# Application of SSD Optimization Algorithm in the Detection of Diseases of Bridge Subaqueous Structures

**Bo Li, Delong Zou, Yuan Liang, Xiangyang Liu**

***Abstract*— At present, the detection methods for bridge subaqueous structures are relatively single. Nowadays, deep convolutional neural networks are increasingly applied to the field of object detection, improving the traditional detection effect. In this paper, an R-MobileNet-SSD network model with lightweight characteristics is built for the detection of diseases of bridge subaqueous structures. The Retinex algorithm is used for processing to improve the brightness and clarity of images, while retaining the natural color and details of images, and the contrast of images is enhanced through histogram equalization. Subsequently, in order to improve the detection accuracy and training quality, a deep convolutional network DCGAN is built to generate virtual samples, and the SSD model is improved and analyzed. The main VGG model in the original SSD is replaced with a MobileNet model to make it more lightweight. Through the detection and analysis in different underwater environments, this paper verifies that the R-MobileNet-SSD model can detect diseases in various underwater environments, and the detection effect is greatly improved compared with the original SSD model.**

***Index Terms*— Image Enhancement, Crack Detection, SSD Network, Lightweight, Small Target Detection**

## I. INTRODUCTION

The SSD algorithm uses feature maps of different scales for target objects of different sizes through a convolutional neural network during the detection process. The characteristic that the size of the object is inversely proportional to the size of the feature map used for detection enables the SSD algorithm to simultaneously handle the detection of large and small targets. Riaño Yorley Dayana Caro[1] et al. introduced two control strategies for the remotely operated underwater vehicle (ROV) using two PID controllers or an LQG controller in the inner loop and a PID controller in the outer loop, which largely reduced the obstacles of the marine environment to the modeling of the imaging equipment and improved the control response through the anti-integral saturation algorithm gain. Zulkarnain O.W [2] et al. designed and developed a specific operating system and algorithm design for the ROV, which can take into account the original functions and improve stability through operation and testing, and supports the use of radio control. Peng Lincong [3] An improved SSD object detection algorithm is proposed. Introduce the improved comprehensive convolutional attention module CCBAM to enhance the network's sensitivity to small targets, construct the hierarchical feature fusion network HFFNet, and use dilated convolution to extract feature information of different scales. Jiang Shuai[4] proposed an SSD object detection algorithm based on stepwise multi-scale feature fusion. To increase the detailed and semantic information contained in the shallow features of the SSD model, two feature layers are introduced in the low-level feature part of the model. Only perform deconvolution operations on the two feature maps at the lower level of the model, and fuse the features of the three feature maps at different scales at the lower level in two steps. Zhou Maojun[5] proposed a MobileNetV3-SSD object detection algorithm, which can effectively detect surface defects of workpieces. By improving the skeleton network of SSD to MobileNetV3-Large, the number of network parameters and computational load can be effectively reduced, further enhancing the detection performance. Zhou Qiuhong[6] proposed a MobileNetV3-SSD object detection algorithm. By improving the skeleton network of SSD to MobileNetV3-Large, the network can capture target information of different scales and shapes, combining semantic segmentation assistance tasks with multi-layer feature fusion.

**Manuscript received September 10, 2025**

**Bo Li**, School of Traffic and Transportation Engineering, Dalian Jiaotong University, Dalian, China

**Delong Zou**, School of Traffic and Transportation Engineering, Dalian Jiaotong University, Dalian, China

**Yuan Liang,** School of Traffic and Transportation Engineering, Dalian Jiaotong University, Dalian, China

**Xiangyang Liu,** School of Traffic and Transportation Engineering, Dalian Jiaotong University, Dalian, China

## II. CONVOLUTIONAL NEURL NETWORK

### A. Basic Models of Convolutional Neural Networks

In recent years, with the continuous development of convolutional neural networks, there are currently two categories centered around convolution. One is more inclined to be used for image classification, and its basic models include fundamental models such as LeNet, AlexNet, and VGG. The other is more inclined to be used for object detection, and its main models are the YOLO and SSD models.

The network of the SSD (Single Shot MultiBox Detector) algorithm consists of a convolutional neural network and a prediction network, as shown in Figure 1.1. The convolutional neural network is used for feature extraction, mainly to convert the input image into a set of feature maps, usually using a pre-trained network . The prediction network is used to predict the location and category of the target, mainly to find the target in the feature map. Its basic idea is to classify and regress a series of prior boxes with specified sizes and aspect ratios at each position in the feature map.

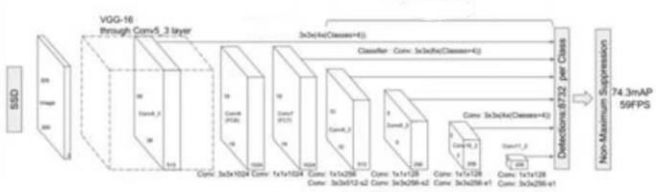



Fig.1.1 SSD Model Network Structure Diagram

### B. Image Enhancement Algorithms

It is a very common method to reflect the specific features of an image through other means.

(1) Histogram Equalization

The image histogram can utilize its own shape characteristics to reflect the gray-level distribution of target pixels. By transforming the shape of the image histogram, its contrast can be changed. The linear relationship expression is shown in formula (1-1). In the histogram, the abscissa represents the gray level ; the ordinate represents the number of pixels corresponding to the gray level.

$p(r_k) = \frac{n_k}{N}$(1-1)

(2) Retinex Algorithm

Retinex is an image processing technology aimed at simulating the ability of the human visual system to adapt to changes in lighting conditions. Through continuous development, it has been applicable to fields such as image enhancement and color correction. Its underlying principle is to separate and weight the color and brightness of an image, thereby improving the clarity and contrast of the image while maintaining natural colors.

## III. Establishment of Crack Disease Database and Image Preprocessing

### A. Image Enhancement Method

(1)Apply the histogram equalization method in gray value enhancement

The first step: Calculate the histogram of the image.

Calculate the number of pixels for each gray level of the image to form the histogram of the image. The number of pixels $n$ of gray level $k$ can be calculated using the following formula (2-1). Here, $I$ represents the pixel value of the image at coordinates (i, j), $M$ and $N$ are the number of rows and columns of the image respectively, and $[I = k]$ is 1 when $I(i,j) = k$ , otherwise 0.

$n_\kappa = \sum_{i=1}^{M}\sum_{j=1}^{N} I(i,j) = [I = k]$ (2-1)

The second step: Calculate histogram equalization.

Calculate the histogram equalization function $h_{eq}(k)$ , which maps the gray level $k$ of the original image to the enhanced gray level $h_{eq}(k)$ . Calculate using formula (2-2). Here, $L$ is the number of gray levels, and $n_j$ represents the number of pixels of the gray level in the image.

$h_{eq}(k) = \frac{(L-1)}{M*N}\sum_{j=0}^{k} n_j$(2-2)

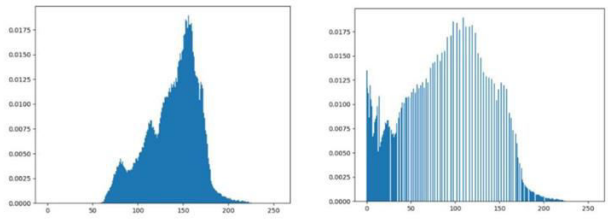
The third step: Apply the histogram equalization function.

It is necessary to apply $h_{eq}(k)$ to each pixel point in the original image to obtain the enhanced image. The enhanced gray value $h_{\text{out}}(i,j)$ can be calculated using formula (2-3).

$h_{\text{out}}(i,j) = h_{eq}(I(i,j))$ (2-3)

After equalization processing, stretch the gray values of the image. In the interval of (0-50) in Figure (b), there exists a pixel image.

The equalized histogram is shown in Figure 2.1.



(a) Histogram of the original image (b) equalized histogram

Fig. 2.1 Schematic Diagram of Histogram Equalization Comparison

Through the specific analysis of the actual image and the comparison before and after applying the equalized histogram, the difference can be clearly seen. Therefore, histogram equalization can indeed improve the clarity and contrast of the image. As shown in Figures 2.2 and 2.3.

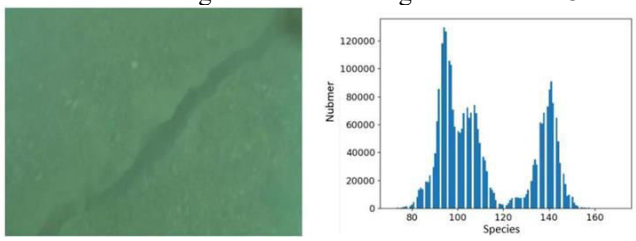



Fig. 2.2 Original Image and Histogram of Underwater Crack

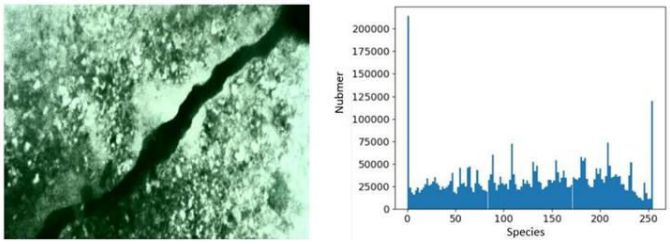



Fig. 2.3 Histogram Equalization Algorithm and Histogram

Processing and Application of Retinex Image Processing Technology.

(1) Reflectance Estimation: In this step, the reflectance distribution of the image is calculated based on the image pyramid at different scales. This step is usually implemented using the local contrast enhancement algorithm in the Retinex algorithm, that is, the local contrast of the image at each scale is calculated, and then the reflectance distribution is obtained through a non-linear function.

$R_i(x,y) = \frac{\log(I_i(x,y)) - \log(I_i(x,y)) * G_i(x,y)}{\log(I_i(x,y))}$(2-4)

Among them, $I_i(x,y)$ represents the pixel value at the coordinate (x, y) in the $i$ -th layer image, $G_i(x,y)$ is the Gaussian filter, and $R_i(x,y)$ represents the reflectance.

(2) Brightness Restoration: In this step, the brightness distribution of the image is calculated based on the image pyramid at different scales. This step is usually implemented using the global contrast enhancement algorithm in the Retinex algorithm, that is, the global contrast of the images at all scales is calculated, and then the brightness distribution is obtained through a non-linear function.

$L_i(x,y) = \frac{1}{N}\sum_{j=1}^{N} R_j(x,y)$(2-5)

Among them, $N$ is the number of scales in the image pyramid, and $L_i(x,y)$ represents the brightness distribution.

(3) Combination of Reflectance and Brightness: In this step, the reflectance and brightness information are combined to obtain the enhanced image. This step is usually implemented using the maximum reflectance and minimum brightness algorithm in the Retinex algorithm, that is, the reflectance and brightness are normalized respectively, and then the maximum reflectance and minimum brightness are taken for combination.

$$E_i(x,y) = R_i(x,y) * \left(I_i(x,y) - L_i(x,y) + L_m(x,y)\right) \quad (2\text{-}6)$$

Among them, $L_m(x,y)$ is the maximum brightness in all sizes, and $E_i(x,y)$ represents the enhanced image. The Retinex image processing technology can enhance the clarity and contrast of the image while retaining the natural color and details of the image. As shown in Figure 2.4.

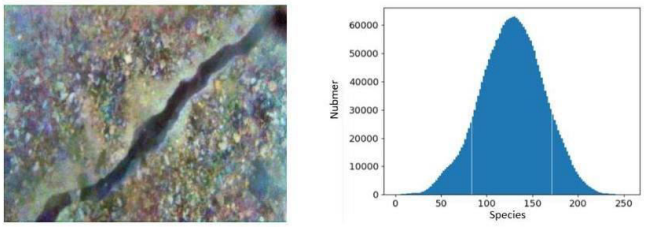



Fig. 2.4 Retinex Algorithm and Histogram

B.Combined Analysis Based on Two Image Enhancement Methods

(1) First, the Retinex algorithm [7]is applied to the original image for correction and illumination level, then histogram equalization is performed on the image processed by Retinex to enhance the contrast of the image, and finally the quality of the enhanced image is checked. The result is shown in Figure 2.5.

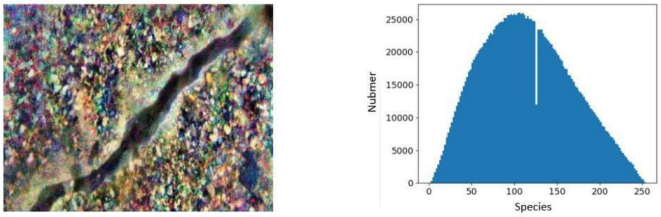



Fig. 2.5 Image after Histogram Equalization of Retinex Enhanced Image

(2) Experimental Analysis

Next, the special evaluation index for underwater images is used to evaluate the image quality. The commonly used underwater image evaluation index is UCIQE. The UCIQE value can be used to quantitatively evaluate the color distortion, blur and low contrast of underwater images. Its expression is:

$$\text{UCIQE} = c_1 \times \delta_c + c_2 \times \text{con}_1 + c_3 \times \mu_s \quad (2\text{-}7)$$

In the formula, $c_1$ 、 $c_2$ 、 $c_3$ are the weight coefficients of the three feature components, and their values are usually taken as 0.468, 0.2745, and 0.2576 respectively. $\delta_c$ represents the standard deviation of the image colorfulness, $\text{con}_l$ represents the contrast of the image, and $\mu_s$ represents the average value of the image saturation. The weight coefficients and the UCIQE values are shown in Table 2.1.

Table 2.1 UCIQE Value Table of Underwater Images

| Enhanced algorithm | $\delta_c$ | $\text{con}_1$ | $\mu_s$ | UCIQE value |
|---|---|---|---|---|
| Original underwater image | 0.0134 | 0.1507 | 0.3284 | 0.1322 |
| Histogram equalization | 0.0181 | 0.9383 | 0.3987 | 0.3688 |
| Retinex | 0.2680 | 0.7441 | 0.2194 | 0.3862 |
| Combined method | 0.2513 | 0.8674 | 0.3482 | 0.4454 |

| Enhanced algorithm | $\delta_c$ | $\text{con}_1$ | $\mu_s$ | UCIQE value |
|---|---|---|---|---|
| Original underwater image | 0.0134 | 0.1507 | 0.3284 | 0.1322 |
| Histogram equalization | 0.0181 | 0.9383 | 0.3987 | 0.3688 |
| Retinex | 0.2680 | 0.7441 | 0.2194 | 0.3862 |
| Combination method | 0.2513 | 0.8674 | 0.3482 | 0.4454 |

It can be observed from the above table that the combined enhancement algorithm of the two methods studied in this paper can greatly improve the quality of underwater images, and the UCIQE value is significantly higher than that of other individual algorithms.

C.Underwater crack data augmentation based on CGAN

The data samples were preprocessed by applying Retinex image equalization. The DCGAN network was improved as follows:

(1) The pooling layer in the original network was replaced with fractional-strided convolution [8]to complete the upsampling operation. Fractional-strided convolution is opposite to normal convolution. In the case of normal convolution, the width and height of the feature layer will shrink after convolution, while fractional-strided convolution will make the width and height of the feature layer continuously increase. The schematic diagram of fractional-strided convolution is shown in Figure 2.6.

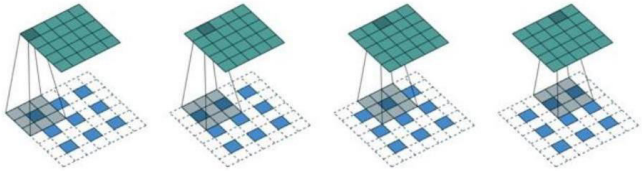
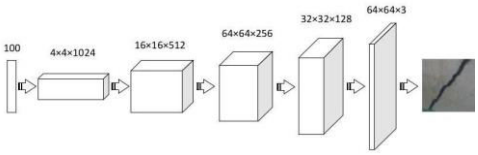


Figure 2.6 Schematic diagram of fractional-strided convolution

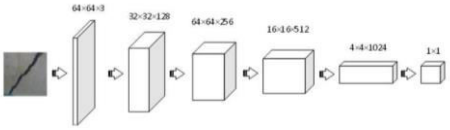
(2) The BN layer was added to the network layer to avoid the problem of network collapse and make the gradient descent more stable, reducing the possibility of gradient disappearance. The first few layers of the generator network used the ReLU function as the activation function, and the last network layer used the tanh function as the activation function. The first few layers of the discriminator network used the leakyrelu function, and the last layer used the sigmoid function.

The structure diagram of the DCGAN network is shown in Figure 2.7. First, a 4*4*1024 image was obtained through a single fully connected layer, and then through 4 convolutional layers with fractional-strided convolution, the scale of the feature layer was gradually increased, and finally a 64*64*3 matrix-expanded output crack image was obtained. As shown in Figure 2.7 (b), the input image first underwent 4 consecutive samplings to perform convolution extraction on the eigenvalues of each layer, shrinking the image, and finally passing through the fully connected layer

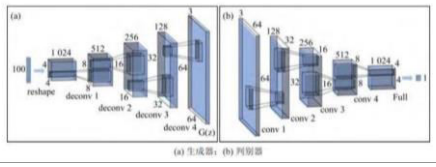
to obtain an output probability for judging the authenticity of the image.




(a) Generator network structure diagram



(b) Discriminator network structure diagram



(c)U-Net network structure diagram

Fig. 2.7 Structure diagram of the DCGAN network

D.Improvement and Optimization of the SSD Algorithm

In this paper, the lightweight MobileNet was used as the backbone feature extraction network. Compared with VGG, the accuracy decreased by 9%, but the model parameters were only 1/3 of VGG, which could further reduce the computational amount at the cost of sacrificing a small amount of accuracy.

(1) The MobileNet structure makes the whole lighter

By performing layer-by-layer convolution, depthwise separable convolution, and pointwise convolution in the deep convolutional neural network respectively, the computational amount can be greatly reduced. The structure of the deep convolutional neural convolution is shown in Figure 2.8.

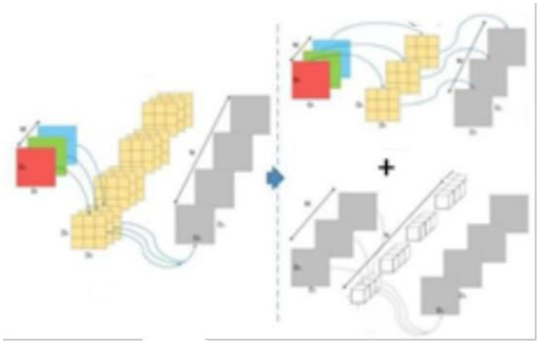


Fig. 2.8 Structure diagram of depthwise convolution

The computational amount formula of its depthwise separable convolution (DW) is shown in Equation 2-8, and the computational amount formula of the standard convolution is shown in Equation 2-9. The ratio formula is shown in Equation 2-10. Among them, $D_F$ is the height and width of the input feature matrix, $D_K$ is the size of the convolution kernel, $M$ is the depth of the input feature matrix, and $N$ is the depth of the output feature matrix.

$$D_K \times D_K \times M \times D_F \times D_F + 1 \times 1 \times M \times N \times D_F \times D_F \quad (2\text{-}8)$$

$$D_K \times D_K \times M \times N \times D_F \times D_F \quad (2\text{-}9)$$

$$\frac{D_K \times D_K \times M \times D_F \times D_F + 1 \times 1 \times M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{{D_K}^2} \quad (2\text{-}10)$$

The improvement of the original convolution process by MobileNet is shown in Figure 2.9.

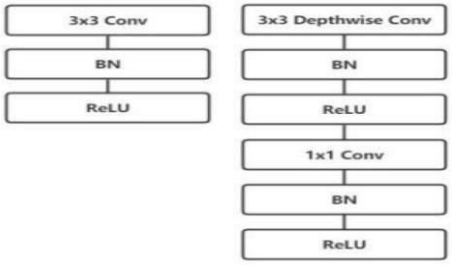



Fig.2.9 Diagram of the improvement process of the MobileNet network

After a standard convolution of $3 * 3$, depthwise separable convolution is used instead of standard convolution. In order to reduce the computational amount, convolution kernels with a stride of 2 are used instead of average pooling layer operations, which not only extracts features but also reduces the computational amount at the same time. As can be seen from Table 2.2, the calculation of the MobileNet network mainly focuses on 1*1 convolution calculation, and the parameters also account for 74.59%. Table 2.3 can reflect the performance comparison between MobileNet and VGG.

Table 2.2 Distribution table of the computational amount and parameters of the MobileNet network

| Type | Mult-Adds | Parameters |
|---|---|---|
| Conv 1x1 | 94.86% | 74.59% |
| Conv DW 3x3 | 3.06% | 1.06% |
| Conv 3x3 | 1.19% | 0.02% |
| Fully Connecte | 0.18% | 24.33% |
| Type | Mult-Adds | Parameters |
| Conv 1x1 | 94.86% | 74.59% |
| Conv DW 3x3 | 3.06% | 1.06% |
| Conv 3x3 | 1.19% | 0.02% |
| Fully Connecte | 0.18% | 24.33% |

Table 2.3 Comparison Table of the Performance of MobileNet and VGG

| Model | ImageNet Accuracy | Million Mult-Adds | Million Parame |
|---|---|---|---|
| MobileNet | 70.6% | 5690 | 4.2 |
| VGG | 71.5% | 15300 | 138 |
| Model | ImageNet accuracy | Millions of multiplicative cumulative operations | Millions of parameters |
| MobileNet | 70.6% | 5690 | 4.2 |
| VGG | 71.5% | 15300 | 138 |

Improve the SSD algorithm. Incorporate MobileNet into SSD in the way that the MobileNet-SSD network extracts feature layers according to the method from the VGG feature layers to the fully connected layer. This improvement method is to extract 6 convolutional layers respectively. After prior box processing through Conv11, Conv13, Conv14_2, Conv15_2, Conv16_2, and Conv17_2, connect them to the fully connected layer for classification and regression tasks. The number of prior boxes for these 6 extracted feature layers is 4,

6, 6, 6, 4, and 4 respectively. After training, obtain the predicted result sizes corresponding to the best all feature layers of these 6 convolutional layers, as shown in Table 2.4.

Table 2.4 Feature Size and Prior Box Size Table of MobileNet-SSD Network

| Effective | The number of | The feature size of the prior box | Predict the feature size of the classification |
|---|---|---|---|
| Conv11 | 4 | (38,38,16) | (38,38,84) |
| Conv13 | 6 | (19,19,24) | (19,19,126) |
| Conv14_2 | 6 | (10,10,24) | (10,10,126) |
| Conv15_2 | 6 | (5,5,24) | (5,5,126) |
| Conv16_2 | 4 | (3,3,16) | (3,3,84) |
| Conv17_2 | 4 | (1,1,16) | (1,1,84) |
| Effective feature layer | Number of prior boxes | Feature size of prior boxes | Feature size for predicting classification |
| Conv11 | 4 | (38,38,16) | (38,38,84) |
| Conv13 | 6 | (19,19,24) | (19,19,126) |
| Conv14_2 | 6 | (10,10,24) | (10,10,126) |
| Conv15_2 | 6 | (5,5,24) | (5,5,126) |
| Conv16_2 | 4 | (3,3,16) | (3,3,84) |
| Conv17_2 | 4 | (1,1,16) | (1,1,84) |

After embedding the pre-trained MobilelNet model into SSD, in this step, it is necessary to connect the feature extraction layer of the MobileNet model to the detection head of the SSD algorithm, and then use its feature map (including image features and location information) as the input of the SSD algorithm. In the feature extraction layer of MobileNet-SSD, a set of convolutional layers usually follow to form the detection head, and the output of the feature extraction needs to be connected to the input of the detection head.

(2) The receptive field module RFB enhances the detection of small targets

The receptive field module RFB is a multi-branch convolutional module similar to the inception module, which obtains receptive fields of various sizes through different multi-branch convolutional kernels. The multi-branch convolutional module replaces the 3*3 convolutional kernel with 3*1 and 1*3 convolutional kernels. The number of parameters of the three models, namely the original SSD-VGG model, the MobileNet-SSD model, and the MobileNet-SSD model with the receptive field module with fused features, are calculated and compared. As shown in Table 2.5.

Table 2.5 Comparison of the computational amount and the number of parameters of each model

| Model | calculation parameters | quantity | model memory |
|---|---|---|---|
| SSD-VGG | 15300M | 138M | >500MB |
| MobileNet-SSD | 1300M | 8.8M | 32MB |
| R-MobileNet-SSD | 1450M | 9.6M | 33.8MB |
| Model | Computational complexity | Number of parameters | Model memory |
| SSD-VGG | 15300M | 138M | >500MB |
| MobileNet-SSD | 1300M | 8.8M | 32MB |
| R-MobileNet-SSD | 1450M | 9.6M | 33.8MB |

## IV. SURFACE CRACK DETECTION BASED OBJECT DETECTION

Since this model first enhances the image, it has the characteristics of simplicity, convenience and high economic efficiency compared with other detection methods. Next, the specific detection effects of each network model of the image are analyzed.

(1) In a clear environment, the detection effects of each model are compared as shown in Figure 3.1.

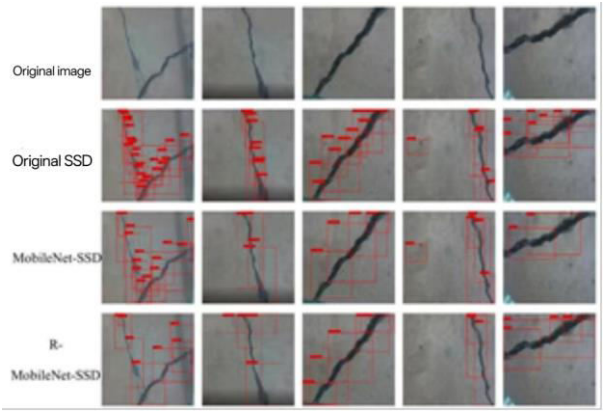
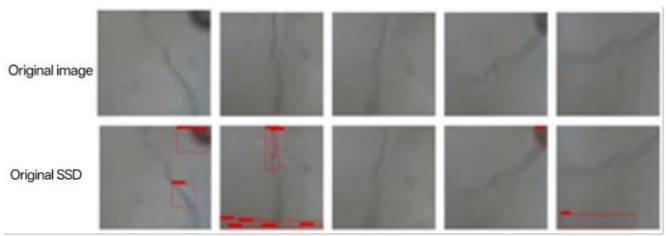



Fig. 3.1 Comparison Diagram of Model Detection in Clear Water Environment

The original SSD has the situation of over-detecting cracks. The lightweighting of the original SSD network by MobileNet is also specifically reflected. The reduction in the number of prediction boxes generated can reduce the over-detection phenomenon. After adding the feature fusion receptive field module on this basis, the detection of small targets has also been improved.

(2) In a turbid water environment, the detection effects of each model are compared as shown in Figure 3.2.



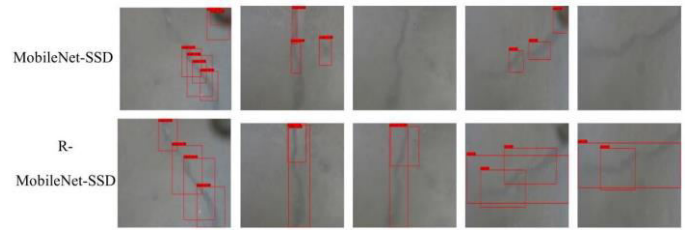



Fig. 3.2 Comparison Diagram of Model Detection in Turbid Water Environment

As can be seen from the above figure, after improving the network, the crack detection has been significantly improved. And after applying the network in this paper, while the detection of small targets has been improved, the underwater turbid cracks can be predicted and judged.

(3) In a deep water environment, the detection effects of each model are compared as shown in Figure 3.3.

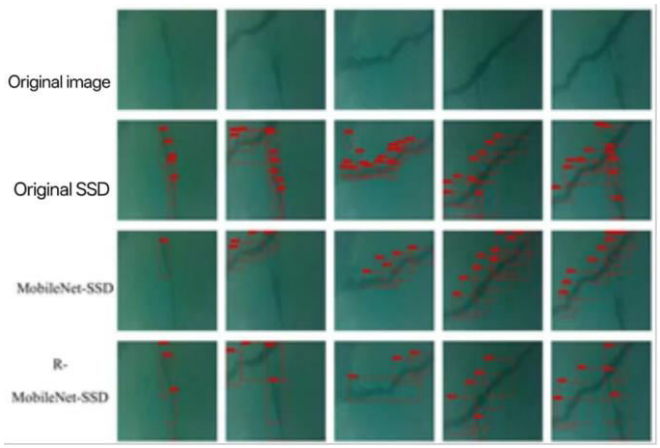



Fig.3.3 Comparison Diagram of Model Detection in Deep Water Environment

As can be seen from the above figure, in the deep water environment. The original SSD network can identify cracks, but there are many prediction boxes. The network after improving the main network can reduce the number of prediction boxes of the original SSD network, but there is a phenomenon that the prediction boxes are not completely wrapped and there are omissions. After adding the feature fusion receptive field model on the basis of the improvement of the text method, the omissions and the detection of small targets are supplemented. It meets the detection in the deep water environment.

## V. CONCLUSION

Aiming at the requirements of bridge detection for speed and accuracy, based on the original SSD network model, a more lightweight R-MobileNet-SSD network model is built. The R-MobileNet-SSD network model replaces the VGG basic structure in the original SSD model with the MobileNet structure, which is 65% smaller than the original model, and the computational amount is greatly reduced. The receptive field RFB module for fusing features is added to the R-MobileNet-SSD model, which enhances the ability to detect small targets. Through experiments under three different underwater environmental conditions of clear water environment, turbid water environment and deep water environment, it is proved that in different underwater environments, the R-MobileNet-SSD model enhances the ability of the original SSD network to detect small targets. The results show that the R-MobileNet-SSD network model can detect the underwater crack diseases of bridges with higher accuracy and faster detection speed.

## REFERENCES

[1] Riaño Yorley Dayana Caro, Prada Sebastián Roa. Development of Two Control Strategies for Tracking the Trajectory of An Unmanned Underwater Structure Inspection Vehicle[J]. Journal of Physics: Conference Series,2022,2224(1).

[2] J. Guerrero, J. Torres, V. Creuze, A. Chemori, E. Campos. Saturation based nonlinear PID control for underwater vehicles: Design, stability analysis and experiments[J]. Mechatronics, 2019, 61.

[3] Peng Lincong, Wang Kerui, Zhou Hao, et al. Target detection algorithm based on improved SSD [J]. Laser journal, 2024, (11) : 71-76.

[4] Jiang Shuai, Xue Bo SSD Object Detection Algorithm Based on Stepwise Multi-scale Feature Fusion [J]. Computer and Digital Engineering, 24,52(10):2972-2976.

[5] Zhou Maojun, Hu Jiangtao, Wang Junjie, et al. Application of Improved SSD Algorithm in Workpiece Surface Defect Detection [J]. Combined Machine Tool & Automated Machining Technology,2024,(08):139-144.

[6] [1] Zhou Qiuhong. Research on Road Disease Detection Based on Improved SSD Model [J]. Heilongjiang Transportation Science and Technology,2023,46(04):30-32.

[7] Xie Wengao, Zhang Yixiao, Liu Airong, et al. Detection method of concrete structure surface cracks based on underwater robot and digital image technology[J]. Engineering Mechanics, 2022, 39(S1): 64 - 70.

[8] Zhao Chunyan. Research on the detection of surface defects of underwater structures based on image processing technology[D]. Shenyang Jianzhu University, 2021.