IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 December 2023, accepted 8 January 2024, date of publication 11 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3352601

RESEARCH ARTICLE

Object Detection of Remote Sensing Image Based on Multi-Scale Feature Fusion and Attention Mechanism

ZUOQIANG DU[1] AND YUAN LIANG[2]
[1]School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China
[2]Jinan Inspur Data Technology Company Ltd., Jinan 250000, China

Corresponding author: Yuan Liang (yuanliang_hrb@163.com)

This work was supported in part by the Natural Science Foundation of China under Grant 606750192, in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province under Grant UNPYSCT-2020212, and in part by the Science Foundation of Harbin Commerce University under Grant XL0095.

ABSTRACT In view of the small size and dense distribution of remote sensing image targets, this paper adds a detection head P2 specifically for small-scale targets on the basis of the three detection layers of the original YOLOv5 model, and involves the shallow high-resolution feature map in the subsequent multi-scale feature fusion module. The problem of losing the key feature information of the small-scale target in the process of multiple downsampling is effectively avoided. Firstly, an enhanced multi-scale feature fusion pyramid network DSI-FPN is designed. The FPN+PAN network is optimized by using DepthwiseSparable Convolution and Involution operators with fewer parameters and computations, as well as a spatial attention mechanism to generate feature graphs with richer information for network detection tasks. Secondly, we propose an adaptive channel spatial attention mechanism SCBAM, which introduces a self-attention mechanism into CBAM module to add non-local information to the interaction that originally had only local information, breaks the convolution kernel limit, expands the model receptive field, and improves the feature expression ability of the model. Thirdly, in order to solve the problem of insufficient computing power when deploying the target detector for equipment, we propose a network knowledge distillation framework for joint teachers based on the feature layer. The distillation loss of teacher is designed, and the trend of student online learning is adjusted dynamically by balancing the contributions of teacher network and truth value. The detection accuracy of the student network is obviously improved, and the parameters and model size of the network are effectively reduced. Finally, Comparing with other remote sensing image object detection methods, the experimental results show that the approach presented has better detection effect for small-scale targets of remote sensing images under different lighting conditions. The detection accuracy reached 43.9%, and 7.4% higher than that of the original model. After knowledge distillation, the model parameters are reduced to 1/3 of the original, and the detection accuracy is 40.2%.

INDEX TERMS Deep Learning, object detection, remote sensing image, multi-scale feature fusion pyramid network, adaptive channel spatial attention mechanism, joint teacher knowledge distillation.

I. INTRODUCTION

Remote sensing image analysis detection is to use the image data obtained by remote sensing technology to classify ground objects and detect changes, so as to understand the status of natural and human activities on the earth's surface [1], [2]. The feature information of the object in the image, such as contour, texture and color are manually extracted by traditional remote sensing image processing [3], [4]. Due to certain challenges in diverse application scenarios and the handling of massive data, most remote sensing image analyses currently employ methods based on Machine

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

Learning (ML) and Deep Learning (DL). These methods leverage large volumes of sample data for training, enabling the automatic extraction of ground object information and reducing manual intervention [5], [6].However, the remote sensing image analysis has a strong dependence on sample data. In addition, because the existing methods need to carry out complex processes such as model training and parameter adjustment, they need a lot of computing resources and time, and the operation efficiency is low [7].

Object detection is a key task in the field of computer vision, which aims to automatically identify and locate objects in an image. Haar–Like features [8], Histogram of Oriented Gradient (HOG) [9], Local Binary Pattern (LBP) [10] and Support Vector Machines (SVM) [11] are included the early object detection techniques. The traditional methods are to use sliding window for region selection, which not only wastes time, but also produces a lot of useless features. The operators used for feature extraction are manually set and unchanged, and the less semantic features should be extracted [12]. In 1998, Yann Lecun et al. [13] proposed the LeNet-5 Convolutional Neural Network (CNN) model, which included a 7-layers network structure and used the alternating structure of convolutional layer and pooling layer to extract image features, replacing the traditional manual feature extraction method. With the emergence of ImageNet dataset [14], DL has received unprecedented attention in the field of computer vision, and a large number of neural network models have emerged. In the visual recognition challenge based on the ImageNet dataset, DL methods with ultra-high precision have also emerged. Among them, AlexNet [15] won the first place with an overwhelming advantage in the image classification competition of ImageNet 2012, which was highly valued by the academic community.

Object detection algorithms based on DL can generally be divided into two categories: the two-stages object detection algorithm represented by Faster R-CNN and the single-stage object detection algorithm represented by YOLO and SSD series. Two-stages object detection algorithm usually has higher target recognition and location accuracy, while single-stage object detection algorithm achieves higher inference speed. Girshick et al. [16] proposed an algorithm named R-CNN, which divided object detection into two stages: selective search for generating candidate regions and Non-Maximum Suppression (NMS) for border regression strategy. In the classical CNNs, the size of the input image is fixed, which makes it no flexible enough to process images of different sizes. In order to solve this problem, He et al. [17] introduced a Spatial Pyramid Pooling (SPP) technology into the VGG network in 2015. After the convolution pooling of input images of any sizes is processed, the SPP method is used to obtain the same size feature layer. It allows CNN to process input image of any sizes while maintaining the same feature dimensions. In the same year, Girshick et al. [18] proposed an improved Fast R-CNN algorithm. On the basis of R-CNN, the introduction of RoI pooling layer and shared convolutional feature map technology are introduced for optimization. Fast R-CNN integrates feature extraction, target classification and border regression into a network, which greatly improves the speed and accuracy of detection. The proposal of Fast R-CNN lays a foundation for the practical application of two-stages object detection algorithm. In 2016, Ren et al. [19] proposed a more efficient Faster R-CNN algorithm to realize end-to-end object detection, and its most notable feature is the introduction of a Regional Proposal Network (RPN). Candidate boxes are firstly generated using the RPN module, which is classified and regressed. All the methods above need to generate candidate regions in advance to help the NN to extract features, and then classify and regress the features. Therefore, the two-stages object detection algorithm limits the speed of object detection to a certain extent.

After the two-stages object detection algorithm achieved excellent detection results, researchers paid attention to the balance between the accuracy and speed, and tried to obtain the category and location information of the target by directly processing the image, which is also called single-stage object detection algorithm. The You Only Look Once (YOLO) series [20], [21] and Single Shot MultiBox Detector (SSD) are the representative algorithms. In 2016, YOLOv1 was proposed as the first version to realize object detection by dividing input images into $S \times S$ grids and predicting B bounding boxes on each grid. In the same year, Liu W et al. proposed the SSD algorithm, which was an improvement based on the YOLO model. The SSD model optimizes the model during training using cross entropy losses and smooth L1 losses. Cross entropy losses are used for the classification of the training model, and smooth L1 losses are used for the regression of the training model. By using them, the SSD model can optimize both classification and regression tasks for better detection accuracy. Subsequently, YOLOv2 [22] was proposed to improve the network structure and loss function and introduce multi-scale feature fusion technology, further enhancing detection accuracy and speed. The latest YOLOv3 which optimized the network structure and loss function, and introduced multi-scale prediction and multi-scale feature fusion technology, greatly improved the detection accuracy and speed [23]. In addition, in order to better adapt to different scene requirements, the YOLO series has also derived some variant algorithms, such as YOLO-Tiny [24], YOLOv4 [25] and so on.

The anchoring mechanism used in the single-stage object detection method will adversely affect the accuracy of position regression under the variable perspective. Zhou et al. [26] proposed Center Net (CN), which adopts the regression method without anchoring points. After positioning the object, a regression-based method is adopted to determine the size of the object. Based on this principle, a Dynamic Reflection Network (DRN) predicting rotation angles was proposed by Pan et al. [27]. Compared with the conventional candidate frame generation model, the object detection algorithm without anchor frame has a better application prospect.

Liu et al. [28] firstly optimized the Resblock in the YOLOv3 network by concatenating two ResNet units with the same width and height, and used residual networks and jump connections to fuse different scale feature layers. Lv et al. [29] proposed a scale adaptive balance mechanism based on anchor-free frame for small target detection, and designed a new loss function to adapt the detection model for different target scales, thus alleviating the unbalance of target scales. There is a high correlation between semantic segmentation and object detection. Mask R-CNN [30], Mask Lab [31] and HTC [32] took into accounttwo tasks in the field of machine vision, and obtain great results on both tasks. Li et al. [33] proposed an image candidate region extraction based on semantic segmentation to filter the noise in the image. By fusing multi-scale features and conducting ASPP [34], candidate regions are obtained and positioning accuracy ofsemantic features are improved.

Aim at the little feature information for small target identification in the image, we make the contributions as follow:

(1) Compared with other object detection algorithms, YOLOv5 has been significantly improved in terms of speed and accuracy, achieving better performance. Based on the three detection layers of the original YOLOv5 model, a detection head P2 is added specifically for small targets, and the shallow high-resolution feature map is involved in the subsequent multi-scale feature fusion, effectively avoiding the problem of the network losing the key feature information of small-scale targets in the process of multiple downsampling.

(2) An enhanced DSI-FPN multi-scale feature fusion pyramid network is designed. In order to avoid excessive increase in computational load, the FPN+PAN pyramid network structure is optimized by Depthwise Separable Convolution (DSC), involution operator and spatial attention mechanism, which fully integrates high-level semantic information and shallow detail information.

(3) An adaptive channel spatial attention mechanism is proposed. Self-attention mechanism is introduced into CBAM module, and non-local information is added to the interaction with local information, which breaks the restriction of convolution kernel.

(4) The knowledge distillation of joint teacher network based on the feature layer is applied to the lightweight design of the model. The optimized feature extraction network and residual network are combined as teacher networks, and the intermediate feature layer information and output layer information of the united teacher networks are transferred to guide the learning of the student network.

(5) The ablation experiments are executed to verify the effectiveness of the proposed model in this paper. Compared with other remote sensing image object detection methods, the experimental results shown that the approach presented has better detection effect for small scale targets in remote sensing images and targets under different lighting conditions.

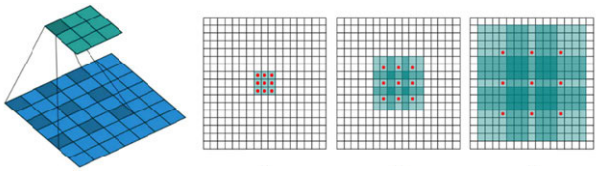


FIGURE 1. (a) A Dilated Convolution with an expansion rate of 2, (b) ERF of a dilated convolution with a convolution kernel size of 3 × 3 and an expansion rate of 1, (c) ERF of a dilated convolution with a convolution kernel size of 3 × 3 and an expansion rate of 2, (d) ERF of a dilated convolution with a convolution kernel size of 3 × 3 and an expansion rate of 4.

II. OBJECT DETECTION TECHNOLOGY

A. DILATED CONVOLUTION

Dilated Convolution [35] is a special kind of convolution operation that increases the Effective Receptive Field (ERF) of the convolution kernel, allowing for better capture of long-distance dependencies in the image.

The structure of Dilated Convolution is shown in Figure 1. 1(a) shows the Dilated Convolution with an expansion rate of 2, and 1(b) shows that the ERF of both a Dilated Convolution with a convolution kernel size of 3×3, an expansion rate of 1 and the ordinary convolution with a convolution kernel size of 3×3 are both 3×3. 1(c) indicates that the ERF of a Dilated Convolution with a convolution kernel size of 3×3 and an expansion rate of 2 is 5×5, and 1(d) indicates that the ERF of a Dilated Convolution with a convolution kernel size of 3×3 and an expansion rate of 4 is 9×9. Therefore, the function of empty convolution is to increase the receptive field and obtain the feature information of multi-scale context [36]. In the convolution operation, the larger the receptive field is, the larger the range of original features it comes into contact with, which also means that it can obtain features with a larger range and higher semantic level [37].

B. DEPTHWISE SPARABLE CONVOLUTION

The core idea of MobileNet is to use the Depthwise Separable Convolution (DSC) [38] to replace the traditional convolution for reducing the number and amount of calculation model.

In DSC, the convolution operation is broken down into two steps: Depthwise Convolution and Pointwise Convolution. Deep convolution means that the convolution operation is carried out on each channel, that is, different convolution is used to check each channel of the input feature graph for convolution operation. Assume that the size of the input feature map is (H, W, C_in), the size of the convolution kernel is $C(k, k, C\_in, 1)$, then the size of the output feature map after depth convolution is $(H-k+1, W-k+1, C\_in)$.

The output feature map of each channel is convolved point-by-point after Depthwise Convolution, the operation we call it Pointwise Convolution. Suppose that the size of the output feature graph of the Depthwise Convolution is $(H-k+1, W-k+1, C\_in)$, the size of the convolution kernel for Pointwise convolution is $(1, 1, C\_in, C\_out)$, and the size of the output feature graph is $(H-k+1, W-k+1, C\_out)$. DSC

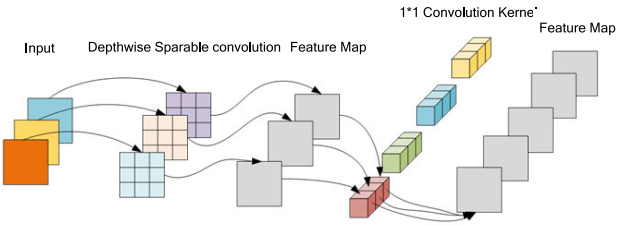


FIGURE 2. Operation of depthwise separable convolution.

can be implemented by a combination to achieve the original convolution operation.

Compared with the$k^2C_inC_out$ parameters of primitive convolution, DSC only requires $C\_inC\_out+k^2C\_in$ parameters. In addition, DSC can also improve the robustness and generalization ability of the model to better adapt to a variety of scenarios and tasks. The detailed operation of DSC is shown in Figure 2.

C. INVOLUTION POERATOR

The kernel of Involution is generated based on a single pixel. The kernel of the feature vector $\boldsymbol{X}_{i,j}$ of a pixel in input feature map is shown as follow.

$$\sigma = \mathrm{Re}LU + BN \tag{1}$$

$$y_{reduce} = \sigma\left(\mathbf{W}_0\mathbf{X}_{i,j}\right) y_{reduce} \in R^{1\times1\times C/r} \tag{2}$$

$$y_{span} = \mathbf{W_1} y_{reduce} y_{span} \in R^{1\times1\times k\times k\times G} \tag{3}$$

$$\mathrm{H}_{i,j} = \phi\left(\mathbf{X_{i,j}}\right) = y_{span} \tag{4}$$

where $\boldsymbol{W}_0$ be a linear transformation matrix which is implemented by a 1×1 convolution operation to reduce the channel dimension to C/r, and C is the number of channels input to the feature map, r is the compression ratio. σ means a batch normalization operation is performed to $\boldsymbol{W}_0X_{i,j}$, and the ReLU activation function is used again. $\boldsymbol{W}_1$ be a linear transformation matrix, and y_{span} is the output after expanding the number of channels. ϕ represents series of operations for $\boldsymbol{X}_{i,j}$,, and $\mathrm{H}_{i,j}$ is the generating kernel [39].

Figure 3 shows the specific operation process of involution operator. Because each feature point of the feature map corresponds to the $K\times K$ parameters of kernel, it cannot be calculated directly. Unfold the feature map first, expand each feature point to its $K\times K$ neighborhoods, and then Multiply-Add the feature map and kernel to get the output feature, as shown in (5) and (6).

$$X_{unfold} = unflod\left(X_{i,j}\right), X_{unfold} \in R^{1\times1\times G\times C/G} \tag{5}$$

$$Y = sum\left(mul\left(\mathrm{H}_{i,j}, X_{unfold}\right)\right), Y \in R^{1\times1\times C} \tag{6}$$

where $\boldsymbol{X}_{unfold}$ is the output feature map after *unfold* operation of $\boldsymbol{X}_{i,j}$, and *mul* indicates that the product operation is performed on the obtained kernel and the output feature graph. *sum* represents the sum of each channel of the feature graph obtained after *mul* operation, C is the number of channels of the input feature graph, and G be the number of groups divided into the input feature graph.

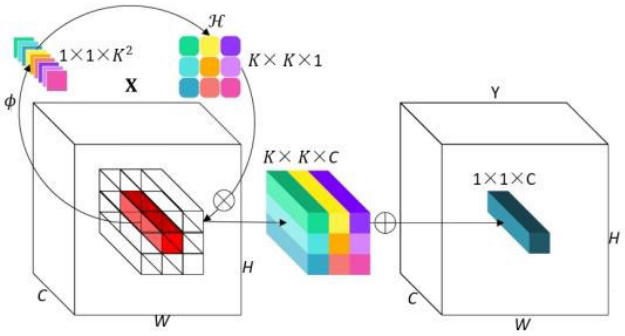


FIGURE 3. Operation of involution.

Compared to conventional convolution operations, the parameters of Involution is $\left(C^2 + C\times G\times K^2\right)/r$, and calculated amount is $H\times W\times K^2\times C$, which is linearly related to the number of channels C. However, the parameters of Convolution is $K^2\times C^2$, and calculated amount is $H\times W\times K^2\times C^2$ which is square relation with the number of channel C. It will not significantly increase the amount of computation with a larger kernel because Involution channels are shared. Using a larger kernel will increase the receptive field, which is more conducive to summarizing context information.

D. DEFORMABLE CONVNETS V2

Influenced by Dilated Convolution, the Deformable Convnets v2 (DCNv2) proposed by Zhu et al. [40]. It adds displacement variables to the traditional convolution and introduces the ability to learn the geometric deformation of the target space. DCNv2 does not change the calculation operation of the traditional convolution, but adds a learnable parameter, which we call two-dimensional offset. The shape of the sampling area can freely be changed, which makes the network more sensitive to the shape of the object.

Taking a 3×3 convolution kernel as an example, the position distribution of sampling points for traditional convolution and DCNv2 are shown in Figure 4. Figure 4(a) represents the sampling points of traditional convolution kernel, 4(b) is the sampling points of convolution kernel with a random offset vector, and 4(c) and 4(d) are the special cases of DCNv2, where the sampling points of convolution kernel introducing regular offset vector are introduced. It is shown that deformable convolution can be used as a special case of scale change, proportional change and rotation change.

The mathematical expression of DCNv2 operation is expressed as follow.

$$y\left(P_0\right) = \sum_{P_n\in R} \omega\left(P_n\right)\times x\left(P_0 + P_n + \Delta P_n\right) \tag{7}$$

where y is the output feature matrix, x is the input feature matrix, and R defines the size of the receptive field and covers the entire input feature matrix. P_n is the pixel at any position on R, and ω is the weight matrix of the sample value. $\omega\left(P_n\right)$ is the weight at the position of the pixel P_n, ΔP_n is the position offset of the pixel P_n, which is generally not an integer. The

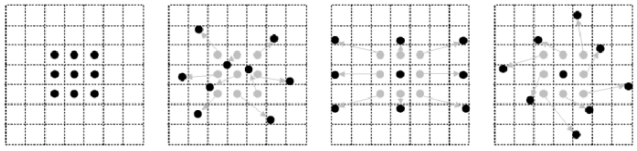


FIGURE 4. (a) Sampling points of traditional convolution kernel, (b) Sampling points of convolution kernel with a random offset vector, (c) Sampling points of convolution kernel with regular offset vector I, (d) Sampling points of convolution kernel with regular offset vector II.

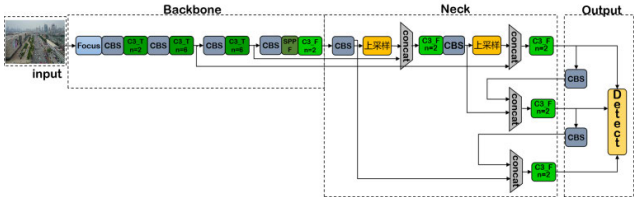


FIGURE 5. Structure of YOLOv5 model.

value of sampling point after migration is determined by bilinear interpolation.

E. SINGLE-STAGE TARGET DETECTION ALGORITHM

YOLOv5 adopts the improved CSPDarkNet53 [41] as the backbone network to extract features from the input image. Firstly, the Focus module is introduced to divide each 2 × 2 adjacent pixels into a patch, and the pixels in the same position for each patch are splicing to obtain four independent feature layers. Then these feature layers are stacked, and the channel dimension is expanded to four times and the width and height dimension is reduced to one quarter of the original. The Focus operation resets tensor dimension, which allows downsampling without losing image information, while reducing computational cost and memory usage of cuda. Some Conv modules in Bottleneck CSP are deleted to obtain the C3 module. YOLOv5 contains two kinds of C3 modules, which are divided into C3_T and C3_F according to set residual edge of Bottleneck. Neck network is located behind Backbone one, and it can further improve the diversity and robustness of features and make better use of image features extracted by Backbone network. YOLOv5 uses Spatial Pyramid Pooling (SPP) module, Feature Pyramid Network (FPN) module and Path Aggregation Network (PAN) [42] module to complete information transmission and fusion. The Output layer utilizes the previously extracted features to make final predictions about the objects in the image. In addition, the weighted NMS approach is used in YOLOv5 to screen the boundary box, which enhances the detection ability of occluded and overlapping targets.

III. OBJECT DETECTION APPROACH BASED ON MULTI-SCALE FEATURE FUSION AND ATTENTION MECHANISM

In order to improve the detection ability of the original YOLOv5 model on small-scale target, we introduce an object detection approach based on multi-scale feature fusion and attention mechanism.

A. LAYER EXTENSION OF SMALL SCALE OBJECT DETECTION

Small object detection depends on deep semantic information and low level detail information, however, simple feature fusion method is difficult to greatly improve its detection accuracy. The backbone network of YOLOv5 model will lose more details after multiple downsampling, and the feature map extracted by the deep network is difficult to learn the feature information conducive to small-scale object detection. However, after multiple downsampling, the feature information of the target is too sparse to detect such small target through original three-scale feature map. By extending multi-scale sensing fields, the extraction of global and local features of image is conducive to target inference. Therefore, a new detection layer specifically for small-scale targets on the basis of the original three detection layers is introduced, and the shallow feature maps after fusion with the deep feature map are detected, so as to improve the detection accuracy of small targets.

There are only three detection layers in the original YOLOv5 network structure, and the feature maps of 8, 16 and 32 times downsampling are adopted to predict the target samples of small, medium and large scales respectively. When the input image size is 640 × 640, three detection layers of 80 × 80, 40 × 40, 20 × 20 with different resolutions can be obtained after feature extraction in the 17th, 20th, and 23rd layer of the multi-scale feature fusion pyramids. The three detection layers can detect targets larger than 8 × 8, 16 × 16, 32 × 32 respectively. When there are a large number of small scale objects smaller than 8 × 8 in the input image, the problem of missing detection is easy to occur.

In order to solve the problems above, a small-scale object detection layer to detect smaller target is added. The improved network structure is shown in Figure 6. Features are enhanced in the second layer of the backbone network, and an upsampling operation is added after the 18th layer. In the detection head module, the 160 × 160 feature map obtained from the 19 layers by upsampling is fused with the feature map of the second layer in the backbone network to obtain four detection layers of different sizes: 20 × 20, 40 × 40, 80 × 80, 160 × 160. The additional 160 × 160 detection layer can detect small targets as low as 4 × 4. Since the feature map of layer 2 has higher resolution and retains more feature information about small-scale targets, it provides more sufficient target information in the fusion process of feature information transmission, thus strengthening the learning ability of the network. The implementation of this approach is very simple and effective, only need to modify the yaml model file of YOLOv5 to add a small scale object detection layer.

B. DESIGN OF MULTI-SCALE FEATURE FUSION NETWORK

At present, FPN+PAN is used for image multi-scale feature fusion to solve the problem of contradiction between the

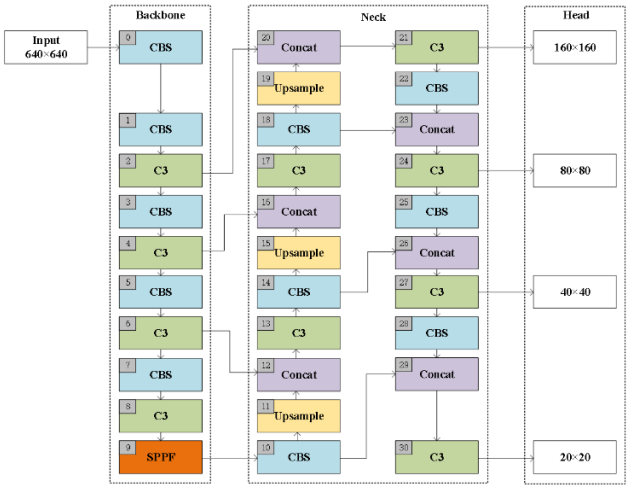


FIGURE 6. Network structure of YOLOv5+P2.

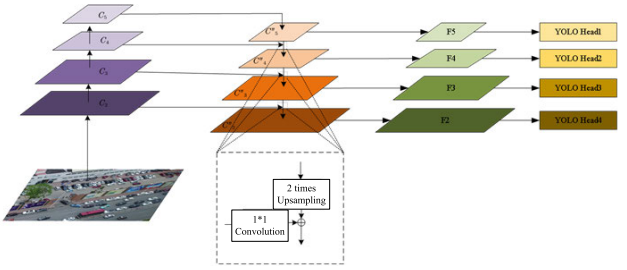


FIGURE 7. FPN network.

details of shallow features and the semantic information of deep features, so as to generate a more abundant and effective image feature layer for network detection.

As shown in Figure 7, FPN uses upper and lower paths and horizontal connections to fuse feature maps of different sizes. Two paths are mainly divided by bottom-up network and top-down one. The bottom-up path is the feed forward calculation of CNNs, and the input image generates feature maps of different resolutions through feature extraction network. Let the feature layer of the input network be $< C_2, C_3, C_4, C_5\}$, which has <4,8,16, 32} times downsampling compared to the original image.

$$C_i = f_i\left(C_{i-1}\right) = f_i\left\{f_{i-1}\left[\cdots f_1(I)\right]\right\} \tag{8}$$

where i=<1,2, 3,4} be number of the feature layer, and $I$and f represent the input image and the corresponding convolution operation respectively.

The top-down path starts from C_5, and C_5' is obtained by 1×1 convolution to adjust the number of channels. Then C_5'' is obtained by double up-sampling of the feature figure C_5', and C_4'' is obtained by using 1×1 convolution to adjust the number of channels for C_4, and C_4'' and C_5'' have the same resolution and number of channels, and C_4' is obtained by direct addition of each element. The same operation is carried out for C_4, C_3 and C_2, and the multi-scale feature structure is obtained by

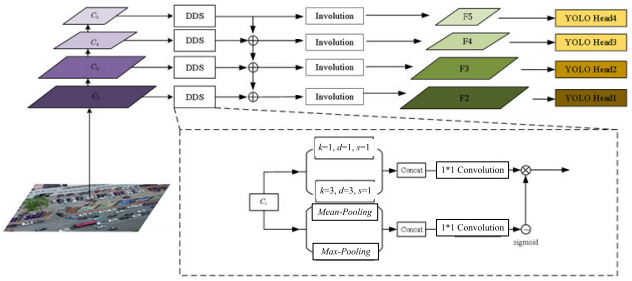


FIGURE 8. DSI-FPN network.

random 3×3 convolution to eliminate aliasing effect.

$$F_i = g\left(C_i, C_{i+1}\right) = f^{3\times3}\left\{f^{1\times1}\left[C_i\right] + f^{1\times1}\left[v\left(C_{i+1}\right)\right]\right\} \tag{9}$$

where v be deep feature upsampling operation, and g represents the fusion operation of adjacent features.

As shown in Figure 8, a DSI-FPN network structure is designed in this paper, and the structure of FPN is enhanced on the basis of minimizing additional computation and parameter number. Based on FPN + PAN structure, we introduce Depthwise Separable Convolution, Dilated Convolution, the mechanism of SAM [43] and Involution operator to enhance the ability of the network to pay attention to effective information.

The feature graph $< C_2, C_3, C_4, C_5\}$ extracted through the backbone network is input into the DSI-FPN network, firstly, the feature refinement and fusion are operated by the DDS network structure. The feature map is simultaneously carried out two depth-separable convolution operations, one with kernel_size=1, dilation rate=1, stride =1 of Dilated Convolution, and the other with kernel_size =3, dilationrate =3 and stride =1 of Dilated Convolution, the feature graphs after two convolution are fused into channels, and then a 1×1 convolution is performed again to change the number of channels. Then, a spatial attention mechanism is introduced to carry out an average pooling operation and a maximum pooling operation respectively on the feature graphs input into the DDS network. After channel fusion of the output after two pooling operations, a 1×1 convolution operation is used to unify the number of channels. Finally, the feature maps obtained by the two branches are output after pixel dot multiplication. The feature image obtained after DDS network processing is sampled twice by the nearest neighbor interpolation algorithm, and then the input is fused with the feature image of the previous layer after DDS network processing. The fused feature image is exported after a 3×3 Involution operator operation.

The enhanced multi-scale feature fusion pyramid network DSI-FPN uses deep separable convolution with different expansion rates in the DDS module structure to improve the prediction ability of small targets and dense targets by increasing the sensitivity field, while preserving spatial features better and enhancing the context connection. The spatial attention mechanism helps the model focus on the important

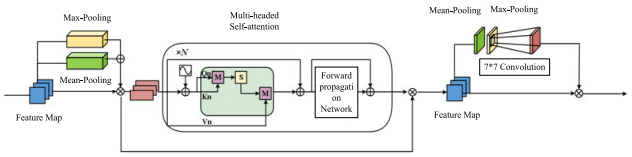


FIGURE 9. Adaptive spatial channel attention mechanism.

part of the input data, and improves the feature representation ability of the model in different spatial locations, so that the feature information more favorable to the object detection can be selected more efficiently.

C. ATTENTION MECHANISM OF ADAPTIVE CHANNEL SPATIAL

Aim to the problems such as unclear semantic information and larger receptive field requirement for small-scale object detection, CBAM attention mechanism [44] is generally used to solve the problems. In the CBAM attention mechanism, 7×7 convolution is performed on the pooled images after using the spatial attention mechanism to search for contextual semantic information. However, such the receptive field is still not large enough for small target tasks. The number of parameters that the network can learn will be increased if the size of the convolutional kernel is enlarged, which will be bring unnecessary redundant information. We improve the CBAM attention mechanism, and adds non-local information to the interaction that originally had only local information. The restriction of convolution kernel will be broken, the receptive field of the model will be expanded, and the appropriate position coding will be given the important information.

As shown in Figure 9, the output feature picture of channel attention network is set as $F \in R^{1\times1\times2C}$. Firstly, it is mapped to the feature picture $F_{forward} \in R^{1\times1\times d}$ by 1×1 convolution to improve the feature characterization ability of the network. Then, the one-dimensional vector after dimensional pooling is processed into blocks and M real numbers are divided into a group. If the feature graph F is used directly to calculate the similarity, the semantic similarity is directly reflected, and certain limitations will be generated in calculating the attention weight. After mapping to the new space, the diversity of similarity calculation between inputs is increased, not only in semantic similarity, but also in context attention.

A phase characteristic vector is mapped to query vector $\boldsymbol{Q_n} \in R^{1\times M\times d/M}$, key vector $\boldsymbol{K}_n \in R^{M\times1\times d/M}$ and value vector $\boldsymbol{V_n} \in R^{1\times M\times d/M}$ by 3×n matrix $\boldsymbol{W_q}$, $\boldsymbol{W_k}$, $\boldsymbol{W_v}$, where n represents the number of the self-attention mechanism heads. Then, the attention score matrix $A \in R^{N\times N}$ is calculated as follows.

$$\boldsymbol{A} = \text{softmax}\left(\frac{\boldsymbol{QK}^{\text{T}}}{\sqrt{d}}\right) \tag{10}$$

The self-attention mechanism interprets the correlation between feature graphs well by calculating the correlation between vectors to obtain contextual information in the object detection task.

The forward propagation network is essential, and within the internal structure of the self-attention mechanism calculation, linear transformation takes place. However, the learning ability of linear transformation is not as strong as that of nonlinear change. Although the attention output uses the self-attention mechanism to learn a new expression form of each feature, the expressive ability of this representation may not be strong. By activation function, the larger number part can be strengthened, and the smaller part can be suppressed, so that the expression effect of the relevant part is better. At the same time, in the fully connected layer, the process of first mapping data to a high-dimensional space and then to a low-dimensional space can learn more abstract features and prevent the network from overfitting.

D. DESIGN OF NETWORK OVERALL STRUCTURE

In this paper, the model structure of YOLOv5 was optimized based on the three methods above, and the improved model structure is shown in Figure 10.

The C3 module in the CSPDarkNet53 backbone network uses the SCBAM to replace the last convolutional layer in the original module, so that the network can focus its attention on the key area of the target in the process of image feature extraction, where improving the recognition ability of small-scale target.

The original multi-scale feature fusion module FPN+PAN is replaced by the enhanced feature fusion pyramid network DSI-FPN structure designed in this paper. In order to ensure the detection speed of the model, the depth-separable convolution and Involution operators with relatively few parameters and computations are used. At the same time, it further enhances the ability of the network to pay attention to the effective information and to express the features of multi-scale target. The 160×160 small-scale object detection head is added to the detection head module, corresponding to 4 times of the original image (640×640). Compared with the detection head with 80×80, 40×40 and 20×20 resolutions, the detection head added in this paper has a higher resolution. Secondly, the shallower feature maps in the backbone network are involved in the subsequent multi-scale feature fusion network, which retains more feature information about small-scale targets and provides more adequate target information in the fusion process of feature information transmission, thus strengthening the learning ability of the network.

IV. MODEL LIGHTWEIGHT DESIGN OF JOINT TEACHER KNOWLEDGE DISTILLATION BASED ON FEATURE LAYER

In order to improve the real-time property of image detection algorithm, we propose a lightweight object detection algorithm of the joint teacher network based on the knowledge distillation. By designing a more powerful teacher network and a more reasonable distillation loss, the student network are guided to learn more global output.

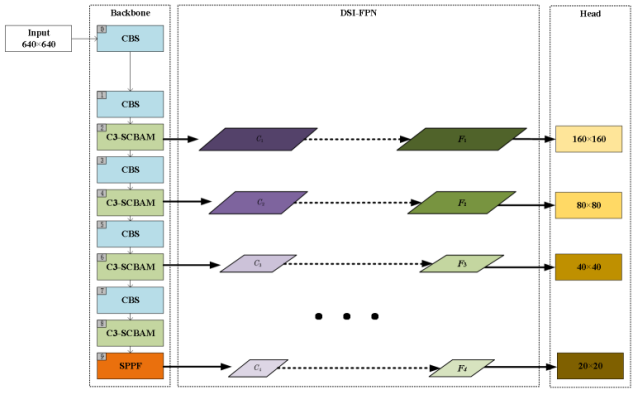


FIGURE 10. Overall network structure based on multi-scale feature fusion and attention mechanism.

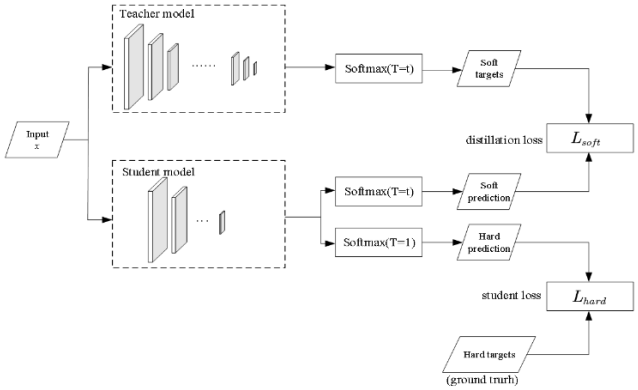


FIGURE 11. Knowledge distillation.

A. KNOWLEDGE DISTILLATION

The basic principle of knowledge distillation is to train the student network by generating Soft target on the teacher network, which refers to the probability distribution vector of the teacher network's output rather than individual category labels.

As shown in Figure 11, a large teacher network is used to train a high-precision model, so that the teacher network can fully learn the knowledge contained in the data. Then, Soft target are generated on training data by teacher network. Finally, the generated Soft target is used to guide a small student network for training, and the student neural network is equivalent to obtain prior information about the dataset from the teacher one.

B. ARCHITECTURE DESIGN OF JOINT TEACHER KNOWLEDGE DISTILLATION BASED ON FEATURE LAYER

Knowledge distillation with a single teacher network can compress a large NN into a small one, which reduces the computational and storage cost. However, since the knowledge distribution learned by the student network is obtained by the teacher network training, it is easy to overfit in some cases. Meanwhile, if the selected teacher model is not suitable for the current task, the performance of the student network will be little improved, and it may even cause the performance degradation. In order to solve the problems above, we select the joint teacher network based on the feature layer to build the knowledge distillation framework. The improved YOLOv5m network is used as the main teacher network (Teacher1), and ResNet50 is selected as the auxiliary teacher network (Teacher2), and the key feature layers of Teacher1 and Teacher2 network are merged to guide student network learning. The network response to the key feature layer of the student network is forced to approximate the network response to the corresponding feature layer of the joint teacher one. The residual structure of the ResNet network uses a cross-layer connection to add the input directly to the output of the convolutional layer. Therefore, the function of ResNet50 as the auxiliary teacher network is to supplement the feature layer extracted by Teacher1 and prevent the important feature information of small-scale targets from being omitted in the feature extraction process.

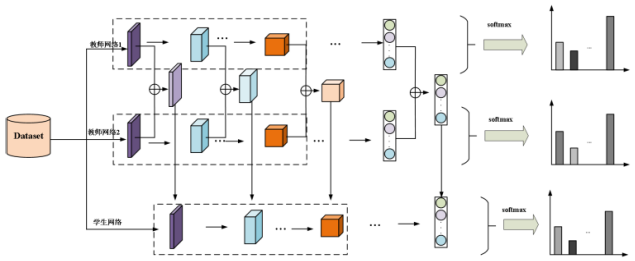


FIGURE 12. Joint teacher knowledge distillation architecture based on feature layer.

As shown in Figure 12, the joint teacher network distillation architecture based on feature layer is mainly divided into three stages.

Stage 1: Teacher1 and Teacher2 pre-train independently and save models weights;

Stage 2: Channel splicings of key feature layers in Teacher1 and Teacher2 networks are carried out, a 1×1 convolution operation is used to carry out feature fusion, and effective image feature extraction is realized through weight sharing.

Stage 3: The characteristic information obtained from Stage 2 guides the student network YOLOv5s to learn.

C. LOSS DESIGN OF JOINT TEACHER KNOWLEDGE DISTILLATION BASED ON FEATURE LAYER

In knowledge distillation, Soft target is used as the label of the student network, and the training process is to fit the probability of the whole model output.

Soft target contains the category probability distribution of the target and the location information of the detection box, which has more knowledge and information than the Hard target. In order to adjust the contribution degree of positive and negative samples, the distillation temperature coefficient T is introduced into Sigmoid activation function, as shown in (11).

$$p_i = \frac{1}{1 + e^{-z_i/T}} (0 < p_i < 1) \tag{11}$$

where z_i is the output value of the last fully connection layer, and p_i is the probability of the input belonging to category i. The larger the temperature coefficient T, the smoother the output probability distribution of the function, and the information of negative samples will be relatively enlarged. Specifically, when $T \rightarrow \infty$, all categories have the same probability, and when $T \rightarrow 0$, the Soft target reverts to Hard target.

In the knowledge distillation, another loss function needs to be defined because of the error between the large model and the small model considered. The loss function of knowledge distillation is weighted by distill loss in teacher network (corresponding to Soft target) and student loss (corresponding to Hard target). The mathematical expression is shown as (12):

$$L_{Total\ loss} = \alpha L_{soft} + \beta L_{hard} \tag{12}$$

where α and β are used to adjust the contributions of teacher loss and student loss.

The loss function of the knowledge distillation architecture of the joint teacher network based on feature layer consists of three parts: the distillation loss of knowledge transfer in the middle feature layer $L_{feature}$, the distillation loss of the teacher network L_{Soft} and the student network L_{Hard}.

$$L_{Total\ loss} = \varphi L_{feature} + \alpha L_{soft} + \beta L_{hard} \tag{13}$$

The loss of the middle feature layer $L_{feature}$ uses KL divergence to measure the difference between the student and teacher models. Specifically, assuming that the middle layer feature of teacher model and student model be $T(x)$ and $S(x)$ respectively, then the KL divergence between them can be defined as:

$$L_{feature} = L_{KL}(x) = \frac{1}{HWC} \sum_{h,w,c} T(x)_{h,w,c} \log \frac{T(x)_{h,w,c}}{S(x)_{h,w,c}} \tag{14}$$

where H, W and C are the height, width and channels' number of feature layer respectively, $S(x)_{h,w,c}$ and $T(x)_{h,w,c}$ be the characteristic values in the specific locations and channels of student and teacher model respectively. Unlike the Euclidene distance, KL divergence can better handle cases with large distribution differences, which can improve the generalization ability of the model.

In knowledge distillation, L_{Soft} is primarily used to optimize the predictions of the student model to more closely match the teacher one, and the Cross-entropy loss function is used for calculation.

Assuming that y_T^i and y_S^i represent the predicted results of the ith sample in the teacher and the student model respectively, and N be the number of samples.

$$L_{soft} = -\frac{1}{N} \sum_{i=1}^{N} y_T^i \log \left(y_S^i\right) \tag{15}$$

Student network loss function L_{Hard} is used to measure the error before the student network and the true value, and y^i is the true label of the dataset.

$$L_{hard} = -\frac{1}{N} \sum_{i=1}^{N} y^i \log \left(y_S^i\right) \tag{16}$$

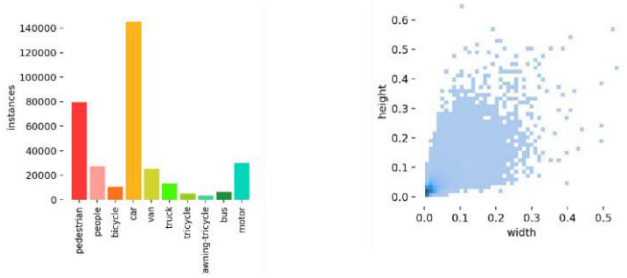


FIGURE 13. (a) Numbers of annotations, (b)Distribution of the targets sizes.

V. EXAMPLE ANALYSIS

In order to verify the effectiveness of the image object detection algorithm proposed in this paper based on multi-scale feature fusion and attention mechanism, on the validation set of VisDrone-2019 dataset, the experiments analyses are conducted on the small-scale object detection layer P2, the enhanced feature pyramid network DSI-FPN and the adaptive channel spatial attention mechanism. Compared with other image object detection algorithms, the results are analyzed to verify that the proposed method has better detection effect.

A. DATA SET

In this paper, VisDrone-2019 which is an extremely challenging object detection dataset is selected for experiment and verification. VisDrone-2019 is a large dataset collected by multiple drone platforms in various situations such as different weather and lighting conditions. The images were captured by cameras mounted on the drone, including locations (taken from 14 different cities in China), environments (urban and rural), objects (pedestrians, vehicles, bicycles, etc.), and densities (sparse and crowded scenes). The images used in the object detection task in the dataset are divided into 6471 training images, 1610 test set images and 548 validation set images, and the targets are divided into 10 categories: pedestrian, human, bicycle, car, van, truck, tricycle, awning tricycle, bus and motorcycle (<0:pedestrian, 1:people, 2. Bicycle, 3:car, 4:van, 5:truck, 6:tricycle, 7:awning-tricycle, 8:bus, 9:motor}).

Figure 13(a) shows the numbers of annotations for all categories in the dataset. The distribution of the targets sizes in the dataset relative to the original image is shown in Figure 13(b).

Figure 14 shows a partial sample of this dataset. It can be seen that the size of most targets are small, which make the model face great challenges in dealing with the detection task on small-scale targets. The dataset contains a large number of complex scenes, great influence of light environment, wide coverage of target size, variable shooting angle, and a large number of small targets with occlusion phenomenon.

FIGURE 14. A partial sample of the dataset.

B. EXPERIMENTAL EVALUATION INDICATORS

In order to prove the contribution of various improvement points in this paper to the improvement of model performance, four evaluation indicators commonly adopted will be used [45]: Precision, Recall, Mean Average Precision (mAP@0.5) for all classes with Intersection over Union (IoU) of 0.5 and Mean Average Precision (mAP@0.5:0.95) for all classes with IoU between 0.5 and 0.95.

The Precision and Recall are calculated with TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative), where TP refers to the number of positive cases predicted by the model, FP represents the number of negative cases predicted by the model, TN refers to the number of negative cases predicted by the model, and FN represents the number of positive cases predicted by the model.

Recall represents the proportion of positive samples with correct predictions to all true positive samples. The higher the Recall is, the more positive samples the model can identify. Accuracy refers to the proportion of positive samples predicted by the model. The higher the Accuracy is, the more accurate the results predicted by the model.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$\text{Re}call = \frac{TP}{TP + FN} \tag{18}$$

A crossover ratio threshold is selected, and a P-R curve is drawn with Recall under the IoU threshold as the horizontal axis and Precision as the vertical axis. The Average Precision (AP) of this category can be obtained by averaging the accuracy on the P-R curve. The average AP of all categories N under this threshold is called the average accuracy mAP.

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{19}$$

C. EXPERIMENTAL PARAMETERS SETTING

The models are trained and verified by remotely connecting the supercomputing cloud server. The input image resolution size is 640×640, the batch-size of each training round is set to 16, the number of training round is set to 300, the initial value of learning rate is 0.01, the termination learning rate is set to 0.0005. The optimization algorithm adopts AdamW, and the momentum is set to 0.937, and weight_decay is 0.0005.

TABLE 1. Comparisons of performance between YOLOv5m and YOLOv5m_P2.

Model	Precision	Recall	mAP@0.5	mAP@0.5: 0.95
YOLOv5m	49.2	36.3	36.5	21.0
YOLOv5m_P2	**51.8**	**38.2**	**39.4**	**22.4**

TABLE 2. Comparisons of performance between YOLOv5m_P2+FPN and YOLOv5m_P2+DSI-FPN.

Model	Precision	Recall	mAP@0.5	mAP@0.5: 0.95
YOLOv5m_P2 +FPN	51.8	38.2	39.4	22.4
YOLOv5m_P2 +DSI-FPN	**55.9**	**39.7**	**41.5**	**24.5**

At the same time, the parameters of ending areset. The training is end when the loss value of the verification set does not change in 20 consecutive rounds. Based on the VisDrone2019 dataset, the K-Means++ clustering algorithm is used to determine the size of 12 initial anchor boxes: (2,4), (3,6), (5,10), (8,9), (10,14), (16,13), (15,20), (12,31), (25,27), (22,36), (38,46), (60,72), which are successively from small to large.

D. EXPERIMENTAL RESULTS BASED ON MULTI-SCALE FEATURE FUSION AND ATTENTION MECHANISM

In order to verify the effect of the extended small-scale object detection layer P2 on the model detection performance, the ablation experiments of YOLOv5m network and YOLOv5m_P2 network are first conducted in this paper. The experimental results are shown in Table 1. It can be seen that the detection accuracy of the model is significantly improved after adding a small-scale object detection layer.

Precision of the YOLOv5m_P2 model is 51.8%, which is 2.6% higher than that of the original YOLOv5m model. Recall value is 38.2%, which is 1.9% higher than YOLOv5m one. The value of mAP@0.5 is 39.4%, which is 2.9% higher than that of the YOLOv5m model. The value of mAP@0.5:0.95 is 22.4%, which is 1.4% higher than YOLOv5m one. Through the comparison of the four evaluation indicators above, it can be seen that the newly added small-scale object detection layer P2 can significantly improve the detection performance of small targets.

So as to verify the effect of DSI-FPN on model detection performance, ablation experiments are conducted on YOLOv5m_P2+FPN and YOLOv5m_P2+DSI-FPN networks. The comparisons are shown in Table 2.

It can be seen that compared with YOLOv5m_P2+FPN network, Precision of the YOLOv5m_P2+DSI-FPN model is 55.9%, which is 4.1% higher than the original model. Recall value is 39.7%, which is 1.5% higher than YOLOv5m_P2+FPN. The value ofmAP@0.5 is 41.5%, which is 2.1% higher than the original one. The value

TABLE 3. Comparison of different attention mechanisms.

Model	Precision	Recall	mAP@0.5	mAP@0.5: 0.95
YOLOv5m	49.2	36.3	36.5	21
YOLOv5m_SE	49.6	36.6	36.6	21
YOLOv5m_ECA	**50.9**	37.1	37.1	21.1
YOLOv5m_CBAM	50.6	**37.7**	**38.1**	**22.0**

of mAP@0.5:0.95 is 24.5%, which is 2.1% higher than YOLOv5m_P2+FPN.

For testing the effects of SE, ECA and CBAM attention mechanisms on small object detection performance, YOLOv5m_SE, YOLOv5m_ECA and YOLOv5m_CBAM three models are built in this paper and compared with the original model of YOLOv5. The experimental results of four types of models are compared as shown in Table 3.

Compared with the original YOLOv5m model, after introducing the attention mechanism SE, ECA and CBAM, the Precision value is increased by 0.4%, 1.7% and 1.4% respectively. Recall is increased by 0.3%, 0.8% and 1.4%, respectively. mAP@0.5 increased by 0.1%, 0.6% and 1.6%, respectively. YOLOv5m_ECA has no increase on mAP@0.5:0.95, YOLOv5m_ECA has an increase of 0.1% on mAP@0.5:0.95, and YOLOv5m_CBAM has an increase of 1.0% on mAP@0.5:0.95.

The experimental results show that compared with SE and ECA attention mechanisms, the introduction of CBAM attention mechanism has the most significant improvement effect on the model. Except Precision value, the remaining three indicators are at the highest level, indicating that the performance of YOLOv5m_CBAM model is the best. Therefore, the subsequent experiments in this paper choose CBAM attention mechanism to improve.

In this paper, added small-scale object detection layer P2, DSI-FPN and SCBAM are successively introduced, and the ablation experiments are conducted successively. The comparisons are shown in Table 4.

As shown from the table, the image object detection approach based on multi-scale feature fusion and attention mechanism has significantly improved the detection accuracy. Precision value is 57.2%, which is 8% higher than the original YOLOv5m model. Recall value is 40.9%, which is 5.7% higher than the original one. mAP@0.5 is 43.9%, an increase of 7.4%, and mAP@0.5:0.95 is 26.2%, an increase of 5.2%.

In order to verify the impact of each module on small object detection performance, we add each module in sequence in the designed object detection method based on multi-scale feature fusion and attention mechanism, and compare it with the original YOLOv5m model on mAP@0.5 of ten categories in VisDrone-2019 dataset in Table 5.

Experimental results show that the proposed network has good detection accuracy in most categories. mAP@0.5 has the most obvious improvement effect on truck, car and tricycle, which is 8.1%, 6.8% and 6.6% higher than YOLOv5m, respectively. The total mAP@0.5 value of the model is 7.4% higher than the original YOLOv5m model.

FIGURE 15. (a) Detection results of YOLOv5m, (b) Detection results of ours.

As shown in Figure 15, the visual contrast of the image object detection algorithm designed in this paper based on multi-scale feature fusion and attention mechanism is compared with the detection effect of the original YOLOv5m model.

Figure 15 (a) is the detection results of the original YOLOv5m model, and Figure 15 (b) is the detection results of the model we presented. The red box is the area with obvious contrast of the same image. The method proposed in this paper has better detection effect for small scale targets under different lighting conditions.

E. EXPERIMENTAL RESULTS OF JOINT TEACHER KNOWLEDGE DISTILLATION BASED ON FEATURE LAYER

In this paper, the knowledge distillation approach is adopted to carry out lightweight design. The optimized YOLOv5m is used as the main teacher network, and ResNet50, MobileNetv3 and ShuffleNetv2 are respectively used as the auxiliary teacher network. The combined teacher network jointly guides the student network for training and learning. The student network adopts the YOLOv5s model with fewer parameters and shallower depth. The comparisons of

TABLE 4. Comparisons of module ablation experiments.

YOLOv5m	P2	DSI-FPN	SCBAM	Precision	Recall	mAP@0.5	mAP@0.5:0.95
√				49.2	36.3	36.5	21.0
√	√			51.8	38.2	39.4	22.4
√	√	√		55.9	39.7	41.5	24.5
√	√	√	√	57.2	41.9	43.9	26.2

TABLE 5. Comparisons of classification accuracy of each module superposition.

Category	YOLOv5m	+P2	+DSI-FPN	+SCBAM	Increasing rate
Pedestrian	42.8	43.7	43.5	47.9	5.1
People	35.5	36.7	37.6	40.5	5.0
Bicycle	15.4	16.5	17.8	19.2	3.8
Car	74.7	75.3	79.6	**81.5**	6.8
Van	36.2	37.6	38.9	40.7	4.5
Truck	53.7	55.2	59.0	**61.8**	8.1
Tricycle	23.0	24.6	27.2	**29.6**	6.6
Awning-tricycle	12.4	13.8	15.8	17.1	4.7
Bus	46.9	47.5	49.7	53.1	6.2
Motor	42.3	43.4	46.6	47.7	5.4
All	36.5	39.4	41.5	**43.9**	7.4

FIGURE 16. Example of detection effect.

knowledge distillation are shown in Table 6. The distillation effect of ResNet50 as the auxiliary teacher network is best, and the detection accuracy of the distilled student network is increased by 7.8% and 4.6% compared with YOLOv5s model and the improved one. As shown in Figure 16, it can be seen that the student network after distillation still has a good detection effect.

TABLE 6. Comparisons of knowledge distillation.

Category	YOLOv5s	Improved YOLOv5s	Joint ShuffleNetv2	Joint Mobile Netv3	Joint ResNet 50
Pedestrian	37.5	41.3	42.6	44.2	45.3
People	31.1	33.6	33.8	34.1	38.6
Bicycle	11.4	13.2	14.2	15.4	16.2
Car	71.4	74.6	75.3	78.2	79.4
Van	31.8	34.9	35.1	38.4	39.3
Truck	30.7	33.5	34.1	35.3	40.8
Tricycle	19.5	22.3	24.3	26.2	26.3
Awning-tricycle	10.7	13.5	14.8	16.2	16.8
Bus	43	46.7	48.5	50.6	52.4
Motor	37.1	42.8	44.3	45.4	46.9
All	32.4	35.6	36.7	38.4	**40.2**

Table 7 is the comparisons of the lightweight model based on the improved YOLOv5 (Ours) and a variety of existing models on the evaluation indicators of mAP@0.5 in ten categories of Visdron-2019 dataset. It can be seen that the detection accuracy of the proposed method is quite good in most categories, which is reaching 45.3% in Pedestrian category, which is about 17.1% higher than Faster R-CNN two-stage detection algorithm. The detection accuracy of Car category is 79.4%, which is about 3.7% higher than that of YOLOv4. The detection accuracy in the Truck category reached 40.8%, which is about 5.6% higher than the TPH-YOLOv5 algorithm. It can be seen that compared with other algorithms, the proposed method surpasses most categories, among which the detection accuracy of Car category is the highest, which proves the effectiveness of the proposed method.

VI. CONCLUSION

In view of the small size and dense distribution of targets captured by remote sensing instrument, this paper adds a detection layer specifically for tiny targets on the basis of the three detection layers of the original YOLOv5 model, and involves the shallower feature layer in the subsequent multi-scale feature fusion. The problem of losing the key feature information of the small-scale target in the process

TABLE 7. Comparison of lightweight models.

Model	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor	All
FCOS	21.17	11.33	10.02	50.39	31.02	24.30	20.41	13.79	37.12	30.51	28.08
VC-YOLO	31.50	26.2	5.44	72.0	31.30	19.2	15.7	7.46	41.4	33.9	28.4
YOLOv3_ReSAM	27.28	15.78	28.70	74.31	**53.77**	32.96	20.48	24.38	52.35	41.57	33.15
FPN-Cascade	20.3	19.5	**29.7**	65.3	48.5	34.8	23.8	13.6	55.1	44.2	34.48
ST-YOLOv5	35.4	23.2	19.4	66.8	43.5	35.9	26.9	15.8	54.2	33.6	35.5
YOLOv4	44.03	28.18	23.14	74.67	32.67	34.31	21.42	**26.72**	**55.36**	45.87	35.72
TPH-YOLOv5	29	16.75	15.69	68.94	49.79	35.16	27.33	24.72	51.80	30.90	37.32
Faster R-CNN	38.17	24.44	14.11	69.98	40.02	32.41	23.11	17.50	51.36	41.00	39.97
Ours	**45.3**	**38.6**	16.2	**79.4**	39.3	**40.8**	**26.3**	16.8	52.4	**46.9**	**40.2**
Motor	21.17	11.33	10.02	50.39	31.02	24.30	20.41	13.79	37.12	30.51	28.08
All	31.50	26.2	5.44	72.0	31.30	19.2	15.7	7.46	41.4	33.9	28.4

of multiple downsampling is effectively avoided. Meanwhile, an enhanced DSI-FPN multi-scale feature fusion pyramid network is proposed, which can generate more abundant and effective image feature layers for network detection.

In this paper, an adaptive channel spatial attention mechanism is put forward. A self-attention mechanism is introduced into CBAM module to add non-local information to the interaction that originally had only local information, so as to break the convolution kernel limitation, expand the receptive field, and improve the feature expression ability of the model. The model was trained, validated, and tested on a public dataset VisDrone2019. The detection accuracy of the improved model based on YOLOv5m on the publicly available remote sensing image dataset VisDrone-2019 reached 43.9%, and 7.4% higher than that of the original model. Experiments show that the object detection algorithm based on multi-scale feature fusion and spatial attention mechanism is significantly superior to most of the existing remote sensing image object detection methods.

In order to improve the real-time performance of the model detection, the model light weight design is carried out by using the knowledge distillation of joint teacher network based on the feature layer. The optimized YOLOv5m is proposed as the main teacher network, and ResNet50 is introduced as the auxiliary teacher network to supplement the important characteristic information. The experimental results show that the student network after distillation can achieve good detection accuracy under the condition of few parameters and calculation.

There are still some follow-up works for improvement in this research. The focus of this study is to improve the detection accuracy of small-scale targets by optimizing the structure of CNNs, and no additional data enhancement operations are carried out on the original images. Moreover, the weather of images in the VisDrone2019 aerial image data set is mostly sunny, and there is a lack of special weather image samples such as rain, snow and haze. Therefore, new image preprocessing methods can be developed for small-scale targets in future research. For example, how to use DL to automate the image preprocessing process, or add different noises to improve the robustness of small object detection.

REFERENCES

[1] Y. Alghamdi, A. Munir, and H. M. La, "Architecture, classification, and applications of contemporary unmanned aerial vehicles," *IEEE Consum. Electron. Mag.*, vol. 10, no. 6, pp. 9–20, Nov. 2021.

[2] R. Perz and K. Wronowski, "UAV application for precision agriculture," *Aircr. Eng. Aerosp. Technol.*, vol. 91, no. 2, pp. 257–263, Feb. 2019.

[3] A. Saif and Z. R. Mahayuddin, "Moving object detection using semantic convolutional features," *J. Inf. Syst. Technol. Manag.*, vol. 7, pp. 24–41, Dec. 2022.

[4] A. F. M. S. Saif, E. D. Wollega, and S. A. Kalevela, "Spatio-temporal features based human action recognition using convolutional long short-term deep neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, pp. 1–15, 2023.

[5] Q. Wang, W. Li, and Z. Jin, "Review of text classification in deep learning," *Open Access Library J.*, vol. 8, no. 3, pp. 1–8, 2021.

[6] T. Jain, V. K. Verma, A. K. Sharma, B. Saini, N. Purohit, H. Mahdin, M. Ahmad, R. Darman, S.-C. Haw, S. M. Shaharudin, and M. S. Arshad, "Sentiment analysis on COVID-19 vaccine tweets using machine learning and deep learning algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, pp. 32–41, 2023.

[7] K. Wang, Z. Meng, and Z. Wu, "Deep learning-based ground target detection and tracking for aerial photography from UAVs," *Appl. Sci.*, vol. 11, p. 8434, Sep. 2021.

[8] A. Panning, A. K. Al-Hamadi, R. Niese, and B. Michaelis, "Facial expression recognition based on Haar-like feature detection," *Pattern Recognit. Image Anal.*, vol. 18, pp. 47–52, Sep. 2008.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 886–893.

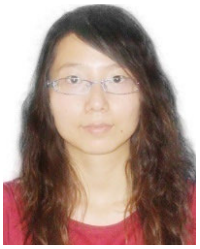
[10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[11] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[12] C. Sun, Y. Ai, S. Wang, and W. Zhang, "Mask-guided SSD for small-object detection," *Int. J. Speech Technol.*, vol. 51, no. 6, pp. 3311–3322, Jun. 2021.

[13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami Beach, FL, USA, Jun. 2009, pp. 248–255.
[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[18] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Comput. Vis. (ICCV)*, Santiago, MN, USA, Dec. 2015, pp. 1440–1448.
[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2016, pp. 779–788.
[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.
[22] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525.
[23] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[24] D. Padilla Carrasco, H. A. Rashwan, M. Á. García, and D. Puig, "T-YOLO: Tiny vehicle detection based on YOLO and multi-scale convolutional neural networks," *IEEE Access*, vol. 11, pp. 22430–22440, 2023.
[25] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[26] Z. Xingyi, D. Wang, and P. Krahenbuhl, "Objects as points," 2019, *arXiv:1904.07850*.
[27] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, C. Ma, and C. Xu, "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11207–11216.
[28] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "UAV-YOLO: Small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, p. 2238, Apr. 2020.
[29] R. Lv, X. Wang, and T. Yang, "Small object detection with scale adaptive balance mechanism," in *Proc. 15th IEEE Int. Conf. Signal Process. (ICSP)*, vol. 1, Dec. 2020, pp. 361–365.
[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[31] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "MaskLab: Instance segmentation by refining object detection with semantic and direction features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4013–4022.
[32] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4969–4978.
[33] C. Li, C. Xu, Z. Cui, D. Wang, Z. Jie, T. Zhang, and J. Yang, "Learning object-wise semantic representation for detection in remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 20–27.
[34] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[35] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[36] X. Su, Y. Zhang, C. Wang, H. Liang, and S. Li, "Multi-scale object detection algorithm based on faster R-CNN," in *Proc. Int. Conf. Business Intell. Inf. Technol.*, 2022, pp. 379–391.
[37] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.
[38] H. Park, L. Lowe Sjösund, Y. Yoo, J. Bang, and N. Kwak, "ExtremeC3Net: Extreme lightweight portrait segmentation networks using advanced C3-modules," 2019, *arXiv:1908.03093*.
[39] D. Li, J. Hu, C. Wang, X. Li, Q. She, L. Zhu, T. Zhang, and Q. Chen, "Involution: Inverting the inherence of convolution for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12316–12325.
[40] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9300–9308.
[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.
[42] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11534–11542.
[43] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 496–500, 2020.
[44] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.
[45] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

ZUOQIANG DU received the M.D. degree from the College of Computer Science and Technology, Harbin Engineering University, China, in 2005. He is a Professor with the School of Computer and Information Engineering, Harbin University of Commerce. He has published with the identity of the first author and a corresponding author more than ten articles indexed by SCI/EI in journals. His main research interests include the application of quantum computing, information processing, artificial intelligence, and machine learning.



YUAN LIANG received the M.D. degree from the College of Automation, Harbin Engineering University, China, in 2011. She is a Researcher with Jinan Inspur Data Technology Company Ltd., and responsible for the research and development of multiple host security products. She has published with the identity of the first author and a corresponding author more than ten articles indexed by SCI/EI in journals. Her main research interests include EDR technology, eBPF technology, artificial intelligence, and machine learning.