IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 November 2022, accepted 20 November 2022, date of publication 24 November 2022,
date of current version 7 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224594

RESEARCH ARTICLE

Detection of Surface Defects on Railway Tracks Based on Deep Learning

MAOLI WANG, KAIZHI LI, XIAO ZHU, AND YINING ZHAO
School of Computer Science, Qufu Normal University, Rizhao 276800, China

Corresponding author: Maoli Wang (wangml@qfnu.edu.cn)

This work was supported in part by SD2019NJ007 from the Department of Agriculture and Rural Affairs of Shandong Province, and in part by the Natural Science Foundation of Shandong Province under Grant ZR20111260301.

ABSTRACT The detection of rail surface defects is very important in railway transportation. However, the edge defects on both sides of the rail and the multi-scale variation between different types of defects both pose challenges to the detection of rail surface defects. In order to solve the above problems, this paper proposes a novel rail surface defect detection network, YOLOv5s-VF. First, we design a sharpening functional attention mechanism (V-CBAM) that contains two key components: adaptive channel attention (F-CAM) and sharpened spatial attention (SSA). In F-CAM, we use one-dimensional convolution with adaptive convolution kernels for cross-channel connections, which reduces the number of parameters of the attention mechanism without affecting its performance. In SSA, we design a sharpening filter suitable for spatial attention, which is used to enhance the attention to the edge position defects of railway tracks and enhance the detection effect of the network on edge defects. Second, we construct a microscale adaptive spatial feature fusion (M-ASFF), which adds a high-resolution feature extraction layer to enhance the details of the underlying features of tiny defects. At the same time, in order to prevent the loss of detailed information and the excessive increase of the parameters of the model, the low-resolution feature layer is removed. Combined with adaptive spatial feature fusion, it can prevent the semantic conflict caused by the fusion of features at different scales. Finally, given the lack of labeled public rail surface defect datasets, this paper is based on the collection of real rail images and manually labels defects to train an object detection network and open source it. The experimental results show that YOLOv5s-VF outperforms the existing rail surface defect detection methods with a detection accuracy of 93.5% and a detection speed of 114.9 fps.

INDEX TERMS YOLOv5, attention mechanism, adaptive spatial feature fusion, rail surface defect.

I. INTRODUCTION

In In recent decades, the rapid development of high-speed railways has made railways one of the foremost essential modes of transportation for Chinese citizens [1]. The rail is an important support for the railway track, and its role is to ensure that the train runs forward and bears the extrusion of the wheels. With the aggravation of railway transportation tasks, the negative pressure on railways is also increasing, as is the harsh environment and the ageing of materials. These are the things that cause defects on the rail surface. Therefore, timely detection of the health status of the rail surface is essential for preserving the security of the train. In traditional rail surface defect detection, the inspection methods are mostly ultrasonic [2], eddy current [3], and magnetic particle [4] methods. Although these methods can detect rail surface defects, they require much time.

Based on traditional machine vision techniques, researchers combine imaging systems with defect detection. These methods usually go through manual analysis of rail surface defect images to design manual features or predefined features and classify defects by a classification network. In [5], defects are captured and segmented through an automatic visual inspection system. In [6], a local Weber-like contrast (LWLC) algorithm was proposed to enhance track images. In addition, in [7], the original data were converted into three-dimensional point cloud patterns, and the digital rail surface defects were reconstructed. In [8], morphological operations were combined with defect detection for the detection and shape extraction of rail defects. In [9], the

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

inhomogeneous illumination of the rail surface is eliminated by partitioned edge features (PEFs). In [10], the rail image is divided into three scales and filtered and segmented by the coarse and fine models.

A method for detecting surface defects based on 3D laser reconstruction was proposed in [11]. In practice, these methods have proven to be effective for rail defect detection. However, their common disadvantage is that the accuracy and recall of the detection results are usually low. Some defects, such as cracks, dents, and spalling, are challenging to detect and categorize.

With the rapid development of deep learning, we combine deep learning with rail images to achieve more accurate detection of rail defects. Existing deep learning-based defect detection strategies can be broadly categorized as follows:

Image classification methods, such as hybrid detection methods consisting of wavelet packet transforms (WPTs), kernel principal component analysis (KPCA) and SVMs, are proposed in [12]. For a limited data sample, the defect images are treated as sequential data, and pixel lines were classified by [13] using a one-dimensional convolutional neural network to extract features. These studies are prospective for identifying rail damage but are unable to detect and localize multiple defects on a single image.

Pixel segmentation uses a classification network to pixelate defects [14], [15] or large pixels [16], [17]. A local pixel inhomogeneity factor (LPIF)-based image enhancement method was proposed in [18] to enhance the contrast pairs of defective images and to segment defects by the maximum interclass difference method (Otsu). A pixel-level segmentation network based on deep feature fusion was proposed in [19] to improve defect segmentation accuracy by combining a multibranch decoder and the multibranch structure of the attention module to reply with defect details. The method segments the defect contours at a high level, while pixel classification is more sensitive to greyscale changes in the background. In addition, the fixed large prime number is not conducive to the scale adaptation of defect segmentation.

For sliding windows, the original image is divided into several subimages for detection [20]. In [21], the use of three different scales of sliding windows is proposed, and different computational methods are established to cope with the variations of different scales of defects. In [22], the size of the sliding window is obtained by the least squares method to address the need for traditional sizes that are difficult to adapt to the detection target. A temporal spectrogram was obtained by [23] using a sliding window to scan the morphological feature signals of the defect. However, fixing the size of the sliding window can, to some extent, lead to localization errors in multiscale defects.

For defect detection based on anchor frames, the field uses Faster-RCNN [24], represented by two stages, and YOLOv3 [25]. To address the low detection accuracy and large number of network parameters in rail defect detection in this field, many scholars have proposed different improvement strategies. The recurrent neural network (CRF-RCNN) proposed in [26] is a two-stage extractor combining bilateral convolutional networks and conditional random fields, which helps to smooth out constraints or obtain fine-grained inspection results. An improved single-shot multibox detector (SSD) is proposed in [27], which adds a full convolutional compression and excitation (FCSE) module. The attentional neural network based on joint intersection consistency (IoU)-guided centroid estimation (CCEANN) proposed in [28] achieves high accuracy in defect detection. In [29], researchers use MobileNetv3 as the backbone network of YOLOv4 to extract image features and simultaneously apply depthwise separable convolutions, enabling lightweight networks and real-time detection of railway surfaces. In [30], the researchers used the fuzzy C-means algorithm to re-cluster the anchor boxes based on YOLOv4 and added a shallow feature layer to solve the problem of occlusion of hanging insulators and power components. In [31], contextual information is integrated into the backbone of the Swin Transformer, and skip-connected BiFPN is used to improve detection of small objects.

To sum up, in the area of defect detection based on deep learning, a large number of researchers have conducted research on problems such as small targets for defect detection and proposed effective improvement methods. However, in the above detection methods, the models are generally large (greater than 50 MB), which is not conducive to porting them to mobile devices, and the detection speed is low. Therefore, we need to explore a new model that can achieve a balance in detection accuracy, detection speed, and model size so that it has the characteristics of being fast (greater than 90 FPS), highly precise, and small model(size below 20 MB).

Most rail surface defects are caused by rolling fatigue contact (RFC) and can be classified into the following categories depending on the texture characteristics: cracks, dents, spalling and transverse fractures [32]. Although the above methods have played a positive role in the detection of rail surface defects, some unresolved problems still exist due to the complexity of the railway environment. The challenges of computer vision-based rail surface defect detection are as follows.

(1) Rail surface defects are multiscale and have uneven foreground and background. The number of different types of defects varies, and some defects have a small sample size, which creates an imbalance of defect categories and makes it difficult to target them. Defects of the same type are multiscale in nature; for example, spalling and concave have extreme aspect ratios.

(2) Variations in the reflective properties of the track surface: The brightness and contrast between the track surface and defects in the image will change due to variations in natural light and different weather conditions in the railway environment. Moreover, the contrast between defects and

wheel-rail contact areas is high, but the contrast between defects and background in rough metal areas is low, which results in uneven illumination for defect detection on the track surface.

(3) Interference in complex environments: The debris on both sides of the rails, fasteners and surface stains, wear and tear increases the difficulty of computer vision-based defect detection. In addition, as rails are exposed to the external natural environment, they are affected by sunshine, shadows and rain, resulting in reduced imaging quality and hence detection effectiveness.

Aiming at the above problems, this paper proposes a new detection framework for the detection of concave and exfoliation defects dominated by small objects and multi-scale objects. Its core contributions are as follows: (1) In order to solve the problem of difficult and effective detection of edge defects, we propose a hybrid attention mechanism (V-CBAM) with a sharpening function that enhances the attention mechanism by constructing a sharpener suitable for the spatial attention module. Focus on edge defects so that the network can effectively locate them.At the same time, the one-dimensional convolution of the adaptive convolution kernel is used in the channel attention module for cross-channel connection to reduce the amount of parameters in the attention module. Compared with other attention modules, this module can effectively locate edge defects. (2) Aiming at the situation that the detailed features of tiny defects will be ignored in multi-scale feature fusion, we propose a microscale adaptive spatial feature fusion (M-ASFF). By adding a feature extraction layer for small defects, the detailed features of small defects are enhanced, and the low-resolution feature layer is removed to prevent the loss of information about the underlying features. At the same time, adaptive spatial feature fusion is used to adaptively assign weights to features of different scales to prevent semantic conflicts caused by fixed weight fusion. (3) Given the lack of labeled datasets of rail surface defects, we constructed a rail surface defect dataset to train convolutional neural networks based on real rail images and published it to the outside world.

The remaining portions of the article are organized as follows: Section II presents pertinent prior research, while Section III introduces the methodology; Section IV describes the construction, comparison experiments, and ablation experiments of the rail surface defect dataset; and Section V concludes the paper.

II. RELATED WORK

This section introduces the current mainstream attention mechanisms, including ECANet [33], SENet [34], CBAM [35] and other modules, as well as adaptive spatial feature fusion (ASFF) and YOLO target detection network. Among them, the application of the YOLO network in defect detection is analyzed, which lays the foundation for the construction of the track surface defect detection network YOLOv5s-VF.

A. ATTENTION MECHANISM

The visual attention mechanism is a brain signal processing mechanism unique to human vision that enables humans to find salient regional locations in complex natural environments [36]. Inspired by this, the attention mechanism was introduced to computer vision, which draws on the attention mode of human vision and has been widely used [37]. Attention mechanisms can be simply divided into three categories: channel attention, spatial attention, and coordinate attention mechanisms. SENet [33] introduced the first effective channel attention mechanism, which adopts the squeeze and excitation structure to adaptively recalibrate the channel feature response and shows good performance in DCNN. As an improved version of SENet, ECANet [34] replaces the fully connected layer (MLP) in SENet with a one-dimensional convolution with adaptive convolution kernels to achieve cross-channel interaction. DANet [38] proposes location attention and channel attention mechanisms to enhance the correlation between global feature fusion and semantic feature quality. CBAM [35] is a hybrid attention mechanism that combines channel and space, where channel attention is used to learn what to pay attention to, while spatial attention is used to learn where to pay attention. In CBAM, global pooling or maximum pooling structures are no longer used, and instead, a combination of the two is used, using the form of addition in the channel and the form of stacking in the space. This paper proposes a novel lightweight attention mechanism to strengthen the attention of CBAM to image edge features.

B. ADAPTIVE SPATIAL FEATURE FUSION (ASFF)

The main problem solved by the FPN network is the insufficiency of target detection in dealing with multiscale changes. It performs multiscale feature fusion to improve the richness of features. However, this fusion is carried out in a fixed way; that is, in the detection branch, it is suitable to detect the low-level features of small objects, the high-level features of large objects, and the middle-level features. Merging occurs in the form of direct splicing or direct addition, which causes conflicts between features at different scales. This conflict is mainly manifested when the target is detected in a feature map of a certain scale and regarded as a positive sample, and the feature maps of other scales are regarded as the background in the corresponding area when the area contains both large and small objects. The information carried between the feature layers of different scales for detecting large and small objects is contradictory.

To address this issue, Songtao et al. [39] proposed ASFF in 2019 and applied it to YOLOv3 with outstanding results. This method can funnel features of various scales and retain only valid features. For the features of a certain scale, we first adjust the features of other scales to the same size and then find the best fusion weight coefficient through training. In this paper, three-layer ASFF is used.

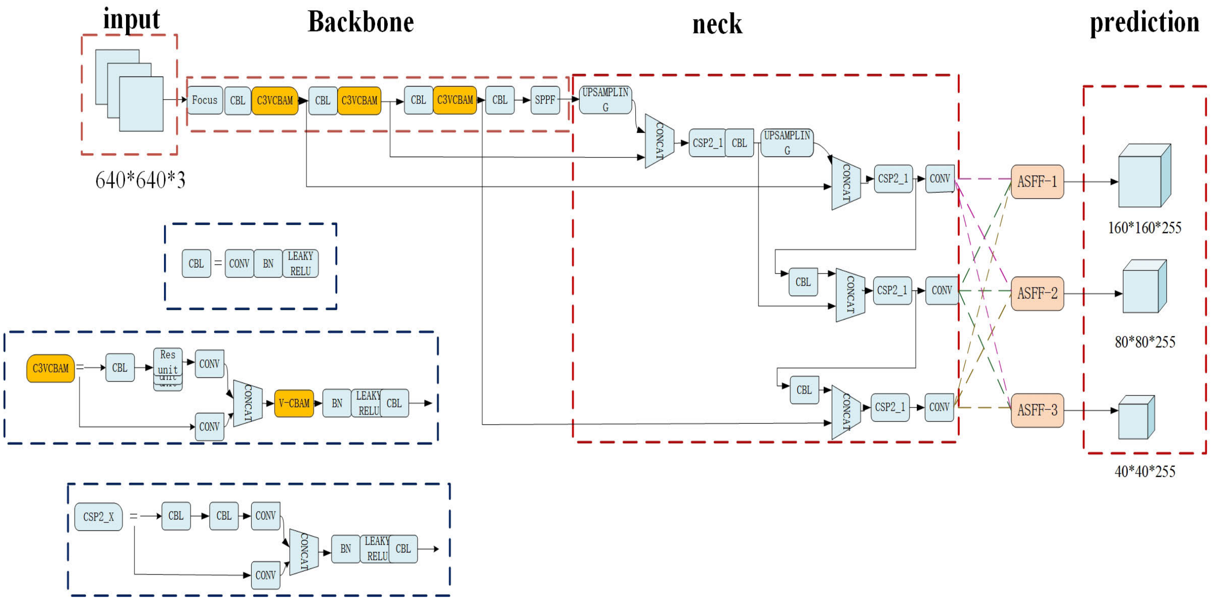


FIGURE 1. YOLOv5s-VF model.

C. YOLO TARGET DETECTION NETWORK

In this subsection, we introduce the application of the YOLO series network in defect detection. In [40], the authors use a YOLOv2-based network to detect void defects in airport runways, combined with incremental random sampling (IRS) and ResNet 18. The localization of hole defects is enhanced, and the recall rate of defect detection is improved.

In [41], researchers used a YOLOv3-based network to detect bridge surface defects (cracks and exposed steel bars), and using transfer learning and data enhancement, the mAP of bridge surface defect detection was increased by 6–10%.

In [42], researchers based on YOLOv4 network tunnel lining defect detection. After using EfficientNet and depthwise separable convolution, the detection average accuracy and F1 of tunnel lining defects are improved to 81.84% and 81.99%, respectively.

In [43], the researchers detected insulator defects based on the YOLOv5 network. The F1 value of insulator defect detection was 96.2% when the channel attention mechanism SE was combined.

In summary, the YOLO target detection network is widely used in defect detection. With the update of the YOLO series of networks, the performance of defect detection has been greatly improved. However, there are still some issues that need to be resolved, as follows: (1) Some defects are distributed in the edge part of the image, and the gray value of the defect is the same as the gray value of the edge, so it is difficult to be detected. (2) The scale of defects varies greatly, and the fusion method of fixed weights will lead to the loss of the underlying detail features, which will make the detection effect of small target defects worse.

III. OUR METHOD

In this paper, YOLOv5s [44] is used as the benchmark model, and the constructed sharpening attention mechanism V-CBAM and microscale adaptive spatial feature fusion M-ASFF are applied to the model to improve the detection performance of small defects and multi-scale defects. The improved YOLOv5s method is named YOLOv5s-VF, and Fig. 1 shows the overall structure of the method.

A. SHARPENING ATTENTION MECHANISM (V-CBAM)

In Part A of the related work, we introduced the characteristics and working principle of the CBAM attention mechanism, which has a relatively good performance in the field of object detection, but when we applied the CBAM attention mechanism to the detection of rail surface defects, we did not achieve a big improvement. We analyze the reason because, because the rail surface contains many defects combined with the edge of the rail surface, as shown in Fig. 2, the edges of these defects are attached to the side of the rail, and the gray value of the defect is the same as the gray value of the side, it is difficult to effectively localize these defects using the CBAM attention mechanism. Therefore, to address the above problems, we construct a sharpening filter to enhance the edge details of defects. At the same time, in order to reduce the number of parameters brought about by the introduction of CBAM, we use one-dimensional convolution with adaptive convolution kernels for cross-channel connections.We name the new attention mechanism V-CBAM. Through the visualization of the heat map shown in Fig. 3, we can clearly see that our V-CBAM can pay more attention to the defects that fit the rail surface than the source network and CBAM and is more sensitive

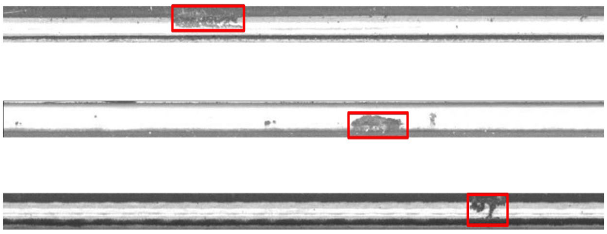


FIGURE 2. Defects combined with rail edges.

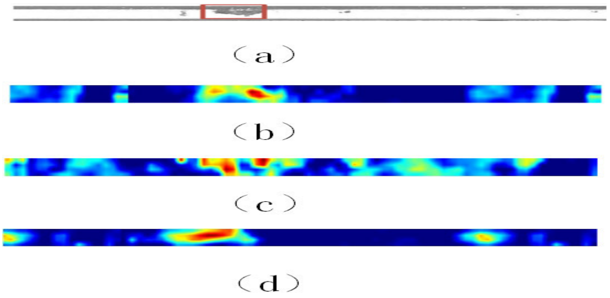


FIGURE 3. Defect heatmap visualization? (a) original image (b) YOLOv5 (c) CBAM (d) V-CBAM.

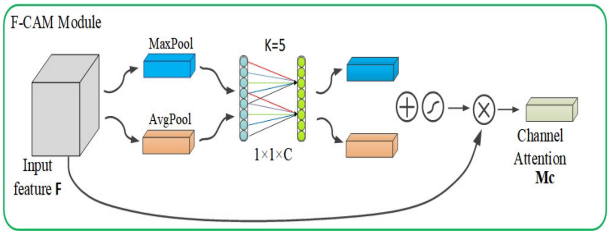


FIGURE 4. F-CAM structure diagram.

to the edge portion of the defect. The specific workflow of V-CBAM is as follows:

First, in the channel attention mechanism (CAM), the fully connected layers in the CAM are replaced with 1D convolutions with adaptive convolution kernels. The inherent effect of one-dimensional convolution is that it is not fully connected. Each convolution process only works with part of the channel, that is, to achieve appropriate cross-channel interactions instead of full-channel interactions such as those of the fully connected layer.

It is empirically shown that using 1D convolution instead of fully connected layers can significantly reduce model complexity while maintaining model detection accuracy. The improved CAM is named F-CAM. The structure of F-CAM is shown in Fig. 4.

(1) The given feature map is first made subject to Max Pool and Avg Pool in producing two [1,1,C] vectors. F1 and F2 are the features remaining after global maximum pooling and global average pooling. The working process of F-CAM is as follows:

(2) The two feature vectors are subjected to a one-dimensional convolution with a convolution kernel length of K to aggregate the information of the k channels in the channel neighbourhood. The size of K is adaptively determined by the number of input channels and calculated using Formula 1:

$$k = \left| \frac{l_b C}{2} + \frac{1}{2} \right|_{odd} \tag{1}$$

k represents the size of the convolution kernel, C represents the number of channels of the input feature map, the base is indicated, and 1 is added if the result is even. The size of the convolution kernel can be altered, which is an advantage of the adaptive convolution kernel. The convolution kernel will grow correspondingly as the number of channels increases.

(3) The two features are connected after convolution according to the corresponding elements and converted into probability values (normalized) between 0 and 1 through the sigmoid function. A channel of attention is generated.

(4) The generated channel attention is then broadcast and expanded to H×W×C along two dimensions in space and then dotted with the original feature map to output a final feature map of channel attention.

Second, we construct a sharpening filter and apply it in spatial attention in order to enhance the recognition of object edges by the spatial attention module, focusing on the "location" and "how much" of the object edge to strengthen the edge for better localization, which is a complementary enhancement to the target. The sharpening filter is constructed as follows:

(1) Define a 5×5 initialization kernel. Since the defined kernel is a 2-dimensional list, it cannot directly participate in the operation as a parameter of convolution. It needs to be converted into one that satisfies (batch, width, height, channel) through dimension transformation. Only four-dimensional tensors can participate in operations. Therefore, first convert it to a tensor tensor using the FloatTensor function in Pytorch and expand it to 4 dimensions using 2 times unsqueeze (0).

(2) In order to adaptively learn and change the sharpening kernel according to the characteristics of the input image to meet the learnability of the training parameters, the parameter function is used to convert them into trainable parameters so that for different input features, the adaptive learns the most effective sharpening kernel.

(3) In the forward propagation, the 0 and 1 channels of the feature map are extracted from the input feature map, and X1 and X2 are defined to perform convolution operations on the extracted channels, and the results of the convolution output by the convolution kernels X1 and X2 are in the column direction splicing and compress the number of channels by 3×3 convolutions as the result.

We embed the constructed sharpening filter into spatial attention and name it SSA. The SSA module structure is shown in Fig. 5, and the specific implementation steps are as follows:

(1) First, the output feature map of the channel attention module is made subject to Max pool and average pool to generate two weight vectors of [H,W,1], namely, maximum pooling and average pooling by channel. The number of

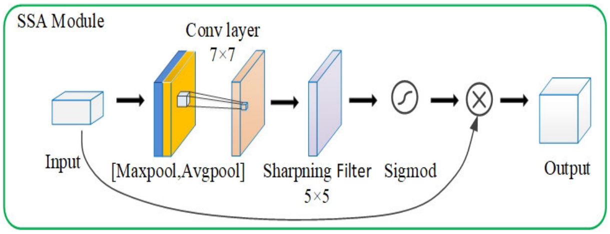


FIGURE 5. SSA structure diagram.

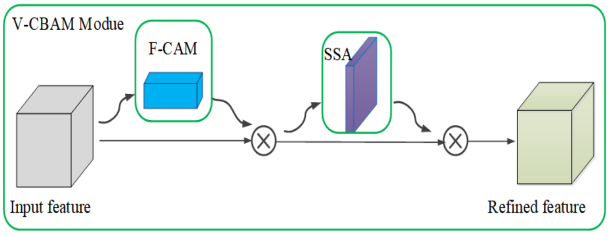


FIGURE 6. Structure diagram of V-CBAM.

channels is changed from [H,W,C] to [H,W,1], pooling all channels of the same feature point.

(2) The generated feature map is spliced into a feature map of [H, W, 2] based on the number of channels. Then, after a 7×7 convolution operation, the dimension is reduced to 1 channel number, that is, [H×W×1]. Form the spatial feature weights of [H,W,1].

(3) Pass the obtained feature map through a 5×5-order sharpening filter, and stack the convolution output results X1 and X2 in the column direction to form a [H×W×1] feature map. And through the sigmoid function, it is converted into a probability value between 0 and 1 (normalized).

(4) The obtained spatial weight [H,W,1] is multiplied by the original feature map [H,W,C] so that each [H, W] point on the feature map is assigned a weight, and the weight represents the importance of this area, which allows the network to adaptively focus on areas with larger weights.

The calculation method of our sharpened spatial attention module is as follows:

$$\begin{aligned} M_S(I) &= \delta(S_{5\times5}(f^{7\times7}([Maxpool(I); Avgpool(I)]))) \\ &= \delta(\mathrm{S}_{5\times5}(\mathrm{f}^{7\times7}(\mathrm{I}_{\max};\mathrm{I}_{avg}))) \end{aligned} \tag{2}$$

Formula 2, δ is the sigmoid function, where $S_{\mathrm{n}\times n}$ represents the convolution for sharpening when the sharpener size is n and $f^{7\times7}$ is a convolution with a parameter of 7×7, whose channel is equal to the channel of the feature map.

The F-CAM module and the SSA module are combined to form the V-CBAM module, as shown in Fig. 6. V-CBAM can be expressed by the following Formula 3 and Formula 4:

$$I_C = M_c(I) \otimes I \tag{3}$$

$$I_{sc} = M_s(M_c(I)) \otimes I \tag{4}$$

M_C and M_S represent the F-CAM model and the SSA model, respectively. The dot product is presented by elements. The precise results of the two parts are I_c and I_{sc}.

Finally, we explore the different ways in which the attention mechanism can be inserted. As the attention mechanism is a plug-and-play module, it can be adapted to any part of the YOLOv5 network in principle, but the introduction of the attention mechanism will inevitably bring in some parameters. Embedding too many parameters will lead to an overly large number of model parameters and an overly complex network model, making it difficult to reach the fitted state in a short time during training.

To satisfy the need for a lightweight model, we consider adding the attention mechanism only at the backbone of the YOLOv5 network because the scope of the attention mechanism is global, so adding it only at the backbone will also have an impact on the whole network. In Fig. 7, on the left, attention is added at the last layer of the backbone, and on the right, it is added to the csp residual module. The first method requires modifying the connections and number of channels of the entire network layer, which needs to be adjusted manually when performing experiments; the second method is integrated with the C3 module, which does not require modifying the number of network layers and channels and is convenient for conducting experiments. In this paper, we use the second addition method, adding V-CBAM to the C3 module to form a new C3VCBAM module and replacing all the C3 modules in the backbone.

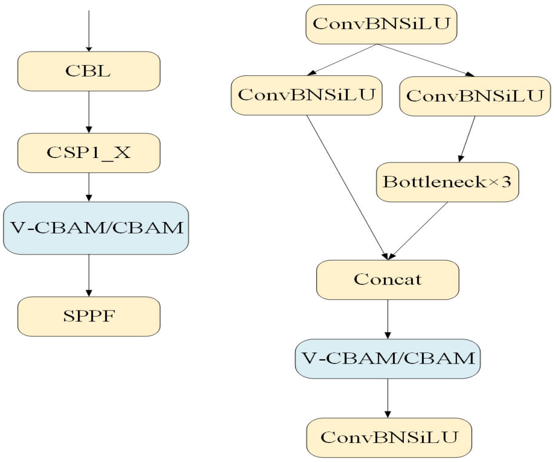


FIGURE 7. Insertion position structure of the attention mechanism.

Our construction process is as follows: in the common.py file of YOLOv5, we define the F-CAM and SSA classes and the C3VCBAM class and call the F-CAM and SSA classes in C3VCBAM; in yolo.py, we register our modified C3VCBAM class; and in the yaml file, we replace the original C3 module with C3VCBAM.

B. MICROSCALE ADAPTIVE SPATIAL FEATURE FUSION(M-ASFF)

To perform feature fusion on the features extracted by the backbone network, YOLOv5s adopts Feature Pyramid (FPN) and Path Aggregation Network (PANet). However, this fusion is a fixed-weight fusion and adopts a direct splicing method, which will lead to the loss of low-scale features containing more location information. At the same time, as shown in

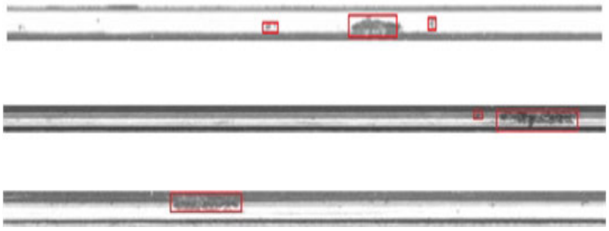


FIGURE 8. Multiscale defects.

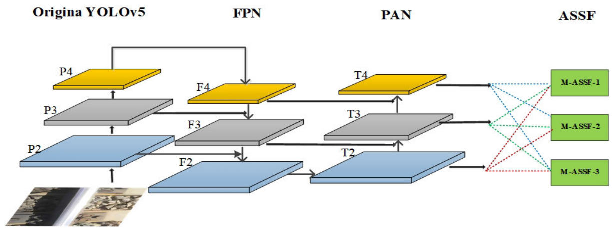


FIGURE 9. M-ASFF structure diagram.

Fig. 8, the larger size of the defects on our rail surface is 26 × 260 pixels, and the smaller size of the defects is 19 × 22 pixels. Since the YOLOv5 source network uses three scales of detection heads (with 640×640 size input as an example): are 20×20, 40×40, 80×80, corresponding to the detection of 32×32, 16×16, 8×8 size targets, the smaller the size of the detection head, the larger the corresponding receptive field, which can extract richer semantic information for detecting large objects; on the contrary, the smaller the receptive field, the more position and detail information can be extracted for detecting small objects. So even if our larger-sized object becomes a 1×8-scale feature after downsampling by a factor of 32, it will be treated as a pixel in the 20×20-scale detection layer and ignored by the network. For small-sized defects, due to their small pixel values, their own feature information will be lost after multi-layer convolution operations; even the 80×80-scale detection layer is not easy to detect. Therefore, in order to realize the feature extraction of small defects and fully fuse the semantic information of high-level features with the location information of low-level features, we construct a microscale adaptive spatial feature fusion (M-ASFF). Fig. 9

depicts the M-ASFF structure.

First, we output a 160×160-scale feature layer after the first C3 module in the backbone network and remove the last layer of convolution in the backbone network and the corresponding output layer in the neck part, which corresponds to removing 20×20 detection. As shown in Fig. 10, the 160×160 detection head can be used to detect tiny objects with a size of 4×4 pixels.In this way, it can meet the needs of feature extraction for small defects, and at the same time, removing redundant 20×20 detection heads can reduce the loss of details in defect features and position information and at the same time prevent the excessive increase of network parameters that results in a complex network structure. From the visualization of the feature map in Fig. 11, it can be seen that the P2 layer can obtain more defect features than the P3 layer, and the defect shape is clearer. Therefore, adding a 160×160-scale detection layer can effectively improve the feature extraction of micro-sized defects.

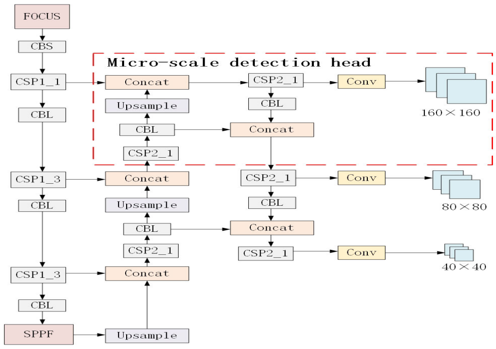


FIGURE 10. Micro-scale detection head.

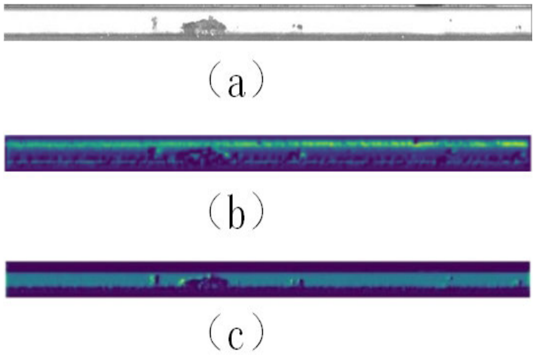


FIGURE 11. Feature map visualization: (a) original image (b) P3 layer (c) P2 layer.

Then, a three-layer ASFF is added after the output three-scale feature layer, which we name MASFF-YOLOv5s. Three-layer ASFF is capable of adaptively studying weights and combining multiscale data for adaptive feature fusion. M-ASFF then performs weighted fusion after adjusting the T2, T3,and T4 layers to have identical numbers of channels and resolutions. The entire procedure consists of the first step, feature size adjustment, followed by the second step, adaptive fusion.

Since the three different scale feature layers of YOLOv5s have distinct channel counts and resolutions, the upsampling and downsampling techniques of each scale must be modified. For upsampling, we compress the number of channels of features to level 1 using a 1×1 convolution, and then, we use interpolation to increase the resolution. For 1/2 proportion downsampling, we use 3×3 convolutional layers, which simultaneously modify the number of channels and the resolution. Before the convolution for 1/4 proportion, a two-step max pool layer is added. M-ASFF-2 is taken as an example. First, the channel counts of T3 and T4 are equalized through convolution, and after interpolation processing, the size is adjusted to the same ratio as T2. Then, M-ASSF-2 is weighted and fused through the obtained weights. The whole process of obtaining M-ASFF consists of the following

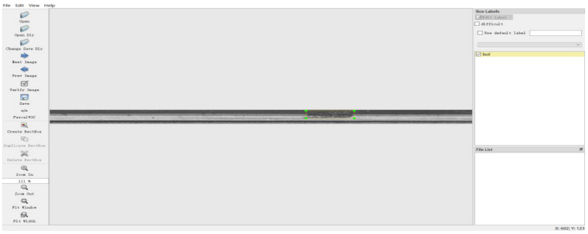


FIGURE 12. **Marking process.**

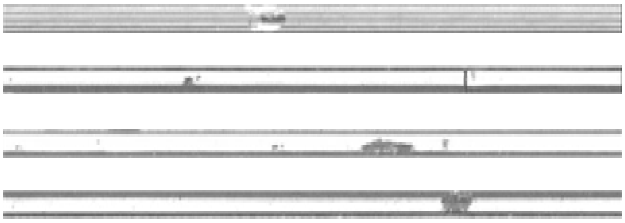


FIGURE 13. **Partial picture of the dataset.**

formula 5.

$$\begin{aligned} M-ASFF_{ij}^{l} &= \alpha_{ij}^{l} \times I\text{nter}polate(Conv(T_{ij}^{4\rightarrow l},1,1),4) \\ &\quad + \beta_{ij}^{l} \times Inter-polate(Conv(T_{ij}^{3\rightarrow l},1,1),2) \\ &\quad + \gamma_{ij}^{l} \times Conv(T_{ij}^{2\rightarrow l},1,1) \qquad (5) \end{aligned}$$

$M-ASFF_{ij}^{l}$ represents feature map $M-ASFF_{ij}^{l}$ eigenvectors at (i,j), $T_{ij}^{n\rightarrow l}$represents the feature vector adjusted from level n to level l on the feature map divided by position (i, j) on the PAN network, Interpolate(I, i) indicates that the step size is i, and the interpolation value is I. α_{ij}^{l}, β_{ij}^{l}, and γ_{ij}^{l} represent the adaptively learned spatial weighting factors of the feature space from three levels to the l-level. α_{ij}^{l}, β_{ij}^{l}, and γ_{ij}^{l} can be simple scalar variables shared across all channels, $\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1$, $\alpha_{ij}^{l}\beta_{ij}^{l}$ and γ_{ij}^{l} are ∈[0,1], and the defined as Formula 6:

$$\alpha_{ij}^{l} = \frac{e^{\lambda_{\alpha ij}^{l}}}{e^{\lambda_{\alpha ij}^{l}} + e^{\lambda_{\beta ij}^{l}} + e^{\lambda_{\gamma ij}^{l}}} \qquad (6)$$

α_{ij}^{l}, β_{ij}^{l}, and γ_{ij}^{l} are defined by using $\lambda_{\partial_{ij}}^{l}$, $\lambda_{\beta_{ij}}^{l}$, and $\lambda_{\gamma_{ij}}^{l}$ as the softmax function of the control parameters, but $\lambda_{\partial_{ij}}^{l}$, $\lambda_{\beta_{ij}}^{l}$, and $\lambda_{\gamma_{ij}}^{l}$ are defined through the changed feature map $T_{ij}^{4\rightarrow l}$, $T_{ij}^{3\rightarrow l}$ $T_{ij}^{2\rightarrow l}$ obtained by 1×1 convolution. Therefore, they can be learned by standard backpropagation.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The effectiveness of V-CBAM and M-ASFF in ablation experiments is demonstrated in this section. In addition, the method is then compared to the target detection algorithm to validate its efficacy on the track defect dataset, and then, the experimental conclusions are presented.

A. DATASET

To evaluate the efficacy and robustness of YOLOv5s-VF, a dataset consisting of real rail inspection video supplied by the Chinese Academy of Railway Sciences was created. On the track, high-speed cameras with a resolution of 1920 × 1080 were utilized to record forty 100-minute videos from various sections of the railway site. With these acquisitions, the video of the railroad tracks was converted to 1250 × 55 pixel stills using frame-by-frame interception, and the images were saved in PNG format. Using the LablImage annotation tool, the generated images were marked. To enhance the capacity of the YOLOv5 network to detect flaws, we utilized the minimum outer rectangle method for marking, with the goal of including the defects while framing the background as little as possible. Fig. 12 depicts the marking process.

The tagged files are in XML format, and the names of the original images are maintained. The dataset contains a total of 5027 images of the concave and exfoliation classes studied in this paper, and a representative example of the dataset 9is shown in Fig. 13. There are approximately 2604 images in the concave category and 2423 in the exfoliation category. In the exfoliation category, severe exfoliation samples and small exfoliation samples account for approximately 15% and 34%, while in the concave category, large concave samples and small concave samples account for approximately 15% and 19%, respectively. In this dataset, 4022 images are utilized for training and 1005 for testing. All noted flaws must be verified by technicians.

B. EVALUATION STANDARD

The assessment index serves as a crucial foundation for assessing the effectiveness of the target detection model. The evaluation indicators include precision (P), recall (R), average precision (AP), average category precision (mAP), frame processing speed (FPS), and F1 score. In our experiments, we utilized AP, mAP, F1, frame processing speed (FPS) [45], and model size.

The ideal state for the target detection model is when both accuracy and recall are relatively high, but in reality, an increase in accuracy will result in a decrease in recall, and vice versa [9]. Consequently, the PR curve and F1 score are utilized to analyse the model's performance from a global perspective. The PR curve sorts all detection targets within each category based on their scores and calculates the precision and recall from greatest to least. The curve formed by connecting various points along the coordinate axis is known as the PR curve. The F1 value is the weighted harmonic average of accuracy and recall. When there is a discrepancy between the P and R indicators, the F1 value can counterbalance the anomaly between them. The calculation process is shown in Formula 7:

$$F1 = \frac{2*P*R}{P+R} \qquad (7)$$

In general, AP and mAP indicators are used in multicategory detection tasks. A particular variety of AP refers to the region encompassed by the PR curve introduced previously. mAP is

TABLE 1. Comparison of YOLOv5 models with different network depths.

Methods	AP		mAP	F1	FPS	Model
	Neg	Bol				size(MB)
YOLOv5n	86.9%	85.2%	86.1%	83.0%	186.9	3.85
YOLOv5s	89.3%	87.9%	88.6%	88.3%	139	13.6
YOLOv5m	92.1%	88.7%	90.4%	90.3%	92.9	40.6
YOLOv5l	92.7%	90.4%	91.5%	91.5%	63.2	80.9
YOLOv5x	94.9%	91.2 %	93.1 %	93.0 %	44.64	165

TABLE 2. YOLOv5 models with different network depths add V-CBAM attention module comparison.

Methods	AP		mAP	F1	FPS	Model
	Neg	Bol				size(MB)
YOLOv5n-VCBAM	89.8%	87.8%	88.8%	88.0%	177.6	4.23
YOLOv5s-VCBAM	91.9 %	90.4 %	91.2 %	91.0 %	129.4	13.9
YOLOv5m-VCBAM	92.0%	89.2%	90.6%	90.5%	80.9	41.48
YOLOv5l-VCBAM	90.4%	86.5%	89.5%	88.4%	51.35	84.62
YOLOv5x-VCBAM	92.1%	89.2 %	90.6 %	89.0 %	33.3	168.7

the average of all AP categories. The calculation process is shown in Formulas 8 and 9:

$$AP = \int_0^1 PRdR \tag{8}$$

$$mAP = \frac{1}{C}\sum_{c_i \in C} AP_{(C_i)} \tag{9}$$

In addition to detection accuracy, the speed of a target detection algorithm is an important evaluation factor. Real-time detection can only be achieved when the speed is high [46]. FPS is a metric that measures the rate of target detection. It indicates how many frames (images) per second the network can process (detect). Assuming that it takes the target detection network 0.02 seconds to process one image, the frame rate is 1/0.02 = 50.

C. PARAMETER SETTING

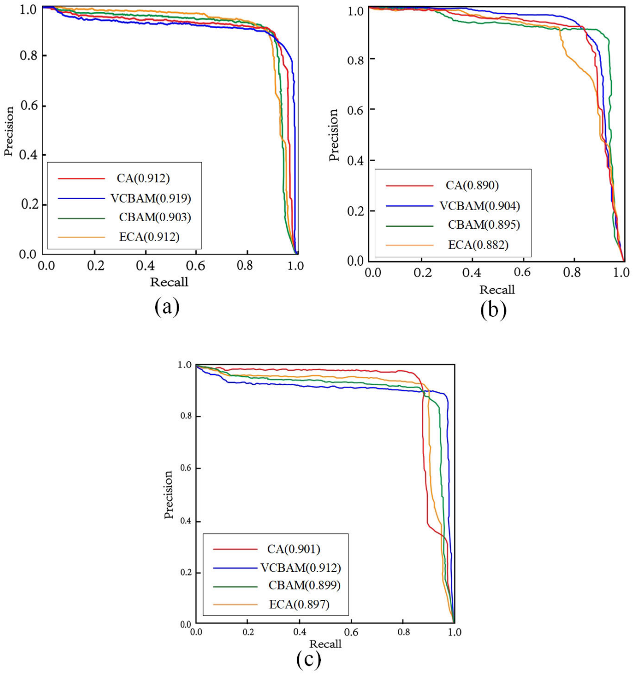
All experiments are conducted on a server running Ubuntu 16.04 with an Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40 GHz, an NVIDIA RTX 3090 GPU, and 24G video memory using the PyTorch framework. Note that none of the parameters in the experiments were loaded with pretrained models. A total of 300 epochs were trained in the experiment, and the batch size was set to 8. The initial learning rate was set to 0.001, and the NMS threshold was set to 0.5.

D. ABLATION EXPERIMENT

To evaluate the functionality of V-CBAM and M-ASFF, we quantitatively evaluate and analyse the results of different settings of YOLOv5s.

1) THE EFFECT OF V-CBAM

In this subsection, we explore the impact of V-CBAM on the task of rail surface defect detection using a self-made rail surface defect dataset. Since the introduction of the attention mechanism will increase the number of parameters, it is not appropriate to add too many attention mechanism modules. In this experiment, we only added the attention mechanism to the backbone to verify its impact. We first tested the detection effects of YOLOv5 models with different depths, and then verified the effects of different depth models after adding V-CBAM by introducing V-CBAM. Meanwhile, we conduct ablation experiments on the V-CBAM module to find the best use of V-CBAM. Finally, we compare the detection effects of different attention mechanisms and verify the effectiveness of the improved attention mechanism in defect detection.



(a)

(b)

(c)

FIGURE 14. (a) Concave PR curve (b) Exfoliation PR curve (c) mAP-PR curve.

All parameters were kept stable during the experiment. The YOLOv5 network models of different depths are shown in Table 1. It can be seen that as the network depth increases, the detection accuracy continues to rise, but the speed also decreases. From Table 2, we can conclude that YOLOv5s-VCBAM has the highest mAP value among YOLOv5 models

TABLE 3. Comparison of different attention mechanisms.

Methods	AP		mAP	F1	FPS	Model
	Neg	Bol				size(MB)
YOLOv5s	89.3%	87.9%	88.6%	88.3%	139	13.6
YOLOv5s+CBAM	90.3%	89.5%	89.9%	90.0%	127.7	14.0
YOLOv5s+F-CAM	89.5%	88.5%	89.0%	89.1%	134	13.7
YOLOv5s+SSA	0	0	0	0	0	0
YOLOv5s+(F-CAM+SSA)	91.9 %	90.4 %	91.2 %	91.0 %	129.4	13.9

TABLE 4. Comparing the effects of different edge detection operators on V-CBAM.

Methods	Kernel 3*3	Kernel 5*5	AP		mAP
			Neg	Bol	
YOLOv5s			89.3%	87.9%	88.6%
YOLOv5s+V-CBAM	Yes		89.9%	90.2%	90.0%
YOLOv5s+V-CBAM		Yes	91.9%	90.4%	91.2%

TABLE 5. Comparison of different attention mechanisms.

Methods	AP		mAP	F1	FPS	Model
	Neg	Bol				size(MB)
YOLOv5s+CBAM	90.3%	89.5%	89.9%	90.0%	127.7	14.0
YOLOv5s+ECA	91.2%	88.2%	89.7%	89.6%	133.6	13.7
YOLOv5s+CA	91.2%	89.0%	90.1%	89.0%	131.9	13.82
YOLOv5s+VCBAM	91.9 %	90.4 %	91.2 %	91.0 %	129.4	13.9

with V-CBAM attention modules embedded at different depths. The mAP value of YOLOv5n and YOLOv5s both increased by about 2.6% after embedding the V-CBAM module, and the improvement effect was obvious. YOLOv5m increased the mAP value by 0.4% after using the V-CBAM module, but the V-CBAM module was used in YOLOv5l and YOLOv5x. After that, the AP, mAP, and F1 values of the concave and exfoliated types all decreased to different degrees, and the deeper the network, the more severe the decrease. This is because, with the increase of network depth, the model complexity and parameter volume gradually increase, the convergence speed gradually decreases, and there is also an effective problem of gradient propagation, which will make it difficult to fit the parameters of the attention module during training, good result. Therefore, our attention module is more suitable for lightweight models. Since the basic detection accuracy of YOLOv5n is low, even if the V-CBAM attention module is added, its mAP value fails to reach more than 90%, so we choose YOLOv5s as the benchmark model. The detection accuracy of YOLOv5s after using the V-CBAM module is comparable to that of YOLOv5l.

Table 3 show that V-CBAM using the combination of F-CAM+SSA has achieved the highest index value, indicating that V-CBAM is better than CBAM, especially since V-CBAM has achieved 91.2% compared to that of the source model, the mAP increased by 2.6%, the exfoliation AP increased by 2.5%, and the mAP increased by 2.6%.

In Table 3, we found an interesting phenomenon: when only the spatial attention mechanism SSA module is used, all indicators are 0. We speculate that the SSA module is not suitable for use alone because the edge enhancement module in SSA is directly placed into the feature extraction network when the feature map is not squeezed or stimulated by the channel attention mechanism, which would induce the weight of the contour segment to fluctuate, resulting in considerable loss and the failure to successfully achieve convergence during the training process. Therefore, the SSA module is not suitable for use on its own.

We compared the effects of using different edge detection operators on V-CBAM, as shown in Table 4. Compared to the effects of 3×3 order and 5×5 order initialize the kernel on V-CBAM, we found that a 5×5 sharper with a higher order can produce better results. Because the 5×5 initialize the kernel is larger than 3×3 and has a large receptive field, more feature information can be captured. Therefore, for this paper, we chose 5 × 5 initialize the kernel.

Through Table 5, comparing the channel attention mechanism ECA and the coordinate attention mechanism CA, it can be concluded that the AP, mAP, and F1 values of our V-CBAM attention module in Neg and Bol are higher than those of the ECA module and the CA module. The degree of mAP was higher by 1.5% and 1.1%, and the F1 value was higher by 1.4% and 1.2%, respectively. As shown in Fig. 14, it can be concluded that the area enclosed by the PR curve of V-CBAM is larger than the area enclosed by the contrasting attention modules.

2) INFLUENCE OF M-ASFF

In this section, we explore the impact of micro adaptive feature fusion (M-ASFF) on the model. Since the main goal of M-ASFF is to achieve adaptive fusion of features at different scales, we selected comparative models of different feature fusion methods, mainly including YOLOv3 using

TABLE 6. M-ASFF comparison results.

Methods	AP		mAP	F1	FPS	Model
	Neg	Bol				size(MB)
YOLOv3	82.2%	78.2%	80.2%	80.0%	67.3	206.0
YOLOv5s	89.3 %	87.9%	88.6%	88.3%	139	13.6
YOLOv5s+TBIFPN	92.2 %	90.2%	91.2%	91.1%	103.3	18.7
YOLOv5s+MASFF	92.6%	90.8%	91.7%	91.5%	126.4	14.32

TABLE 7. Detection results of different feature layers combined with ASFF.

Methods	AP		mAP	F1
	Neg	Bol		
P345	89.3 %	87.9%	88.6%	88.3%
P345+ASFF	90.8 %	90.2%	90.5%	90.1%
P234	90.3 %	89.7%	90.0%	89.4%
P234+ASFF	92.6%	90.8%	91.7%	91.5%

TABLE 8. YOLOV5s-VF comparison results.

Methods	AP		mAP	F1	FPS	Model
	Neg	Bol				size(MB)
YOLOv5s	89.3 %	87.9%	88.6%	88.3%	139	13.6
Grid RCNN	90.4 %	90.4%	90.4%	90.2%	28.7	280.3
CCEANN	93.0%	92.2%	92.6%	92.8%	45	270.0
YOLOv5s-VF	94.3%	92.7%	93.5%	93.2%	114.9	14.8
Faster RCNN	88.2%	85.5%	86.9%	86.4%	11.1	159.5
Yolov4	91.2%	89.7%	90.4%	89.2%	35.8	244.0
SSD	90.0%	86.9%	88.5%	85.7%	43.6	203.7

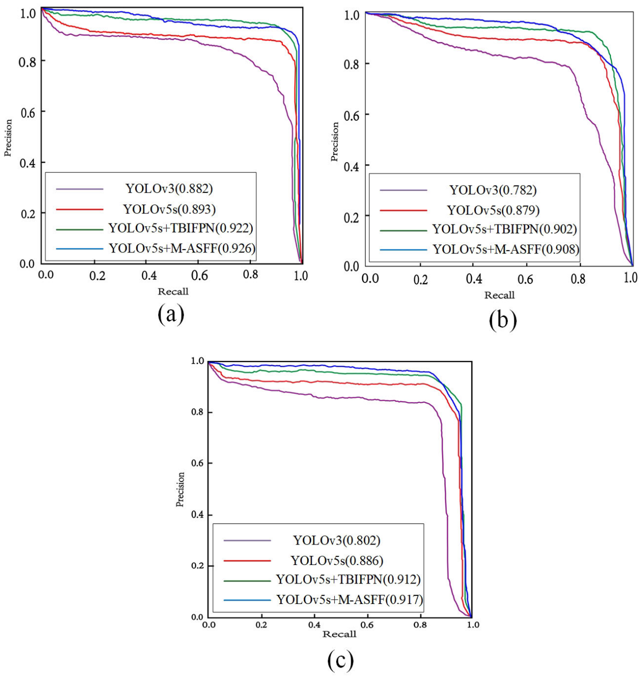


FIGURE 15. (a) Concave PR curve (b) Exfoliation PR curve (c) mAP-PR curve.

only the FPN structure, the YOLOv5s source model using FPN+PANet, and a combination of The Swin Transformer's Weighted Bidirectional Feature Pyramid Network (TBIFPN) [31]. From Table 6, we can conclude that the YOLOv5s model with the addition of the M-ASFF module performs the best on the rail surface defect dataset. On the basis of the source YOLOv5s, M-ASFF only increases the model size by 0.72 MB, the mAP is increased by 3.1%, and the AP of the concave and exfoliation types is increased by 3.3% and 2.9%, respectively. The effect is significantly improved.Compared with TBIFPN, our M-ASFF has a 0.4% higher mAP in detection results and 0.4% and 0.6% higher AP in concave and exfoliated categories, respectively; however, our model is faster than TBIFPN in detection speed out of 23 fps, the model is smaller.It can be seen that the feature fusion method of FPN+PANet+M-ASFF has a better detection effect on the surface defects of the rail. Through Fig. 15, the area enclosed by M-ASFF in the PR curve is larger than that of other comparison models, which can also reflect that the performance of the YOLOv5s model using M-ASFF is better than the three compared models.

We also conducted an experimental analysis of the impact between the micro-object detection layer and adaptive spatial feature fusion. Through Table 7, we compared the performance of different scale feature layers combined with ASFF and found that the combination of micro-scale detection layer P2, small-scale detection layer P3, and medium-scale detection layer P4 combined with ASFF has the best detection effect. Compared with the combination of P3, P4, and P5 layers combined with ASFF, our combination method improves the mAP by 1.2%, and the AP of concave and exfoliation types increases by 1.8% and 0.6%, respectively. Our analysis is that the defect size on the surface of the rail is small, so it cannot be detected in the P5 layer. The P4 and P3 layers actually play the role of detection. However,

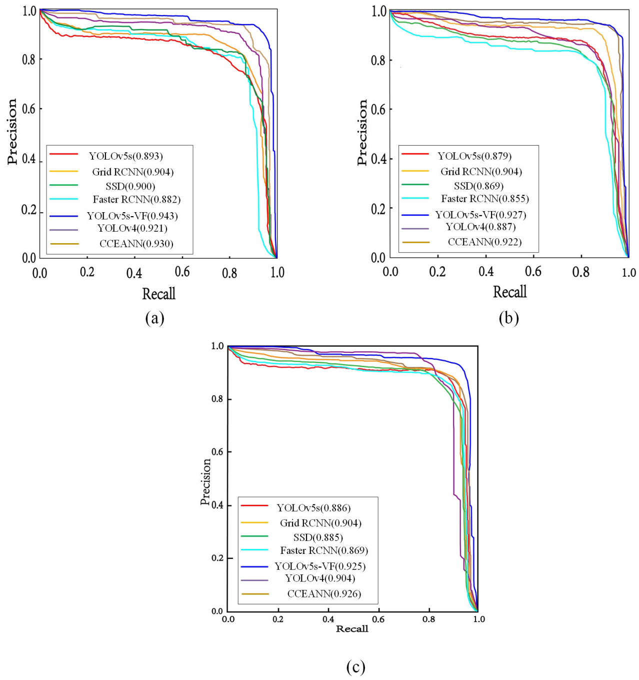


FIGURE 16. **(a) PR curve representing concave (b) PR curve representing exfoliation (c) PR plot of mAP.**

some small defects are due to their small pixels. When downsampling to extract features, it will be ignored as a pixel, so adding a P2 layer can better extract the features of this part of the defect, and after weighting by ASFF, the multi-scale features are further fused.

The above experimental results show that the performance of the model is improved after adding the micro-detection layer, but the feature fusion of YOLOv5s is of a fixed scale, so the performance is not optimal. By using ASFF to adjust the scale of the feature map, the performance can be further improved. This experiment shows that M-ASFF can perform weighted fusion of multi-scale feature information more efficiently, thereby improving detection accuracy. In conclusion, the use of ASFF in combination with a micro-detection layer has a positive impact on the detection of rail surface defects.

E. COMPARISON WITH RELATED FRAMEWORKS

We compare YOLOv5s-VF with five current mainstream detection networks based on deep learning, including the two-stage target detector Grid RCNN [47], Faster RCNN [24] and the improved superposition model hourglass network CCEANN [39], as well as the single-stage target detector SSD [48], YOLOv4 [49]. Table 8, shows the values of various indicators of these detection frameworks on the rail surface defect detection data set. From the table, we can see that our YOLOv5s-VF model achieves the highest average detection accuracy compared to other detection models. Compared with CCEANN, our model has a 0.9% higher average detection accuracy, the model size is only 1/18 of the CCEANNN model, and the detection speed is about 70 FPS faster, so our model is more suitable for deployment on mobile terminals and mobile micro-development boards, thereby saving the human resources of the railway system. Compared with the source network YOLOv5s, we achieved a large improvement in detection

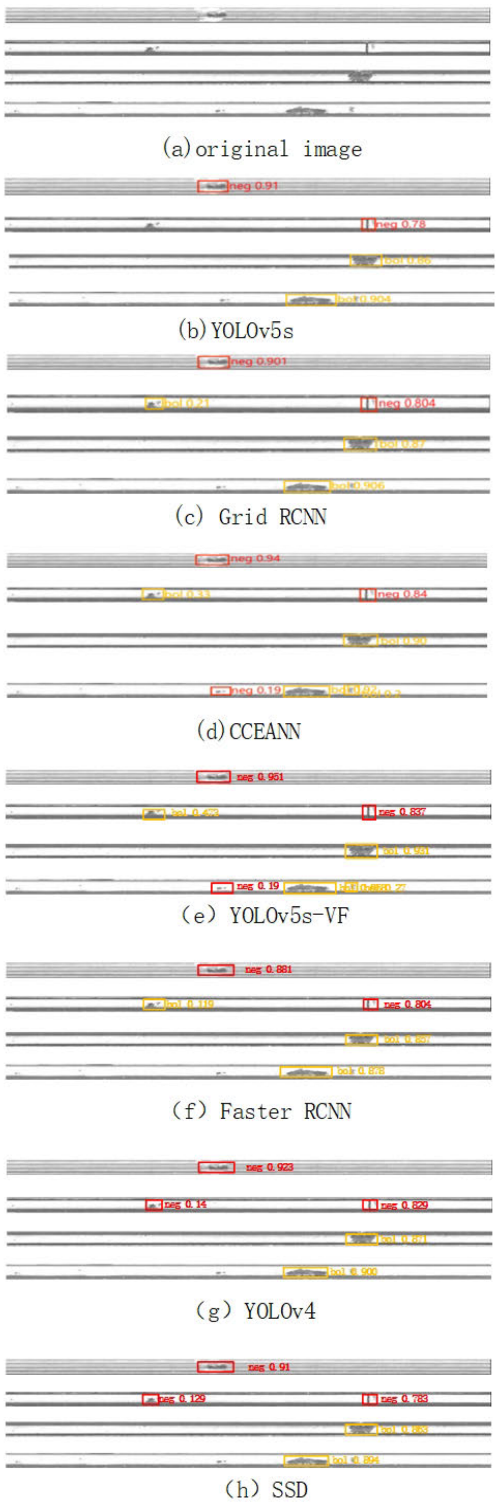
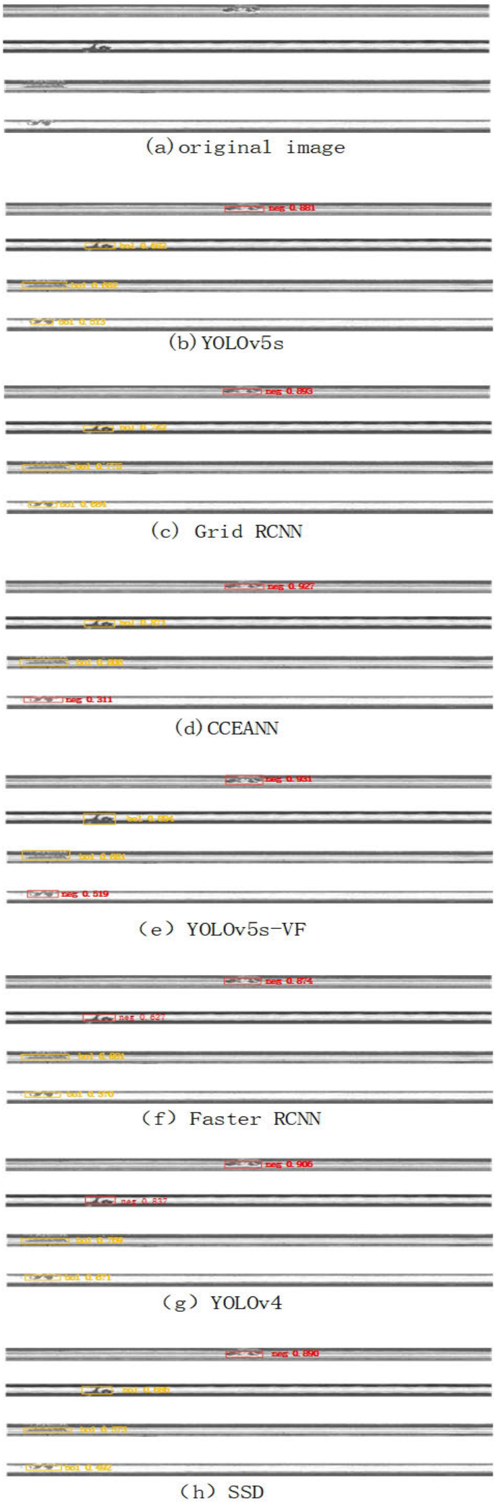


FIGURE 17. **Comparison of YOLOv5s-VF detection results.**



(a) original image

(b) YOLOv5s

(c) Grid RCNN

(d) CCEANN

(e) YOLOv5s-VF

(f) Faster RCNN

(g) YOLOv4

(h) SSD

FIGURE 18. Comparison of YOLOv5s-VF detection results.

accuracy when the model only increased by 1.2 MB, the detection accuracy of our model in the concave category is improved by 5%, the exfoliation category is improved by 4.8%, the mAP is improved by 4.9%, and the F1 is improved by 4.9%. At present, in the actual engineering application of rail surface defect detection, the detection speed of the rail mobile detection terminal is required to be 60–90 FPS. Therefore, although our YOLOv5s-VF detection model is about 20 FPS lower than the detection speed of the source network, it can still meet the actual requirements, and our model is faster than other detection models in terms of detection speed. For Grid RCNN, Fast RCNN, YOLOv4 and SSD, the four models do not exceed our models in terms of detection accuracy and speed. The Fig. 16, shows the PR curves of the concave type and the exfoliation type. From the area enclosed in the figure, the superiority of the YOLOv5s-VF model is more verified.

With Fig. 17 and Fig. 18, we can more clearly see the actual detection effect of different models on the rail surface defect dataset. For the source networks YOLOv5s, Fast RCNN, and SSD, there are multiple missed detections in YOLOv4, while our model is able to detect small defects due to the use of an attention mechanism with sharpening.Also, our model is able to locate defects in the complete edge portion. For SSD, there is error detection in YOLOv4, and our model uses microscale adaptive spatial feature fusion, which enhances the feature extraction ability of small defects while allowing the network to better learn the features of concave and exfoliation classes, so that when classifying objects, it can better distinguish between large-scale spalling and small-scale concave.

V. CONCLUSION

In order to solve the problems that edge position defects cannot be effectively located in rail surface defects, information about small size defects is lost during feature extraction, and semantic conflicts are generated when the features of multi-scale defects are fused, this paper proposes a rail surface defect detection framework, YOLOv5s-VF, with a sharpening attention mechanism (V-CBAM) and microscale adaptive spatial feature fusion (M-ASFF). First, we design a sharpening filter for the spatial attention mechanism to strengthen the localization of edge defects by the network and use 1D convolution with adaptive convolution kernels for cross-channel connections to reduce the parameters of the attention mechanism. Second, we add a micro-object detection layer to the detection head to enhance the feature extraction of micro-scale defects and remove low-resolution feature layers to reduce the loss of local details and the amount of network parameters. Then, ASFF is used to fuse the extracted features to satisfy the adaptive fusion of features of different scales while retaining the underlying fine-grained features to the greatest extent. Finally, we created a dataset of 5024 labeled rail surface defects based on real rail videos for training and testing.

The experimental results show that in the rail surface defect dataset, YOLOv5s-VF achieves better detection performance than other deep learning-based detection frameworks in terms of average detection accuracy (93.5%) and detection speed (114.9 fps), which verifies model validity and has potential for practical application in non-destructive testing of railway tracks.

Although our model can effectively detect the surface defects of rails, there are still some problems that we need to solve further. First, the net structure will be further improved to improve the detection of occlusion defects. Second, consider optimizing the loss function to accelerate the convergence of the model, thereby reducing the time for model training.

REFERENCES

[1] I. Aydin, E. Akin, and M. Karakose, "Defect classification based on deep features for railway tracks in sustainable transportation," *Appl. Soft Comput.*, vol. 111, Nov. 2021, Art. no. 107706.

[2] H. Ge, D. C. K. Huat, C. G. Koh, G. Dai, and Y. Yu, "Guided wave–based rail flaw detection technologies: State-of-the-art review," *Struct. Health Monitor.*, vol. 21, no. 3, pp. 1287–1308, May 2022.

[3] H. Li, B. Gao, L. Miao, D. Liu, Q. Ma, G. Tian, and W. L. Woo, "Multiphysics structured eddy current and thermography defects diagnostics system in moving mode," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2566–2578, Apr. 2020.

[4] J.-H. Ye, R.-H. Ni, and Q.-C. Hsu, "Image feature analysis for magnetic particle inspection of forging defects," *Proc. Inst. Mech. Eng., B, J. Eng. Manuf.*, vol. 236, pp. 1923–1929, Dec. 2021.

[5] J. Gan, Q. Li, J. Wang, and H. Yu, "A hierarchical extractor-based visual rail surface inspection system," *IEEE Sensors J.*, vol. 17, no. 23, pp. 7935–7944, Dec. 2017.

[6] Y. Wu, Y. Qin, Z. Wang, and L. Jia, "A UAV-based visual inspection method for rail surface defects," *Appl. Sci.*, vol. 8, no. 7, p. 1028, Jun. 2018.

[7] X. Cao, W. Xie, S. M. Ahmed, and C. R. Li, "Defect detection method for rail surface based on line-structured light," *Measurement*, vol. 159, Jul. 2020, Art. no. 107771.

[8] M. Nieniewski, "Morphological detection and extraction of rail surface defects," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6870–6879, Sep. 2020.

[9] X. Ni, H. Liu, Z. Ma, C. Wang, and J. Liu, "Detection for rail surface defects via partitioned edge feature," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5806–5822, Jun. 2021.

[10] H. Yu, Q. Li, Y. Tan, J. Gan, J. Wang, Y.-A. Geng, and L. Jia, "A coarse-to-fine model for rail surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 3, pp. 656–666, Aug. 2018.

[11] L. Hua, Y. Lu, J. Deng, Z. Shi, and D. Shen, "3D reconstruction of concrete defects using optical laser triangulation and modified spacetime analysis," *Autom. Construction*, vol. 142, Oct. 2022, Art. no. 104469.

[12] Y. Jiang, H. Wang, G. Tian, Q. Yi, J. Zhao, and K. Zhen, "Fast classification for rail defect depths using a hybrid intelligent method," *Optik*, vol. 180, pp. 455–468, Feb. 2019.

[13] D. Zhang, K. Song, Q. Wang, Y. He, X. Wen, and Y. Yan, "Two deep learning networks for rail surface defect inspection of limited samples with line-level label," *IEEE Trans. Ind. Informat.*, vol. 17, no. 10, pp. 6731–6741, Oct. 2020.

[14] H. Zhang, X. Jin, Q. M. J. Wu, Y. Wang, Z. He, and Y. Yang, "Automatic visual detection system of railway surface defects with curvature filter and improved Gaussian mixture model," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 7, pp. 1593–1608, Jul. 2018.

[15] D. Zhang, K. Song, J. Xu, Y. He, M. Niu, and Y. Yan, "MCNet: Multiple context information segmentation network of no-service rail surface defects," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2020.

[16] J. Wang, Q. Li, J. Gan, H. Yu, and X. Yang, "Surface defect detection via entity sparsity pursuit with intrinsic priors," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 141–150, Jan. 2019.

[17] M. Niu, K. Song, L. Huang, Q. Wang, Y. Yan, and Q. Meng, "Unsupervised saliency detection of rail surface defects using stereoscopic images," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2271–2281, Jun. 2020.

[18] Z. Lin, Z. Yingjie, D. Bochao, C. Bo, and L. Yangfan, "Welding defect detection based on local image enhancement," *IET Image Process.*, vol. 13, no. 13, pp. 2647–2658, Nov. 2019.

[19] J. Cao, G. Yang, and X. Yang, "A pixel-level segmentation convolutional neural network based on deep feature fusion for surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2020.

[20] L. Zhuang, L. Wang, Z. Zhang, and K. L. Tsui, "Automated vision inspection of rail surface cracks: A double-layer data-driven framework," *Transport. Res. C-Emer.*, vol. 92, pp. 258–277, Jul. 2018.

[21] Y. Wen, K. Fu, Y. Li, and Y. Zhang, "A sliding window method to identify defects in 3D printing lattice structure based on the difference principle," *Meas. Sci. Technol.*, vol. 32, no. 6, Jun. 2021, Art. no. 065008.

[22] J. Deng, J. Liu, C. Wu, T. Zhong, G. Gu, and B. W.-K. Ling, "A novel framework for classifying leather surface defects based on a parameter optimized residual network," *IEEE Access*, vol. 8, pp. 192109–192118, 2020.

[23] D. Ma, P. Jiang, L. Shu, and S. Geng, "Multi-sensing signals diagnosis and CNN-based detection of porosity defect during al alloys laser welding," *J. Manuf. Syst.*, vol. 62, pp. 334–346, Jan. 2022.

[24] 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. 2015.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[26] Z. Zhang, M. Liang, and Z. Wang, "A deep extractor for visual rail surface inspection," *IEEE Access*, vol. 9, pp. 21798–21809, 2021.

[27] H. Xie, Y. Zhang, and Z. Wu, "An improved fabric defect detection method based on SSD," *AATCC J. Res.*, vol. 8, pp. 181–190, Sep. 2021.

[28] X. Ni, Z. Ma, J. Liu, B. Shi, and H. Liu, "Attention network for rail surface defect detection via consistency of intersection-over-union(IoU)-guided center-point estimation," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1694–1705, Mar. 2021.

[29] T. Bai, J. Gao, J. Yang, and D. Yao, "A study on railway surface defects detection based on machine vision," *Entropy*, vol. 23, no. 11, p. 1437, Oct. 2021.

[30] H. Zhao, C. Wang, R. Guo, X. Rong, J. Guo, Q. Yang, L. Yang, Y. Zhao, and Y. Li, "Autonomous live working robot navigation with real-time detection and motion planning system on distribution line," *High Voltage*, vol. 2022, pp. 1–13, Jun. 2022.

[31] W. Xu, C. Zhang, Q. Wang, and P. Dai, "FEA-swin: Foreground enhancement attention swin transformer network for accurate UAV-based dense object detection," *Sensors*, vol. 22, no. 18, p. 6993, Sep. 2022.

[32] L. Kou, "A review of research on detection and evaluation of the rail surface defects," *Acta Polytechnica Hungarica*, vol. 19, no. 3, pp. 167–186, 2022.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[34] 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. (CVPR)*, Jun. 2020, pp. 2235–2239.

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[36] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Mar. 2021.

[37] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 2022, pp. 1–38, Mar. 2022.

[38] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[39] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

[40] J. Zhang, Y. Lu, Z. Yang, X. Zhu, T. Zheng, X. Liu, Y. Tian, and W. Li, "Recognition of void defects in airport runways using ground-penetrating radar and shallow CNN," *Autom. Construct.*, vol. 138, Jun. 2022, Art. no. 104260.

[41] S. Teng, Z. Liu, and X. Li, "Improved YOLOv3-based bridge surface defect detection by combining High- and low-resolution feature images," *Buildings*, vol. 12, no. 8, p. 1225, Aug. 2022.

[42] Z. Zhou, J. Zhang, and C. Gong, "Automatic detection method of tunnel lining multi-defects via an enhanced you only look once network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 37, no. 6, pp. 762–780, May 2022.

[43] Z.-D. Zhang, B. Zhang, Z.-C. Lan, H.-C. Liu, D.-Y. Li, L. Pei, and W.-X. Yu, "FINet: An insulator dataset and detection benchmark based on synthetic fog and improved YOLOv5," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–8, 2022.

[44] D. Thuan, "Evolution of Yolo algorithm and YOLOv5: The state-of-the-art object detection algorithm," M.S. thesis, Dept. Inf. Technol., Oulu Univ. Appl. Sci., Oulu, Finland, 2021, pp. 1–61.

[45] X. Zhang and T. Wang, "Elastic and reliable bandwidth reservation based on distributed traffic monitoring and control," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 12, pp. 4563–4580, Dec. 2022.

[46] X. Zhang, Y. Wang, G. Geng, and J. Yu, "Delay-optimized multicast tree packing in software-defined networks," *IEEE Trans. Services Comput.*, early access, Aug. 20, 2021, doi: 10.1109/TSC.2021.3106264.

[47] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 7363–7372.

[48] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[49] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

MAOLI WANG received the bachelor's degree in automation from the College of Engineering, Qufu Normal University, in 2004, and the Doctorate degree in control theory and control engineering from Harbin Engineering University, in 2008. He is currently the Dean and a Professor at the Cyberspace Security College, Qufu Normal University. His research interests include edge computing, machine learning, and deep learning.

KAIZHI LI was born in Liaocheng, Shandong, China, in 1996. He is currently pursuing the master's degree with Qufu Normal University, China. He has participated in many provincial and national scientific research projects at the School of Computer Science. His research interests include computer vision, deep learning, and target detection.

XIAO ZHU was born in Jining, Shandong, China, in 1998. He is currently pursuing the master's degree with Qufu Normal University, China. He has published many papers in his joint training at the Computer School and West Lake University. His research interests include computer vision and image processing.

YINING ZHAO was born in Tai'an, Shandong, China, in 1997. She is currently pursuing the master's degree with Qufu Normal University, China. She applied for several patents in her joint training at the Computer College and the Shandong Academy of Sciences. Her research interests include deep learning and natural language processing.