems of the region-free object detection algorithm. Aiming at solving the abovementioned problems, an improved object detection method using feature map refinement and anchor optimization is proposed. Firstly, the reverse fusion operation is performed on each of the object detection layer, which can provide the lower layers with more semantic information by the fusion of detection features at different levels. Secondly, the self-attention module is used to refine each detection feature map, calibrates the features between channels, and enhances the expression ability of local features. In addition, the anchor optimization model is introduced on each feature layer associated with anchors, and the anchors with higher probability of containing an object and more closely match the location and size of the object are obtained. In this model, semantic features are used to confirm and remove negative anchors to reduce search space of the objects, and preliminary adjustments are made to the locations and sizes of anchors. Comprehensive experimental results on PASCAL VOC detection dataset demonstrate the effectiveness of the proposed method. In particular, with VGG-16 and lower dimension 300×300 input size, the proposed method achieves a mAP of 79.1% on VOC 2007 test set with an inference speed of 24.7 milliseconds per image.

Keywords—Anchor optimization, Feature expression, Feature map refinement, Reverse fusion operation, Self-attention module, Semantic information.

I. Introduction

Object detection has achieved important advances in recent years, with the development of deep neural networks. The current detectors of state-of-the-art are mainly based on convolutional neural networks, they can be divided into two categories: the region-based methods, including [1,2,19,20,21], and the region-free methods, including [3-6]. In the region-based methods, a sparse set of high-quality candidate boxes are first generated, and then a region-wise subnetwork is designed to classify and refine these candidate boxes [23]. The region-based methods have been achieving very high accuracy with the lower computation speed. The region-free methods make full use of the thought of regression, without region proposals, for a given image, the bounding box coordinates and object categories probabilities are regressed directly at multiple locations in this image [7]. The main advantage of these methods is high computation efficiency. However, the detection accuracy of these methods is usually behind that of the region-based methods.

The region-free methods attract much more attention recently due to their high efficiency, and how to improve the accuracy of these methods has become a research hotspot [8]. It was found that the lack of feature expression ability in the backbone network is one of the main reasons for the low detection accuracy [9]. Some region-free methods, such as SSD, use multiple convolutional layers to detect objects, smaller objects are detected by lower layers while larger objects are detected by higher layers. However, due to the lack of semantic information of the objects in the shallow feature maps, small objects may not be detected well. In order to address this problem, Liu et al. [10] propose the DSSD algorithm, which uses the deconvolution module to extract more contextual information, and continuously combines contextual information to improve the expression ability of features. Although this algorithm has some improvements in the detection accuracy, it has a large amount of calculation and cannot meet the real-time requirements. Shen Z et al. [11] propose the DSOD algorithm, which uses the DenseNet network to replace the original VGG network. This algorithm only improves the SSD algorithm from the perspective of pre-training and does not emphasize the expression ability of features. How to better improve the expression ability of features in the network is still a problem to be solved.

Researchers found that the foreground-background class

This work is supported by the Natural Science Foundation of Hebei Province under Grant F2021201012 and the Natural Science Foundation of China under Grant 6217070810.

imbalance problem is another main reason for the low accuracy of region-free methods [12]. This imbalance problem has been solved in the region-based methods. The region proposal stage (Selective Search, RPN, etc.) could reject most of the simple background samples, and the search space of the objects is greatly reduced [13]. The region-free methods rely on a large number of consistent and densely distributed anchors to detect objects. Many of these anchors correspond to the background areas that are irrelevant to the objects of interest. The existence of a large number of simple background samples not only brings about serious foreground-background class imbalances, but also makes the detector difficult to be optimized [24]. There are some methods in the region-free approach are proposed to address the class imbalance problem [25].

Kong et al. [14] use the objectness prior constraint on convolutional feature maps, and the search space is reduced effectively by guiding the searching locations of objects. Lin et al. [15] propose a new loss function Focal loss to dynamically adjust the weight of each anchor, such that it can reduce the loss assigned to well-classified examples and put more focus on a sparse set of hard, misclassified examples. Chen et al. [16] propose a new anchor generation method (Guided Anchoring), which generates sparse and arbitrary-shaped anchors by predicting the locations where the centers of the objects may exist, the scales and aspect ratios of objects at different locations.

In order to further improve the detection accuracy of the region-free approach, an improved object detection method using feature map refinement and anchor optimization is proposed. The main contributions are summarized as follows.

(1) The reverse fusion operation is applied on each of the object detection source layer in the feature extraction network, which can further enhance the semantic information of the lower detection layers by the fusion of feature map information.

(2) The self-attention module is introduced on each detection feature map of different levels, and the feature maps can be refined from the two dimensions of channel and space. By this method, the information that needs special attention on each feature map can be obtained.

(3) The anchor optimization model is designed, which uses semantic features to filter and adjust anchors. The negative anchors can be confirmed and removed by predicting the probability that each anchor contains an object so as to reduce search space of the objects. At the same time, preliminary adjustments are made to the locations and sizes of anchors.

II. Methodology

In this section, the proposed method is presented in detail. First, the overall structure of this method is illustrated specifically. Then, how to enhance the expression abilities of detection features is described. Finally, the anchor optimization model is introduced to obtain high-quality anchors by filtering and adjusting the original anchors.

A. Architecture

The overall structure of the proposed object detection model is shown in Fig. 1. Aiming at the problem that the region-free object detection algorithm uses multi-scale detection feature maps to cause poor prediction results, improvements are made from two aspects: the enhancement of feature expression abilities and the optimization of original anchors. First, the reverse fusion operation is applied on each detection feature map to fuse the feature map information, which can provide the lower layers with more semantic information; At the same time, the self-attention module is designed to refine each detection feature map to enhance the expression of effective features and suppress background interference. Then, the anchor optimization model is designed, which uses semantic features to filter and adjust anchors on each detection feature map to obtain the final high-quality anchors. The refined detection feature maps and the high-quality anchors are used for the final classification and regression operations.

B. The Enhancement of Feature Expression Abilities

The insufficient of feature expression abilities is a common problem in object detection. SSD uses detection feature maps of different convolutional layers for prediction, and the detection effect of small objects is poor due to the lack of semantic information in the lower detection layers. The reverse fusion operation and feature map refinement are applied to each detection feature map in this detection method. The reverse fusion operation is used to provide the lower layers with more semantic information, and feature map refinement is used to improve the attention of each detection feature map to its own information. These two operations are used to effectively enhance the expression abilities of detection features.

1) Semantic strengthen at lower layers

SSD uses lower object detection feature layers with high resolution and low semantic information to detect small objects, which results in poor detection effect of small objects. In order to obtain more sufficient semantic information, the reverse fusion operation is applied to multiple detection feature layers of different scales in the backbone network, and the information of the higher layers is continuously fused with the information of the lower layers. The reverse connection operation is shown in Fig. 2. The reverse fusion feature map (marked as rf-map) rf-map n+1 is up-sampled to the same size as the feature map (marked as f-map) f-map n in the backbone network using bilinear interpolation; Then the 1×1 convolution operation is used to change the number of channels of the rf-map n+1, and ensure that the number of channels of this feature map is the same as the number of channels of the f-map n; Finally, the two feature maps are merged by element-wise addition to obtain the reverse fusion feature map rf-map n, which corresponding to the feature map f-map n. In this way, the reverse fusion feature maps corresponding to original detection feature maps are obtained, and these reverse fusion feature maps will replace the original detection feature maps for subsequent feature map refinement, classification and regression operations.

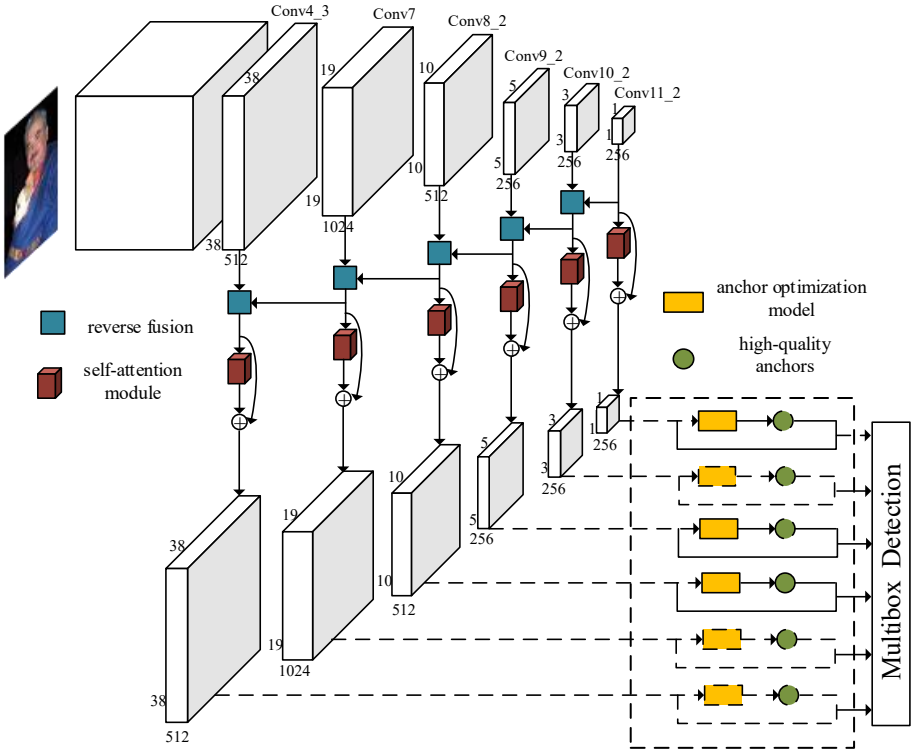


Fig. 1. The object detection model using feature map refinement and anchor optimization

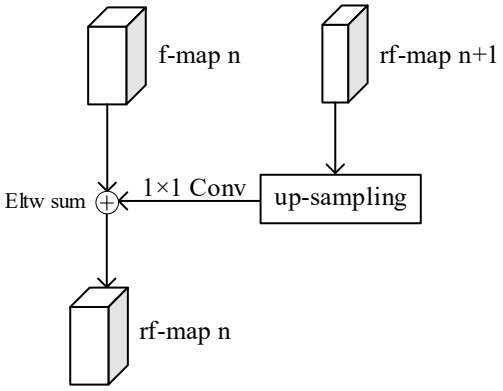


Fig. 2. The reverse fusion operation

2) Feature map refinement

In Non-local Neural Networks [17], the self-attention mechanism is used to consider the correlation of spatial features in the feature maps, and new features are obtained according to this correlation. In Squeeze-and-Excitation Networks [18], network performance is improved by considering the relationship between feature channels. Based on these ideas, the self-attention module is designed to refine each detection feature map, improving the attention of the detection feature map to its own information, so that the features that contain important information can be obtained and the irrelevant features can be suppressed selectively. The self-attention module is shown in Fig. 3, which includes channel-wise attention and spatial-wise attention.

The channel-wise attention can prune the spatial information of each feature map, focusing on the relationship between channels and object classes, as shown in Fig. 3 (a). It consists of three phases: global average-pooling, channel-wise learning and broadcasted multiplying. Specifically, given an input $X \in \mathbb{R}^{C\times W\times H}$, the global average-pooling operation is performed along spatial dimension $W\times H$ to aggregate spatial information and generate channel-wise descriptors $Z \in \mathbb{R}^{C}$. The i-th element of a channel descriptor can be expressed as:

$$Z_i = \frac{\sum_{h,w} X_{ihw}}{HW} \tag{1}$$

The channel-wise learning stage is to use two fully connected layers (a convolution layer of 1×1 is used to replace the fully connected layer), RELU nonlinear activation function and Sigmoid activation function to adaptively model the correlation between each channel, and generate activation features $S \in \mathbb{R}^{c\times 1\times 1}$, the formula is as follows:

$$S = Sigmoid(W_2 \cdot \mathrm{Re}LU(W_1 Z)) \tag{2}$$

where $W_1 \in \mathbb{R}^{c/r\times c}$, $W_2 \in \mathbb{R}^{c\times c/r}$ and r is the reduction ratio.

In the broadcasted multiplying stage, S is used to activate X to get $X' \in \mathbb{R}^{C\times W\times H}$, where:

$$X'_{ihw} = X_{ihw} \cdot S_i \tag{3}$$

Finally, the X' will replace the original X to get the spatial-wise features.

The spatial-wise attention can model the spatial relationship of features and improve the expression of information in specific spatial regions. As shown in Fig. 3 (b), it also consists of three stages: average-pooling and max-pooling, convolution operation and broadcasted multiplying. First, average-pooling and max-pooling operations are utilized along the channel axis to aggregate channel information, generating two spatial context

descriptors:

$$AvgPool(X'), MaxPool(X') \in \mathbb{R}^{1\times W\times H} \quad (4)$$

These two descriptors are cascaded to generate a more effective feature descriptor. Then a convolution operation with a Sigmoid function is applied on this feature descriptor to generate a 2D activation feature $M \in \mathbb{R}^{1\times W\times H}$. This process is represented as:

$$M = Sigmoid(conv[AvgPool(X'); MaxPool(X')]) \quad (5)$$

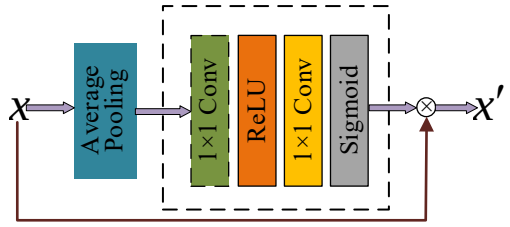
where *conv* represents a 7×7 convolution operation.

In the broadcasted multiplying stage, M is used to activate X' to get the final refined feature map:

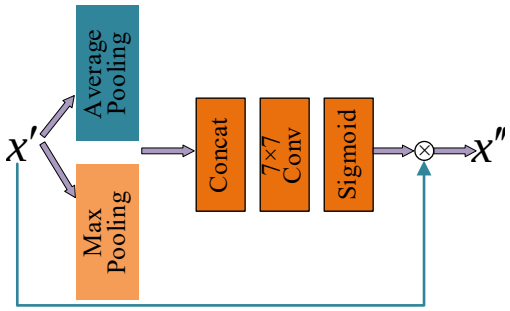
$$X'' = M \times X' \in \mathbb{R}^{C\times W\times H} \quad (6)$$

Finally, the refined feature map is added as a residual branch to the original reverse fusion feature map, the obtained feature map Y is used for subsequent classification and regression operations.

$$Y = X + X'' \quad (7)$$



(a) channel-wise attention



(b) spatial attention

Fig. 3. The self-attention module

C. Anchor Optimization Model

The sliding window is a simple and widely adopted anchoring scheme. By this scheme, there are a set of anchors with predefined scales and aspect ratios being deployed at every location of a feature map. On the one hand, only a small number of anchors contain the objects, most of anchors corresponding to the background areas which are irrelevant to the objects. On the other hand, this method does not consider the actual sizes of the objects in an image, resulting in poor detection effect for small objects or objects in special scenes. The anchor optimization model is designed and applied to the multi-scale detection feature maps. In this model, the semantic features are used to confirm and remove negative anchors, and preliminary adjustments are made to the locations and sizes of the anchors. The anchor optimization model is shown in Fig. 4.

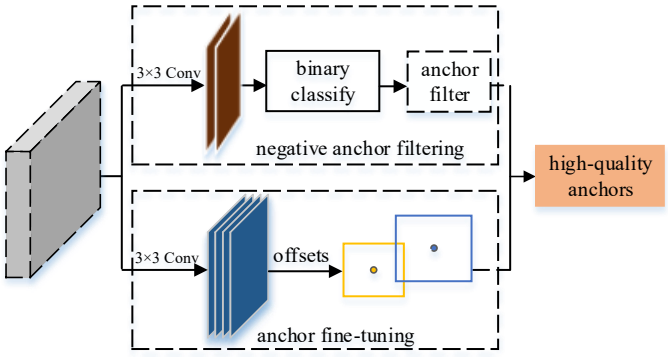


Fig. 4. The anchor optimization model

1) The filtration of negative anchors

In this anchor optimization model, semantic features are used to filter negative anchors. Confirm and remove negative anchors that are well-classified before detection can reduce search space for the objects and effectively alleviate the imbalance problem in object detection. The essence of this filtering operation is to determine the probability that each anchor contains an object. Specifically, a 3 × 3 convolution operation with 2c convolution kernels is added to each refined detection feature map, where c represents the number of anchors at each location of these detection feature maps, and 2 represents the two confidence scores predicted for each anchor to indicate whether this anchor contains an object. The Softmax function is added after the convolution operation to get the binary classification probability of each anchor as foreground or background. Given a negative confidence threshold value θ, the original anchors are selected according to this value. If the negative confidence probability np of an anchor is greater than this threshold value, it means that this anchor is a negative anchor which is well-classified and does not contain an object, and this anchor will be discarded in the subsequent detection task. Otherwise, this anchor will be reserved for the subsequent object detection.

2) Anchor fine-tuning

Current region-free methods rely on one-step regression to predict the locations and sizes of objects, for the detection of some small objects or objects in special scenes, one-step regression affects the accuracy of the detection. This is because in these methods, the anchors are sampled uniformly over the spatial domain with a predefined set of scales and aspect ratios. The scales and aspect ratios of these anchors are too fixed, and the actual sizes of the objects in an image are not considered. For the detection of small objects or objects with extreme irregular aspect ratios (such as very tall or very wide objects), the matching effect of the anchors with these objects will be very poor. The detection effect will be greatly affected if these initialized anchors are used as regression references directly. Therefore, the anchor optimization model makes a preliminary adjustment to the locations and sizes of anchors so as to provide high-quality anchors that more closely match the locations and sizes of the objects for subsequent object detection.

Take the actual sizes of the objects as references, the locations and sizes of the initial anchors are adjusted before the

object detection to make the locations and sizes of the anchors match the objects better. Refer to the regression process in object detection, in this anchor fine-tuning operation, the locations and sizes of anchors on the detection feature maps are adjusted by regression method. Specifically, a 3×3 convolution operation with $4c$ convolution kernels is added to each detection feature map. Among them, c represents the number of anchors at each location of a detection feature map, and 4 represents predicted four offsets of the adjusted anchors relative to the original anchors. After convolution operation, the original anchors are transformed according to the corresponding four offsets to obtain the new anchors which can match the locations and sizes of the objects more closely.

The high-quality anchors obtained after negative anchor filtering and anchor fine-tuning are combined with the corresponding detection feature maps for further object detection which can generate the scores for object classes and shape offsets relative to the adjusted anchor box coordinates. Use the adjusted anchor boxes as input for further object detection can make the detection results more accurate.

III. Results and Discussion

A. Experimental Settings

PASCAL VOC dataset is used in this experiment, which contains 20 categories including people, animals (such as cats, dogs, etc.), vehicles (such as cars, boats, planes, etc.), and furniture (such as chairs, tables, etc.). In this experiment, the VOC2007 and VOC2012 datasets are used for training, and the VOC2007 dataset is used for testing. The VOC2007 and VOC2012 training sets consist of 16,551 images and 40058 sample frames. The average number of sample frames for a single image in the training sets is 2.4. The VOC2007 test set has 4952 images and 12032 sample frames.

The experimental platform is a computer with a CPU of Intel Xeon Silver 4215 and a memory of 64GB. The experimental environment is windows 10, and Nvidia GTX 2080 GPU is used for training.

In this experiment, the pre-trained VGG-16 model [26] is used as the backbone network. The initial learning rate is set to 10^{-3} for the first $80k$ iterations, then it is attenuated to 10^{-4} and 10^{-5} for training another $20k$ and $20k$ iterations, respectively. The Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005 are used for network optimization. Several data augmentation strategies presented in [27] are used to randomly expand and crop the original training images to generate the training samples, and the hard negative mining strategy is used to further alleviate the extreme foreground-background class imbalance.

B. Experimental Results

In this paper, all models are trained on the VOC 2007 and VOC 2012 training sets, and these models are tested on the VOC 2007 test set. Mean Average Precision (mAP) is used to evaluate the accuracy of these models, and FPS is used to evaluate the detection rate of these models. In this article, the mAP is calculated when IoU=0.5.

1) Comparison with some state-of-the-art methods

The test results of the proposed method compared with some state-of-the-art methods such as SSD300, DES300 [23], DSSD321, DFPR300 [28] are shown in Table 1. Although the detection accuracy of the proposed method is slightly lower than that of the R-FCN [20], DES300 and DFPR300, compared with the most object detection methods, the proposed method has achieved a higher detection accuracy. It can be seen from Table 1 that the detection accuracy of the proposed method is 1.6% higher than that of the SSD300 and 1.4% higher than that of the DSOD300. Compared with RON384, YOLOv2 and other detectors, the detection accuracy of the proposed method is also improved significantly.

In terms of detection rate, the detection rate of the proposed method far surpasses the region-based object detection detectors such as Fast R-CNN [21] and Faster R-CNN [22], and it is also faster than most detectors of the same type such as RON384, DES300 and DFPR300. In summary, although the detection accuracy of the proposed method is slightly lower than DES300 and DFPR300, its detection rate is faster than these methods. Therefore, the proposed method achieves a better balance between detection rate and detection accuracy, which means that it has a better application prospect.

Table 1. Comparison with some state-of-the-art methods

Method	mAP (%)	FPS
Fast R-CNN	70.0	0.5
Faster R-CNN	73.2	7.0
R-FCN	80.5	9.0
SSD300	77.5	46.0
YOLOv1	63.4	45.0
YOLOv2	78.6	40.0
DSSD321	78.6	9.5
RON384	77.6	15.0
DSOD300	77.7	17.4
DES300	79.7	32.9
DFPR300	79.6	39.5
Ours	79.1	40.5

Table 2. Comparison of the proposed method with some methods in different object categories

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
SSD300	77.5	79.5	83.9	76.0	69.6	50.5	**87.0**	85.7	88.1	60.3	81.5	77.0	86.1	87.5	84.0	79.4	52.3	77.9	79.5	87.6	76.8
DSSD321	78.6	81.9	84.9	**80.5**	68.4	53.9	85.6	86.2	88.9	61.1	83.5	**78.7**	86.7	**88.7**	**86.7**	**79.7**	51.7	78.0	**80.9**	87.2	**79.4**
RON384	77.6	**86.0**	82.5	76.9	69.1	59.2	86.2	85.5	87.2	59.9	81.4	73.3	85.9	86.8	82.2	79.6	52.4	78.2	76.0	86.2	78.0
Ours	79.1	84.2	**85.7**	78.3	**72.9**	**59.8**	84.0	**86.3**	**89.1**	**62.2**	**85.6**	78.1	**88.3**	87.2	84.7	77.9	**52.6**	**78.3**	79.9	**88.1**	78.9

The comparison results of the proposed method with the SSD300, DSSD321 and RON384 in different categories are shown in Table 2. In the comparison of detection accuracy of 20 categories, the proposed method has achieved the best detection effect in 11 categories. In these 20 categories, the detection effect of small objects such as 'boat' and 'bottle' is generally poor. Compared with other methods, the proposed method has improved the detection effect of these small objects significantly. The detection accuracy of the proposed method for 'boat' is 72.9%, which is an increase of 3.3% compared to the SSD300. The detection accuracy for 'bottle' is 59.8%, which is 9.3% higher than SSD300. In addition, the proposed method has also achieved good detection results of 'cow' and 'dog'.

2) Ablation study

In order to demonstrate the effectiveness of different components in this proposed method, four variants are constructed and evaluated, the experimental results are shown in Table 3. For a fair comparison, all models are evaluated based on the same setup parameters and input size.

Table 3. Comparison of the effectiveness of different components

anchor fine-tuning	negative anchor filtering	feature map refinement	reverse fusion operation	mAP (%)
✓	✓	✓	✓	79.1
	✓	✓	✓	78.4
		✓	✓	77.9
			✓	77.6
				76.6

Anchor fine-tuning. In order to demonstrate the effectiveness of the anchor fine-tuning operation in the proposed method, when optimizing the anchors, only the negative anchors are filtered, the anchor fine-tuning operation which can adjust the locations and sizes of the original anchors is omitted. The experimental results show that removing the anchor fine-tuning leads to 0.7% drop in mAP (*i.e.*, 79.1% *vs.*78.4%). It can be seen that the anchor fine-tuning help promote the performance effectively. The reason is that adjusting the original anchors by a regression manner can make our method to adapt to different scales and aspect ratios of objects better. The adjusted anchors can match the locations and sizes of the objects more closely.

The filtration of negative anchors. To verify the effectiveness of the negative anchor filtering operation in this method, the process of optimizing the anchors can be omitted, and only the original anchors are used for detection. Specifically, the negative confidence threshold value is set to 1, which can ensure that the negative confidence probability of any anchor is less than this value. It means that none of the negative anchors are filtered out during this process. The mAP is reduced from 78.4% to 77.9%. That's because the class imbalance problem affects the detection accuracy, but the filtering of these well-classified negative anchors can alleviate the class imbalance problem to a certain extent.

Feature map refinement. On the basis of omitting the process of optimizing anchors, the process of using the self-attention module to refine the detection feature maps is cut off, and the mAP is reduced by 0.3%. This operation proves the effectiveness of using the self-attention model to refine each detection feature map. The reason is that the self-attention module used on each detection feature map can calibrate the features between channels and enhance the expression ability of local features. In this way, the information that needs special attention on each detection feature map can be captured.

Reverse fusion operation. Finally, by removing the reverse fusion operation performed on the multi-scale detection feature layers, the mAP of the proposed method is reduced from 77.6% to 76.6%, which is a direct reduction of 1%. This shows that in this paper, the reverse fusion operation is the most effective component to improve the detection accuracy. The reason is that the lower feature maps have poor detection effect on small objects due to lack of semantic information, but the reverse fusion operation can provide the lower layers with more semantic information by the fusion of detection features at different levels.

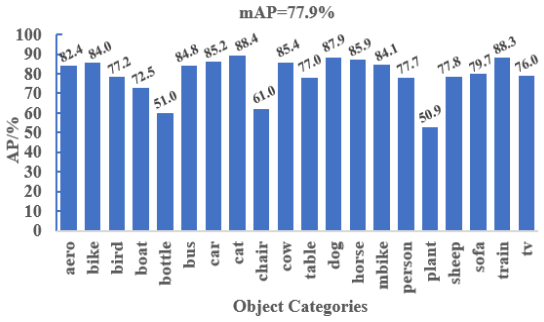
C. Discussion

In this paper, three experimental models are designed based on SSD300, including Model A which adds the reverse fusion operation and feature map refinement to enhance the expression ability of detection features, Model B which adds the anchor optimization model for negative anchor filtering and anchor fine-tuning to optimize the original anchors, and Model C with two improved schemes including detection feature enhancement and anchor optimization.

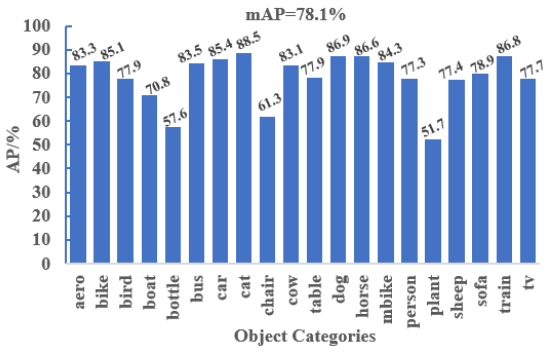
When IoU = 0.5, the mAP of three experimental models and the prediction results of each category are shown in Fig. 5. It can be seen from Fig. 5 that, the mAP of the three models has reached more than 77.5%, and the detection effect of Model C is

the best, with the highest mAP of 79.1%. It can be seen that enhancing the detection features and optimizing the original anchors, both of which can improve the detection effect. Compare Model A and Model B, the detection accuracy of Model B is 0.2% higher than that of Model A, which shows that in the proposed method, optimizing the anchors can get a little better detection effect than enhancing the detection features. Although the detection rate of Model C with two improved schemes is slightly lower than that of Model A and Model B which contains only one improvement, the experimental results show that the detection rate of Model C is 24.7 milliseconds per image, which can still meet the requirement of real-time detection.

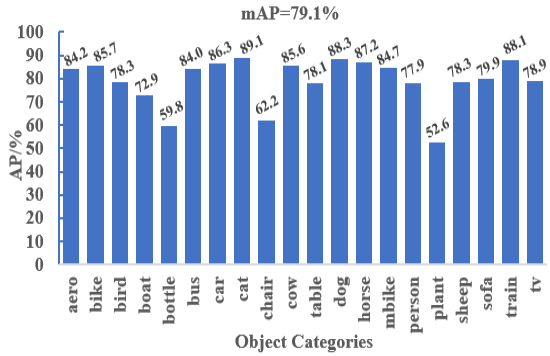
Table 4 is designed to further justify the effectiveness of the reverse fusion operation and feature map refinement in Model A. It can be seen that both reverse fusion operation and feature map refinement can improve the detection effect, and the reverse fusion operation can get better detection effect than feature map refinement (*i.e.*, 77.6% vs.77.1%). The best detection effect can be obtained by refining the detection feature maps after the reverse fusion operation, whose mAP is 77.9%. It is invalid that using reverse fusion operation after refining the detection feature maps. In comparison with the corresponding studies of other papers, for example, the self-attention mechanism is used in DES [23] and CBAM [24] to enhance the detection ability, and the deconvolution operation is used in DSSD to extract more contextual information, our scheme is more specific and accurate. In our scheme, the detection features are enhanced by two improvements of reverse fusion operation and feature map refinement. The results show that the detection performance can be improved effectively by fusion and refinement the detection features.



(a) mAP and object categories prediction results of Model A



(b) mAP and object categories prediction results of Model B



(c) mAP and object categories prediction results of Model C

Fig. 5. mAP of three models and the prediction results of each category

Table 4. Comparison of different components in Model A. F stands for the reverse fusion operation and R stands for the feature map refiment

Method	mAP (%)
SSD300	76.6
SSD300 + F	77.6
SSD300 + R	77.1
SSD300 + F + R	77.9
SSD300 + R + F	invalid

Table 5 is designed to illustrate the role of the negative anchor filtering and anchor fine-tuning in Model B. It can be seen that only filtering the negative anchors, the mAP is increased from 76.6% to 77.4%; in comparison, fine-tuning the anchors can get better detection accuracy whose mAP is 77.7%. The highest mAP is 78.1% while connecting the negative anchor filtering and anchor fine-tuning in parallel in the anchor optimization model. Compared with the other corresponding studies, such as Kong et al. [14] use the objectness prior constraint to reduce the search space; Lin et al. [15] propose the Focal loss to reduce the loss assigned to well-classified examples. Both of these schemes only consider the imbalance problem, but the matching problem between anchors and objects is not considered. This matching problem is considered in the anchor optimization model. In this model, the original anchors are adjusted by regression method, so that these anchors can match the locations and sizes of the objects more closely. Experimental results prove that the anchor fine-tuning operation can improve the detection effect significantly.

Table 5. Comparison of different components in Model B which contains negative anchor filtering and anchor fine-tuning

Method	mAP (%)
SSD300	76.6
SSD300 + negative anchor filtering	77.4
SSD300 + anchor fine-tuning	77.7
SSD300 + negative anchor filtering & anchor fine-tuning	78.1

IV. Conclusion

In this paper, we analysed the problems when the region-free object detection algorithm was used in multi-scale feature maps detection. Then an improved object detection method using

feature map refinement and anchor optimization is proposed. Aiming at the problem of insufficient expression abilities of detection features, the detection feature layers are fused reversely to improve the semantic information of lower layers, and the self-attention module is used to refine each detection feature map. The anchors generated by sliding window method are optimized through the designed anchor optimization model, which can confirm and remove the negative anchors to reduce search space of the objects and fine-tune the locations and sizes of the anchors to obtain high-quality anchors.

Experimental results show that the proposed method achieves the-state-of-the-art detection accuracy with high efficiency. But this method still has some limitations. 1) Although the reverse fusion operation can enhance the semantic information of the lower detection feature maps to some extent, the detection effect of small objects is still poor. 2) The effect of filtering the well-classified negative anchors through the anchor optimization model is not outstanding, and the imbalance issue is still obvious. Next, we plan to make some improvements based on these limitations. 1) we plan to introduce the dilated convolution structure to down-sample the lower detection feature maps to improve the receptive field of these feature maps and further improve the detection accuracy of small objects. 2) we plan to integrate GIoU and Focal loss for the proposed method to further alleviate the imbalance problem by optimizing the loss function.

We can also make some improvements according to several future research directions in the object detection area, such as: 1) Although object detection has been converted from anchor-based methods to anchor-free methods (e. g. methods based on key-points), there are still some limitations to improve, and thus designing a more efficient and effective proposal generation strategy is necessary. 2) The current basic networks of object detection detectors come from image classification, which can cause a learning bias due to the differences between classification and detection. How to learn object detectors from scratch becomes a major concern. In general, there are still many problems to be solved in the future research filed of object detection.

References

[1] Girshick, R., Donahue, J., Darrell, T., & Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587, 2014.

[2] He, K., Gkioxari, G., Dollár, P., & Girshick, R., "Mask r-cnn," Proceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017.

[3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., "You only look once: Unified, real-time object detection," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, 2016.

[4] Redmon, J., & Farhadi, A., "YOLO9000: better, faster, stronger," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271, 2017.

[5] Redmon, J., & Farhadi, A., "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[6] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C., "Ssd: Single shot multibox detector," European conference on computer vision, Springer, Cham, pp. 21-37, 2016.

[7] Zhiqiang, W., & Jun, L., "A review of object detection based on convolutional neural network," 2017 36th Chinese Control Conference (CCC). IEEE, pp. 11104-11109, 2017.

[8] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M., "Deep learning for generic object detection: A survey," International journal of computer vision, vol. 128, no. 2, pp. 261-318, 2020.

[9] Zhang, Z., Qiao, S., Xie, C., Shen, W., Wang, B., & Yuille, A. L., "Single-shot object detection with enriched semantics," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5813-5821, 2018.

[10] Fu, C. Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A. C., "Dssd: Deconvolutional single shot detector," arXiv preprint arXiv:1701.06659, 2017.

[11] Shen, Z., Liu, Z., Li, J., Jiang, Y. G., Chen, Y., & Xue, X., "Dsod: Learning deeply supervised object detectors from scratch," Proceedings of the IEEE international conference on computer vision, pp. 1919-1927, 2017.

[12] Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E., "Imbalance problems in object detection: A review," IEEE transactions on pattern analysis and machine intelligence, 2020.

[13] Wu, X., Sahoo, D., & Hoi, S. C., "Recent advances in deep learning for object detection," Neurocomputing, 396: 39-64, 2020.

[14] Kong, T., Sun, F., Yao, A., Liu, H., Lu, M., & Chen, Y., "Ron: Reverse connection with objectness prior networks for object detection," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5936-5944, 2017.

[15] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P., "Focal loss for dense object detection," Proceedings of the IEEE international conference on computer vision, pp. 2980-2988, 2017.

[16] Wang, J., Chen, K., Yang, S., Loy, C. C., & Lin, D., "Region proposal by guided anchoring," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2965-2974, 2019.

[17] Wang, X., Girshick, R., Gupta, A., & He, K., "Non-local neural networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794-7803, 2018.

[18] Hu, J., Shen, L., & Sun, G., "Squeeze-and-excitation networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141, 2018.

[19] Girshick, R., "Fast r-cnn," Proceedings of the IEEE international conference on computer vision, pp. 1440-1448, 2015.
[20] Ren, S., He, K., Girshick, R., & Sun, J., "Faster R-CNN: towards real-time object detection with region proposal networks," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 6, pp. 1137-1149, 2016.
[21] Dai, J., Li, Y., He, K., & Sun, J., "R-fcn: Object detection via region-based fully convolutional networks," Advances in neural information processing systems. pp. 379-387, 2016.
[22] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S., "Cbam: Convolutional block attention module," Proceedings of the European conference on computer vision (ECCV), pp. 3-19, 2018.
[23] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X., "Object detection with deep learning: A review," IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212-3232, 2019.
[24] Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D., "Libra r-cnn: Towards balanced learning for object detection," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 821-830, 2019.
[25] Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z., "Single-shot refinement neural network for object detection," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4203-4212, 2018.
[26] Simonyan, K., & Zisserman, A., "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[27] Henriques, J. F., Carreira, J., Caseiro, R., & Batista, J., "Beyond hard negative mining: Efficient detector learning via block-circulant decomposition," proceedings of the IEEE International Conference on Computer Vision, pp. 2760-2767, 2013.
[28] Kong, T., Sun, F., Tan, C., Liu, H., & Huang, W., "Deep feature pyramid reconfiguration for object detection," Proceedings of the European conference on computer vision (ECCV), pp. 169-185, 2018.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Wenzhu Yang did the study conceiving and methodology.
Yuxia Wang carried out the experiment and original draft preparation.
Tongtong Yuan was responsible for review and editing.
Qian Li was responsible for visualization and supervision.

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)