IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 17, 2021, accepted August 25, 2021, date of publication August 27, 2021, date of current version September 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108398

Research on Object Detection Method Based on FF-YOLO for Complex Scenes

CHEN BAOYUAN, LIU YITONG, AND SUN KUN

[1]Higher Educational Key Laboratory for Measuring and Control Technology and Instrumentation of Heilongjiang Province, Harbin University of Science and Technology, Harbin 150080, China
[2]National Experimental Teaching Demonstration Center for Measurement and Control Technology and Instrumentation, Harbin University of Science and Technology, Harbin 150080, China
[3]School of Measurement-Control Technology and Communications Engineering, Harbin University of Science and Technology, Harbin 150080, China

Corresponding author: Sun Kun (sunkun1982@126.com)

This work was supported by the Fundamental Research Foundation for Universities of Heilongjiang Province under Grant LGYC2018JC045.

ABSTRACT YOLO v3 has poor accuracy in target location recognition, and the detection effect needs to be improved in complex scenes with dense target distribution and large size differences. To solve this problem, an improved multi-scale target detection algorithm based on feature fusion (FF-YOLO) is proposed in this paper. Firstly, the residual structure in Darknet53 backbone of YOLO v3 is replaced by the optimized dense connection network FCN-DenseNet, and features are extracted effectively through feature reuse, and the problem of vanishing gradient is further alleviated. Secondly, based on the three-scale prediction mechanism of YOLO v3, we added a fourth detection scale to make the network learn more shallower location information. Finally, Spatial Pyramid Pooling (SPP) module is added before each detection layer to make the local feature information deeply fused. It increases the receptive field of the backbone network and significantly isolates the most important contextual feature receptive. Experiments show that FF-YOLO can effectively improve the detection accuracy of multi-scale targets in complex scenes. On Pascal VOC2007 data set, the mAP of FF-YOLO is 5.8% higher than that of YOLO v3. At the same time, the mAP of FF-YOLO for medium and small targets are 1.5% and 2.2% higher than that of YOLO v3 on MS COCO data set.

INDEX TERMS Machine learning, object detection, feature fusion, complex scenes.

I. INTRODUCTION

Target detection aims to detect the object of interest in the image and determine its category and location. It has been widely used in X-ray image detection [1], traffic sign recognition [2], intelligent recognition monitoring [3], industrial detection [4] and other fields. With the breakthrough of deep learning algorithm in image classification [5], [6], target detection algorithm has developed from the traditional algorithm based on manual feature extraction [7]–[9] to the algorithm based on deep learning [10]–[12]. The target detection algorithm based on deep learning overcomes the problems of relying on artificial experience and poor robustness in traditional methods, and its detection accuracy has been significantly improved.

If your paper is intended for a conference, please contact your conference editor concerning acceptable word processor formats for your particular conference. According to the detection steps, the object detection algorithm based on deep learning can be divided into two-stage detection algorithm based on proposed region [13] and one-stage detection algorithm based on regression. Two-stage target detection algorithm is represented by R-CNN [14], Fast RCNN [15], and Faster RCNN [16]. This kind of algorithm searches out some candidate regions of possible objects from the image, and recognizes the objects of each candidate region. The one-stage detection algorithm extracts features directly from the network to predict the category and location of the target, which has high detection efficiency and realizes end-to-end real-time detection. Compared with previous versions, YOLO v3 [17] uses a better classification network and adds a multi-scale prediction mechanism, which is the most widely used and most versatile one-stage target detection algorithm.

Alexey Bochkovskiy *et al.* [18] improved YOLO v3, and proposed YOLO v4 target detection algorithm. Data augmentation method named Mosaic is used to preprocess the

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

training input to enrich the detection data set. Moreover, tricks such as Cross Stage Partial Network (CSPNet) [19], Feature Pyramid Network (FPN) [20] and Path Aggregation Network (PAN) [21] are integrated into the network structure to enhance the learning ability of the network and further improve the detection performance of small targets. Ju *et al.* [22] added adaptive receptive field fusion module in YOLO v3 to increase the contextual information around the targets. And they used spatial attention mechanism to determine the relationship among different regions, which can strengthen the correlation and compactness among different regions. Gong *et al.* [23] proposed a ship detection method for complex remote sensing images, combining YOLO v3 and Inception structure to realize dimension reduction transition and enhance feature transmission. The above researches have improved the YOLO v3 from training strategy, the perspectives of feature prediction and feature transfer, which improved the detection effect of multi-scale targets. However, there still need some improvement for multi-scale target detection in complex scenes. Firstly, YOLO v3 does not make full use of the position information from the shallow feature map when it integrates the features between the adjacent scales in the deep layer, so it cannot accurately locate the overlapping targets; Secondly, the common way of feature fusion is to series the features of different scales on the channel dimension, which cannot reflect the correlation and importance of features between different channels, so the feature information cannot be utilized in multiple dimensions.

To solve the above problems, this paper focuses on strengthening the fusion and reuse of features, and proposes a multi-scale target detection model FF-YOLO (YOLO with enhanced feature-fusion) for complex scenes. FF-YOLO retains the advantages of YOLO v3, such as simplicity and easy transplantation, and strengthens the utilization of different levels of features. Firstly, the backbone network with dense connection structure is used for feature extraction to alleviate the vanishing gradient problem and enhance the transmission and reuse of features. Then, based on the idea of feature pyramid, a four scale multi-scale prediction mechanism is used to extract the location information in the shallower feature map. Finally, the spatial pyramid pooling (SPP) module is added in front of the four detection layers to strengthen the fusion effect of features in the same feature map, so that a feature map has rich deep semantic information and shallow location information at the same time. In this way, the improved network model improves the recognition accuracy of complex overlapping target locations, and improves the problems of missed detection and misdetection caused by inaccurate positioning of multi-scale targets.

II. RESEARCH ON COMPLEX TARGET DETECTION METHOD BASED ON YOLO V3

A. NETWORK STRUCTURE OF YOLO V3

YOLO v3 removes the full connection layer and the last pooling layer in YOLO v1 and adopts the full convolution network (FCN) structure [24], so that the network extracting features can obtain higher resolution features. Its network is mainly composed of backbone DarkNet53 and three-branch prediction structure. Among them, Darknet53 is the basic network for feature extraction. By adding residual structure between the convolutional layer and the lower sampling layer [6], the loss caused by gradient disappearance can be reduced and the learning ability of the network can be enhanced. In addition, the network draws on the idea of FPN, uses three branches to fuse and predict three feature layers of different scales, and fuses the feature layers of different sizes obtained by down sampling, thus enhancing the fusion and reuse of feature information of different layers. The network structure of YOLO v3 is shown in Figure 1.

B. OPTIMIZATION OF TARGET DETECTION METHOD FOR COMPLEX SCENES

This paper improves the structure of YOLO v3 convolution network: In the feature extraction part, the ResNet jump-layer connection structure in the backbone network of YOLO v3 was replaced by DenseNet [25] dense connection structure to realize the efficient transmission of features. Then, the feature map was deeply fused between different dimensions, and the detection layer of the fourth size was added to optimize the multi-scale prediction mechanism. In addition, the spatial pyramid pooling module is added before the detection layer, and the feature map containing rich semantic information and location information is finally sent into the detector. The improved network FF-YOLO structure is shown in Figure 2.

1) REPLACEMENT AND OPTIMIZATION OF BACKBONE NETWORK CONNECTION STRUCTURE

Compared with skip-connection of ResNet, dense connection of DenseNet can make the transmission of features and gradients more effective. Therefore, replacing Residual module with DenseNet structure can enhance the reuse of features in the network, make the network learn more feature information, and improve the accuracy of target detection.

The network structure of DenseNet contains a large number of Dense blocks. Each Dense Block is composed of a different number of convolutional layers, and each Dense Block is connected by a transition layer. The output of the first layer is represented by x_l, and the output of a nonlinear transformation process is represented by H_l. That is, the input of H_1 is x_0 and the output is x_1, and the input of H_2 is x_0 and x_1, meaning that the input of each layer comes from the output of all previous layers.

The expressions of X_l in DenseNet and ResNet are shown in (1) and (2).

$$X_l = H_l([x_0, x_1, \ldots, x_l]) \tag{1}$$

$$X_l = H_l(x_{l-1}) + x_{l-1} \tag{2}$$

It can be seen from formula (1) and formula (2) that in DenseNet, the output of layer l is obtained through nonlinear transformation (H_l) after the merger of channels from layer

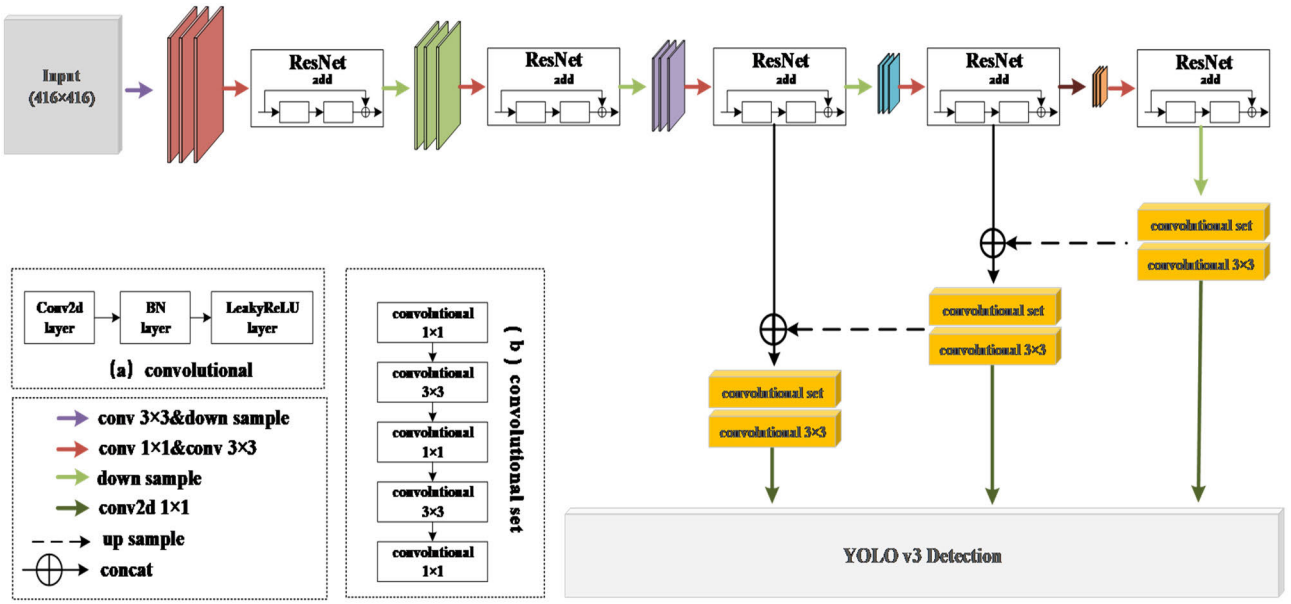


FIGURE 1. YOLO v3 network structure.

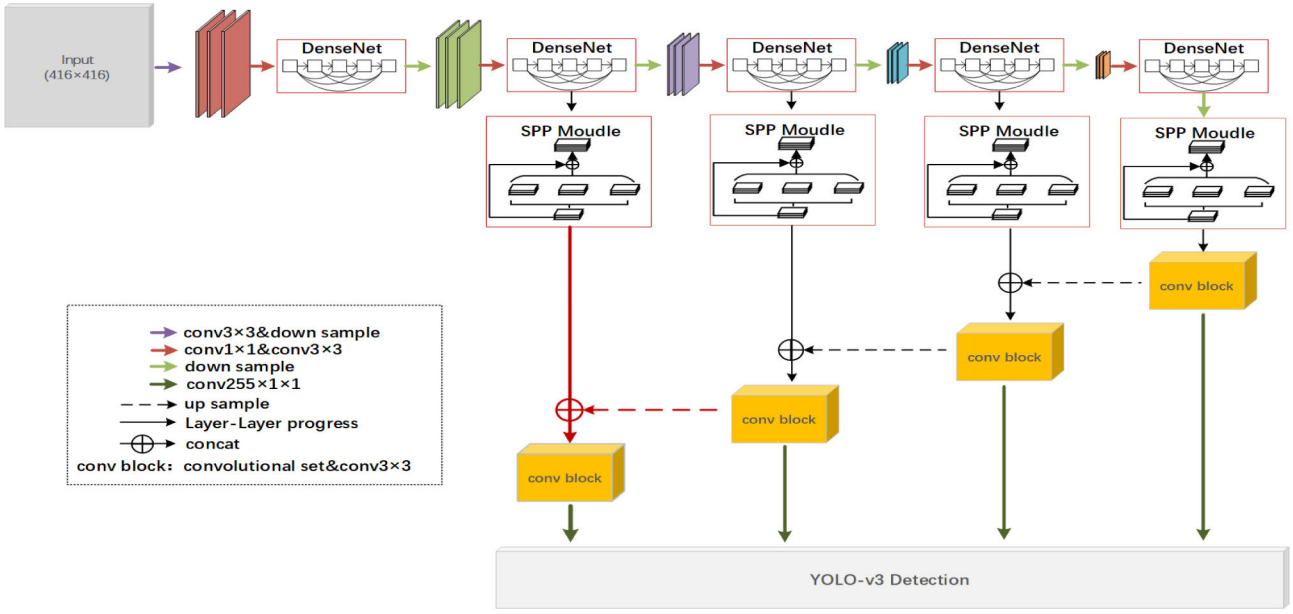


FIGURE 2. FF-YOLO network structure.

0 to layer *l*-1, while in ResNet, the output of layer *l* is obtained by adding values of the output of layer *l*-1 and the nonlinear transformation output of layer *l*-1. It can be seen that ResNet has obvious redundancy, and only a small number of features are extracted from each layer. In contrast, DenseNet does not need to learn redundant feature mapping and realizes feature reuse directly through the connection of features on the channel. The picture in the middle of Figure 4 of Reference26 shows the comparison of DenseNet with bottleneck layer and compression module and ResNet. With the same error, Densenet-BC has much less parameter complexity. The figure to the right of Figure 4 in Reference26 also clearly reflects that Densenet-BC-100 requires only a few parameters to achieve the same results as ResNet-1001. As a result, DenseNet is able to achieve better performance than ResNet with fewer parameters and calculations.

Therefore, dense connection structure was used in this paper to replace the skip-layer connection structure of ResNet in the backbone network of YOLO v3, and Densenet-121 [26] was selected for dense connection network. The backbone

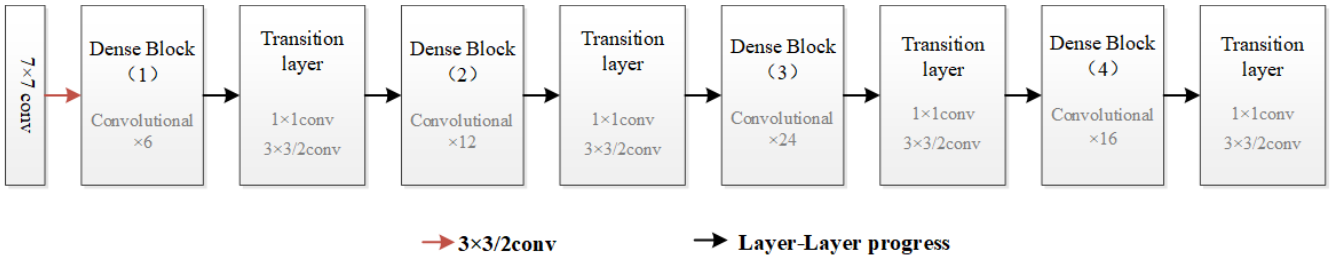


FIGURE 3. FCN-DenseNet network structure.

network with dense connected structure ensures the maximum information transmission between layers, and combines the channels of the previous convolutional layers, which alleviates better and makes more Feature information can be better passed down, reducing the number of parameters.

On this basis, in order to alleviate the problem of information loss in the pooling operation, this paper further optimizes DenseNet-121 to obtain the FCN-DenseNet. The specific optimization process is as follows: Four Dense Blocks of FCN-DenseNet are formed by connecting different numbers of convolutional modules (as shown in Figure 1 (a)); The structure of FCN-DenseNet is adjusted by using the method of adjusting step size to realize down sampling in YOLO v3. The pooling layer in the transition layer is replaced by a 3×3 convolution layer with step size of 2, which effectively solves the problem of loss of useful feature information in the pooling layer. The network preserves 1×1 convolution layer, which not only reduces the number of input feature graphs but also realizes the fusion of features of each channel. The structure of FCN-DenseNet is shown in Figure 3.

2) IMPROVEMENT OF MULTI-SCALE PREDICTION MECHANISM

In complex scenarios, one of the main reasons for the problem of missing and false detection is the target positioning error. Therefore, considering that the larger shallow feature map contains more location information and the smaller deep feature map contains more semantic information, this paper optimizes the multi-scale prediction mechanism of the YOLO v3 network, adds the multi-scale prediction mechanism of the fourth detection layer. By concatenating the small-size feature layer after up sampling and the larger-size feature layer to integrate the semantic information and location information in different feature layers deeply, so that the accuracy of the network's recognition of the target location is further improved. The specific implementation is as follows: First, the 52×52 feature map obtained by the third downsampling is up-sampled twice, and then merged with the 104×104 feature map obtained from the second downsampling to form a fourth-scale detection layer.

The backbone network structure after replacing the Residual structure with the FCN-DenseNet and adding the fourth detection layer is shown in Figure 4. The improved backbone network uses the FCN-DenseNet dense-connection structure to extract features, and the sizes of four feature layers for multi-scale prediction are 104×104, 52×52, 26×26 and 13×13.

3) SPATIAL PYRAMID POOLING MODULE

With the addition of dense connection structure and the optimization of multi-scale prediction mechanism, their operation objects are the whole feature map of a certain convolution layer, and lack of mining the potential relationship between local features in the same feature map. Therefore, this paper continues to introduce the spatial pyramid pooling module to enhance the fusion ability of the model for local features in the feature map.

Spatial Pyramid Pooling [27] is a pooling method that can map local features to different dimensional spaces and merge them, so that the network structure can adapt to image inputs of different scales and sizes. Applying the spatial pyramid structure to the target detection network can improve the multi-scale feature fusion ability of the network, and realize the multi-scale local feature fusion of the feature image, which is conducive to improving the detection effect when the size of the complex target in the image to be detected is large difference. The structure of the SPP is shown in Figure 5.

In Figure 5, H represents the height of the feature graph, W represents the width, and C represents the number of channels. Firstly, the SPP module performs 3 times of block pooling with different convolution kernel sizes on the input $H \times W \times C$ feature map, so as to extract feature information from different sizes of receptive fields. Second, padding the feature map obtained by the pooling operation to normalize the size of the feature map. Finally, the three feature maps and the original feature maps are merged on the channels to get the $H \times W \times 4C$ feature map. In this way, the local features and global features are fused through the feature-level fusion.

In this paper, the SPP module is added before the four detection layers, the size of the largest pooling core in each SPP module is set to the same size as the feature map that needs to be pooled, and the sizes of the remaining two cores are sequentially reduced by 4: In the SPP module connected to the 13×13 feature layer, the pooling core size is set to 13,

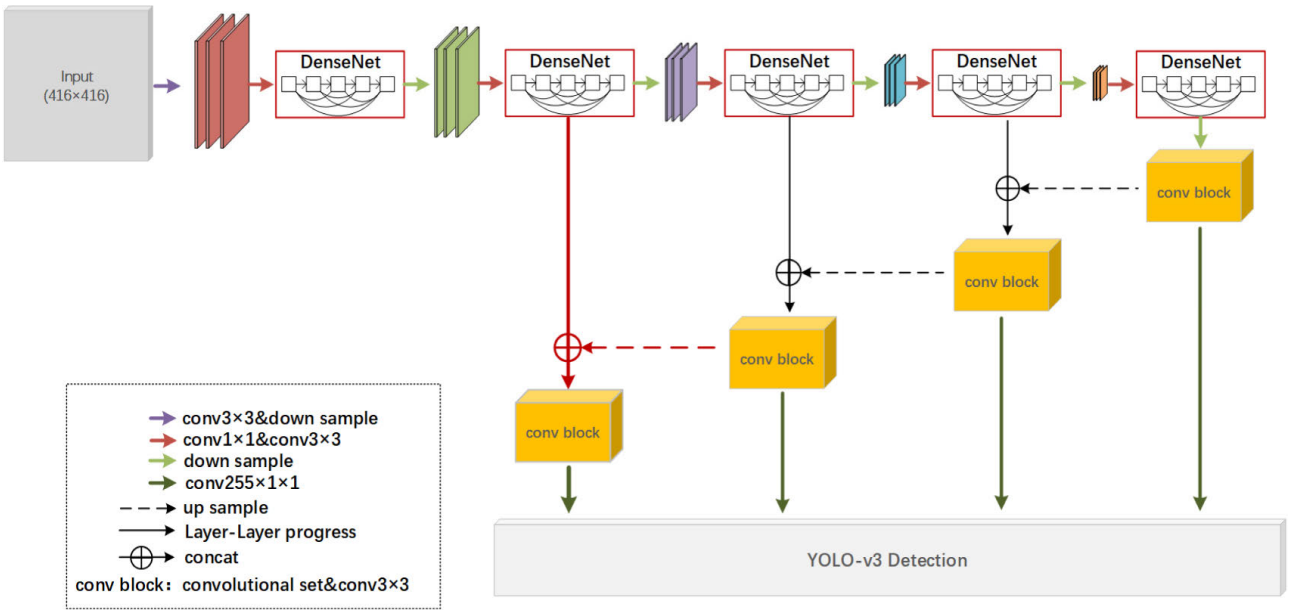


FIGURE 4. Improved backbone network structure.

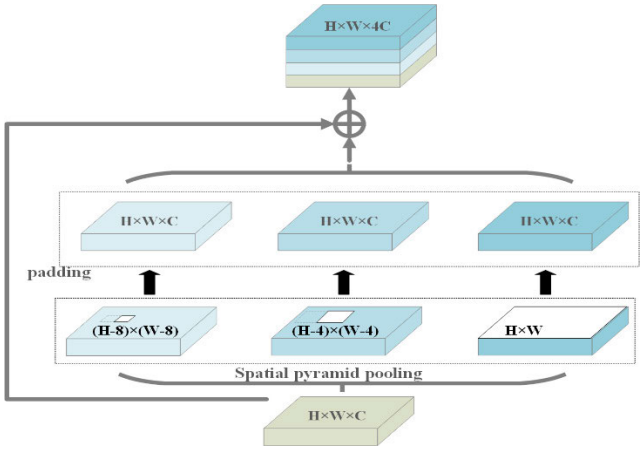


FIGURE 5. Spatial pyramid pooling network structure.

9, 5; in the SPP module connected to the 26×26 feature layer, the pooling core size is set to 26, 22, 18; In the SPP module connected after the 52×52 feature layer, the pooling core size is set to 52, 48, 44; In the SPP module connected after the 104×104 feature layer, the pooling core size is set to 104, 100, 96. After the spatial pyramid pooling operation with different sizes of pooled cores for the four feature maps, larger dimension feature maps containing more local and global information can be obtained.

III. EXPERIMENT

Firstly, in order to verify the effectiveness of each stage improvement of the algorithm and the overall performance of the proposed algorithm, this paper trained and tested each stage improvement on the Pascal VOC2007 [28] data set respectively, and compared it with the MAP and f1-score of YOLO v3 and several improved YOLO v3 algorithms. Then, this paper tests the targets of different sizes on the MS COCO [29] data set, and compares the detection effects with YOLO v3 and SSD (300) [30] to test the detection effect of the algorithm on different sizes of targets. Finally, in order to test the detection effect of the algorithm in the actual complex scene, this paper selects 3 sets of detection images with different target conditions for testing, and compares the detection effect with YOLO v3.

A. EXPERIMENT PLATFORM

All experiments in this paper were carried out on the Ubuntu 16.04 system. The hardware configuration of the system is Intel Core i5 9400 CPU, NVIDIA GTX1080 GPU, and 16G RAM. The software platform for this experiment includes Anaconda3, TensorFlow1.14.0, Keras2.1.5, Opencv-python 3.4.2.16 and Opencv-contrib-python 3.4.2.16.

B. PARAMETERS SETTING AND TRAINING PROCESS

We take into account that momentum can increase stability and make the model learn faster. In training experiment of this paper, the random gradient descent method of momentum (SGD) is used to update the weight, and the momentum parameter is set to 0.9, and the weight attenuation regular term is 0.0005.According to the experience and conclusion of YOLO series algorithms and the memory usage of GPU, we set the batch size to 32.The total number of iterations is set as 50000, and the segment learning rate is dynamically set according to the number of iterations: the initial learning rate is set as 0.001, and the exponential attenuation mechanism is adopted to achieve better convergence effect and faster completion of training by using a smaller learning rate

at the later stage of training. When the training iterations reach 40000 times, the learning rate is set to 1/10 of the initial learning rate, that is, 0.0001. When the training iterations reach 45000 times, the learning rate is set to 0.00001. The loss curve is shown in Figure 6. The setting of learning rate makes the training process get normal convergence effect.

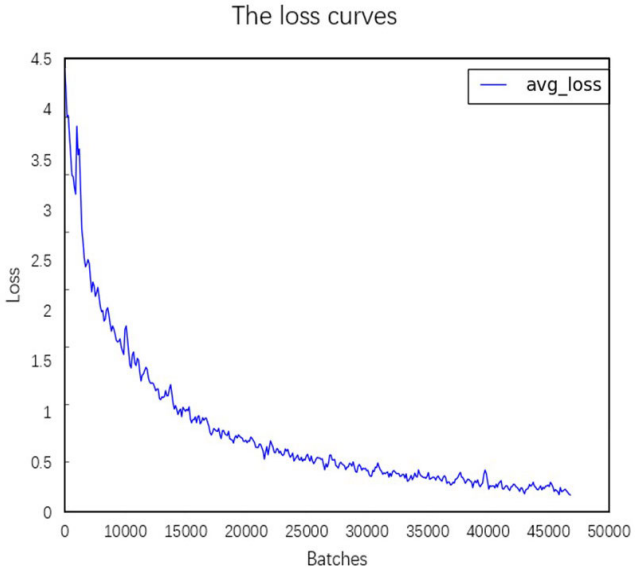


FIGURE 6. The average loss curve.

The training process is divided into two stages: The first stage: Freeze part of the network and train only the underlying weights. During training, the completed model weights are stored every 3 epochs, and the storage mode is set to *save_weights_only* and *save_best_only*; The second stage: use the network weights that have been trained to continue training, and set all the weights to be trainable. The *fit* data of the above two stages both start from the 50th epoch and train to the 100th epoch. When the condition is triggered, it terminates early. After the second stage of training is completed, the output network weight is the final model weights.

C. VALUATION INDEX

This paper uses the commonly used model evaluation indicators for target detection: mean average precision (mAP) and average precision (AP), Precision (P), Recall(R) and comprehensive evaluation index f1-score. Among them, the calculation formulas of P and R are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

In (3) and (4), Precision indicates the accuracy of classification, that is, the proportion of samples classified as positive cases that are actually positive cases; Recall represents the proportion of all positive samples that are correctly predicted; TP represents the number of positive samples predicted to be positive, FN represents the number of positive samples predicted to be negative. Further, on the basis of Precision and Recall, we introduce f1 to comprehensively measure the classification capability of the model. Its calculation formula is as follows:

$$f_1 = \frac{2P \cdot R}{P + R} \tag{5}$$

For a certain category C, the Precision of C is represented by P_C, the number of pictures containing category C is represented by N_C *(Total Images)*, and the total number of categories is represented by N *(classes)*. Then the calculation formulas for AP_C and mAP are:

$$AP_C = \frac{\sum P_C}{N_C\ (Total\ Images)} \tag{6}$$

$$mAP = \frac{\sum AP}{N\ (classes)} \tag{7}$$

D. EXPERIMENT AND RESULT ANALYSIS

1) PERFORMANCE VERIFICATION AND COMPARISON OF THE IMPROVED ALGORITHM

In order to verify the reliability of the network after adding FCN-Densenet (D-YOLO), adding the fourth detection scale (M-YOLO) and adding SPP module (FF-YOLO), we tested them separately and compared them with YOLO v3, Reference23 and Reference 32. This group of experiments uses the Pascal VOC2007 data set, which includes 5011 training sets and 4952 test sets. The types of objects in Pascal VOC2007 are divided into four categories: Vehicle, Household, Animal, and Person, and 20 small types. In order to more accurately evaluate the classification capability and detection performance of the model, f1-score and mAP were used as evaluation indexes in this experiment. The test results are shown in Table 1.

TABLE 1. Experimental results of improved algorithm in each stage.

Detection algorithm	Modules			Results	
	FCN-DenseNet	Optimization of prediction mechanism	SPP	mAP	f1-score
YOLO v3	—			78.3%	86.58
Reference23	—			82.7%	—
Reference32	—			81.5%	—
D-YOLO	√			80.1%	88.04
M-YOLO	√	√		81.8%	88.67
FF-YOLO	√	√	√	84.1%	90.23

Table 1 shows that: The algorithm proposed in this paper at each stage can effectively improve the detection average accuracy and f1-score; Compared with YOLO V3, D-YOLO's mAP increased by 1.8% and F1score by 1.46%. It can be seen that the addition of dense connection structure enriches the expression ability of convolution features and improves the classification ability of the model; After further

optimizing the multi-scale prediction mechanism, the network structure (M-YOLO) increased by 1.7% and 0.63% compared with D-YOLO's mAP and f1-score, respectively. The mAP and f1-score of the multi-scale prediction model FF-YOLO were significantly improved compared with those of YOLO v3 and improved models of YOLO v3.

The experimental results show that the improvement of the algorithm at each stage can improve the learning and prediction ability of the model, which conform to the theoretical analysis and is effective.

Based on the above experiments, the AP and mAP of FF-YOLO in 20 categories of objects in the Pascal VOC2007 data set were calculated. The specific results are shown in Figure 7. It can be seen from the results that the recognition accuracy of 11 categories is higher than the average level of the whole 20 categories. The algorithm has a high recognition accuracy for common targets such as buses, cars and bicycles in daily life. For targets with relatively small size, such as cats and dogs, the detection accuracy is also high. Through the data analysis of the two types of objects with low detection accuracy of chairs and potted plants in the data set, it is found that these two types of images lack obvious texture and edge information, the color depth is not obvious, and the target is similar to the background, which leads to the decline of the algorithm's ability to recognize these two types of objects.

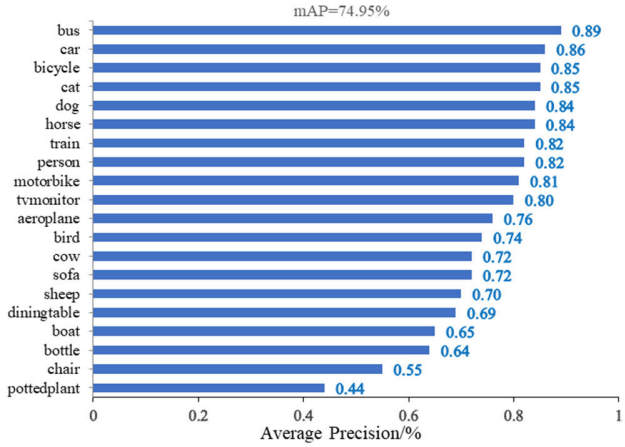


FIGURE 7. Test results of FF-YOLO on Pascal VOC2007.

2) VERIFICATION EXPERIMENT OF DETECTION EFFECT OF DIFFERENT SIZE TARGETS

The COCO data set is used in this group of experiments, which contains more than 140,000 images. The target classification reaches more than 80 categories, and there are more than 5,000 instance objects in most categories. Compared with Pascal VOC data set, images in COCO data set contain richer semantic information and location information. Therefore, COCO data set is used to conduct experiments to verify the detection effect of the proposed algorithm on targets of different sizes.

In this paper, the detection performance of FF-YOLO is compared with the classical one-stage target detection algorithms SSD (300) and YOLO v3. We used FF-YOLO, YOLO v3 and SSD (300) to train 50000 batches on the images in trainval of COCO data set, and run *compute_ map. py* to calculate the mAP. Mark a point every 5000 times, and get the comparison results as shown in Figure 8.

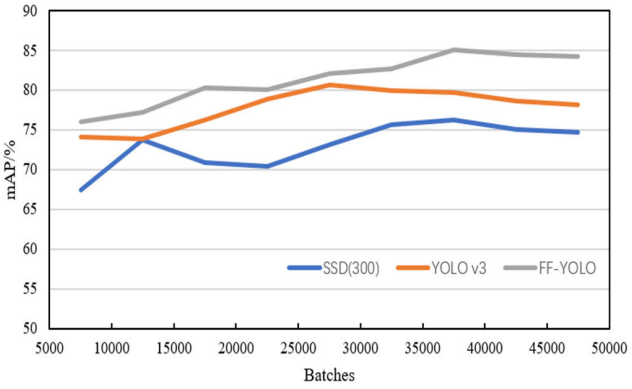


FIGURE 8. mAP of SSD (300), YOLO v3 and FF-YOLO.

In Figure 8, we can see the optimal training weight batches of each algorithm. The overall mAP of FF-YOLO is better than the other two algorithms. Save the weight file with the highest mAP among 50,000 batches as the FF-YOLO optimal model for subsequent follow-up experimental verification.

Further, the API tools in coco data set are used to calculate the AP of FF-YOLO, SSD (300) and YOLO v3 for S, M and L size targets respectively. Among them, "S" means the target frame is smaller than 32×32 pixels, "M" means the target frame is larger than 32×32 and smaller than 96×96 pixels, and "L" means the target frame is larger than 96×96 pixels. The experimental results are shown in Figure 9.

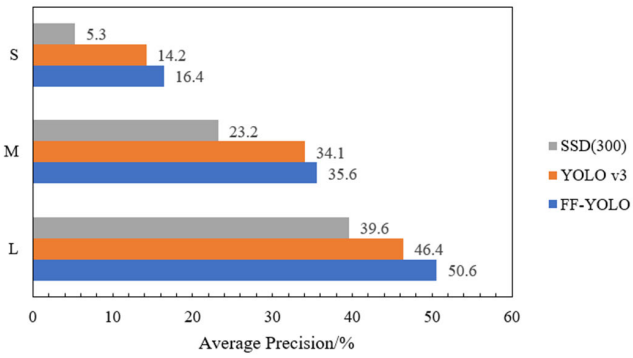
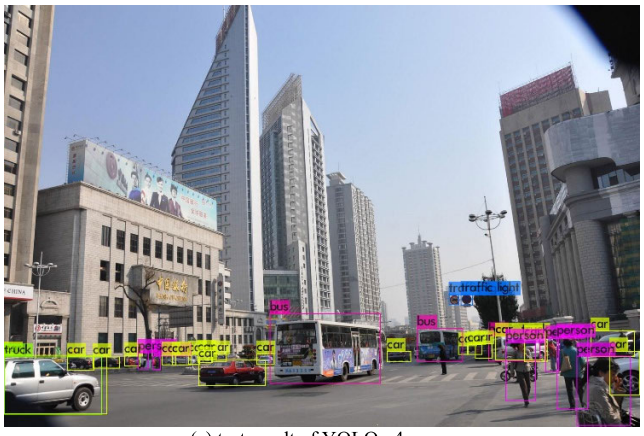


FIGURE 9. AP of SSD (300), YOLO v3 and FF-YOLO for targets with different sizes.

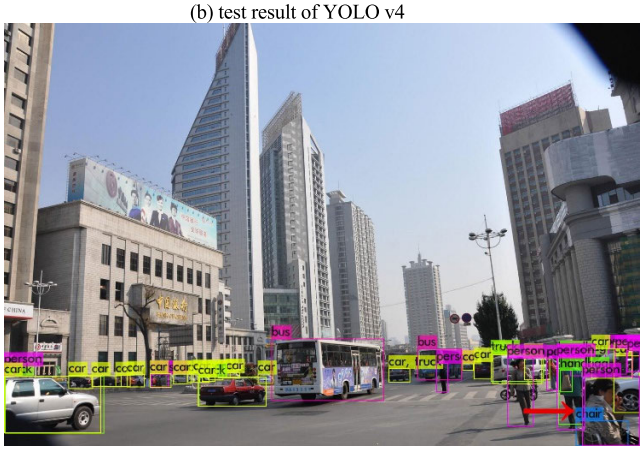
It can be seen from Figure 9 that FF-YOLO has significantly improved the recognition accuracy of targets of different sizes. The reasons are as follows:

SSD (300) extracts feature images of different scales for multi-scale detection, but only uses the large-scale shallow



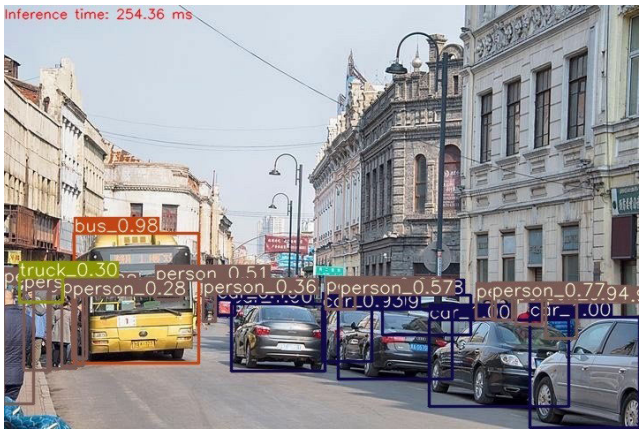
(a) test result of YOLO v4

(b) test result of YOLO v4

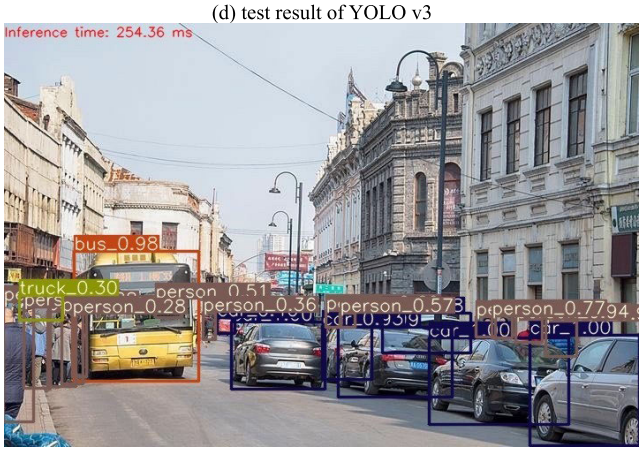


(c) test result of FF-YOLO

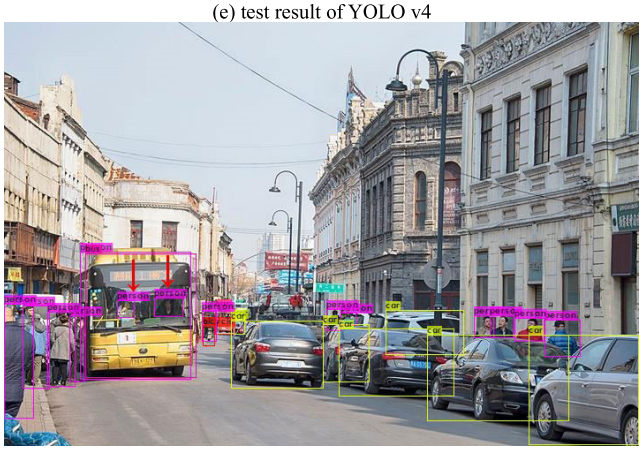
FIGURE 10. Detection results of street view of urban roads (1).



(d) test result of YOLO v3



(e) test result of YOLO v4



(f) test result of FF-YOLO

FIGURE 11. Detection results of street view of urban roads (2).

feature maps to detect small targets. The semantic information in the large-scale shallow feature images is not rich, which leads to the poor detection effect of SSD (300) on small targets.

YOLO v3 uses the multi-scale prediction mechanism based on the idea of feature pyramid, and uses the feature layer which combines the deep feature and the shallow feature to predict. However, the extraction of the location information in the shallow feature map is insufficient, and there is a large positioning error in the process of small target detection, which affects the detection effect of the algorithm for small target;

The network FF-YOLO proposed in this paper has a strong multi-scale target detection capability, which is mainly due to the network structure of deep feature fusion, which integrates the position information in the shallow feature map and the semantic information in the deep feature map. By deepening the feature extraction network, the detection performance of the model for large targets is improved. In addition, FF-YOLO overcomes the problem that SSD (300) and YOLO

TABLE 2. The detection results of instance images.

Detection algorithm	Figure 10			Figure 11			Figure 12		
	YOLO v3	YOLO v4	FF-YOLO	YOLO v3	YOLO v4	FF-YOLO	YOLO v3	YOLO v4	FF-YOLO
Number of detected targets	33	38	38	23	23	24	9	10	11
Detection accuracy	84.8%	86.8%	89.5%	82.6%	95.7%	91.7%	88.9%	100%	100%

(g) test result of YOLO v3

(h) test result of YOLO v4

(i) test result of FF-YOLO

FIGURE 12. Detection results of campus street view.

v3 have insufficient extraction of semantic information and location information in the detection process, and enhances the detection ability of targets of different scales through multi-scale feature fusion model. As a result, the average accuracy of the model is significantly improved despite the different target sizes.

3) THE VERIFICATION EXPERIMENT OF THE DETECTION EFFECT OF THE ACTUAL COMPLEX SCENE

In order to more intuitively verify the detection effect of FF-YOLO on targets of different scales in complex scenes, this group of experiments uses the COCO data set for training and selects life images with complex scenes, including urban road street scenes and school street scene images for testing. The detection images of these actual scenes have the characteristics of different background brightness and different target sizes, which can well verify the robustness of the improved network. Compare the detection effects of YOLO v3, YOLO v4, and FF-YOLO in these detection images, and the detection results are shown in Figure 10 to Figure 12.

Among them, Figure 10 and Figure 11 are the detection comparisons of urban road street view maps, and Figure 12 is the detection effect of campus street view map. After manual confirmation, the test results of each example picture are summarized as shown in Table 2. Among them, the detection accuracy represents the proportion of correct results in the number of all detected targets.

Firstly, the images with dense target distribution and overlapping targets are selected for detection. The results are shown in Figure 10. It can be seen that FF-YOLO can also accurately detect the less obvious target object in the picture, such as the target *chair* in the lower right corner. In the case of serious target overlapping, the detection accuracy of FF-YOLO is 4.7% and 2.7% higher than that of YOLO v3 and v4. Then, the detection accuracy of this algorithm is 9.1% higher than that of YOLO v3 when the distribution of objects to be detected is relatively close. As shown in Figure 11, FF-YOLO can accurately identify the fuzzy targets *person* in the bus without being disturbed by the complex background. Finally, as shown in Figure 12, FF-YOLO can accurately identify the target *fire-hydrant* incorrectly identified by YOLO v3 and the overlapping target *car* not identified by YOLO v4, and the detection accuracy reaches 100%.

From the above 3 groups of comparative experiments, we can see that compared with YOLO v3 and YOLO v4, FF-YOLO has a higher sensitivity to small targets, which reduces the missed detection rate of small targets and provides more accurate detection results in the case of

overlapping or similar targets. That is because it integrates richer feature information in the learning process and has better adaptability to different detection target sizes, so as to achieve accurate recognition of multi-scale targets.

IV. CONCLUSION

Aiming at the problem of low recognition accuracy of YOLO v3 in the case of large target size difference or overlap, this paper effectively combines dense connection structure, multi-scale prediction and spatial pyramid pooling with YOLO v3.In this way, without changing the high generalization ability of YOLO V3, the structure of the model is improved, so that the deep fusion of features in the network is strengthened, semantic information and location information can be extracted and transmitted more effectively, so as to improve the overall performance of the model. This paper conducted experiments on PASCAL VOC and COCO data sets and made quantitative and qualitative comparisons. The experimental results show that the f1-score and mAP of FF- YOLO are significantly improved compared with that of YOLO v3, indicating that the improved network model has better robustness. Through the detection experiment of example images, it can be seen that FF-YOLO has better detection effect for overlapping and small targets in complex scenes than YOLO v4.

However, the improvement of this paper only starts from the perspective of network structure, and does not consider the problems of loss function and other aspects. In fact, the large difference of target size also has a certain influence on the regression accuracy of coordinates in training process. So, using mean-square error (MSE) to calculate the loss function of coordinate regression cannot solve the scale sensitivity problem in target detection. In the future, the model will be improved in flexible selection of loss function and data enhancement through amplification of data sets, so as to further improve the detection performance of the network in complex scenes.

REFERENCES

[1] M. Choudhary, "Automatic target detection and tracking in forward-looking infrared image sequences using morphological connected operators," *J. Electron. Imag.*, vol. 13, no. 4, p. 802, Oct. 2004.

[2] Y. Jin, Y. Fu, W. Wang, J. Guo, C. Ren, and X. Xiang, "Multi-feature fusion and enhancement single shot detector for traffic sign recognition," *IEEE Access*, vol. 8, pp. 38931–38940, 2020.

[3] A. Talukder, T. Sheikh, and L. Chandramouli, "Real-time intelligent pattern recognition, resource management and control under constrained resources for distributed sensor networks," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Budapest, Hungary, Jul. 2004, pp. 1321–1326.

[4] S. Shin, T. Kwon, G.-Y. Jo, Y. Park, and H. Rhy, "An experimental study of hierarchical intrusion detection for wireless industrial sensor networks," *IEEE Trans. Ind. Informat.*, vol. 6, no. 4, pp. 744–757, Nov. 2010.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, vol. 4, pp. 1–14, Sep. 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[7] 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, Oct. 2005, pp. 886–893.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[9] M. Tan, G. Pan, Y. Wang, Y. Zhang, and Z. Wu, "L1-norm latent svm for compact features in object detection," *Neurocomputing*, vol. 139, pp. 56–64, Sep. 2014.

[10] H. X-H, J. G-D, and T. L-N, "Survey of ship detection in SAR images based on deep learning," *Laser Optoelectron. Prog.*, vol. 58, no. 4, pp. 53–64, 2021.

[11] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang, "New generation deep learning for video object detection: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 3, 2021, doi: 10.1109/TNNLS.2021.3053249.

[12] Z. Zheng, L. Lei, H. Sun, and G. Kuang, "A review of remote sensing image object detection algorithms based on deep learning," in *Proc. IEEE 5th Int. Conf. Image, Vis. Comput. (ICIVC)*, BeiJing, China, Jul. 2020, pp. 34–43.

[13] J. Jiang, H. Xu, S. Zhang, Y. Fang, and L. Kang, "FSNet: A target detection algorithm based on a fusion shared network," *IEEE Access*, vol. 7, pp. 169417–169425, 2019.

[14] 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.

[15] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1440–1448.

[16] 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.

[17] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[18] A. Bochkovskiy, W. C-Y, and H. Y. MarkLiao, "YOLOv4: Optimal speed and accuracy of object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Apr. 2020, pp. 1–17.

[19] 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)*, Seattle, WA, USA, Jun. 2020, pp. 390–391.

[20] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 936–944.

[21] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8759–8768.

[22] J. M-R, L. H-B, L. G-Q, and L. Y-P, "Infrared dim and small target detection network based on spatial attention mechanism," *Opt. Precis. Eng.*, vol. 29, no. 4, pp. 843–853, Apr. 2021.

[23] M. Gong, L. Y-Y, and L. G-N, "A ship detection method for remote-sensing images based on improved YOLOv3," *Electron. Opt. Control*, vol. 27, no. 5, pp. 102–107, 2020.

[24] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 379–387.

[25] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.

[27] 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.

[28] M. Everingham, S. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: Aretrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[29] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[30] X. Hua, W. X-Q, D. Wang, M. Z-Y, and S. F-M, "Multi-objective detection of traffic scenes based on improved SSD," *Acta Opt. Sinica*, vol. 38, no. 12, pp. 221–231, 2018.

[31] T. Xu, G.-J. Tang, and Q.-P. Liu, "Improved YOLOv3 based on dilated convolution and Focal Loss," *J. Nanjing Univ. Posts Telecommun.*, vol. 40, no. 6, pp100-108, Dec. 2020.

CHEN BAOYUAN was born in Harbin, Heilongjiang, China, in 1970. He received the B.S. and M.S. degrees in measurement technology and instrument from Harbin University of Science and Technology, Harbin.

Since 2006, he has been an Assistant Professor in instrument science and technology and information and communication engineering with the School of Measurement-Control Technology and Communications Engineering, Harbin University of Science and Technology. He is currently the Vice Director of the Experimental Center at the Higher Educational Key Laboratory for Measuring and Control Technology and Instrumentation of Heilongjiang Province. He has authored two books and six articles. His research interests include manufacture of network communication protocol conversion module, detection and identification of defects in transparent film, voice activity detection, and vision location algorithm.

LIU YITONG was born in Harbin, Heilongjiang, China, in 1996. She received the B.S. degree in electronic information engineering from Northeast Agricultural University, in 2018. She is currently pursuing the M.S. degree in information and communication engineering with Harbin University of Science and Technology.

She has authored one article and holds five patents. Her research interests include multi scale target detection, convolutional neural network model, and pattern recognition.

SUN KUN was born in Harbin, Heilongjiang, China, in 1982. He received the B.S. degree in measurement-control technology and instrument from Harbin University of Science and Technology, in 2004, and the M.S. and Ph.D. degrees in instrument science and technology from the Harbin Institute of Technology, Harbin.

Since 2014, he has been an Assistant Professor in instrument science and technology with the School of Measurement-Control Technology and Communications Engineering, Harbin University of Science and Technology. He has authored more than 20 articles. His research interests include image processing, pattern recognition, spectral analysis, and multi-spectral temperature measurement.