Dual Attention Based Image Pyramid Network for Object Detection

Xiang Dong, Feng Li, Huihui Bai* and Yao Zhao
Institute of Information Science, Beijing Jiaotong University
Beijing, 100044 - China
[e-mail :xiangdong@bjtu.edu.cn, l1feng@bjtu.edu.cn, hhbai@bjtu.edu.cn, yzhao@bjtu.edu.cn]
*Corresponding author: Huihui Bai

Received February 19, 2021; revised September 22, 2021; accepted November 20, 2021; published December 31, 2021

Abstract

Compared with two-stage object detection algorithms, one-stage algorithms provide a better trade-off between real-time performance and accuracy. However, these methods treat the intermediate features equally, which lacks the flexibility to emphasize meaningful information for classification and location. Besides, they ignore the interaction of contextual information from different scales, which is important for medium and small objects detection. To tackle these problems, we propose an image pyramid network based on dual attention mechanism (DAIPNet), which builds an image pyramid to enrich the spatial information while emphasizing multi-scale informative features based on dual attention mechanisms for one-stage object detection. Our framework utilizes a pre-trained backbone as standard detection network, where the designed image pyramid network (IPN) is used as auxiliary network to provide complementary information. Here, the dual attention mechanism is composed of the adaptive feature fusion module (AFFM) and the progressive attention fusion module (PAFM). AFFM is designed to automatically pay attention to the feature maps with different importance from the backbone and auxiliary network, while PAFM is utilized to adaptively learn the channel attentive information in the context transfer process. Furthermore, in the IPN, we build an image pyramid to extract scale-wise features from downsampled images of different scales, where the features are further fused at different states to enrich scale-wise information and learn more comprehensive feature representations. Experimental results are shown on MS COCO dataset. Our proposed detector with a 300×300 input achieves superior performance of 32.6% mAP on the MS COCO test-dev compared with state-of-the-art methods.

Keywords: Dual attention mechanism, Adaptive feature fusion module, Progressive attention fusion module, Image pyramid network, Multi-scale object detection.

This work was supported in part by National Natural Science Foundation of China (No. 61972023) and Fundamental Research Funds for the Central Universities (2019JBZ102, 2018JBZ001).

http://doi.org/10.3837/tiis.2021.12.010 **ISSN : 1976-7277**

1. Introduction

Recently, deep learning technology has gradually become the popular research direction. Relying on deep learning technology, there are many methods that have been proposed for object detection, which have achieved remarkable performance. These methods can be divided into two categories: one-stage methods and two-stage methods. The two-stage detection methods [1, 2, 3, 4, 5, 6, 7] generate region proposals from input images via traditional approaches or neural networks. Then all these proposals are sent to the classifier for object classification and location.

In contrast, one-stage detection methods [8, 9, 10, 11] directly regress the coordinates and category probabilities of the objects, and the final result can be predicted only through a single detection. In [8], Liu *et al.* propose SSD which introduces a multi-scale prediction into the network architecture and balances the detection accuracy and speed. However, due to the lack of high-level context exploration, this method shows poor performance on small objects. Some works [12, 13, 14, 15, 16] intergrate both low-level and high-level feature representations using top-down and bottom-up structures to predict the regression position boxes and category probability. Though these methods fully interact high-level semantic information with low-level detailed information, the loss of the feature information in the massive convolution process is irreversible. LRF [17] and EFIP [18] utilize SSD [8] with VGG-16 [19] as the backbone network and introduce new auxiliary networks which supplement more spatial information at the each detection stages of the backbone network. Nevertheless, they only perform a simple fusion operation between the spatial information provided by the auxiliary network and the features of the SSD backbone network, which fail to emphasize the importance of the information flow between these two parts. Besides, they also ignore the interaction of multi-scale meaningful features that are contributed to object detection within the backbone network.

According to above analysis, in this paper, we consider the object detection from three aspects: 1) the shallow and deep feature information under different spatial sizes make different contributions during the object classification and location process. It is necessary to fully incorporate low-level features into high-level space to enrich the feature representations. 2) The pretained backbone and auxiliary network provide complementary multi-level information for the observed objects. We need to conduct more discriminative fusion to enhance the network ability for accurate detection. 3) When multi-category object detection is performed, the features corresponding to these objects are diverse due to the different types and sizes of objects. Therefore, only relying on single-scale image information for feature extraction can limite the detection accuracy, espacially on medium and small objects.

To this end, we propose an image pyramid network, called DAIPNet, which builds an image pyramid to enrich the spatial information while emphasizing multi-scale informative features based on dual attention mechanism for object detection. The proposed method mainly consists of the SSD [8] as pretrained backbone and another auxiliary network to combine the advantages of these two networks. Specfically, in DAIPNet, the dual attention mechanism is designed to focus on the similarities and differences between features, thereby achieving effective feature extraction and interaction. The proposed dual attention mechanism includes two modules: adaptive feature fusion module (AFFM) and progressive attention fusion module (PAFM). Considering that feature information under different levels in the SSD backbone network makes different contributions that are relevant to object detection, AFFM adaptively learns the weight coefficients between the features produced by the backbone and auxiliary network through attention exploration, and then performs weighted fusion of these features.

PAFM progressively concatenates features from multiple scales and automatically learns the channel-wise feature correlations to propagate more important features for better discriminative ability of our network. In our auxiliary network, we build an image pyramid by progressively downsampling the input image and conduct feature extraction along scale dimension. Such obtained features are further fused at different states to enrich scale-wise information and learn more comprehensive feature representations. Experimental results are shown on MS COCO dataset. Our DAIPNet achieves superior performance on MS COCO compared with other one-stage algorithms.

Our contributions can be summarized as follows:

- We propose an image pyramid network based on dual attention mechanism (DAIPNet) for one-stage object detection. Compared with state-of-the-art methods, experiments demonstrate the superiority of our DAIPNet.
- We design a dual attention mechanism that includes an adaptive feature fusion module (AFFM) and a progressive attention fusion module (PAFM). The AFFM aims balancing the information flow produced by the backbone and auxiliary network. The PAFM bridges the connections among different scales and model the channel-wise feature correspondences to pay attention to more informative features for object detection.
- We build an image pyramid within the constructed auxiliary network and extract the features from the downsampled images of different spatial resolution. By this way, we can further enrich the scale-wise information at different states to learn more comprehensive feature representations.

The remainder of this article is organized as follows. In Section 2, the related work on object detection is introduced. The proposed DAIPNet which includes dual attention mechanism and image pyramid network is presented in Section 3. The experimental results and comparisons with other methods are demonstrated in Section 4. Finally, the conclusion of this article is presented in Section 5.

2. Related Work

Early traditional object detection algorithms generally use sliding windows to capture the target areas on the image, and then classify them according to the extracted features from observed input images, which include shape, textures, and color *et al*.. Based on these features, some methods [20, 21, 22, 23] have been proposed for feature extraction to help the detection task. However, these methods involve critical manual intervention and thus lead to low detection accuracy. Recently, inspired the remarkable improvements of R-CNN [1] in object detection, many algorithms employ convolutional neural networks (CNN) to perform feature extraction and detection, which show obvious superiority than traditional methods. To improve the training and inference speed of R-CNN, in [2], Girshick proposes Fast R-CNN that adopts ROI pooling layer to fix the size of the features and multi-task learning strategy to constrain the bounding box regression. Moreover, Fast R-CNN feeds the entire image rather than only region proposals into the network for feature extraction, which requires less computation than R-CNN [1]. Faster R-CNN [3] replaces the selective search algorithm in [1, 2] with a region proposal network and integrates the generation of regional proposal, extraction of features and classification into a unified framework, thus realizing fully end-to-end training. There are also some methods that have been presented to further improve the detection performance, such as SNIP [24], R-FCN [25], and Cascade R-CNN [26], REN [27] *et al*..

At the same time, some approaches [8, 9, 10, 11, 28, 29] utilize one-stage framework to regress the coordinates and category probabilities of the objects, which achieve a good trade-off between efficiency and detection accuracy. YOLO [9] removes the candidate areas processing in R-CNN and predicts objects in the final output layer according to the whole image information, which requires relatively lower inference time. SSD [8] combines information from multiple feature maps of different scales to predict objects of various sizes. The authors also eliminate proposal generation and feature resampling stages to encapsulate the computational cost in a single network. Although the detection accuracy and speed are maintained to some extent, it is insufficient for the interaction between the different layers in terms of multi-scale object detection.

Considering the differences of the feature information in different resolution, existing methods [12, 13] use top-down feature pyramid network to capture the context information under different scales. PANet [14] employs a bottom-up structure on the basic of FPN [13] to retain the shallow detail information. Taking into account of the interactivity between multi-scale feature information, RetinaNet [30], ZigZagNet [15], MSPN [16], NETNet [31], WeaveNet [32] are subsequently proposed. In addition, some context modules [33, 34] are also used to enhance multi-scale information for object detection. In deep networks, as the layer becomes deeper, the feature flow will suffer from the loss of the location and spatial information caused by continuous convolution operations. This phenomenon can significantly affect the bounding box regression and thus results in features misalignment. To solve this problem, some methods [17, 18, 35] combine the pretrained detection network and an auxiliary network to inject complementary multi-level information into the detection network. In [17], Wang *et al*. proposes the LRF that utilizes a light-weight scratch network as the auxiliary network to supplement useful shallow information to the detection network. However, the auxiliary network in LRF [17] only downsamples images with a single scale, which ignores the importance and relevance of information from different scales in the process of information propagation. In contrast, our DAIPNet uses an image pyramid network to enrich the spatial information from multiple downsampled images with different scales. Furthermore, we present a dual attention mechanism to exploit the multi-scale features interaction by our bi-directional design between the detection backbone and auxiliary network.

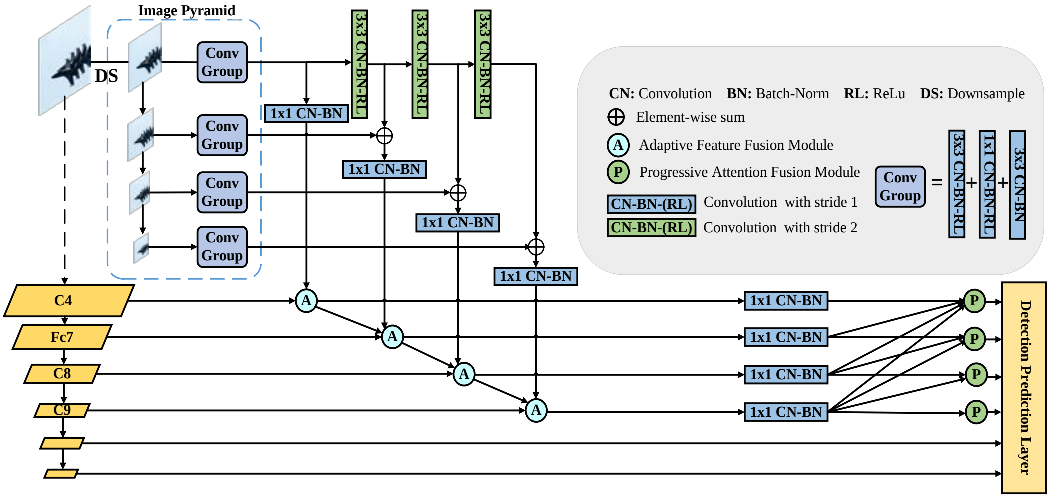


Image Pyramid
DS
Conv Group
Conv Group
Conv Group
Conv Group
3x3 CN-BN-RL
3x3 CN-BN-RL
3x3 CN-BN-RL
1x1 CN-BN
1x1 CN-BN
1x1 CN-BN
1x1 CN-BN
C4
Fc7
C8
C9
A
A
A
A
1x1 CN-BN
1x1 CN-BN
1x1 CN-BN
1x1 CN-BN
P
P
P
P
Detection Prediction Layer
CN: Convolution BN: Batch-Norm RL: ReLu DS: Downsample
⊕ Element-wise sum
A Adaptive Feature Fusion Module
P Progressive Attention Fusion Module
CN-BN-(RL) Convolution with stride 1
CN-BN-(RL) Convolution with stride 2
Conv Group = 3x3 CN-BN-RL + 1x1 CN-BN-RL + 3x3 CN-BN

Fig. 1. The overall architecture of our DAIPNet, which consists of SSD, auxiliary network and dual attention mechanism. The auxiliary network consists of a lightweight image pyramid. The dual attention mechanism includes an adaptive feature fusion module (AFFM) and a progressive attention fusion module (PAFM).

3. Proposed Method

In this section, we introduce the key components in our DAIPNet, which includes the proposed dual attention mechanism and our auxiliary network. **Fig. 1** illustrates the architecture of our DAIPNet. We use SSD with VGG [19] as the backbone which provides feature maps at different scales for prediction. The auxiliary network progressively learns scale-wise feature representations by a constructed image pyramid structure. The extracted features are further combined with the features produced by the backbone network to integrate complementary information. Moreover, we design the dual attention mechanism, which is composed of an adaptive feature fusion module (AFFM) and a progressive attention fusion module (PAFM), to efficiently interact with the information of the SSD backbone network and auxiliary network.

3.1 Dual Attention Mechanism

3.1.1 Adaptive Feature Fusion Module (AFFM)

Generally, the shallow features contain the location information of objects while the high-level features mainly retain the semantic information. When the information at different scales of the high and low layers are fused, they are adjusted to the same resolution and then added together. Nevertheless, in LRF [17], the overall structure integrates not only the contextual information of the SSD backbone network, but also the shallow spatial information provided by the auxiliary network. LRF takes the features provided by the auxiliary network as weights and multiplies them with the corresponding elements of the current prediction layer from the backbone network, which means that the elements of the backbone network are scaled one by one. Although such weight mapping reflects the importance of the information from the backbone network, continuous convolution and pooling operations in the backbone network have made small objects lose details seriously in the deep network.

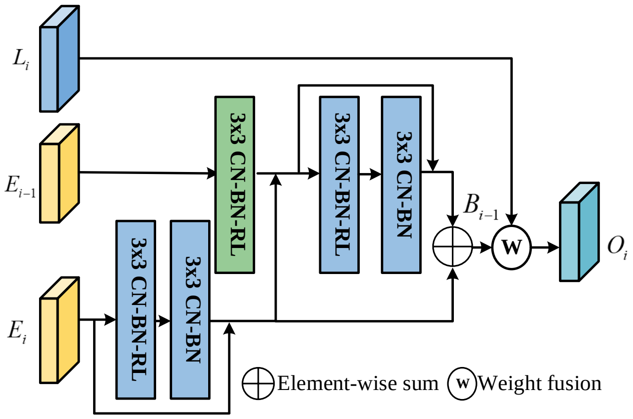


Fig. 2. The architecture of AFFM.

Considering that the auxiliary network is composed of a lightweight network with shallow spatial information, it can make up for the lack of deep semantics of the backbone network to some extent. To this end, we propose an adaptive feature fusion module (AFFM) to automatically balance the information flow with different importance from the SSD backbone and auxiliary network. As can be observed from **Fig. 2**, in AFFM, instead of employing the feature of the auxiliary network as a weight mapping, we try to dynamically treat the features of different layers through the attention exploration, in which network adaptively learns the weight coefficients of the auxiliary network and the SSD backbone network. Then the learned weight coefficients obtained are multiplied with the corresponding feature maps. Here, supposing $L \in \mathbb{R}^{H \times W \times C}$ is the feature from the auxiliary network, while $B \in \mathbb{R}^{H \times W \times C}$ refers to the feature from the SSD backbone network. We perform an element-wise sum operation to obtain the output O_i of the current layer for the i^{th} scale as follows:

$$O_i = \omega_i L_i + (1 - \omega_i) B_i \tag{1}$$

where ω_i is the weight coefficient with an initial value of 0 and adaptively updated during the training process to balance the importance of different layers for feature fusion.

In addition, to strengthen the connection between the high and low layers of the SSD backbone network for the integration of B_i, a residual block is designed to capture features from the mainstream feature E_i for obtaining effective semantic information. Here, $E_i \in \mathbb{R}^{h_i \times w_i \times c_i}$ is the i^{th} feature output by the SSD backbone network. In order to combine the mainstream features E_i with the reference features E_{i-1} from the previous layer of the backbone network, we downsample E_{i-1} to the same scale as E_i, and then capture texture information by a similar residual block. After that, the extracted features and the semantic information from E_i are fused through a short connection to learn more comprehensive feature representations. In our implementation, we explain this process by:

$$B_i = \mathcal{R}(\mathcal{D}(E_{i-1}) + \mathcal{R}(E_i)) + \mathcal{R}(E_i) \tag{2}$$

where $\mathcal{R}$ represents the function of the residual block, $\mathcal{D}(\cdot)$ is a downsampling operation to match features from different scales. In summary, the context feature of the backbone network and the feature of the auxiliary network are interacted through the attention and residual blocks to obtain the final output O_i as:

$$O_i = \omega_i L_i + (1 - \omega_i)\left(\mathcal{R}(\mathcal{D}(E_{i-1}) + \mathcal{R}(E_i)) + \mathcal{R}(E_i)\right) \tag{3}$$

3.1.2 Progressive Attention Fusion Module (PAFM)

Considering that the channel mapping between features is interdependent and different semantic information is also related to each other, we present to enhance the feature representations by emphasizing the channel-wise interdependence through attention mechanism. Therefore, we design a progressive attention fusion module (PAFM) to adaptively learn correlation between channels by progressively fusing multi-scale features. PAFM aligns the features $X \in \mathbb{R}^{H \times W \times C}$ from different scales to a specific scale using upsampling operation, and then performs a concatenation operation on them by a progressive manner to obtain $X^{cat} \in \mathbb{R}^{H \times W \times C}$.

$$X_i^{cat} = Cat\big(\mathcal{U}(X_{i+1}), \mathcal{U}(X_{i+2}), \ldots, \mathcal{U}(X_n)\big), i \in [1, n-1] \tag{4}$$

where i denotes the i^{th} scale and $n = 4$ denotes the number of feature pyramid levels selected for context interaction. $Cat(\cdot)$ and $\mathcal{U}(\cdot)$ denote the concatenation and the upsampling operation, respectively. Then, global average pooling (GAP) is used to obtain the global descriptor $\boldsymbol{z}_i \in \mathbb{R}^{nC}$ from X_i^{cat} with the size of 1 × 1 × nC.

$$z_{i,c} = GAP\big(X_{i,c}^{cat}\big) = \frac{1}{H \times W} \sum_{m=1}^{H} \sum_{n=1}^{W} X_{i,c}^{cat}(m, n) \tag{5}$$

where $X_{i,c}^{cat}$ represents the c^{th} element of X_i^{cat} at the position (m, n).

After that, two 1×1 convolutions, $\mathcal{C}_{1\times1}^{1}$ and $\mathcal{C}_{1\times1}^{2}$, are adopted to form a gate mechanism X_i^g for dimensionality reduction and increasing. Here, $\mathcal{C}_{1\times1}^{1}$ reduce the dimensionality of features and $\mathcal{C}_{1\times1}^{2}$ adjust the feature to the same size as $\boldsymbol{z}_i$:

$$X_i^g = \mathcal{C}_{1\times1}^{2}(ReLU(\mathcal{C}_{1\times1}^{1}(\boldsymbol{z}_i))) \tag{6}$$

where $X_i^g \in \mathbb{R}^{nC}$ is the feature after dimension change through the $\mathcal{C}_{1\times1}^1$ and $\mathcal{C}_{1\times1}^2$. Next, we adopt the sigmoid function σ to get the channel weight mapping and multiply it with the input X_i^{cat} to allocate each channel a weight. The process is shown as following:

$$X_i^{att} = \sigma(X_i^g) \cdot X_i^{cat} \tag{7}$$

where X_i^{att} is the i^{th} output feature processed by the channel-wise multiplication.

During the training processing, the network filters unimportant channel values and captures informative features around objects by adaptively learning the correlation between channels of different scale features. Furthermore, considering that the feature information on the current scale is superior to predicting objects of this scale than other scales, as shown in the **Fig. 3**, we treat the features of the current layer as a separate branch, fuse it with X_i^{cat} to obtain the intermediate result X_i^{inter} though a 1×1 convolution and a concatenation operation, and then X_i^{inter}needs to be concatenated with X_i^{att}, thus ensuring the dominance of the current layer while emphasizing the important channel values.

$$X_i^{inter} = Cat(\mathcal{C}_{1\times1}^3(X_i),\ \mathcal{C}_{1\times1}^4(X_i^{cat})) \tag{8}$$

$$P_i = Cat(\mathcal{C}_{1\times1}^5(X_i^{inter}),\ \mathcal{C}_{1\times1}^6(X_i^{att})) \tag{9}$$

where P_i denotes the i^{th}output feature for final prediction, $\mathcal{C}_{1\times1}^n (n = 3, 4, 5, 6)$ represents the convolution operation that is designed to match the features with different resolution.

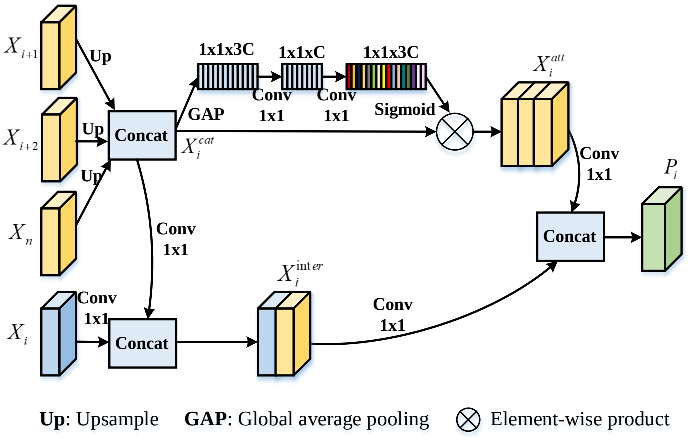


Fig. 3. The architecture of PAFM.

3.2 Image Pyramid Network (IPN)

In object detection, the image pyramid is often used to solve the scale problem. In the early days, the objects of different sizes in the images can be detected by changing the form of sliding window. Recently, it is common to use the same sliding window and change the image size to detect objects with inconsistent scales in the images. Since the information conveyed by the image on each scale is different, previous method [17] merely introduces single-scale images as supplement to enrich the detailed information that is lost in the backbone network. Therefore, in our framework, we build an image pyramid network (IPN) to enhance the spatial information of the auxiliary network. For an input image I as in **Fig. 1**, we continuously downsample it to obtain an image pyramid network $I_p = \{I_p^1, I_p^2, \dots, I_p^i\}$, where i represents the number of the image pyramid network levels. The downsampling process is constructed as:

$$I_p^1 = \mathcal{D}(I_p)$$

$$
\begin{aligned}
I_p^2 &= \mathcal{D}(I_p^1) \\
&\vdots \\
I_p^i &= \mathcal{D}(I_p^{i-1})
\end{aligned}
\tag{10}
$$

where $\mathcal{D}(\cdot)$ denotes the downsampling operation that pools the input I to the same scale as the prediction layers in [8]. Then we adopt a lightweight convolution group $\mathcal{G}$, including two 3×3 convolutions and one 1×1 convolution, to extract shallow features F_p^i from all downsampled images I_p^i as:

$$F_p^i = \mathcal{G}(I_p^i) \tag{11}$$

Furthermore, in order to maximize the contribution of F_p^i, we also use the iterative 3×3 convolution with stride 2 to obtain the features $M_n^1 = \{M_1^1, M_2^1, \dots, M_n^1\}$ of the same size as the SSD prediction layers for the first scale feature of the image pyramid. Here, we set the n as 3:

$$
\begin{aligned}
M_1^1 &= \mathcal{C}_{3\times3}^1(F_p^1) \\
M_2^1 &= \mathcal{C}_{3\times3}^2(M_1^1) \\
M_3^1 &= \mathcal{C}_{3\times3}^3(M_2^1)
\end{aligned}
\tag{12}
$$

where $\mathcal{C}_{3\times3}^n$ represents the convolution with stride 2. Finally, we perform an element-wise sum operation between F_p^i and M_n^1 to achieve the comprehensive feature representation L_i.

$$L_i = \begin{cases} F_p^i, & i = 1 \\ F_p^i + M_n^1, & where\ n = i - 1\ and\ i > 1 \end{cases} \tag{13}$$

4. Experiments

4.1 Datasets

4.1.1 MS COCO

In terms of object detection, MS COCO [36] is the most common dataset that has 143k images and 80 object categories. These images are divided into three parts: 118k images for training, 20k images for testing and 5k images for validation. In addition, CodaLab provides a special evaluation standard for MS COCO dataset and gives 6 Average Precision (AP) indicators according to IOU threshold and object size.

4.1.2 PASCAL VOC

PASCAL VOC 2007 and PASCAL VOC 2012 are two classic datasets in PASCAL VOC [37] which contains 20 different object categories for classification and detection. The models are usually trained on the trainval set with 16k images provided by VOC2007 and VOC2012 and tested on the PASCAL VOC 2007 test set with 5k images. PASCAL VOC provides an official indicator, the mean average accuracy (mAP), to measure the accuracy of the models.

4.2 Implementation Details

All our experiments are implemented based on Pytorch framework and Titan Xp GPUs. During the experiments, we employ VGG-16 pre-trained on ImageNet [38] to initialize our network. We set the initial learning rate to 4×10^{-3}. During the training process, the learning rate decreases by a factor of 0.1 at 90 epochs, 120 epochs, and 140 epochs for the MS COCO dataset. In PASCAL VOC, the same learning rate decreases at 150 and 200 epochs. Besides, to make the later training of the model more stable, we adopt the warmup strategy following [33]. Specifically, the learning rate gradually increases from 10^{-6} to the 4×10^{-3} during the first 6 epochs. We usually set the weight decay as 5×10^{-4}, the momentum as 0.9 and the batch size as 32. Here, for different input sizes, we adjust the value of batch size due to the limitation of GPUs. In other aspects, our loss function, data enhancement and other training details are the same as SSD [8]. We obtain superior detection accuracy when the model is trained to 160 epochs on MS COCO and 250 epochs on PASCAL VOC, respectively.

4.3 MS COCO Dataset

4.3.1 Comparing with State-of-the-Arts

We evaluate our approach on MS COCO test-dev dataset and compare with state-of-the-art object detection approaches. **Table 1** shows the detection results. For a 300×300 input, our DAIPNet obtains the mAP of 32.6, which surpasses about 0.6 compared with the most state-of-the-art method LRF. In addition, the performance for AP_{75} is largely improved by 0.9 while the performance for AP_{50} is improved by 0.4, which demonstrate the effectiveness of DAIPNet. Even on medium and small objects, our method can achieve 1.5 and 0.3 gains in terms of AP_m and AP_s. In contrast, RefineDet [34], EFIP [18] and RFBNet [33] achieve AP scores of 29.4, 30.0 and 30.3, respectively. Our detector obtains more accurate results than these methods under the same backbone. Besides, compared with the recently proposed NETNet [31], our method far surpasses NETNet on medium objects and large objects. Especially for the medium objects, our method is nearly 2.0 better than NETNet.

For a 512×512 input, as shown in **Table 1**, our DAIPNet is still significantly superior than LRF with the same input size and backbone. Besides, though the RetinaNet+AP-loss [39] performs slightly better than our method, due to the large-size input image (832×500) and more complex backbone ResNet101-FPN, it spends twice time cost as our method. Meanwhile, compared with other one-stage detection algorithms, the detection accuracy in **Table 1** demonstrates the superior performance of our DAIPNet on medium objects.

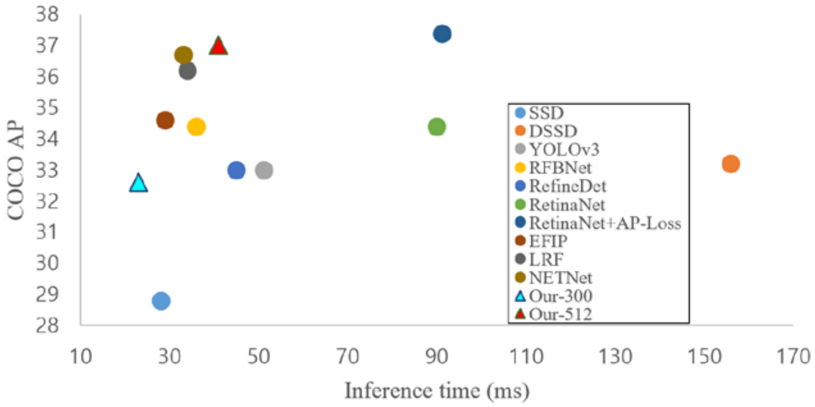


Fig. 4. Speed (ms) vs. accuracy (mAP) on MS COCO test-dev.

Moreover, we show the comparison of inference speed and detection accuracy with other detectors in **Fig. 4**. The blue and red triangles represent the detection results under different input sizes (300×300, 512×512), and the circles of different colors are the results of other methods with an input of 512×512. It can be seen intuitively from **Fig. 4** that our DAIPNet is significantly better than other detectors in balancing detection accuracy and inference speed.

Table 1. Comparison of our DAIPNet on MS COCO test-dev. We test the time on a Titan Xp. * represents the time obtained by testing in the same environment with DAIPNet.

Method	Backbone	size	Time(ms)	AP	AP$_{50}$	AP$_{75}$	AP$_{s}$	AP$_{m}$	AP$_{l}$
Two-stage Method:									
Faster R-CNN [3]	VGG-16	1000×600	147	24.2	45.3	23.5	7.7	26.4	37.1
Mask R-CNN [4]	ResNeXt-101-FPN	1280×800	210	39.8	62.3	43.4	22.1	43.2	51.2
FPN [13]	ResNet-101-FPN	1000×600	240	36.2	59.1	39.0	18.2	39.0	48.2
Cascade R-CNN [26]	ResNet-101-FPN	1280×800	141	42.8	62.1	46.3	23.7	45.5	55.2
One-stage Method:									
SSD [8]	VGG-16	300×300	12	25.3	42.0	26.5	6.2	28.0	43.3
DSSD [12]	ResNet-101	321×321	-	28.0	46.1	29.2	7.4	28.1	47.6
RFBNet [33]	VGG-16	300×300	15(21*)	30.3	49.3	31.8	11.8	31.9	45.9
RefineDet [34]	VGG-16	320×320	26	29.4	49.2	31.3	10.0	32.0	44.4
EFIP [18]	VGG-16	300×300	14 (22*)	30.0	48.8	31.7	10.9	32.8	46.3
LRF [17]	VGG-16	300×300	13 (20*)	32.0	51.5	33.8	12.6	34.9	47.0
NETNet [31]	VGG-16	300×300	18	32.0	51.5	33.6	**13.9**	34.5	46.2
Ours	VGG-16	300×300	23	**32.6**	**51.9**	**34.7**	12.9	**36.4**	**47.6**
SSD [8]	VGG-16	512×512	28	28.8	48.5	30.3	10.9	31.8	43.5
DSSD [12]	ResNet-101	513×513	156	33.2	53.3	35.2	13.0	35.4	51.1
YOLOv3 [11]	DarkNet-53	608×608	51	33.0	57.9	34.4	18.3	35.4	41.9
RFBNet [33]	VGG-16	512×512	33 (36*)	34.4	55.7	36.4	17.6	37.0	47.6
RefineDet [34]	VGG-16	512×512	45	33.0	54.5	35.5	16.3	36.3	44.3
RetinaNet [30]	ResNet-101-FPN	832×500	90	34.4	55.7	36.8	14.7	37.1	47.4
RetinaNet+AP-Loss [39]	ResNet-101-FPN	832×500	91	**37.4**	**58.6**	**40.5**	17.3	40.8	**51.9**
EFIP [18]	VGG-16	512×512	29	34.6	55.8	36.8	18.3	38.2	47.1
LRF [17]	VGG-16	512×512	26 (34*)	36.2	56.6	38.7	19.0	39.9	48.8
NETNet [31]	VGG-16	512×512	33	36.7	57.4	39.2	**20.2**	39.2	49.0
Ours	VGG-16	512×512	41	37.0	57.5	39.7	19.8	**42.3**	48.8

4.3.2 Ablation Studies

In order to verify the effectiveness of the image pyramid network (IPN) and the dual attention mechanism, we conduct a series of experiments on the MS COCO validation set with a 300 × 300 input. We first train a baseline model without any proposed component and adopt LRF as reference for comparisons, where the results are shown in Table 2. We first validate the AFFM in the dual attention mechanism. As we can see, AFFM improves the baseline by 0.3 AP. Among them, the detection performance of large objects increases by 0.9 AP, and the detection performance increases by 0.7 on medium objects. It benefits from that AFFM automatically learns the importance of different information through the attention mechanism and focuses on our contextual information simultaneously. At the same time, it also proves that our AFFM has better feature extraction capabilities, which can make a trade-off of information between the backbone and auxiliary network.

It also can be observed that AFFM makes detection accuracy of the small objects drop slightly. It may be that the network emphasizes the feature of the large objects and weakens the expression of the small objects relatively when using the residual block for context interaction. As for IPN, in **Table 2**, we can observe that the model using IPN can obtain obvious performance improvement on small objects (12.6 to 13.3 AP), which means the complementary information from IPN benefits the feature maps at low levels and improves the feature representation.

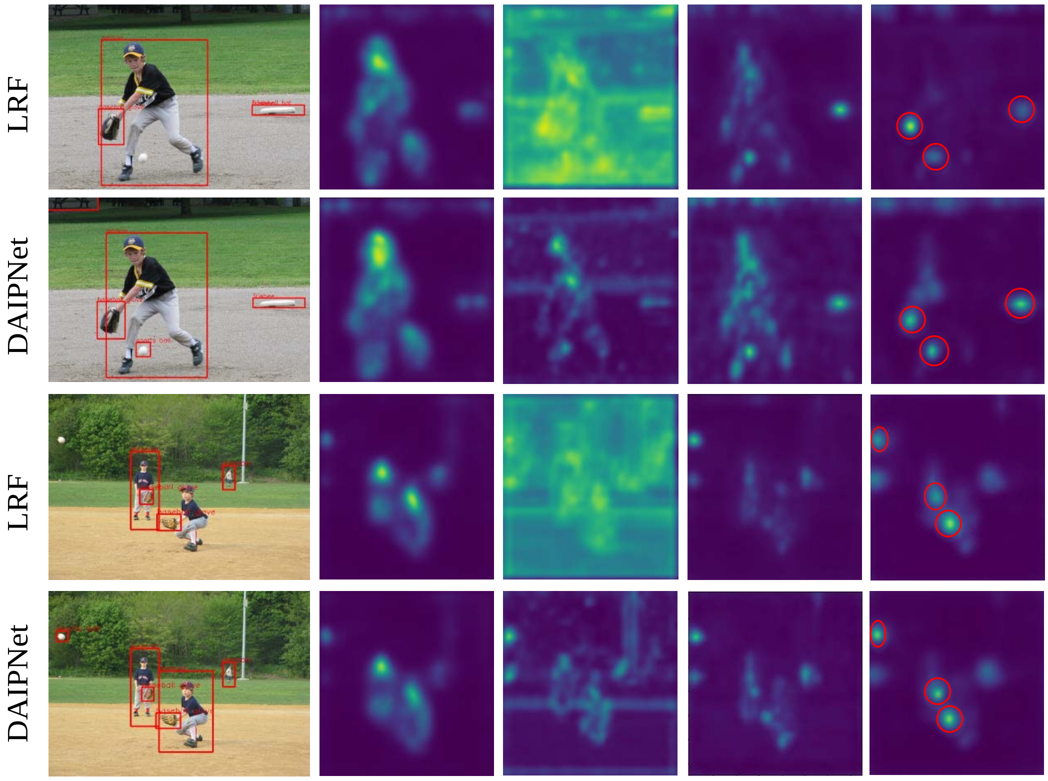
Table 2. Ablation Studies on MS-COCO val.

LRF	AFFM	IPN	PAFM	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
√				31.9	51.4	33.6	13.4	36.3	47.6
√	√			32.2	51.4	34.0	12.6	37.0	48.5
√	√	√		32.2	51.1	34.2	13.3	37.0	48.6
√	√	√	√	32.4	51.5	34.3	13.7	37.3	48.8

Moreover, we visualize the features associated with predicting small objects as shown in **Fig. 5**. All features are extracted and visualized on the same scale. The first column (a) shows the detection results of LRF and our DAIPNet. The second (b) and third (c) columns represent features from the backbone and auxiliary networks, respectively. The fourth column (d) shows the output features via the first AFFM module. The last column (e) represents the output features via the first PAFM module for prediction.

As shown in **Fig. 5**, for the baseball ignored by LRF in the first column (a), it can be seen from the column (e) that the corresponding feature of the baseball is not obvious compared with other features. However, in our DAIPNet, since the feature of baseball is treated equally with the other little features in column (e), the baseball can be accurately detected by our network in column (a). Besides, in our DAIPNet, the features of objects (people, baseball, glove) from the auxiliary network contain abundant spatial information, and the features of the column (d) have clearer outlines and more details than the features of the backbone network. It indicates that the auxiliary network compensates for the tiny features lost by the backbone network, and further proves that IPN can focus on more medium and small objects to improve the feature representation and AFFM can fully integrate the information the backbone and auxiliary network to learns the importance of different information.

On the basis of the AFFM and IPN, the experimental result of employing the progressive attention fusion module (PAFM) has increased by 0.5 AP compared with LRF. In **Table 2**, there is a slight improvement in each metric for calculating AP, which confirms the effectiveness of our PAFM. The detection accuracy is improved because PAFM allows the network to adaptively learn the correlation and importance between channels to some extent. Besides, this attention mechanism enables the high-level features containing more information of semantic details to interact with low-level features which contain more spatial information. Furthermore, **Fig. 6** shows detection examples, which verifies that our DAIPNet has a good detection performance for objects of different size.



(a) Detection results (b) Backbone (c) Auxiliary (d) AFFM (e) Prediction

Fig. 5. The detection results and feature visualization of LRF and our DAIPNet. The first column (a) shows the detection results of LRF and our DAIPNet. The second (b) and third (c) columns represent features from the backbone and auxiliary network, respectively. The fourth column (d) shows the output feature via the first AFFM module. The last column (e) represents the output features via the first PAFM module for prediction.

4.4 PASCAL VOC Dataset

Here, we validate our DAIPNet on PASCAL VOC 2007 dataset [37]. **Table 3** shows the performance comparison with other detection methods on the VOC 2007 test set. For two-stage detectors, R-FCN [25] achieves excellent detection performance due to the larger input size and the deeper backbone ResNet-101. Among one-stage detectors, the detection results of LRF [17] are trained and tested in the same environment with DAIPNet. For a 300 × 300 input, our DAIPNet obtains the mAP of 80.0 that is the same as the detection result of RefineDet [34]. However, the input size of RefineDet is 320 × 320, which is larger than our input size. Besides, when adopting a 512 × 512 input, our DAIPNet is 0.4 better than RefineDet. For a 512 × 512 input, our DAIPNet achieves AP scores of 82.2 that is slightly higher than other detectors, which also verifies the effectiveness of our network to a certain extent. **Fig. 7** shows detection examples for our approach on PASCAL VOC 2007 test set.

Table 3. Comparison of our DAIPNet on PASCAL VOC 2007 test set. * represents the time obtained by training and testing in the same environment with DAIPNet.

Method	Backbone	Input size	mAP
Two-stage Method:			
Faster R-CNN [3]	ResNet-101	1000×600	76.4
R-FCN [25]	ResNet-101	1000×600	80.5
One-stage Method:			
SSD [8]	VGG-16	300×300	77.2
DSSD [12]	ResNet-101	321×321	78.6
RefineDet [34]	VGG-16	320×320	80.0
WeaveNet [32]	VGG-16	320×320	79.7
DES [29]	VGG-16	300×300	79.7
DFPR [28]	VGG-16	300×300	79.6
LRF [17]	VGG-16	300×300	79.8*
Ours	VGG-16	300×300	**80.0**
SSD [8]	VGG-16	512×512	79.5
DSSD [12]	ResNet-01	513×513	81.5
RefineDet [34]	VGG-16	512×512	81.8
DES [29]	VGG-16	512×512	81.7
DFPR [28]	VGG-16	512×512	81.1
RFBNet [33]	VGG-16	512×512	82.1
LRF [17]	VGG-16	512×512	82.0*
Ours	VGG-16	512×512	**82.2**

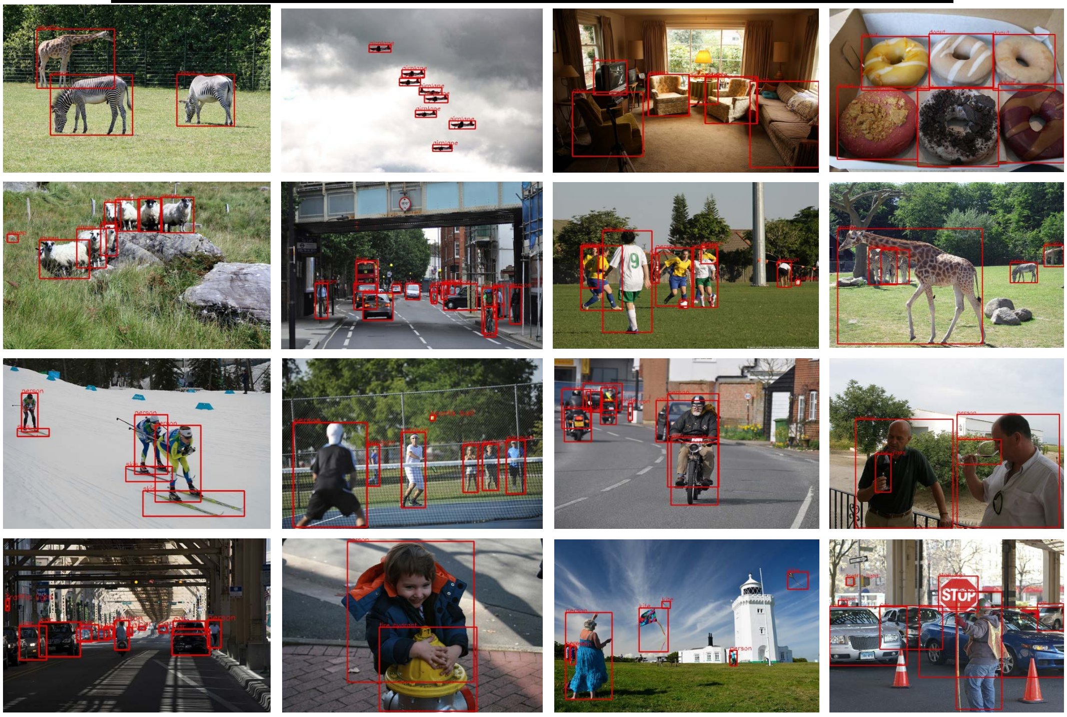


Fig. 6. Qualitative detection results of our DAIPNet on MS COCO test-dev.

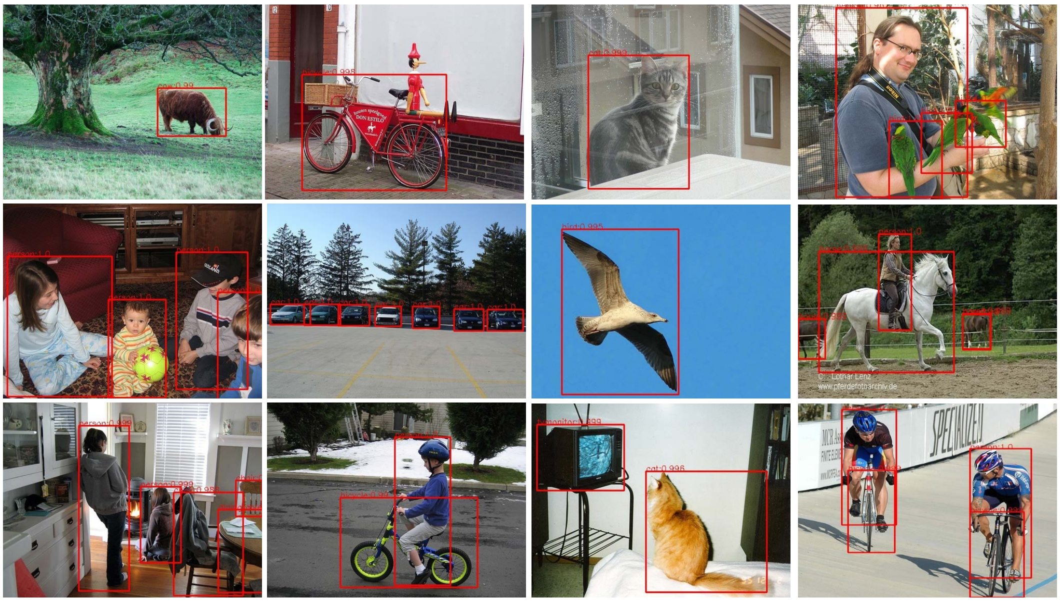


Fig. 7. Qualitative detection results of our DAIPNet on PASCAL VOC 2007 test set.

5. Conclusion

In this paper, we discuss the inherent problems in LRF caused by insufficient interaction of multi-scale features. Based on the observation, we propose an image pyramid network based on dual attention mechanism (DAIPNet) to further explore the effect of different scale features. By integrating dual attention mechanism and the image pyramid structure, DAIPNet improve the detection performance by a large margin on MS COCO. In this work, the proposed method aims at tackling general object detection task. However, compared with the detection results of medium and large objects, the model shows limited improvements on small objects. Furthermore, our study finds out that the loss of feature information is inevitable with the deepening of convolution network and the change between scales. In the future, we will do more attempts on network designing and better feature representation learning for small object detection.

References

[1] R. Girshick, J. Donahue, T. Darrell, et al., “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014. Article (CrossRef Link)

[2] R. Girshick, “Fast r-cnn,” in *Proc. of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015. Article (CrossRef Link)

[3] S. Ren, K. He, R. Girshick, et al., “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, 2017. Article (CrossRef Link)

[4] K. He, G. Gkioxari, P. Dollár, et al., “Mask r-cnn,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386-397, 2020. Article (CrossRef Link)

[5] M. Aamir, Y.-F. Pu, Z. Rahman, W.A. Abro, Z. Hu, F. Ullah, and A. M. Badr, “A Hybrid Proposed Framework for Object Detection and Classification," *Journal of Information Processing Systems 14*, no. 5, 2018. Article (CrossRef Link)

[6] M. Aamir, Y.-F. Pu, Z. Rahman, W.A. Abro, H. Naeem, Z. Rahman, “A hybrid approach for object proposal generation,” in *Proc. of International Conference on Sensing and Imaging*, 506, 251-259, 2017. Article (CrossRef Link)

[7] Y. Guan, M. Aamir, Z. Rahman, A. Ali, W.A. Abro, Z. A. Dayo, M. S. Bhutta, Z. Hu, “A framework for efficient brain tumor classification using MRI images,” *Mathematical Biosciences and Engineering*, 18(5), 5790-5815, 2021. Article (CrossRef Link)

[8] W. Liu, D. Anguelov, D. Erhan, et al., “Ssd: Single shot multibox detector,” in *Proc. of European Conference on Computer Vision*, pp. 21–37, 2016. Article (CrossRef Link)

[9] J. Redmon, S. Divvala, R. Girshick, et al., “You only look once: Unified, real-time object detection,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016. Article (CrossRef Link)

[10] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263-7271, 2017. Article (CrossRef Link)

[11] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv: 1804.02767*, 2018. Article (CrossRef Link)

[12] C.-Y. Fu, W. Liu, A. Ranga, et al., “Dssd: Deconvolutional single shot detector,” *arXiv preprint arXiv:1701.06659*, 2017. Article (CrossRef Link)

[13] T. Y. Lin, P. Dollar, R. Girshick, et al., “Feature pyramid networks for object detection,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117-2125, 2017. Article (CrossRef Link)

[14] S. Liu, L. Qi, H. Qin, et al., “Path aggregation network for instance segmentation,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759-8768, 2018. Article (CrossRef Link)

[15] D. Lin, D. Shen, S. Shen, et al., “Zigzagnet: Fusing top-down and bottom-up context for object segmentation,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7490-7499, 2019. Article (CrossRef Link)

[16] W. Li, Z. Wang, B. Yin, et al., “Rethinking on multi-stage networks for human pose estimation,” *arXiv preprint arXiv:1901.00148*, 2019. Article (CrossRef Link)

[17] T. Wang, R. M. Anwer, H. Cholakkal, et al., “Learning rich features at high-speed for single-shot object detection,” in *Proc. of the IEEE International Conference on Computer Vision*, pp. 1971–1980, 2019. Article (CrossRef Link)

[18] Y. Pang, T. Wang, R. M. Anwer, et al., “Efficient featurized image pyramid network for single shot detector,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7336-7344, 2019. Article (CrossRef Link)

[19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. Article (CrossRef Link)

[20] N. Dalal, B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. of the IEEE* Computer Society *Conference on Computer Vision and Pattern Recognition*, pp. 886-893, 2005. Article (CrossRef Link)

[21] T. Ojala, M. Pietikainen, D. Harwood, “Performance evaluation of texture measures with classification based on kullback discrimination of distributions,” in *Proc. of 12th International Conference on Pattern Recognition*, pp. 582-585, 1994. Article (CrossRef Link)

[22] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Proc. of the Alvey Vision Conference*, pp. 23.1-23.6, 1988. Article (CrossRef Link)

[23] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. Article (CrossRef Link)

[24] B. Singh and L. S. Davis, “An analysis of scale invariance in object detection - snip,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3578-3587, 2018. Article (CrossRef Link)

[25] J. Dai, Y. Li, K. He, et al., “R-FCN: object detection via region-based fully convolutional networks,” *arXiv preprint arXiv:1605.06409*, 2016. Article (CrossRef Link)

[26] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018. Article (CrossRef Link)

[27] Y. Guan, M. Aamir, Z. Hu, W.A. Abro, Z. Rahman, Z.A. Dayo, S. Akram, “A region-based efficient network for accurate object detection,” *Traitement du Signal*, 38(2), 481-494, 2021. Article (CrossRef Link)

[28] T. Kong, F. Sun, C. Tan, H. Liu, and W. Huang, “Deep feature pyramid reconfiguration for object detection,” in *Proc. of the European Conference on Computer Vision*, 2018. Article (CrossRef Link)

[29] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, “Single-shot object detection with enriched semantics,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. Article (CrossRef Link)

[30] T.-Y. Lin, P. Goyal, R. Girshick, et al., “Focal loss for dense object detection,” in *Proc. of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017. Article (CrossRef Link)

[31] Y. Li, Y. Pang, J. Shen, et al., “Netnet: Neighbor erasing and transferring network for better single shot object detection,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13346–13355, 2020. Article (CrossRef Link)

[32] Y. Chen, J. Li, B. Zhou, J. Feng, and S. Yan, “Weaving multi-scale context for single shot detector,” *arXiv preprint arXiv: 1712.03149*, 2017. Article (CrossRef Link)

[33] S. Liu, D. Huang, Y. Wang, “Receptive field block net for accurate and fast object detection,” in *Proc. of the European Conference on Computer Vision*, pp. 404-419, 2018. Article (CrossRef Link)

[34] S. Zhang, L. Wen, X. Bian, et al., “Single-shot refinement neural network for object detection,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4203–4212, 2018. Article (CrossRef Link)

[35] Z. Liu, G. Gao, L. Sun, et al., “Ipg-net: Image pyramid guidance network for small object detection,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1026-1027, 2020. Article (CrossRef Link)

[36] T.-Y. Lin, M. Maire, S. Belongie, et al., “Microsoft coco: Common objects in context,” in *Proc. of European Conference on Computer Vision*, pp. 740-755, 2014. Article (CrossRef Link)

[37] M. Everingham, S. Eslami, L. V. Gool, C. Williams, J. Winn, A. Zisserman, “The pascal visual object classes challenge: a retrospective,” *International Journal of Computer Vision*, 111(1), 98-136, 2015. Article (CrossRef Link)

[38] J. Deng, W. Dong, R. Socher, et al., “Imagenet: A large-scale hierarchical image database,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009. Article (CrossRef Link)

[39] K. Chen, J. Li, W. Lin, et al., “Towards accurate one-stage object detection with ap-loss,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5119-5127, 2019. Article (CrossRef Link)

Xiang Dong received the B.S. degree in Electronic Information Engineering in 2019 from Hebei University. She is currently pursuing the M.S. degree in the Institute of Information Science, Beijing Jiaotong University. She works in object detection and deep learning.

Feng Li received his B.S. degree in Anhui Normal University, China, in 2016. Now, he is pursuing his Ph. D degree in Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests are image and video compression, image and video super resolution, computer vision and deep learning.

Huihui Bai received her B.S. degree from Beijing Jiaotong University, China, in 2001, and her Ph.D. degree from Beijing Jiaotong University, China, in 2008. She is currently a professor in Beijing Jiaotong University. She has been engaged in R&D work in video coding technologies and standards, such as HEVC, 3D video compression, multiple description video coding (MDC), and distributed video coding (DVC).

Yao Zhao received the B.S. degree from Fuzhou University, China, in 1989, and the ME degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the PhD degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), China, in 1996. He became an associate professor at BJTU in 1998 and became a professor in 2001. From 2001 to 2002, he was a senior research fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the editorial boards of several international journals, including as associate editors of IEEE Transactions on Cybernetics, IEEE Signal Processing Letters, and an area editor of Signal Processing: Image Communication (Elsevier), etc. He was named a distinguished young scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He is a senior member of the IEEE.