IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2023, accepted 25 October 2023, date of publication 1 November 2023,
date of current version 10 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329250

RESEARCH ARTICLE

MS-FPN-Based Pavement Defect Identification Algorithm

LEI CHEN, SHIMING AN, SHUANG ZHAO, AND GUANDIAN LI
School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun 130022, China

Corresponding author: Shiming An (2021100794@mails.cust.edu.cn)

This work was supported in part by the Jilin Provincial Department of Science and Technology, and in part by the Jilin Scientific and Technological Development Program under Grant 20210201130GX.

ABSTRACT False or missed detection may happen in the process of pavement defect detection due to cracks with different shapes and sizes and interference from complex pavement background. To solve this problem, we proposed a Cascade R-CNN detection method based on the MS-Feature Pyramid Network (MS-FPN). First, we introduced a deformable convolution module in the backbone network ResNet101, so that it can adaptively change depending on the pavement defect. Second, we utilized the MS-FPN for the cross-scale bi-directional fusion of feature maps output by the backbone network,in which the multi-branch hybrid dilated convolution (MCE) generates feature maps with multi-scale receptive fields while expanding the receptive field. The dual-channel attention fusion algorithm (ST-A) was used to improve the identification between the background and the object, so that more attention was paid to the location and features of the pavement defect object. The improved Cascade R-CNN can better adapt to the detection of various pavement defects. On the open-source dataset and self-built dataset, the detection precision has improved by 5.02% and 5% respectively compared to that of the original Cascade R-CNN.

INDEX TERMS Pavement defect, deformable convolution, dual-channel attention fusion, multi-branch dilated convolution, MS-FPN.

I. INTRODUCTION

With the rapid development of society, individuals are interacting more frequently with each other, cities are increasingly interconnected, and consequently the road traffic system [1] is becoming more and more indispensable in daily life.

China's roads commonly adopt the asphalt concrete pavement with a semi-rigid base. With years of exposure to the weather conditions, rain erosion and repeated rolling by heavy-duty vehicles, the road conditions will get worse and worse, which not only significantly shortens the service life of road, but also directly affects the safety of vehicle driving. Therefore, it is quite necessary to make in-depth research on automatic pavement defects [2] detection technology to strongly support a comprehensive understanding of the existing road conditions and carry out timely maintenance and repair, thus to maintain the safe and smooth road traffic for a long time.

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

In recent years, many scholars at home and abroad have conducted plenty of researches on automatic detection of pavement cracks, and achieved fruitful research results. Based on the traditional algorithm, Peng et al. [3] successively used the improved Otsu thresholding segmentation algorithm to remove the road markings, and then segmented the processed image with the improved self-adaptive iterative thresholding segmentation algorithm to obtain the crack image with significantly reduced image noise. This method, based on double-thresholding image segmentation technologies, better solves the interference from road markings on crack detection. Qingbo [4] removed noise by using gray level transformation and median filtering, processed pothole and alligator crack images with image enhancement methods, and detected their edges by using Roberts, Sobel, Prewiit, Laplace, and Canny operators. Chen et al. [5] proposed a crack detection method based on Local Binary Pattern (LBP) and Support Vector Machine (SVM). The supported algorithm can extract LBP features from each frame of a road video. The dimensionality of the LBP feature space can

then be reduced by Principal Component Analysis (PCA). The simplified samples were trained to determine the type of crack by using SVM. These traditional manual detection algorithms tend to classify defects into transverse crack and longitudinal crack during the detection of pavement defects, since the main structure of these two defects are relatively simple. However, when there are complex cracks with different shapes on the road, they can only play a role of identification and the classification requires additional algorithms. Therefore, only using traditional algorithms inevitably poses low efficiency and high cost, in addition to a difficulty to provide effective support for accurate maintenance of road.

With the continuous development and application of deep learning, multiple different objects in an image can be the classified and located. Compared with traditional algorithms, it has a significant improvement in both efficiency and effect. To maintain traffic safety, Liang Tianjiao et al. proposed an anchor-free lightweight object detector for autonomous driving called ALODAD, which incorporates the attention scheme to GhostNet network, and constructs an anchor-free detection framework, thus reducing the computational cost and providing parameters with high detection precision. This method improves the detection precision while meeting the real-time performance requirements of autonomous driving; An improved Spasrse R-CNN was proposed in the object detection of traffic signs, which integrates the attention mechanism and FPN into the backbone network, so that the extracted features can be focused on useful information, and effectively improve the detection precision of traffic signs. These detection algorithms provide new ideas for the research direction of road defect detection based on deep learning.

In the single-stage series of pavement defect detection algorithms, Chen et al. [6] proposed MANet to detect pavement defects. An encoder-decoder architecture was used in MANet, where the encoder adopted MobileNet as the backbone network to extract pavement defect features. Instead of the original 3×3 convolution, a multi-scale convolution kernel was used for the depthwise separable convolutional layers of the network. Besides, a hybrid attention mechanism was respectively integrated into the encoder and decoder modules to infer the significance of spatial points and inter-channel relationship features on the input intermediate feature map, which provided superior performance compared to traditional algorithms. Wang [7] proposed a pavement crack detection algorithm based on Yolov3 and multi-scale analysis, which de-noises the image with multi-scale analysis before detecting the image with the trained model. The study results showed that the method had a significant improvement for the confidence level, and the mAP value of the Yolov3-based model for the detection of transverse, longitudinal and reticulation cracks was even up to 51.2%. Although the single-stage algorithm is slightly better than the two-stage algorithm in terms of speed, the two-stage network has a greater advantage in the detection precision.

In the two-stage series of pavement defect detection algorithms, Wang [8] implemented a Faster R-CNN pavement defect detection method by using VGG-16 for feature extraction, which achieved a mAP value of 53.86% for the detection of pavement defects. However, during the Faster R-CNN studies, if the IOU [9] threshold is set too low, it will cause noise prediction, and if it is set too high, over-fitting phenomenon may occur. After that, Shen and Nie [10] applied the Cascade R-CNN algorithm to pavement defect detection, using a three-stage cascade detector to gradually increase the IOU threshold, and obtain a new defect classification score and bounding box regression to improve the training effect, which effectively makes up for the shortcomings of the Faster R-CNN algorithm. However, due to the fixed shape proportion and limited size of the Cascade R-CNN receptive field, it is difficult to accurately identify the crack defects with different shapes and sizes, which seriously affects the detection precision.

In view of the problems of the above algorithms, this paper proposed an improved Cascade R-CNN network, which effectively makes up for the disadvantages of traditional CascadeR-CNN such as loss of information [11], missed detection [12], and false detection, and has a stronger adaptive ability for pavement defects with complex shapes and varying sizes, realizing the high-quality detection of complex defects of the pavement.

II. MATERIALS AND METHODS

Figure 1 shows the improved Cascade R-CNN algorithm proposed in this research, which is mainly composed of four parts: backbone network (3D-ResNet101), MS-FPN, RPN and cascade classification regression network. Part 1 of the algorithm flow: The 3D-ResNet101 extracts image features and generates multi-scale feature maps. Part 2: The feature map in the MS-FPN, through MS network and two-stage (bottom-up and top-down) feature fusion mode with cross-scale lateral connection, enhances the ability of network in identification of pavement defects while collecting multi-scale pavement defect information. Part 3: The fused feature map is sent to the RPN, and a large number of bounding boxes are generated by anchor box generation mechanism. In this paper, the aspect ratio of bounding box is set to $\{0.1, 1, 0.9\}$. Part 4: In the detection stage, Cascade RCNN uses cascade detectors for detection, each of which includes ROI Align [13], fully connected layer FC, classification score C and bounding box regression B. During detection, the target area is re-sampled by the bounding box regression B output from the previous detector, and the IOU threshold training is gradually improved to obtain new classification score C and bounding box regression B, ultimately improving the sample quality and network training effect.

A. INTRODUCTION OF DEFORMABLE CONVOLUTION IN BACKBONE NETWORKS

The receptive field of standard convolution is a grid-type rectangle with fixed size, which is a regular rectangular

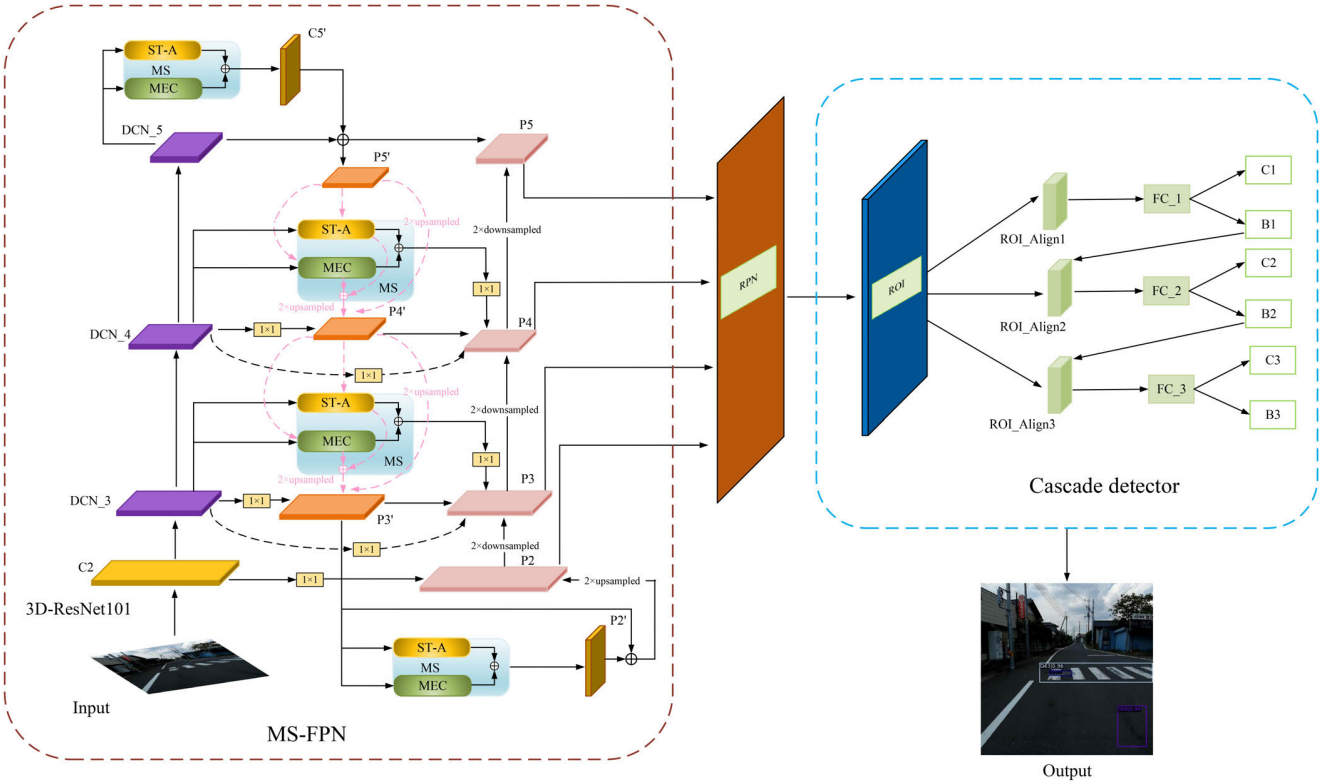


FIGURE 1. Improved Cascade R-CNN network model.

shape when the features are extracted. However, for pavement defects with different scales or complex shapes, it lacks certain adaptive ability, which will lead to the loss of information. To solve the problem of insufficient expression of ResNet101's feature extraction ability, this paper proposed to use deformable convolution [14] to replace the con3–con5 traditional convolutional layers in the last three stages of ResNet101. The network model is shown in Figure 2.

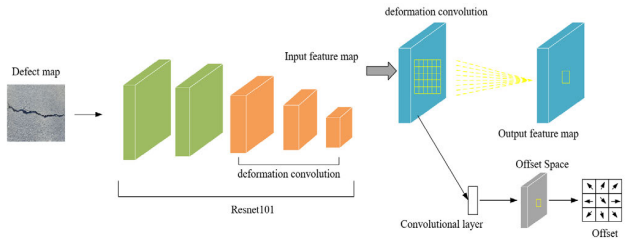


FIGURE 2. Backbone network structure with DCN module.

The traditional ResNet network is mainly divided into five stages, which are composed of the initial stage and the residual module consisting of Identity Block and Conv Block. The initial stage contains convolutional layer, BN layer, activation layer and pooling layer. The equation of definition of its structure is shown in (1):

$$y(p_0) = \sum_{p_i \in R} x\,(p_0 + p_i) \times \omega\,(p_i) \tag{1}$$

Firstly, the convolution kernel R with fixed size is used for sliding-window sampling on the feature map x, and then the sampling points are multiplied by the weights to sum. Where R determines the size and expansion of the receptive field, p_0 represents the pixel point coordinates of the center of the convolution kernel, p_i represents other pixel points except p_0, and R is the image area covered by the convolution kernel. After the deformable convolution DCN is added, it is necessary to add an offset p_i(the offset of the ith sample point) to Equation (1), and this offset can make the sampling position become an irregular deformable area, as shown in the following Equation (2):

$$y\,(p_0) = \sum_{p_i \in R} x\,(p_0 + p_i + p_i) \times \omega\,(p_i) \tag{2}$$

After the deformable convolution is added, the receptive field of sampling can adaptively change according to the pavement defect to be identified, and realize the learning and dynamic adjustment of the irregularly shaped cracks and other types of damage on the pavement, thus adapting to the shape, size and other geometric deformations of different pavement defects and effectively avoiding the loss of information, so there are richer features of pavement defects after the deformable convolution is added. The specific structure of the improved 3D-ResNet101 is shown in Table 1.

TABLE 1. Structure of 3D-ResNet101.

Stage	Input	Output	DCN	3D-ResNet101
Conv1	600×600	300×300	×	7×7，64，stride2 3×3 max pool, stride2
Conv2	300×300	150×150	×	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Conv3	150×150	75×75	√	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
Conv4	75×75	37×37	√	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times23$
Conv5	37×37	18×18	√	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$

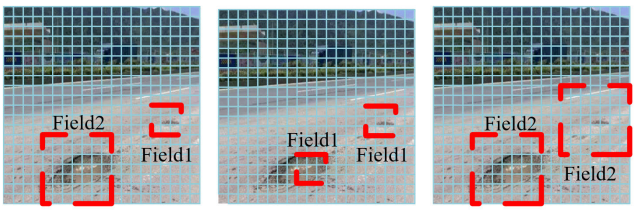
B. MS-FPN

Since the size of pavement defects varies greatly, the traditional model is limited in its ability to capture multi-scale defect features, then it will lead to the loss of information. And there is a lack of information communication between the receptive fields of multi-scale defects, resulting in poor quality of feature map. In contrast, the traditional FPN performs bottom-up and top-down feature fusion, and predicts with the fused feature map with advanced semantic information, which improves the precision to a certain extent. However, this fusion mode is restricted by one-way information flow, which leads to limited precision. Therefore, this paper proposed the MS-FPN, a new FPN architecture, which is guided by MCE [15] module and ST-A [16] attention fusion module in both directions. The MS-FPN effectively solves the problem that traditional FPN is restricted by one-way information flow, Adjustment of the number of channels and size of the feature map by 1×1 convolution and by upsampling and downsampling operations,aggregates the features of different resolutions, and realizes cross-scale fusion. Meanwhile, at each node of the MS-FPN feature fusion, the shared MCE module is utilized to obtain the receptive fields of multi-scale pavement defects, thus improving the ability to capture multi-level features of defects. And ST-A attention fusion module is utilized to obtain stronger semantic information and more accurate positioning information, and enhance the ability in the identification of feature maps. The MS-FPN is shown in Figure 1.

1) MCE MOUDLE

Since the ordinary convolutional layer obtains the feature map by down-sampling, the size of the sampled feature map is getting smaller, and each pixel point of the final feature map corresponds to a larger area of the original image, which also makes the feature response of the whole image become sparse. In contrast, the dilated convolution introduces the parameter of dilation rate (r) on the basis of the ordinary convolutional layer, which is utilized to define the spacing of the data values processed by the convolution kernel during the convolution operation. This method is capable of obtaining a denser spatial response and a larger receptive field without additional computational complexity.

The sizes and shapes of the pavement defects are different, and the receptive fields of different sizes are only beneficial for the detection of defects of different sizes. As shown in Figure 3, the smaller Receptive Field1 in Figure (b) is only applicable to the identification of potholes with small sizes in the image, while it will lead to the loss of information when identifying those with large sizes in the image. The larger Receptive Field 2 is applicable to the identification of potholes with large sizes in the image, but when detecting those with smaller size, the receptive field is mixed with the interference from road lines and other factors that affect the detection results, as shown in Figure (c). Figure (a) shows that receptive fields of different sizes are designed to obtain richer object features according to potholes of different sizes.



(a)Ideal field (b)Field1 (c)Field2

FIGURE 3. Comparison of receptive fields of pavement defects with different scales.

In view of the above drawbacks, this paper proposed a multi-branch hybrid dilated convolution with a convolution kernel of 3×3 [18]. That is, the first branch adopts the dilated convolution with $r = 1$ to expand the first layer of the receptive field, and the cascaded dilated convolution with $r = (1, 2)$ and $r = (1, 2, 3)$ is respectively used in the second and third layers to obtain pavement defect features of different scales. It also effectively solves the grid effect caused by the dilated convolution with a single dilation rate, and avoids the loss of information. The comparison diagram of receptive fields of the dilated convolution is shown in Figure 4.

Finally, the global feature of the feature map are obtained by using average pooling in the fifth layer. The multi-branch hybrid dilated convolution is shown in Figure 5.

2) DUAL-CHANNEL FEATURE FUSION

The process of pavement defect detection is prone to missed or false detection due to the complex road background, since the inconspicuous features of small cracks in pavement defects cannot be highlighted in the complex road background. To pay better selective attention to channel information, so that the key information of small cracks

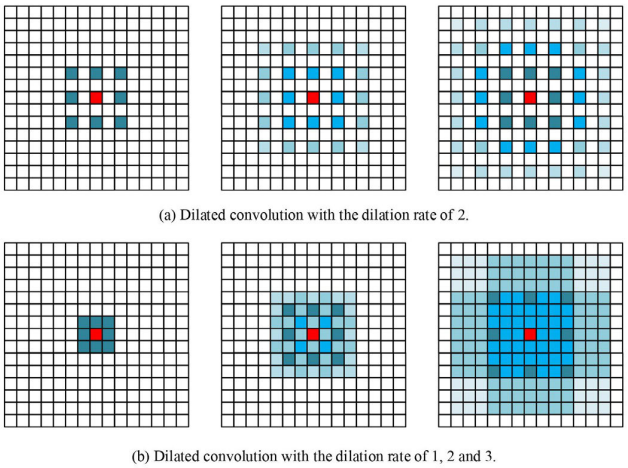


FIGURE 4. Comparison of receptive fields of features at different dilation rates.

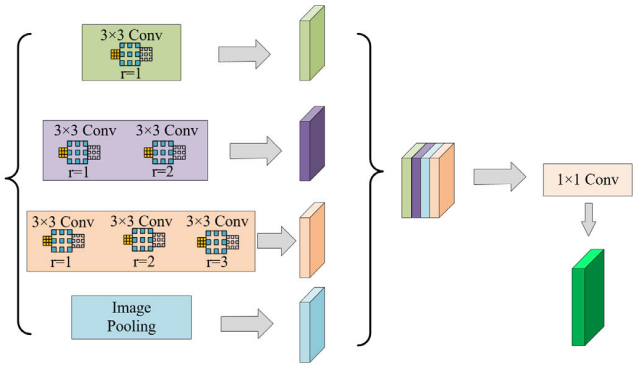


FIGURE 5. Multi-branch hybrid dilated convolution.

can be captured in large quantities, ST-A network was proposed in this paper. The classical Channel Attention Module (CAM) [19] extracts features using only Global Average Pooling (GAP), which may easily lead to equal weighting of key defects and background information within a uniform channel, while weakening the features of small objects. In this paper, the enhanced feature map Q_i was generated by combining scene attention and target attention networks, and finally the two kinds of attention were fused together, so that more attention was paid to the location and features of the target pavement defect, and the information about the location of the cracks as well as the feature differences between the cracks and the background are highlighted, as shown in Figure 6.

The input feature map goes through the scene attention module and the target attention module respectively. In the scene attention module, P_i firstly goes through a remodeling process and a remodeling and transposing process, respectively, and then the two feature maps produced are multiplied to obtain the channel attention feature map through softmax [20]. The feature map is transposed and matrix multiplied with P_i to obtain a weighted value α (the initial value is 0, and higher values are assigned as the network deepens), and then remodeled back to its original shape, and

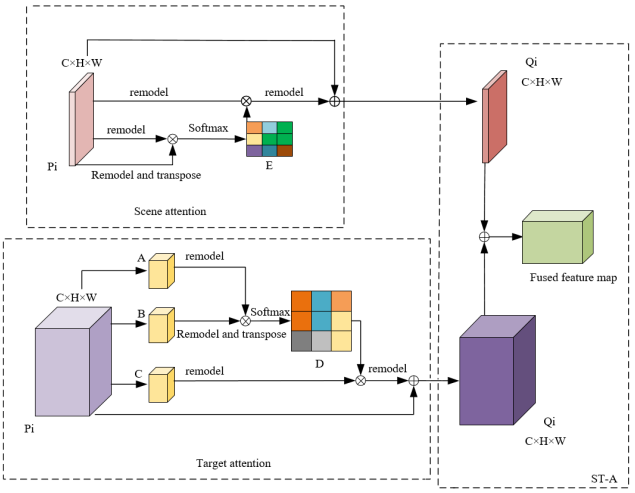


FIGURE 6. Dual-channel feature fusion(ST-A).

finally added to P_i to obtain Q_i. as shown in the following Equation (3):

$$\begin{cases} E_{jk} = \dfrac{\exp\left(Pi_k \times Pi_j\right)}{\sum_{k=1}^{C} \exp\left(Pi_k \times Pi_j\right)} \\ Qi_j = \alpha \sum_{k=1}^{C} \left(E_{jk} \times Pi_k\right) + Pi_j \end{cases} \tag{3}$$

The attention module highlights the feature information of the object by weighting all final channel features and summing them with the original channel features. E_{jk} in the above expression denotes the channel attention matrix to express the influence of Channel k on Channel j, Pi_k denotes the Channel k output from different feature output layers, and Pi_j represents the Channel j output from different feature output layers.

In the target attention module, the concatenated feature map $Pi\,(C \times H \times W)$ goes through three convolutional layers to generate three corresponding feature maps A, B, and C, respectively, and they are remodeled into $(C \times H \times W)$, where B is then transposed. Then B and A are matrix multiplied to obtain the location-weighted feature map D through softmax and then transpose the map. After that, perform matrix multiplication of C and D and assign weight β(the initial value is 0, and higher values are assigned as the network deepens), and remodel it back to the initial shape, and add it to Pi to obtain Qi. The expression for this calculation process is shown in (4):

$$\begin{cases} D_{mt} = \dfrac{\exp\left(A_m \times B_t\right)}{\sum_{t=1}^{H\times W} \exp\left(A_m \times B_t\right)} \\ Qi_m = \beta \sum_{t=1}^{H\times W} \left(D_{mt} \times C_t\right) + Pi \end{cases} \tag{4}$$

The attention module obtains a feature map Qi by weighting all locations, and then sums with the original feature results to reflect a spatial vision of the global context, thus obtaining more accurate object location information. In the above expression, D_{mt} represents the spatial attention matrix, which is used to express the influence of Location t on Location m, Qi_m represents Location m of the feature map from

different feature output layers, and C_t represents Location t of the feature map C. Finally, the enhanced feature maps generated by scene attention and target attention are summed to generate feature maps with richer location and feature information.

The attention feature fusion module combines the dependency between scene and target, and continuously improves the relationship weights of the detection objects, so that its adaptive learning expresses the features of pavement defects more effectively, suppresses the information expression of the rest of the features, and improves the accuracy of identifying the difference between the road background and the defect features more effectively.

III. EXPERIMENTAL PREPARATIONS

A. ESTABLISHMENT OF DATASET

The acquisition of crack image and the establishment of dataset are crucial in the study preparation stage. In order to make the study results more convincing, we used the public dataset RDD2020 and the self-built data QT. To ensure a clearer definition of the categories of pavement defects, we added the labels of three pavement defect categories (D43, D44 and D50) to the original categories of Japanese pavement defects in RDD2020. The resolution of each image is 600×600 and there are 14,758 marks of pavement damage.

TABLE 2. RDD2020 dataset category.

Damage Type	Detailed description	Tags
Longitudinal cracks	Wheel crush/construction joint part crack	D00
Transverse cracks	Equally spaced cracks / cracks in the construction connection area	D10
cracked	Partial or overall cracking	D20
Potholes	Wheel crushing/rainwater washing	D40
Other breakage	Blurred pedestrian lines	D43
	Blurred Road Routes	D44
Manhole cover defects	Uneven manhole cover/unrealistic manhole cover	D50

From Table 2, in general, pavement defects can be divided into four major types: cracks, other breakages, potholes, and manhole cover defects, of which pavement cracks are divided into three types: longitudinal cracks, transverse cracks, and cracks according to their relative position to the road surface, which are represented by labels D00, D10, and D20, respectively. Other damages are categorized into two types: pedestrian identification lines and road identification lines, which are represented by labels D43 and D44, respectively. And manhole cover defects and potholes are characterized more obviously, so each of them accounts for one type of label, D50, D40, respectively.The fan chart of the distribution of the number of each type in the dataset is shown in Figure 7.

In this study, the self-built dataset consists of images taken on the road and downloaded from the Internet (https://image.baidu.com/), involving a variety of natural weather, lighting conditions, and complex backgrounds. We totally sampled 760 images saved in JPG format, each of which has a resolution of 600 × 600. Four class labels

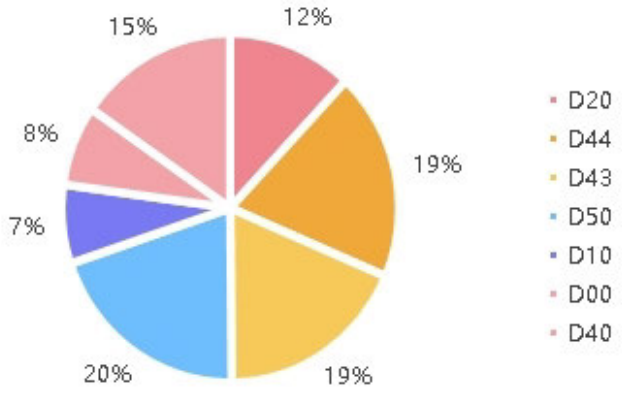


FIGURE 7. The proportion of pavement with the type of damage.

were included, namely D00, D10, D50, and D44 (whose meanings are consistent with those in the RDD2020 dataset). The number of images under each class label is 180, 172, 200, and 208, respectively, and they are labeled according to the PASCAL VOC2007 standard by Labelimg software, and saved as xml files.

Finally, the above two datasets are grouped into training set, validation set and test set in the ratio 6 : 2 : 2. Each type is relatively evenly distributed in the two datasets, which is more adapted to the learning process. Examples of images in the dataset are shown in Figure 8.

B. EXPERIMENTAL PLATFORMS

The experimental platform for the detection method in this paper is in Ubuntu 20.04 system equipped with 2 GPUs ($1080ti \times 2$) and 1 CPU processor (Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz, and Pytorch is used as the deep learning framework. The input image size is 600×600.

C. PARAMETER SETTING PLATFORMS

For a fair comparison, all ablation experiments were conducted under the PyTorch framework and the open-source MMDetection toolkit from ShangTang, which integrates a large number of existing advanced as well as mainstream detection methods, with 2 test images in a batch, 12 consecutive epochs of training, an initial learning rate of 0.002 set, and the other hyper-parameters in MMDetection as the default values.

D. ASSESSMENT OF INDICATORS

We adopted the average precision of individual pavement defect type (AP), the mean average precision of all types (mAP), and the loss of classification (LFocal) as the evaluation indexes for this study, as shown in expression (5).

$$\begin{cases} AP = \int_0^1 P(R)\, dR \times 100\% \\ mAP = \dfrac{1}{n} \sum_{i=1}^{n} AP \times 100\% \\ L_{Focal} = \begin{cases} -\alpha \left(1 - y'\right)^{\gamma} \log y', & y = 1 \\ -(1-\alpha)\, y'^{\gamma} \log\left(1 - y'\right), & y = 0 \end{cases} \end{cases} \tag{5}$$

(a) Examples of images in the self-built dataset

(b) Examples of images in RDD2020 dataset

FIGURE 8. Examples from pavement defect dataset.

IV. RESULTS

A. BACKBONE NETWORK ABLATION EXPERIMENT

In order to prove the effectiveness of the algorithm, this paper firstly carried out ablation studies to the backbone network on the public datasets. In the cascade R-CNN algorithm, the backbone network is replaced by RepLKNet, DenseNet, VGG-16, Inceptionv4, ResNet50, WRN, SeNet, ResNet101, and 3D-ResNet101, in that order. As can be seen from Table 3, embedding ResNet101 into Cascade rcnn is stronger than other backbone networks in detecting the accuracy of seven types of defects in roads, so this paper proposes to improve ResNet101. And after embedding the 3D-ResNet101 network after adding the deformation convolution into Cacade RCNN, the average accuracy is improved by 1.7 percentage points compared with the original ResNet101, and the accuracy is improved in various categories of defects.

In order to more intuitively reflect the effectiveness of the proposed method, this paper shows the original ResNet101 and the 3D-ResNet101 training process visualization results as shown in Figure 9, from the comparison of Figure b and Figure c, it can be clearly seen that, for the irregularly shaped cracks, the 3D-ResNet101 added with deformation convolution obtains more effective feature information through adaptive adjustment.

B. FPN ABLATION EXPERIMENT

In order to verify the performance of Cascade R-CNN before and after the improvement, we embed ResNet101 before and after the improvement into Cascade R-CNN in one-by-one combinations with traditional FPN, PANet [28], Bi-FPN and our proposed MS-FPN, respectively, and conduct ablation studies on public datasets. Table 4 shows that the original cascade R-CNN network model has the lowest mAP value. The PANet with cascaded feature pyramid structure can perform feature fusion on inhomogeneous scales, and its detection accuracy is significantly higher than that of the traditional FPN, while the Bi-FPN with bidirectional feature pyramid structure can utilize both high-level and low-level features, which improves the accuracy of target detection. From the experimental results, Bi-FPN outperforms PANet in the face of complex road defects.And our proposed MS-FPN,The average accuracy of the detection results of integrating the MS-FPN which is guided by MCE and ST-A into Cascade R-CNN algorithm is 2.4% higher than that of Bi-FPN in Cascade R-CNN algorithm. From the experimental results, it can be seen that the detection performance of MS-FPN in Cascade R-CNN is significantly stronger than other feature pyramids. The improved Cascade R-CNN even improves 5.02 percentage points in average accuracy compared to the original Cascade R-CNN.Therefore, the deformation, multi-scale receptive field and attention in the convolution play a key role in the defect detection, and the improved model is more suitable for detection of pavement defects with complex types and different sizes.

The classification loss function for training datasets containing 7 types of pavement defects is shown in Figure 10. It can be seen that the loss decreases faster in the early stage of network training, and the loss curve decreases steadily when the iteration is about 7000 times, and finally converges at about 0.02, which indicates that the improved network is more desirable for the classification of the seven kinds of road defects.

TABLE 3. Comparison of cascade R-CNN detection performance for different backbone networks.

Backbone networks	AP values for different types of pavement defects							mAP
	D20	D44	D50	D43	D10	D00	D40	
RepLKNet[20]	0.549	0.647	0.743	0.751	0.310	0.344	0.450	0.542
DenseNet[21]	0.551	0.653	0.751	0.756	0.314	0.343	0.453	0.546
VGG16[22]	0.558	0.655	0.757	0.752	0.317	0.351	0.455	0.549
Inceptionv4[23]	0.568	0.651	0.756	0.751	0.318	0.356	0.461	0.552
ResNet50[24]	0.565	0.659	0.761	0.758	0.326	0.354	0.458	0.554
WRN[25]	0.571	0.650	0.765	0.760	0.322	0.361	0.472	0.557
SeNET[26]	0.570	0.652	0.764	0.759	0.318	0.360	0.477	0.557
ResNet101	0.574	0.658	0.771	0.762	0.324	0.365	0.475	0.559
Ours(3D-ResNet101)	0.610	0.670	0.776	0.765	0.344	0.374	0.496	0.576

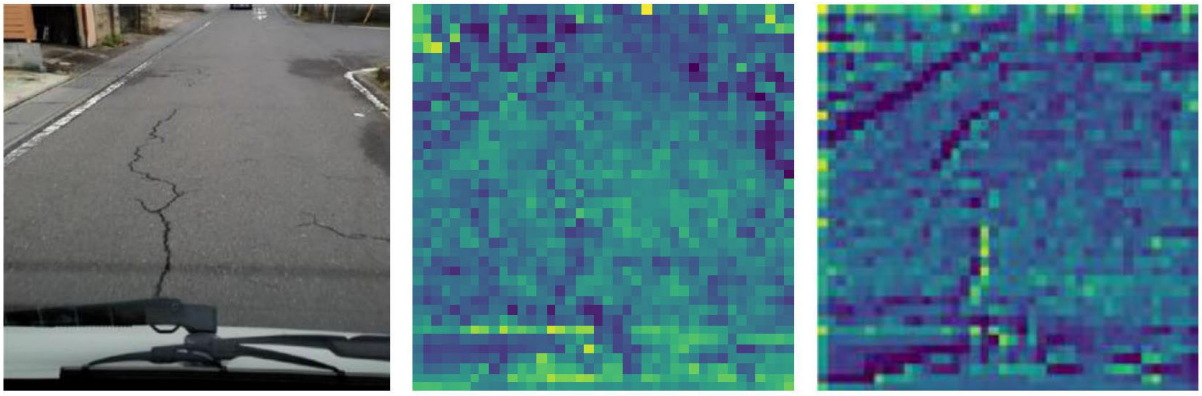


FIGURE 9. Feature map visualization comparison.

TABLE 4. Ablation experiment based on improved Cascade R-CNN detection performance.

Backbone networks	AP values for different types of pavement defects							mAP
	D20	D44	D50	D43	D10	D00	D40	
ResNet101+FPN	0.574	0.658	0.771	0.762	0.324	0.365	0.475	0.559
3D-ResNet101+FPN	0.610	0.670	0.776	0.765	0.344	0.374	0.496	0.576
ResNet101+PANet	0.582	0.662	0.769	0.766	0.332	0.370	0.481	0.566
3D-ResNet101+PANet	0.607	0.665	0.770	0.761	0.353	0.385	0.510	0.579
ResNet101+Bi-FPN	0.599	0.664	0.772	0.763	0.345	0.379	0.487	0.573
3D-ResNet101+Bi-FPN	0.613	0.672	0.774	0.760	0.364	0.388	0.513	0.583
ResNet101+MS-FPN	0.617	0.681	0.782	0.769	0.394	0.412	0.523	0.597
3D-ResNet101+MS-FPN	0.628	0.689	0.791	0.786	0.418	0.427	0.536	0.611

In order to further verify the effectiveness and advancement of the proposed method for pavement defect detection, the models before and after improvement were trained with the public dataset RDD2020, and several typical prediction images were selected from the detection results, as shown in Figure 11. In the detection results of the original model, (a) (c) shows a missed detection and Figure(e) shows a misdetection. In the detection results of the improved model, the disadvantages of leakage and misdetection of the original model are effectively compensated, and the precise localization of the target defects as well as the strong recognition ability are achieved.

C. PERFORMANCE COMPARISON OF DIFFERENT MODELS

We compared the detection performance of this study with that of other advanced one-stage and two-stage network models on the public dataset RDD2020 and the self-built dataset QT respectively. The results are shown in Tables 5 and 6.

From the experimental results, it can be seen that in the public dataset, the cascade-type detectors of the Cascade R-CNN series are significantly better than the Faster R-CNN in detecting defects of each class by improving the Iou of the prediction frame step by step, and also have a significant advantage in the average accuracy compared to the single-stage Yolov5, Yolov7, and Yolov8, but not as good as the SC-R-CNN and Spares.The average accuracy is equal compared to Libra R-CNN, but has a significant disadvantage in the defective categories of D20, D44, D00, and D40. And the improved Cascade R-CNN has a substantial improvement in the detection accuracy of defects in all categories compared to Libra R-CNN, SC-R-CNN

TABLE 5. Performance comparison of different models based on RDD2020 dataset.

Network model	AP values for different breakage types							mAP
	D20	D44	D50	D43	D10	D00	D40	
Yolov5[28]	0.521	0.594	0.794	0.673	0.282	0.302	0.425	0.506
Yolov7[29]	0.542	0.651	0.752	0.698	0.314	0.355	0.453	0.538
Yolov8[30]	0.570	0.655	0.764	0.758	0.321	0.362	0.460	0.556
Faster RCNN	0.545	0.604	0.760	0.716	0.302	0.343	0.465	0.534
Cascade RCNN	0.574	0.658	0.771	0.762	0.324	0.365	0.457	0.559
Libra RCNN[31]	0.580	0.663	0.755	0.751	0.304	0.377	0.482	0.559
SC-RCNN[32]	0.578	0.674	0.763	0.755	0.357	0.370	0.466	0.562
Spares[33]	0.600	0.655	0.777	0.760	0.374	0.402	0.502	0.581
Ours	0.628	0.689	0.791	0.786	0.418	0.427	0.536	0.611

TABLE 6. Performance comparison of different models based on self-built datasets.

Network model	AP values for different breakage types				mAP
	D00	D10	D50	D44	
Yolov5	0.435	0.454	0.780	0.728	0.599
Yolov7	0.443	0.467	0.811	0.765	0.622
Yolov8	0.456	0.462	0.819	0.797	0.634
Faster RCNN	0.451	0.469	0.810	0.777	0.627
Cascade RCNN	0.453	0.478	0.822	0.792	0.636
Libra RCNN	0.452	0.479	0.819	0.768	0.630
SC-RCNN	0.453	0.468	0.826	0.785	0.633
Spares	0.467	0.493	0.841	0.805	0.656
Ours	0.491	0.525	0.885	0.843	0.686

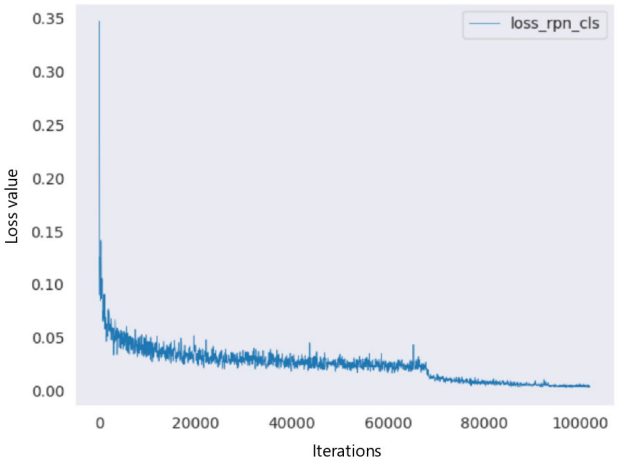


FIGURE 10. The proportion of pavement with the type of damage.

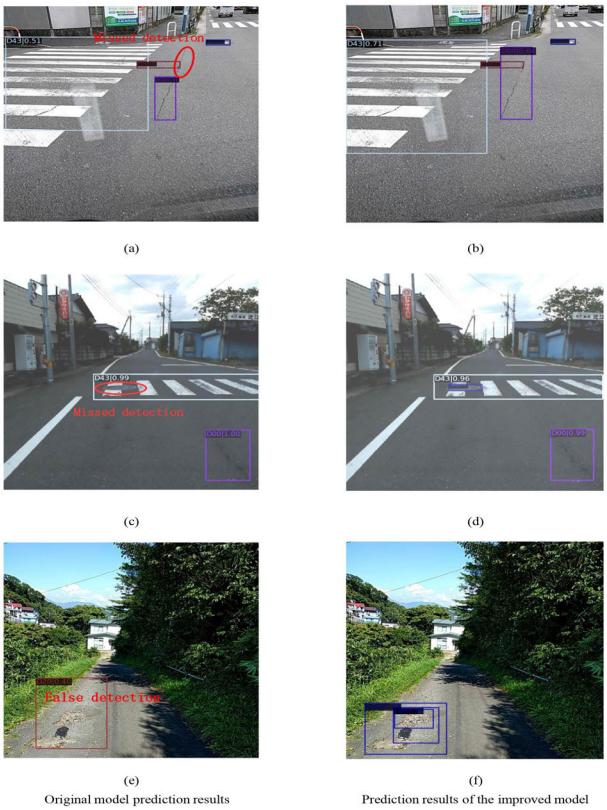
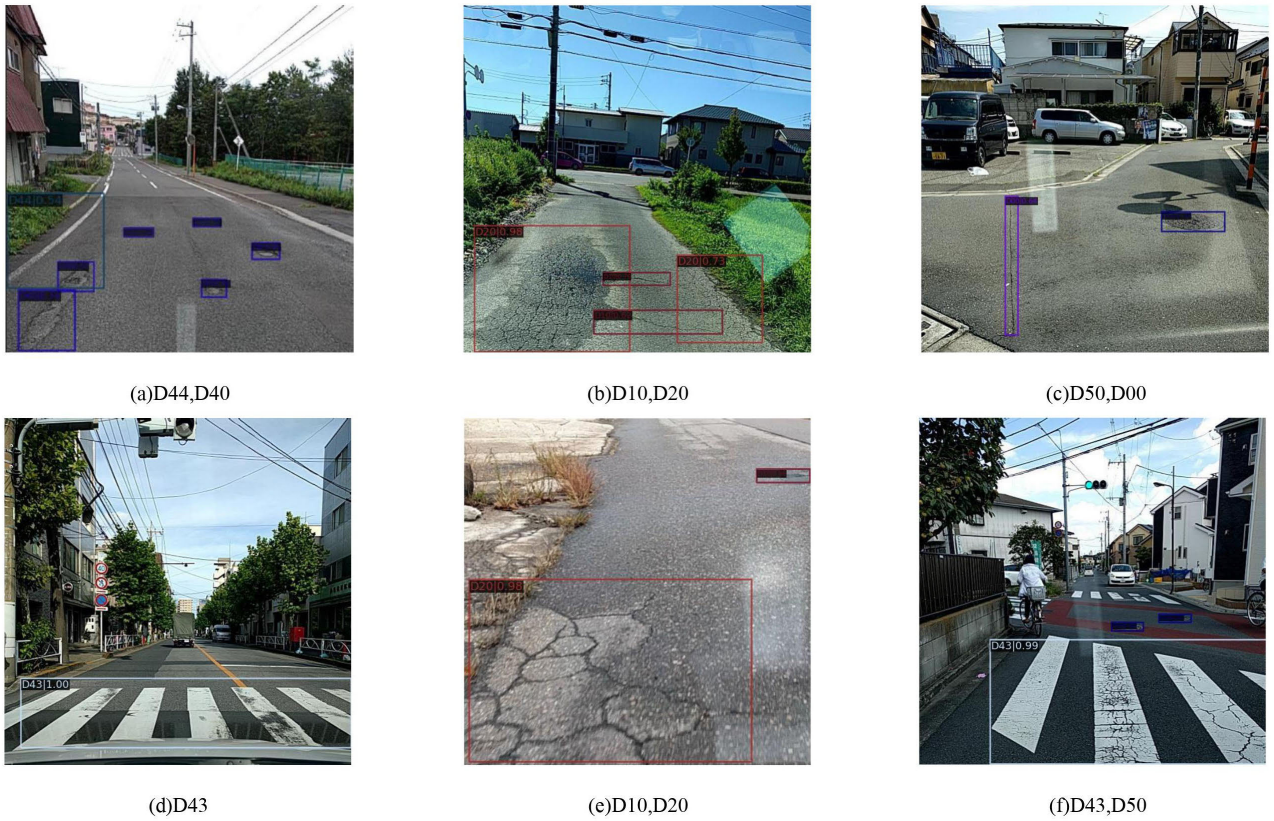


FIGURE 11. Comparison of ablation study results.

and Sparse models, which is 5.02 percentage points higher than the mAP value of the original model, Cascade R-CNN, compared to the original model, Cascade R-CNN.In the self-built dataset, the detection effect of Cascade R-CNN is still better than Yolov5, Yolov7, Yolov8, Faster R-CNN, and maintains a slight advantage over Libra R-CNN and SC-R-CNN algorithms, but there is still a gap with Sparse. The improved Cascade RCNN and other algorithms have significant improvement in each category accuracy, and still improve 5 percentage points compared with the original algorithm. It can be shown that the improved Cascade R-CNN has a stronger adaptive ability to irregular shape and large scale change of pavement defects than the original Cascade R-CNN, and The detection results of the seven categories of road defects based on the improved model are shown in Figure 12.



(a)D44,D40 (b)D10,D20 (c)D50,D00

(d)D43 (e)D10,D20 (f)D43,D50

FIGURE 12. Schematic diagram of detection results of seven types of pavement defects.

V. CONCLUSION

This paper presented a MS-FPN-based Cascade R-CNN pavement defect detection method. In the feature extraction stage, the method incorporated deformable convolution into ResNet101, and adapted to various defects with complex shapes in traffic pavement by learning the offset and weight of center points. Multi-scale receptive fields were obtained through MEC network shared by backbone network and MS-FPN, and the number of parameters and complexity of the model were reduced. And ST-A network was used to fuse the feature map of scene attention and target attention outputs, so that it only paid attention to the key part of the fused feature map and inhibited the information expression of other features. The MS-FPN network for cross-scale bidirectional fusion of feature maps, which enriched the pavement defect information and defect feature expression ability. The experimental results demonstrate that the improved Cascade RCNN algorithm proposed in this paper has a significant improvement in the detection accuracy of defects of all classes on the RDD2020 dataset as well as the self-built QT dataset compared to the original Cascade RCNN as well as some state-of-the-art algorithms. Due to the limited study conditions, the dataset in this study does not contain an abundance of images with large shadows or strong light exposure, which are likely to be encountered in practice. Therefore, we should further strengthen the processing and verification of defect images under special conditions. In addition, the final effect of the algorithm is closely related to the quality of the processed images, so we should enhance the research on image capture technology while improving the algorithm in the future.

REFERENCES

[1] A. Morimoto, A. Wang, and N. Kitano, "A conceptual framework for road traffic safety considering differences in traffic culture through international comparison," *IATSS Res.*, vol. 46, no. 1, pp. 3–13, Apr. 2022.
[2] E. Ranyal, A. Sadhu, and K. Jain, "Road condition monitoring using smart sensing and artificial intelligence: A review," *Sensors*, vol. 22, no. 8, p. 3044, Apr. 2022.
[3] L. Peng, W. Chao, L. Shuangmiao, and F. Baocai, "Research on crack detection method of airport runway based on twice-threshold segmentation," in *Proc. 5th Int. Conf. Instrum. Meas., Comput., Commun. Control (IMCCC)*, Sep. 2015, pp. 1716–1720.
[4] Z. Qingbo, "Pavement crack detection algorithm based on image processing analysis," in *Proc. 8th Int. Conf. Intell. Human-Mach. Syst. Cybern. (IHMSC)*, vol. 1, Aug. 2016, pp. 15–18.
[5] C. Chen, H. Seo, C. H. Jun, and Y. Zhao, "Pavement crack detection and classification based on fusion feature of LBP and PCA with SVM," *Int. J. Pavement Eng.*, vol. 23, no. 9, pp. 3274–3283, Jul. 2022.
[6] J. Chen, Y. Wen, Y. A. Nanehkaran, D. Zhang, and A. Zeb, "Multiscale attention networks for pavement defect detection," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023, doi: 10.1109/TIM.2023.3298391.
[7] C. Wang, "Research and application of pavement crack detection based on deep learning and multi-scale analysis," Wuhan Univ. Technol., Wuhan, China, Tech. Rep., 2019, doi: 10.27381/d.cnki.gwlgu.2019.000361.
[8] K. Wang, "Research and application of pavement distress detection based on deep learning," Wuhan Univ. Technol., Wuhan, China, Tech. Rep., 2019, doi: 10.27381/d.cnki.gwlgu.2019.000218.

[9] S. Lu, H. Lu, J. Dong, and S. Wu, "Object detection for UAV aerial scenarios based on vectorized IOU," *Sensors*, vol. 23, no. 6, p. 3061, Mar. 2023.
[10] T. Shen and M. Nie, "Pavement damage detection based on cascade R-CNN," in *Proc. 4th Int. Conf. Comput. Sci. Appl. Eng.*, Oct. 2020, pp. 1–5.
[11] S. Banerjee and G. K. Singh, "A robust bio-signal steganography with lost-data recovery architecture using deep learning," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–10, 2022.
[12] H. Zhang, F. Chen, Z. Shen, Q. Hao, C. Zhu, and M. Savvides, "Solving missing-annotation object detection with background recalibration loss," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1888–1892.
[13] S. Nam and D. Lee, "Improvement in object detection using multi-scale RoI pooling and feature pyramid network," *J. Comput. Sci. Eng.*, vol. 16, no. 1, pp. 14–24, Mar. 2022.
[14] Y. Zhao, B. Qin, Y. Zhou, and X. Xu, "Bearing fault diagnosis based on inverted Mel-scale frequency cepstral coefficients and deformable convolution networks," *Meas. Sci. Technol.*, vol. 34, no. 5, May 2023, Art. no. 055404.
[15] W. H. Shi and G. H. Bao, "Bamboo end face segmentation and branch position detection method fused with improved ASPP and CBAM," *J. Forestry Eng.*, vol. 8, pp. 1–8, Jul. 2023, doi: 10.13360/j.issn.2096-1359.202211036.
[16] H. Li, X. Y. Wang, Y. Liu, S. J. Fu, and Y. F. Wu, "Detecting underwater objects using multi-scale features fusion and multiple attention," *Trans. Chin. Soc. Agricult. Eng.*, vol. 38, no. 20, pp. 129–139, 2022.
[17] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 10778–10787, doi: 10.1109/CVPR42600.2020.01079.
[18] Z. Qu and C. Y. Wang, "Crack detection of concrete pavement based on attention mechanism and lightweight dilated convolution," *Comput. Sci.*, vol. 50, no. 2, pp. 231–236, 2023.
[19] Z. Li, B. Li, H. Ni, F. Ren, S. Lv, and X. Kang, "An effective surface defect classification method based on RepVGG with CBAM attention mechanism (RepVGG-CBAM) for aluminum profiles," *Metals*, vol. 12, no. 11, p. 1809, Oct. 2022.
[20] T. Li, F. Zhang, G. Xie, X. Fan, Y. Gao, and M. Sun, "A high speed reconfigurable architecture for softmax and GELU in vision transformer," *Electron. Lett.*, vol. 59, no. 5, Mar. 2023, Art. no. e12751.
[21] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31 × 31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11963–11975.
[22] Y. Zhu and S. Newsam, "DenseNet for dense flow," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 790–794.
[23] D. Theckedath and R. R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *Social Netw. Comput. Sci.*, vol. 1, no. 2, pp. 1–7, Mar. 2020.
[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4278–4284.
[25] I. Z. Mukti and D. Biswas, "Transfer learning based plant diseases detection using ResNet50," in *Proc. 4th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Dec. 2019, pp. 1–6.
[26] A. I. Mohammed and A. A. Tahir, "A new optimizer for image classification using wide ResNet (WRN)," *Academic J. Nawroz Univ.*, vol. 9, no. 4, p. 1, Sep. 2020.
[27] O. Hermine, X. Mariette, P. L. Tharaux, M. Resche-Rigon, R. Porcher, and P. Ravaud, "Effect of tocilizumab vs usual care in adults hospitalized with COVID-19 and moderate or severe pneumonia: A randomized clinical trial," *J. Amer. Med. Assoc. Internal Med.*, vol. 181, no. 1, pp. 32–40, Jan. 2021.
[28] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9196–9205.
[29] W. Wu, H. Liu, L. Li, Y. Long, X. Wang, Z. Wang, J. Li, and Y. Chang, "Application of local fully convolutional neural network combined with YOLO v5 algorithm in small target detection of remote sensing image," *PLoS ONE*, vol. 16, no. 10, Oct. 2021, Art. no. e0259283.
[30] D. Wu, S. Jiang, E. Zhao, Y. Liu, H. Zhu, and R. Wang, "Detection of *Camellia oleifera* fruit in complex scenes by using YOLOv7 and data augmentation," *Appl. Sci.*, vol. 12, no. 22, p. 11318, Nov. 2022.
[31] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond," 2023, *arXiv:2304.00501*.
[32] S. B. Xu, L. M. Zheng, and D. C. Yuan, "A method for fabric defect detection based on improved cascade R-CNN," *Adv. Textile Technol.*, vol. 30, no. 2, pp. 48–56, 2022.
[33] L. Liu, J. Zhou, B. Zhang, S. Dai, and M. Shen, "Visual detection on posture transformation characteristics of sows in late gestation based on Libra R-CNN," *Biosystems Eng.*, vol. 223, pp. 219–231, Nov. 2022.
[34] X. Zhou, H. Wang, S. Li, H. Peng, and J. Wu, "Complex traffic scene image classification based on sparse optimization boundary semantics deep learning," *Wuhan Univ. J. Natural Sci.*, vol. 28, no. 2, pp. 150–162, Apr. 2023.

LEI CHEN was born in May 1985. He received the Ph.D. degree. He is currently an Associate Professor and a Master's Supervisor with the Changchun University of Science and Technology. He is mainly engaged in the theoretical research of wireless communication systems, visual sensing systems, and edge computing. He has undertaken the research and development of several major national exploration projects, such as Chang'e and Tianwen, and many of his research results have been applied to the moon exploration project, fire exploration project, and commercial satellites. Several research results have been applied by the Moon Exploration Project, the Fire Exploration Project, and commercial satellites.

SHIMING AN was born in August 1994. He received the bachelor's degree in electronic science and technology from the Chengdu College of University of Electronic Science and Technology (UEST), in 2013, and the master's degree in electronic information from the Changchun University of Science and Technology (Changchun UTech), in 2021. His research interests include electronics and communication engineering.

SHUANG ZHAO was born in November 1981. He received the master's degree from Jilin University, in 2004, the master's and Ph.D. degrees from the Changchun University of Science and Technology, in 2007 and 2013, respectively, and the Ph.D. degree from the University of Western Ontario, in 2020. Since 2020, he has been an Associate Professor and a Master's Supervisor with the Changchun University of Science and Technology.

GUANDIAN LI was born in Shanwei, Guangdong, China, in 1997. He received the B.S. degree in electronic information science and technology from the Changchun University of Science and Technology, China, in 2019. His research interests include semantic segmentation and target detection.