Hindawi
Advances in Materials Science and Engineering
Volume 2021, Article ID 9934250, 10 pages
https://doi.org/10.1155/2021/9934250

Research Article

Crack Identification Method of Steel Fiber Reinforced Concrete Based on Deep Learning: A Comparative Study and Shared Crack Database

Yang Ding,[1,2,3] Shuang-Xi Zhou,[4] Hai-Qiang Yuan,[4] Yuan Pan,[4] Jing-Liang Dong,[4] Zhong-Ping Wang,[1,2] Tong-Lin Yang,[5] and An-Ming She[1,2]

[1]Key Laboratory of Advanced Civil Engineering Materials of Ministry of Education, Tongji University, Shanghai 201804, China
[2]School of Materials Science and Engineering, Tongji University, Shanghai 201804, China
[3]Department of Civil Engineering, Zhejiang University, Hangzhou 310058, China
[4]School of Civil Engineering and Architecture, East China Jiao Tong University, Nanchang 330013, China
[5]College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China

Correspondence should be addressed to An-Ming She; sheanming@tongji.edu.cn

Received 23 June 2021; Revised 8 August 2021; Accepted 23 August 2021; Published 11 September 2021

Academic Editor: Chundi Si

As a common disease of concrete structure in engineering, cracks mainly lead to durability problems such as steel corrosion, rain erosion, and protection layer peeling, and then the building gets destroyed. In order to detect the cracks of concrete structure in time, the bending test of steel fiber reinforced concrete is carried out, and the pictures of concrete cracks are obtained. Furthermore, the crack database is expanded by the migration learning method and the crack database is shared on the Baidu online disk. Finally, a concrete crack identification model based on YOLOv4 and Mask R-CNN is established. In addition, the improved Mask R-CNN method is proposed in order to improve the prediction accuracy based on the Mask R-CNN. The results show that the average prediction accuracy of concrete crack identification is 82.60% based on the YOLO v4 method. The average prediction accuracy of concrete crack identification is 90.44% based on the Mask R-CNN method. The average prediction accuracy of concrete crack identification is 96.09% based on the improved Mask R-CNN method.

1. Introduction

Nowadays, the concrete crack detection is mainly through manual identification [1, 2]. The manual detection method is not only time consuming but also requires a lot of energy from the relevant detection personnel [3, 4]. There are some problems such as low detection accuracy and subjectivity of operators [5, 6]. In addition, cracks in some special areas cannot be detected manually, such as bridge piers, mountainous areas, and high-risk urban areas [7, 8]. These cracks, which are difficult to detect, may cause structural weakness, leading to ductile failure and brittle failure, leading to serious safety accidents [9, 10].

In recent years, the deep learning method has been widely used in the field of civil engineering and has attracted the attention of many researchers [11]. Hinton et al. [12] proposed the deep learning model for the first time. The result showed that the artificial neural network with multiple hidden layers optimizes the network through layer by layer initialization, realizes feature learning, and opens a new era of deep learning. Krizhevsky et al. [13] designed the AlexNet algorithm, which is the first deep neural network model established by convolutional neural network. Girshick [14] proposed a new algorithm based on R-CNN and SPPNet: fast R-CNN. The result showed that the speed and accuracy have been improved, but there is still a long way to go from real end-to-end processing. Ren et al. [15] proposed fast R-CNN algorithm based on fast R-CNN network model and regional recommendation network, which achieved 78.8% detection accuracy on VOC2007 dataset. Lin et al. [16]

designed the feature pyramid network according to the different semantic and target location of different feature maps, which has certain advantages in small target detection. Redmon et al. [17] proposed a regression problem that unifies the classification regression problem into a coordinate frame, that is, Yolo algorithm. The results show that Yolo algorithm has very fast detection speed, but its accuracy is lower than that of the existing R-CNN series algorithm model, and the detection effect is poor when the object is small. Du et al. [18] proposed a new method to detect severe vehicle occlusion, which can be applied to aerial images of weak infrared camera with complex field background. Yu et al. [19] proposed the Mask R-CNN fruit detection model. The results show that the average detection accuracy is 95.78%, the recall rate is 95.41%, and the average intersection rate of instance segmentation is 89.85%. Pang et al. [20] proposed a segmented crack defect segmentation method, which solved the problems of uneven brightness and high noise of dam concrete surface image. Yu et al. [21] proposed a deep learning model YOLOv4-FPM based on the YOLOv4 model. The results show that the average accuracy of YOLOv4-FPM is 0.064 higher than that of original YOLOv4.

This paper takes steel fiber reinforced concrete as the research object, obtains concrete crack pictures through bending test, and expands the crack database based on the transfer learning method. Based on the deep learning algorithm, an automatic crack detection model is established, that is, YOLOv4 and Mask R-CNN. Furthermore, an improved Mask R-CNN concrete crack identification model is proposed based on the Mask R-CNN model.

2. Image Acquisition and Processing

2.1. Materials. Portland cement (42.5) was produced by China United Cement Group Co., Ltd., and its main components are shown in Table 1. Xiamen ISO standard sand is adopted. Steel fiber is a flat copper plated steel fiber with diameter of 0.2 mm and length of 13 mm. Distilled water was used.

Steel fiber concrete with fixed water binder ratio and limestone ratio of 0.4 and 1 : 2 was prepared. In this experiment, 10 batches of steel fiber mortar specimens were prepared, which were 0.1%, 0.3%, 0.5%, 1%, 1.5%, 2%, and 3%, respectively. Each batch was divided into five groups according to the vibration time of 0.5 min, 1 min, 1.5 min, 2 min, and 2.5 min. Firstly, sand and cement are added to dry mix for 1-2 minutes. After mixing evenly, 90% and 10% water are added in turn. When the cementitious material is gradually formed, steel fibers are evenly sprinkled and fully stirred to avoid fiber polymerization at one place of the test block. After the specimen is vibrated, it is placed in the room for 24 hours before demoulding and soaking in water for curing. At the same time, ensure that the water level overflows the specimen. The curing time of the specimens was 90 days. The specimens were dried at room temperature for 12 hours in advance. The concrete bending test is carried out with the size of 100 mm × 100 mm × 400 mm prism specimen. Specifically, the effective span of the beam is 300 mm, the beam height is 100 mm, and the beam width is 100 mm. Based on the CECS 13-2009 standard, the bending test of fiber-reinforced concrete is carried out, and then the pictures of concrete cracks are obtained. Figure 1 shows the initial and final crack pictures of different steel fiber reinforced concrete.

Table 1: Main components of cement.

Materials	Chemical composition (mass ratio (%))							
	CaO	SiO_2	Al_2O_3	Fe_2O_3	MgO	K_2O	SO_3	CaO
Cement	65.87	21.62	5.49	4.08	0.81	0.85	1.28	65.87

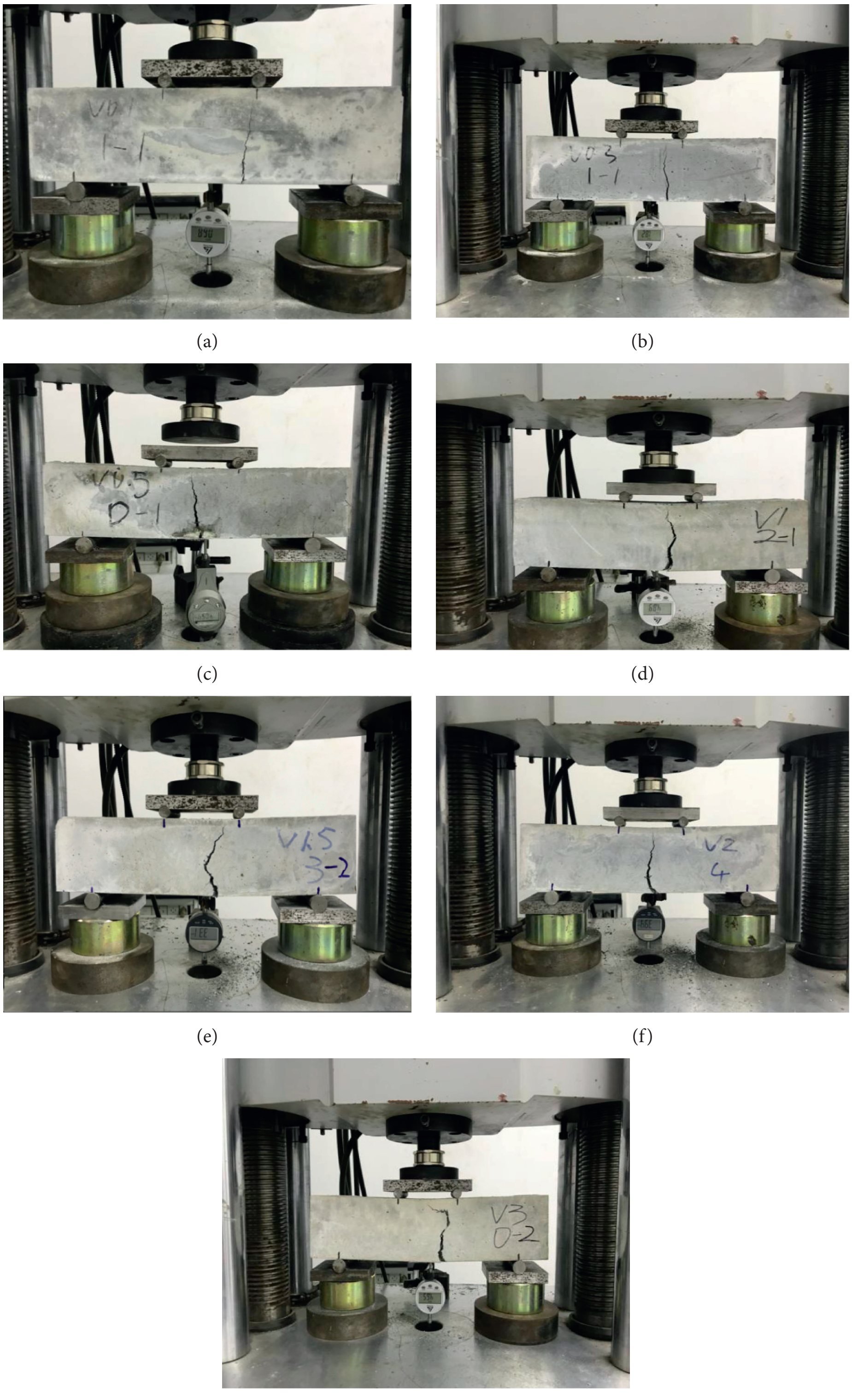
2.2. Image Preprocessing. Because the resolution of the original image is too large, the calculation cost will be too high if the original image is directly input [22]. Therefore, the original image will be cropped to include only the concrete test block image, which is also conducive to better learning the defect features of the model, as shown in Figure 2.

The image input model is transformed into a vector matrix to enter the network, and the latitude of the vector is fixed, so the resolution should be adjusted [23]. In this paper, the image is adjusted to 512 × 512 size, as shown in Figure 3.

Due to the experimental limitations, it is impossible to make enough sample data, so the crack data are enhanced to improve the robustness and generalization ability of the training model [24]. Rotating, blurring, flipping, and noise adding can be seen in Figure 4. Specifically, rotation refers to rotating the image randomly by an angle of 45, 90, and 180 degrees; flipping refers to rotating the image along the horizontal *X* axis or vertical *Y* axis; blurring refers to blurring the image; and adding noise refers to adding salt and pepper noise or Gaussian noise into the crack image. Finally, there are 1200 crack images as the training dataset, 400 crack images as the validation dataset, and 400 crack images as the test dataset.

3. Deep Learning Method

3.1. Model of Object Detection Algorithm for YOLOv4. The YOLOv4 algorithm model not only improves the speed but also improves the detection accuracy [25]. The YOLOv4 network structure includes four parts [26]. (1) The algorithm provides data-enhanced mosaic, cmBN, and SAT self-confrontation training at the input end, which enriches the detection dataset and reduces GPU calculation. (2) In feature extraction network, the activation function uses the Mish activation function to enhance the learning ability of the feature extraction network, ensure the lightweight of the network, reduce the calculation cost, and maintain the accuracy. (3) Neck network consists of SPP module and FPN + PAN structure. (4) In head detection network and loss function, CIoU_Loss is the loss function, which can be expressed by [27]



(a)

(b)

(c)

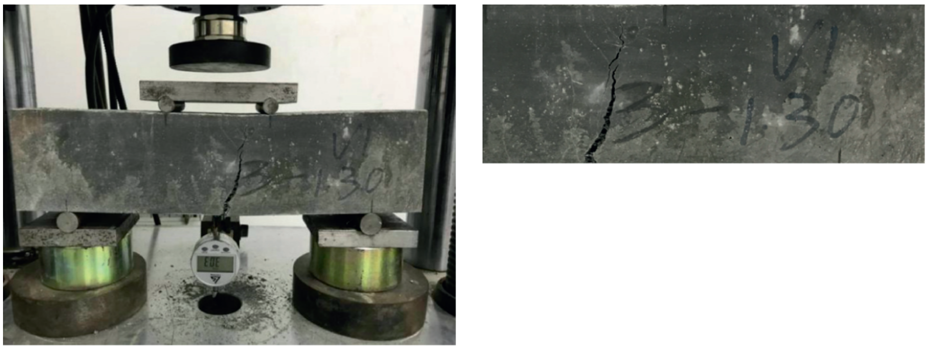
(d)

(e)

(f)

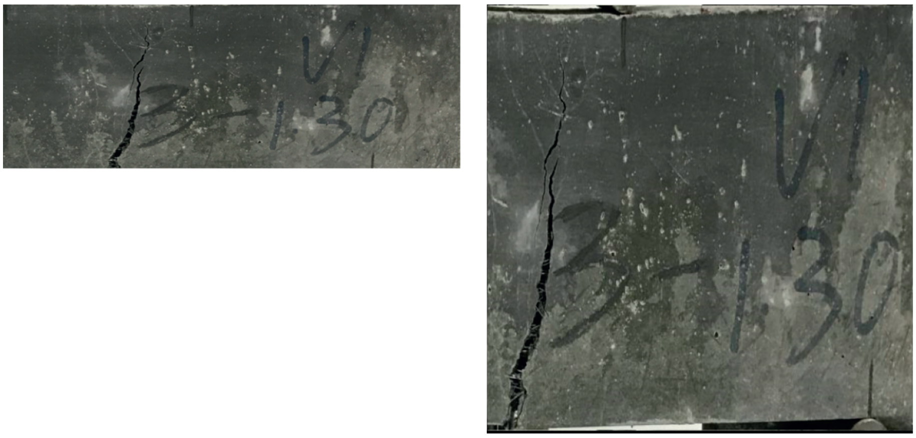
(g)

Figure 1: Crack image acquisition (final crack picture). (a) 0.1% steel fiber. (b) 0.3% steel fiber. (c) 0.5% steel fiber. (d) 1.0% steel fiber. (e) 1.5% steel fiber. (f) 2.0% steel fiber. (g) 3.0% steel fiber.



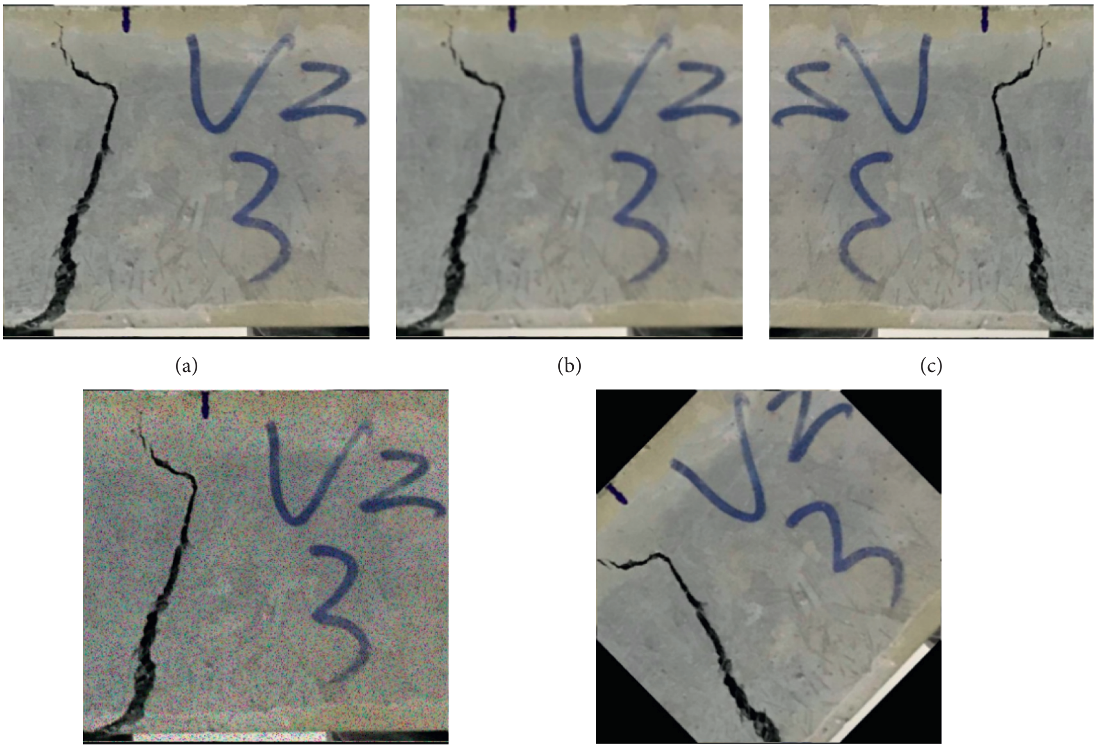
(a) (b)

Figure 2: Before and after cropping. (a) Before cropping. (b) After cropping.



(a) (b)

Figure 3: Resizing to 512 × 512. (a) Original picture. (b) Reconstructed images.



(a) (b) (c)

(d) (e)

Figure 4: Crack image data augmentation. (a) Original picture. (b) Blurring. (c) Flipping. (d) Noise adding. (e) Rotating.

$$\begin{aligned}\text{Loss}_{\text{CIOU}} &= 1 - \text{IOU} + \frac{\rho^2\left(b, b^{\text{gt}}\right)}{c^2} + \text{av},\\ a &= \frac{v}{1 - \text{IOU} + v},\\ v &= \frac{4}{\Pi^2}\left(\arctan\frac{\omega^{\text{gt}}}{h^{\text{gt}}} - \arctan\frac{\omega}{h}\right)^2,\end{aligned} \tag{1}$$

where $\rho^2(b, b^{gt})$ represents the Euclidean distance of the center point of the prediction box and the real box, respectively, and C represents the diagonal distance of the smallest closure region that can contain both prediction box and real box.

YOLOv4 model's parameters are as follows: (1) epoch = 100, that is, 1200 crack image data are trained for 100 times; (2) batch size = 16, that is, one round of 16 image data samples is used for model training; (3) iterations = 75, that is, 1200 pictures, 16 pictures are extracted each time, and there are 75 groups in total, i.e., one epoch is completed; (4) learning rate = 10^{-5}; and (5) momentum = 0.9.

3.2. Model of Object Detection Algorithm for Mask R-CNN.

He et al. [28] proposed the Mask R-CNN algorithm model to complete the task of target detection combined with instance segmentation, and at the same time, the target was segmented at the pixel level, which can be seen in Figure 5.

The Mask R-CNN network structure includes three parts [29, 30]: (1) feature extraction network—the fusion feature map generated by feature extraction network residual network combined with feature pyramid network will cause aliasing effect, and the target detection feature map is obtained by a 3 × 3 convolution; (2) RPN network—3 × 3 × 256 convolution kernel is used to convolute it into 1 × 1 × 256 dimensional feature results, and 2n classification and 4n coordinate regression are obtained through classification layer and regression layer; (3) head detection network and loss function—detection network includes mask branch, prediction category, and frame regression after full connection. The Mask R-CNN model is used to complete classification and location and mask generation, and its loss function is composed of the sum of three loss functions, which can be expressed by [31]

$$\begin{aligned}L &= L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}},\\ L_{\text{cls}} &= \frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}\left(p_i, p_i *\right),\\ L_{\text{box}} &= \lambda\frac{1}{N_{\text{reg}}}\sum_i p_i * L_{\text{CIoU}},\\ L_{\text{mask}} &= -\frac{1}{s}\sum_i\left[s_i * \lg p\left(s_i\right) - \left(1 - s_i^*\right)\lg\left(1 - p\left(s_i\right)\right)\right],\end{aligned} \tag{2}$$

where L_{cls} is the classification loss function; L_{box} is the regression loss function; L_{mask} is the average binary cross entropy; p_i is the probability of predicting the target; p_i^* indicates whether it is a real target; N_{cls} is the number of classification layers; N_{reg} is the number of regression layers; s is the sum of the total number of a category for each pixel; s_i^* is the label of the pixel category; and $p(s_i)$ is the probability of prediction category.

Mask R-CNN model's parameters are as follows: (1) epoch = 100; (2) batch size = 4; (3) iterations = 300; (4) learning rate = 10^{-5}; and (5) momentum = 0.9.

3.3. Model of Object Detection Algorithm for Improved Mask R-CNN.

In order to improve the accuracy of classification and location, the Mask R-CNN algorithm in the crack detection model is improved, which mainly improves the backbone network and enhances its feature expression ability. The main network of Mask R-CNN algorithm in the crack detection model is composed of residual network and feature pyramid network [32]. Based on the repeat layer strategy network of residual network, k–1 cardinal numbers are added to each module. After splitting, the cardinal numbers are decentralized. Each cardinal number is summed and fused by multiple segmentation elements to get the output of feature graph: h, w, and c. In the Cardinal layer, the (1 × 1) network is convoluted into (3 × 3). (3 × 3) The input of the base array is divided into r scattered blocks, and each scattered block is transformed into the distraction module [33]. The elements are added one by one, and the feature graph is fused into the output dimension: $h \times w \times c$. Then, the fusion feature map is pooled globally, and the image, spatial dimension is compressed to output dimension c. The dense c in the weight graph of each scattered block is calculated based on Softmax. The module input characteristic graph and its weight are multiplied to get the cardinality group, and then the output dimension $h \times w \times c$ is weighted and fused [34]. Distractor fuses the corresponding weights calculated from the scatter block feature graph to form ResNeSt unit module, which can be seen in Figure 6.

3.4. Evaluating Indicator.

Average precision can reflect the fracture identification accuracy of the network model, which can be expressed by [35]

$$\begin{aligned}F1 &= \frac{2PR}{P + R},\\ P &= \frac{T_P}{T_P + F_P} \times 100\%,\\ R &= \frac{T_P}{T_P + F_N} \times 100\%,\end{aligned} \tag{3}$$

where $F1$ is the average mean precision; P is the accuracy rate, that is, the proportion of correctly predicted positive case data to predicted positive case data; R is the recall rate, that is, the proportion of the predicted positive case data to the actual positive case data; T_P is the number of positive samples correctly predicted; F_P is the number of negative

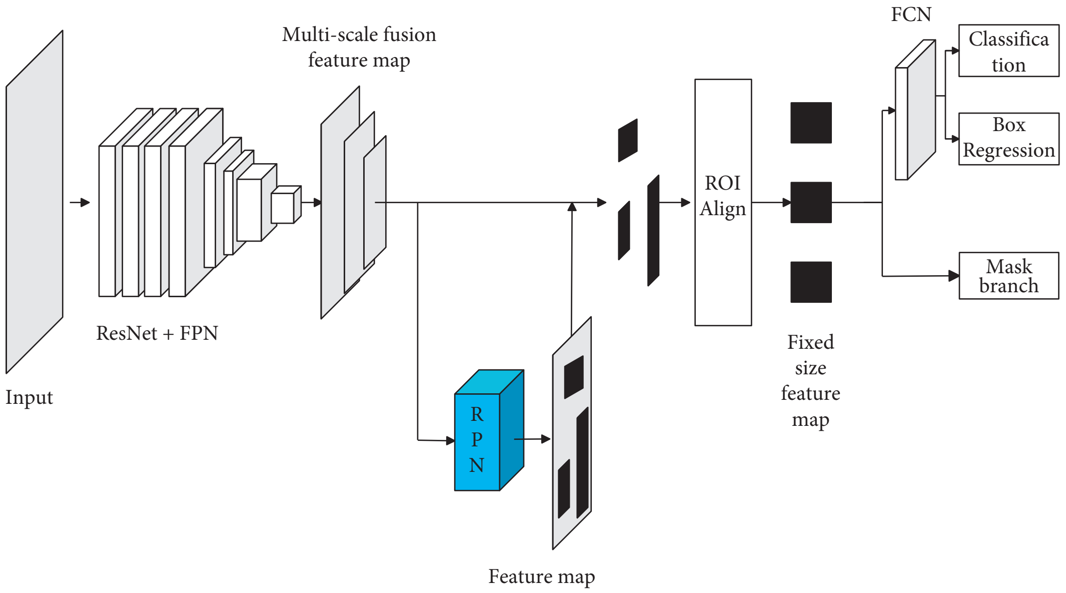


FIGURE 5: Mask R-CNN model frame.

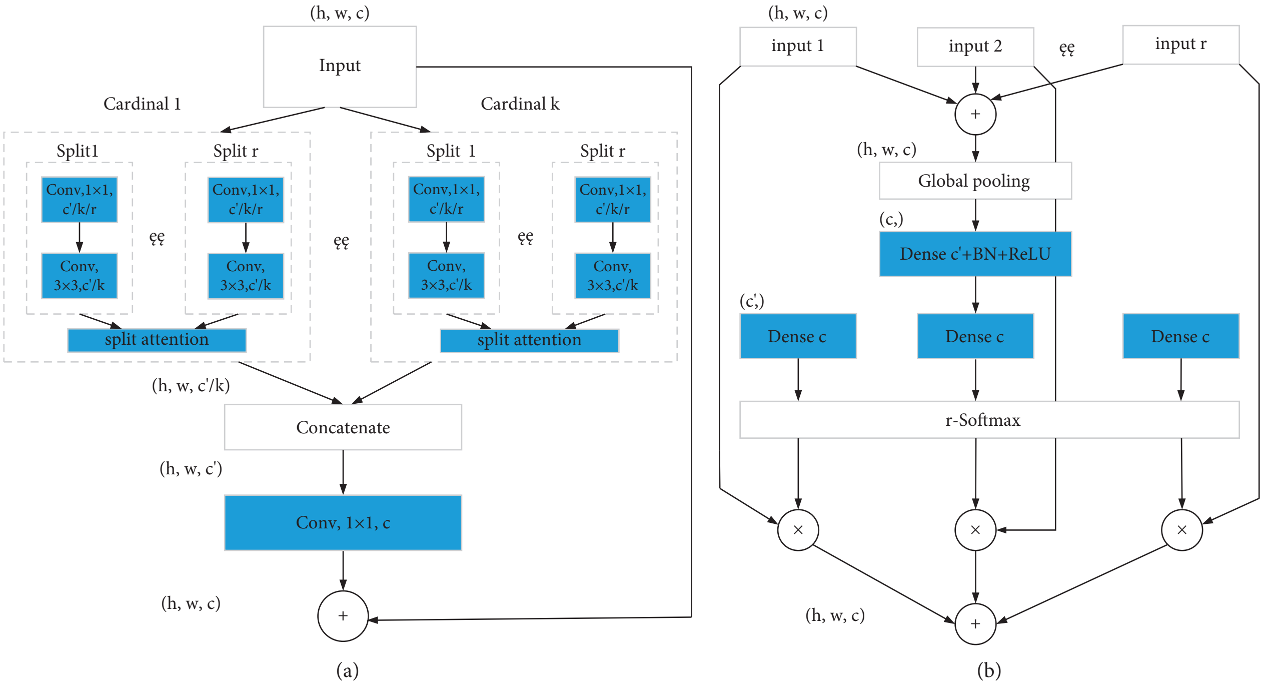


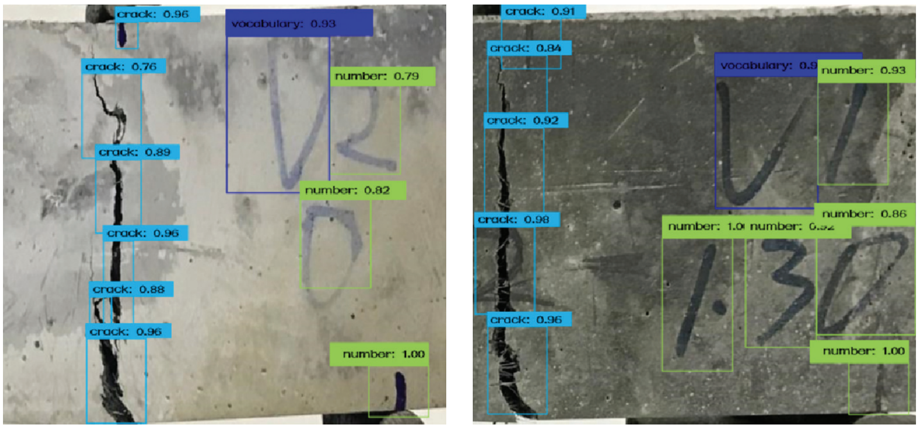
FIGURE 6: ResNeSt block and split attention module. (a) ResNeSt block. (b) Split attention.

samples predicted to be positive samples; and F_N represents the number of negative samples predicted by positive samples.

4. Calculation Results

4.1. Detection Results of YOLOv4. Figure 7 shows the calculation results based on the YOLOv4. The results show that the overall effect of YOLOv4 algorithm in crack detection is better, and the main reason for higher detection accuracy is that the image interference is low, and the object features are relatively simple. It can be seen from Figure 7(a) that the YOLOv4 model has carried out error detection on jamming objects. One is to detect the jamming items as cracks, and the other is to detect the jamming items as substitute numbers. The same error detection occurs in Figure 7(b), but the detection accuracy of other categories is high, which shows that the model has strong robustness.

Furthermore, the detection accuracy and average accuracy of each category are calculated, and the results are shown in Table 2.

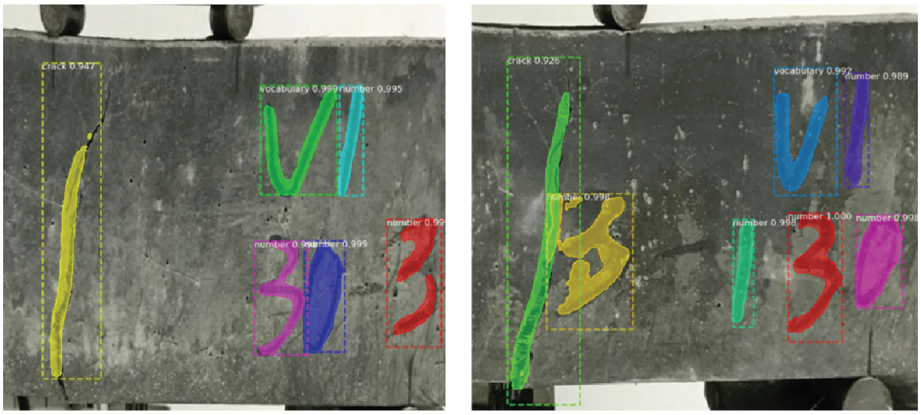


(a) (b)

Figure 7: YOLOv4 results.

Table 2: Detection results of YOLOv4.

Model	Average precision			F1 (%)
	Crack AP (%)	Number AP (%)	Vocabulary AP (%)	
YOLOv4	73.81	84.42	87.96	82.60



(a) (b)

Figure 8: Mask R-CNN results.

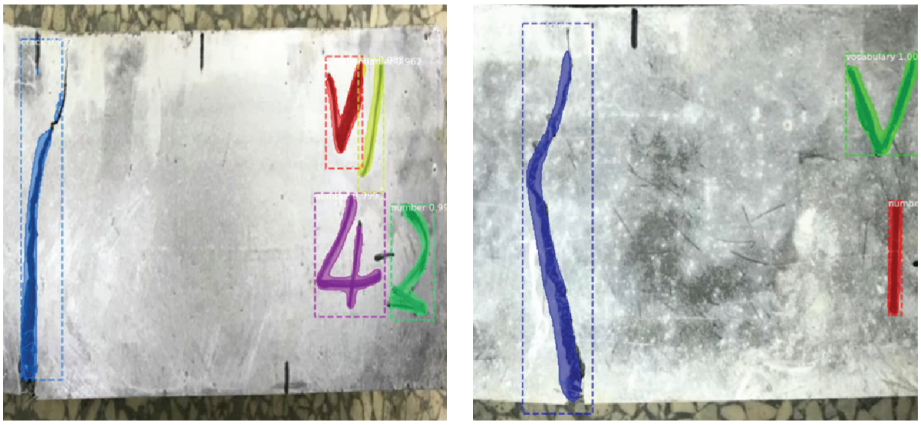
Table 3: Detection results of Mask R-CNN.

Model	Average precision			F1 (%)
	Crack AP (%)	Number AP (%)	Vocabulary AP (%)	
Mask R-CNN	84.32	91.26	95.73	90.44

4.2. Detection Results of Mask R-CNN. Figure 8 shows the calculation results based on Mask R-CNN. Figure 8 shows that the effect of fracture prediction is good, and the accuracy of model detection is still insufficient compared with the other two types. For example, it is difficult to detect and segment the two ends of the crack in the image, which is due to the strong background interference of the predicted image.

Furthermore, the detection accuracy and average accuracy of each category are calculated, and the results are shown in Table 3.

4.3. Detection Results of Improved Mask R-CNN. Figure 9 shows the calculation results based on the improved Mask R-CNN. As can be seen from Figure 9, the improved model



(a) (b)

Figure 9: The test results of improved Mask R-CNN crack detection model.

Table 4: Detection results of improved Mask R-CNN.

Model	Average precision Crack AP (%)	 Number AP (%)	 Vocabulary AP (%)	*F*1 (%)
Improved Mask R-CNN	92.57	97.63	98.08	96.09

can detect and identify cracks well, and the segmentation of cracks is also more accurate.

Furthermore, the detection accuracy and average accuracy of each category are calculated, and the results are shown in Table 4.

5. Conclusion

In order to realize the intellectualization of concrete crack detection and better prevent the occurrence of accidents, in this paper, a crack recognition model of steel fiber reinforced concrete is established based on computer vision and the deep learning method. Therefore, some conclusions are drawn as follows. (1) In this paper, the crack image is obtained through the steel fiber concrete experiment, and the crack database is expanded by using the deep learning data enhancement method. (2) Based on the network of YOLOv4 and Mask R-CNN, the crack recognition model of steel fiber reinforced concrete is established, and the average recognition accuracy is 82.60% and 90.44%, respectively. (3) Based on the traditional Mask R-CNN network, this paper proposes an improved Mask R-CNN network model, and its average recognition accuracy is 96.09%. However, the environment of concrete is very complex, such as shadows, stains, and so on, which will interfere with the accuracy of crack identification in actual engineering. Therefore, we will consider the crack identification of concrete in complex environment and further identify the length and width of cracks in future research.

Data Availability

The crack database data used to support the findings of this study have been deposited in the Baidu online disk repository (https://pan.baidu.com/s/1ozcIOnY4Yl6RzRrQ-IBXUg (password: 093r)).

Ethical Approval

Ethical review and approval were waived for this study because the institutions of the authors who participated in data collection do not require IRB review and approval.

Consent

Not applicable.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yang Ding, Hai-Qiang Yuan, and An-Ming She finished the model. Yang Ding wrote the original manuscript. Tong-Lin Yang and Zhong-Ping Wang supervised the study. Jing-Liang Dong, Yuan Pan, and Shuang-Xi Zhou contributed to manuscript writing. All the authors discussed the results.

Acknowledgments

This study was supported by the National Key R&D Program of China (grant nos. 2019YFC1906203 and 2016YFC0700807), Key R&D Project of Jiangxi Province (grant no. 20171BBG70078), National Natural Science Foundation of China (grant nos. 51108341, 52163034, 51662008, 51968022, and 51708220), and Opening Project of Key Laboratory of Soil and Water Loss Process and Control in Loess Plateau, Ministry of Water Resources (grant no. 201806).

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[2] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. Berg, "Dssd: deconvolutional single shot detector," arXiv preprint arXiv: 1701.06659, 2017.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[4] T. Aoki, A. Yamada, K. Aoyama et al., "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal Endoscopy*, vol. 89, no. 2, pp. 357 e2–363. e2, 2019.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection,"vol. 1, pp. 886–893, in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, Washington, DC, USA, June 2005.

[6] Y. Chen, J. Li, H. Xiao, J. Xiaojie, Y. Shuicheng, and F. Jiashi, "Dual path networks," arXiv preprint arXiv:1707.01629, 2017a.

[7] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Venice, Italy, October 2017.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.

[9] Y. Chen, C. Zhang, T. Qiao, J. Xiong, and B. Liu, "Ship detection in optical sensing images based on YOLOv5," in *Proceedings of the Twelfth International Conference on Graphics and Image Processing (ICGIP 2020)*, p. 117200E, November 2020.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Computer science*, arXiv: 1511.06434v2, 2015.

[11] Y. Yu, K. Zhang, L. Yang, and D. Zhang, "Fruit detection for strawberry harvesting robot in non-structural environment based on Mask R-CNN," *Computers and Electronics in Agriculture*, vol. 163, Article ID 104846, 2019a.

[12] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1097–1105, Long Beach, CA, USA, December 2012.

[14] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on computer vision*, pp. 1440–1448, Santiago, Chile, December 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 91–99, Montreal, QC, Canada, December 2015.

[16] T. Y. Lin, P. Dollár, R. Girshick, H. Kaiming, H. Bharath, and B. Serge, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Venice, Italy, October 2017.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[18] S. Du, P. Zhang, B. Zhang, and H. Xu, "Weak and occluded vehicle detection in complex infrared environment based on improved YOLOv4," *IEEE Access*, vol. 9, pp. 25671–25680, 2021.

[19] Y. Yu, C. Wang, X. Gu, and J. Li, "A novel deep learning-based method for damage identification of smart building structures," *Structural Health Monitoring*, vol. 18, no. 1, pp. 143–163, 2019b.

[20] J. Pang, H. Zhang, C. Feng, and L. Li, "Research on crack segmentation method of hydro-junction project based on target detection network," *KSCE Journal of Civil Engineering*, vol. 24, no. 9, pp. 2731–2741, 2020.

[21] Z. Yu, Y. Shen, and C. Shen, "A real-time detection approach for bridge cracks based on YOLOv4-FPM," *Automation in Construction*, vol. 122, Article ID 103514, 2021.

[22] S. Luan, C. Chen, B. Zhang, C. Xianbin, H. Jungong, and L. Jianzhuang, "Gabor convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018.

[23] Y. J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017.

[24] C. A. Ferreira, T. Melo, P. Sousa, and M. Meyer, "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2," in *Proceedings of the International Conference Image Analysis and Recognition*, pp. 763–770, Springer, Halifax, Canada, July 2018.

[25] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[26] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Venice, Italy, October 2017.

[27] Z. Zheng, P. Wang, W. Liu, L. Jinze, Y. Rongguang, and R. Dongwei, "Distance-IoU loss: faster and better learning for bounding box regression," *AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12993–13000, 2020.

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, October 2017.

[29] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: an improved faster R-CNN approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[30] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[31] H. Yan, H. Lu, M. Ye, K. Yan, Y. Xu, and Q. Jin, "Improved mask R-CNN for lung nodule segmentation," in *Proceedings of the 2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 137–141, Qingdao, China, August 2019.

[32] X. Shao, H. Zhu, D. Guo, R. Zheng, and J. Wei, "Research on detection of large coal blockage at the transfer point of belt conveyor based on improved mask R-CNN," *IOP Conference*

Series: Earth and Environmental Science, vol. 440, no. 5, Article ID 052028, 2020.

[33] L. Zuo, P. He, C. Zhang, and Z. Zhang, "A robust approach to reading recognition of pointer meters based on improved mask-RCNN," *Neurocomputing*, vol. 388, pp. 90–101, 2020.

[34] Y. Wang, J. Wu, and H. Li, "Human detection based on improved mask R-CNN," *Journal of Physics: Conference Series*, vol. 1575, Article ID 012067, 2020.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.