*Computers, Materials & Continua*
DOI: 10.32604/cmc.2022.028571
***Article***

Tech Science Press

# Wall Cracks Detection in Aerial Images Using Improved Mask R-CNN

**Wei Chen[1], Caoyang Chen[1,*], Mi Liu[1], Xuhong Zhou[2], Haozhi Tan[3] and Mingliang Zhang[4]**

[1]College of Civil Engineering, Changsha University of Science and Technology, Changsha, 410114, China
[2]College of Civil Engineering, Chongqing University, Chongqing, 730000, China
[3]Department of Civil and Environmental Engineering, University of Auckland, Auckland, 1010, New Zealand
[4]Hunan Construction Engineering Group Co. Ltd., Changsha, 410004, China
*Corresponding Author: Caoyang Chen. Email: 18574947083@163.com
Received: 12 February 2022; Accepted: 23 March 2022

**Abstract:** The present paper proposes a detection method for building exterior wall cracks since manual detection methods have high risk and low efficiency. The proposed method is based on Unmanned Aerial Vehicle (UAV) and computer vision technology. First, a crack dataset of 1920 images was established using UAV to collect the images of a residential building exterior wall under different lighting conditions. Second, the average crack detection precisions of different methods including the Single Shot MultiBox Detector, You Only Look Once v3, You Only Look Once v4, Faster Regional Convolutional Neural Network (R-CNN) and Mask R-CNN methods were compared. Then, the Mask R-CNN method with the best performance and average precision of 0.34 was selected. Finally, based on the characteristics of cracks, the utilization ratio of Mask R-CNN to the underlying features was improved so that the average precision of 0.9 was achieved. It was found that the positioning accuracy and mask coverage rate of the proposed Mask R-CNN method are greatly improved. Also, it will be shown that using UAV is safer than manual detection because manual parameter setting is not required. In addition, the proposed detection method is expected to greatly reduce the cost and risk of manual detection of building exterior wall cracks and realize the efficient identification and accurate labeling of building exterior wall cracks.

**Keywords:** Exterior wall cracks; object detection; mask R-CNN; DenseNet

## 1 Introduction

The building exterior wall cracks are mainly detected manually in China. The manual detection methods have the disadvantages of high risk, low efficiency, and high cost. Thus, the cracks may not be found in the initial stage, the maintenance and reinforcement of the building may be delayed, and the service life of a building may be affected. The development of Unmanned Aerial Vehicle (UAV) technology provides a solution to overcome these issues. A UAV can carry sensors to safely and efficiently collect samples at high altitudes with a lower cost than a manual detection method.

Note that this data collection method has been well applied in earthquake relief, agricultural plant protection, remote sensing, and other fields [1].

The intelligent methods of recognizing cracks are mainly divided into the image processing based method and the convolutional neural network (CNN) based method. Zhou et al. [2] proposed a crack recognition algorithm, which is suitable for concrete with a small surface color difference. They obtained the geometric feature of cracks through image enhancement, threshold segmentation, and morphological operation. Wang et al. [3] suggested a local texture processing algorithm based on subdivided images. This algorithm addresses the low accuracy of crack detection in complex scenes and weak light environment in tunnels. Wang et al. [4] presented a crack judgment model based on Radial Basis Function-Support Vector Machine algorithm, which can quantify the geometric features of cracks through digital image processing. Then, Wang et al. realized automatic screening of cracks by using a quantitative information training model. Talab et al. [5] proposed a concrete crack detection method based on the Sobel filtering and Otsu threshold segmentation algorithm. Pereira et al. [6] suggested the implementation scheme of the UAV crack detection system based on image detection. The UAV carries a raspberry Pi embedded with a particle filter to assess building facades in real time. The terminal processor receives the crack image transmitted wirelessly by raspberry Pi and uses the Sobel operator to obtain the crack data. It should be noted that all above studies on crack detection methods are based on image processing methods, in which crack geometric information is obtained by edge detection. However, the image processing methods have the disadvantage of artificial threshold setting, and the subjective influence is relatively large. Also, even though a detection method of adaptive threshold [7] and the method of locating cracks with CNN [8] have been proposed in some proposed in some studies, but the disadvantages of limited application scenario, the image noise issue, and manual setting threshold still exist in the crack detection method based on image processing. Thus, this method is unsuitable for the crack detection of exterior walls of buildings with variable backgrounds.

In recent years, the CNN method and its applications in various fields have been extensively studied. Note that this method can provide excellent results for detection in different applications, such as medical aided diagnosis [9–11] and handwritten character recognition [12,13]. Object detection algorithms can be divided into one-stage and two-stage algorithms. The one-stage detection algorithms exemplified by Single Shot MultiBox Detector (SSD) [14], You Only Look Once (YOLO) v3 [15], YOLOv4 [16] can directly predict the anchor boxes without generating proposed regions. Although the one-stage algorithms sacrifice a certain accuracy, they greatly improve the detection speed so that they are suitable for real-time detection. The two-stage detection algorithms are exemplified by Regional Convolutional Neural Network (R-CNN) [17], Faster R-CNN [18], Mask R-CNN [19]. These algorithms classify images twice and have higher accuracy in object detection, but lower detection rate and higher processor requirements. Thus, a two-stage network is more suitable for crack detection when the intelligence requirements, generalization, and accuracy of crack detection are considered.

Since there are few studies on artificial crack detection methods, this paper intends to use UAV and computer vision technology to improve the crack detection methods. As shown in Fig. 1, first, this paper used UAV to collect images of external walls of buildings and establish crack datasets. Second, the crack image features were analyzed, and the crack detection effects of different target detection networks were compared, among which Mask R-CNN has the best detection effect on the cracks. Finally, an improved Mask R-CNN was proposed, which greatly improves the crack detection accuracy and Mask fitting effect. This paper may provide a theoretical and technical basis for intelligent and precise crack detection of building exterior walls.

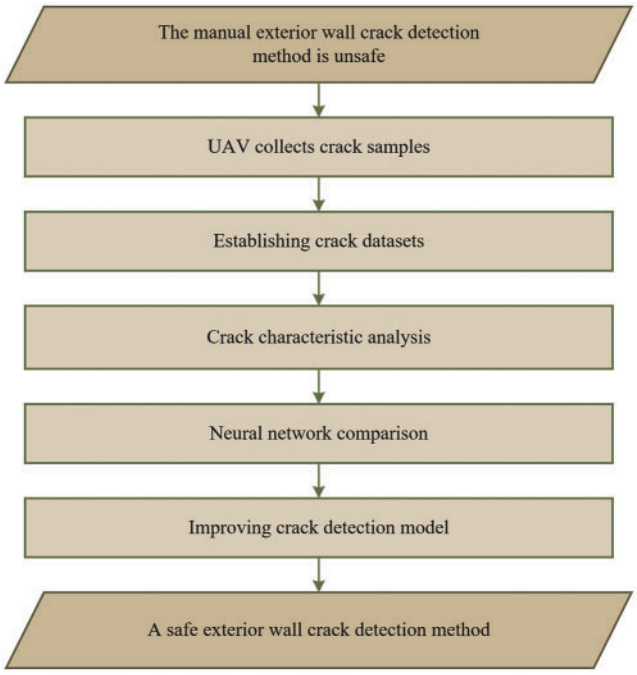



**Figure 1:** Research framework

## 2 Crack Detection Methods

### *2.1 Crack Characteristics Analysis*

Since cracks belong to topological structures, they are easily lost in the feature map when they are pooled. Also, the detection effect of the basic target detection network on cracks is not ideal. In this paper, images of different crack types were randomly selected. The Mask images were transformed into binary images. The total number of pixels in the Masks of different cracks and the total number of pixels in the boundary box were calculated, as shown in Tab. 1.

Tab. 1 shows that the Masks of transverse crack and vertical crack occupy a relatively high proportion, 35.7% and 74.6%, respectively, in the boundary boxes. However, the proportion of boundary boxes in the whole image is relatively low, only 11.8% and 4.5%. The pixel values in the boundary boxes of transverse crack and vertical crack indicate that the cracks are small objects. However, the length or width of the boundary boxes of these two crack types almost cross the whole image. The ratios of Masks of diagonal crack and irregular crack to the ground truth boxes are very low, accounting for only 6.3% and 5.7%, respectively. Most images in the ground truth box are background images, which cause the model to incorrectly learn background features and fail to learn crack features. The proportion of the ground truth box of diagonal crack and irregular crack in the image is very high, reaching 46.5% and 99.5%, respectively, which are not small objects. Therefore, cracks have the characteristics of small objects but are not small objects, which makes crack detection more complex than other objects.

**Table 1:** Comparison of pixel values of different crack images

| Class of crack | Crack image | Mask pixel | Ground truth box pixel | Image pixel | Mask/ Ground truth box | Ground truth box/Image |
|---|---|---|---|---|---|---|
| Transverse crack |  | 20271 | 56729 | 480000 | 0.357 | 0.118 |
| Vertical crack |  | 16049 | 21516 | 480000 | 0.746 | 0.045 |
| Diagonal crack |  | 14033 | 223155 | 480000 | 0.063 | 0.465 |
| Irregular crack |  | 27017 | 477802 | 480000 | 0.057 | 0.995 |

### *2.2 Comparison of Object Detection Network*

Object detection and image segmentation datasets of cracks were made, and SSD, YOLOv3, YOLOv4, Faster R-CNN and Mask R-CNN were used for training. The detection effects and parameters of these models were compared, and the results are shown in Tab. 2.

**Table 2:** Comparison of cracks detection effects of object detection networks

| Model | Object detection | Instance segmentation | Weights MB | AP |
|---|---|---|---|---|
| SSD | √ | × | 90.6 | 0.083 |
| YOLOv3 | √ | × | 59.9 | 0.109 |
| YOLOv4 | √ | × | 244.3 | 0.136 |
| Faster R-CNN | √ | × | 108.1 | 0.064 |
| Mask R-CNN | √ | √ | 244.0 | 0.340 |

Tab. 2 shows that Faster R-CNN has the worst crack detection performance, with Average Precision (AP) of only 0.064. However, Mask R-CNN has the best crack detection effect with AP of 0.340 that 0.276 higher than Faster R-CNN's, and 0.204 higher than YOLOv4. The particularity of cracks leads to poor detection effect of common object detection networks on cracks, while Mask R-CNN not only has the function of classification and regression of targets but also adds the prediction branch of Mask, which has the function of instance segmentation. The loss function of Mask can make the network better fit and it focuses on the characteristics of cracks to improve the detection

accuracy. Therefore, compared with other target detection networks, Mask R-CNN is more suitable for crack detection.

In this paper, multi-wing UAV was used to collect the crack image of building exterior wall and established the crack datasets. Through the analysis and comparison of different target detection networks, Mask R-CNN network was selected as the basic network of crack detection, and the Mask R-CNN network was improved according to the characteristics of the variable backgrounds of the building exterior wall cracks.

### *2.3 Mask R-CNN*

Mask R-CNN is the object detection network proposed by He et al. [19]. Mask R-CNN adds a Mask branch based on Faster R-CNN [18] and adds the instance segmentation function based on object detection. In addition, the Region of Interest (ROI) Align layer is used in Mask R-CNN to replace the ROI Pooling layer in Faster R-CNN, which greatly improves the algorithm performance and poor detection effect of Faster R-CNN on small objects. Fig. 2 shows the overall network structure of Mask R-CNN.

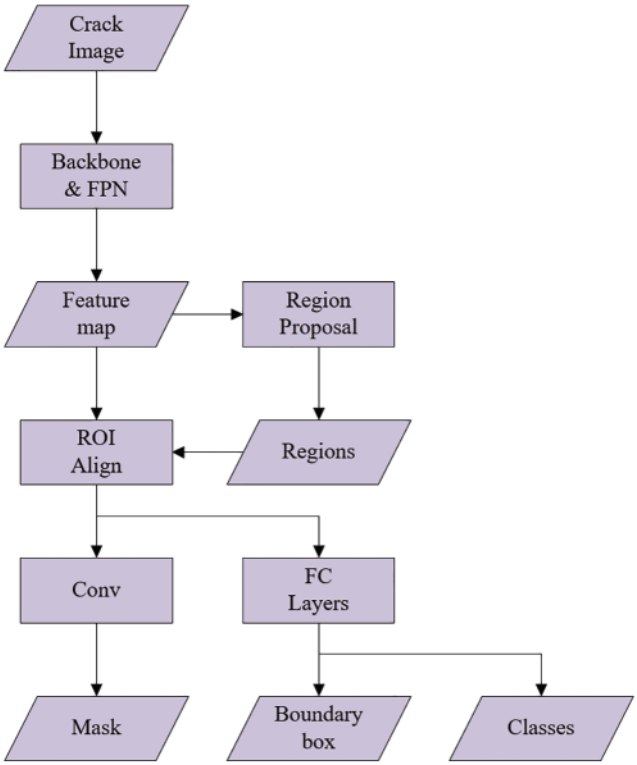



**Figure 2:** Network structure of mask R-CNN

Some researchers have used Mask R-CNN to detect cracks and achieved good results. Attard et al. [20] proposed a method using Mask R-CNN to detect concrete cracks, and the accuracy was reached 93.94%. However, they used a relatively simple crack image background. Thus, their method was only applicable to concrete cracks. Dong et al. [21] presented an application platform based on

Mask R-CNN to automatically extract the number, area, length, and average width of cracks in asphalt pavement. Zhang et al. [22] added a branch in Mask R-CNN to guide loss function to improve the prediction quality of Mask. These researches show that Mask R-CNN is more suitable for crack detection than other object detection networks.

He et al. [23] proposed a convolutional neural network called ResNet, which is also the backbone feature network of Mask R-CNN. Note that increasing the depth of the neural network can improve the effect of the model, but a too deep network will cause network degradation, which results in the decline of the fitting effect. The use of residual blocks in ResNet eliminates the network degradation problem. As shown in Fig. 3, the residual unit uses a short-circuit mechanism to change the network from direct learning mapping to learning residuals of fitting and expected mapping, so that the neural network has a better effect.

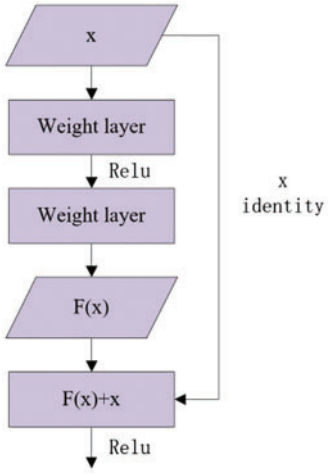



**Figure 3:** Residual block

Note that ResNet enables Mask R-CNN to improve detection accuracy by increasing network depth. However, a real test can show that ResNet cannot effectively retain the shallow crack characteristics in the convolution process, which causes the model to fail to locate the crack position. For example, Fig. 4 shows that when the crack texture is shallow and close to the background pixel, Mask R-CNN cannot locate the crack, resulting in model missed detection. In order to address this deficiency, a network that can better preserve the shallow crack characteristics can be used while the ResNet residual idea is retained and used as the backbone network of Mask R-CNN.

### *2.4 Width and Depth of Network*

The depth and width of a neural network are important factors affecting the performance of the neural network. A deep neural network obtains high-level semantic information of objects by extracting features layer by layer. Theoretically, the deeper the neural network, the better the model performance. It is worth mentioning that the width of a neural network is the number of channels. Zagoruyko et al. [24] proposed a wide ResNet. It should also be noted that by increasing the number of channels, the shallow ResNet achieves the detection accuracy of the deep ResNet and improves the network speed.

Delalleau et al. [25] and Eldan et al. [26] experimentally proved that the width of a shallow neural network needs to be increased exponentially to achieve the effect of a deep neural network.

Narrow and deep neural networks tend to have better performance and stronger generalization ability than wide and shallow networks. Also, a wide and shallow neural network is more consistent with the characteristics of parallel computing, faster computing speed, and easier training. Therefore, an appropriate model is selected by comparing the model detection accuracy under different widths and depths.

**Figure 4:** Shallow crack

## 3 Data Processing and Model Improvement

### *3.1 Image Collection and Processing*

#### *3.1.1 Image Collection*

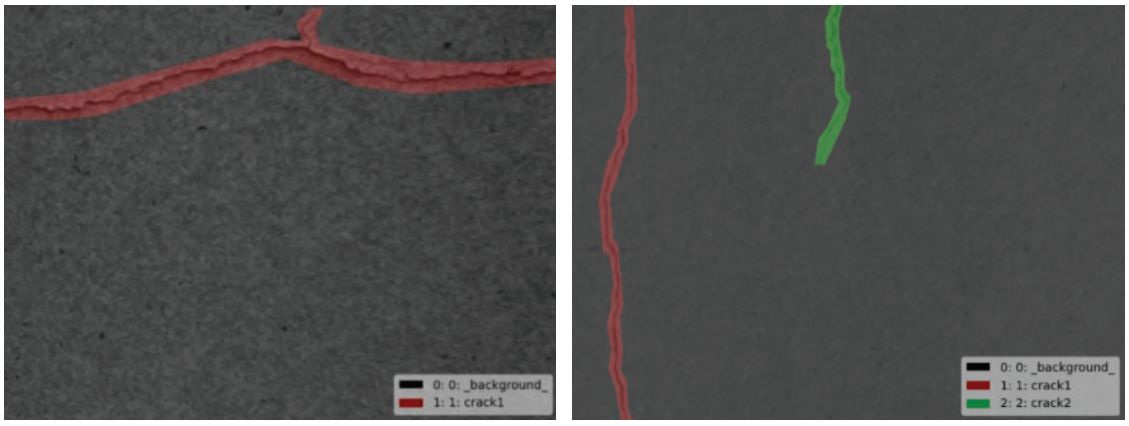
Mask R-CNN has the supervised training method, in which image annotation tools are used to mark the real position of the object in the picture. Note that the crack images in this paper come from a direct shooting of a building's exterior wall by DJI UAV Mavic Air 2. As shown in Tab. 3, the collected crack images include a variety of backgrounds of building facades, which enrich the crack images in different scenes, and enhance the generalization ability of detection models. For Mask R-CNN, the input crack image size is required to be an integer multiple of $2^6$. Considering the problems of Graphics Processing Unit (GPU) memory and detection efficiency, the input image size was set to $512 \times 512$ pixels, and the collected exterior wall image slices were divided into $512 \times 512$ pixels and $1024 \times 1024$ pixels.

**Table 3:** Image of cracks in multiple backgrounds

| Class of background | Brick joints | Window | Pipe blocked | Shadow | Strong light |
|---|---|---|---|---|---|
| Crack image |  |  |  |  |  |

### *3.1.2 Image Annotation*

The images containing cracks after segmentation were selected, and the image annotation tool Labelme was used to annotate the cracks at the pixel level. The annotation results are shown in Fig. 5. The colored part is the positive sample area, and the gray part is the negative sample area.



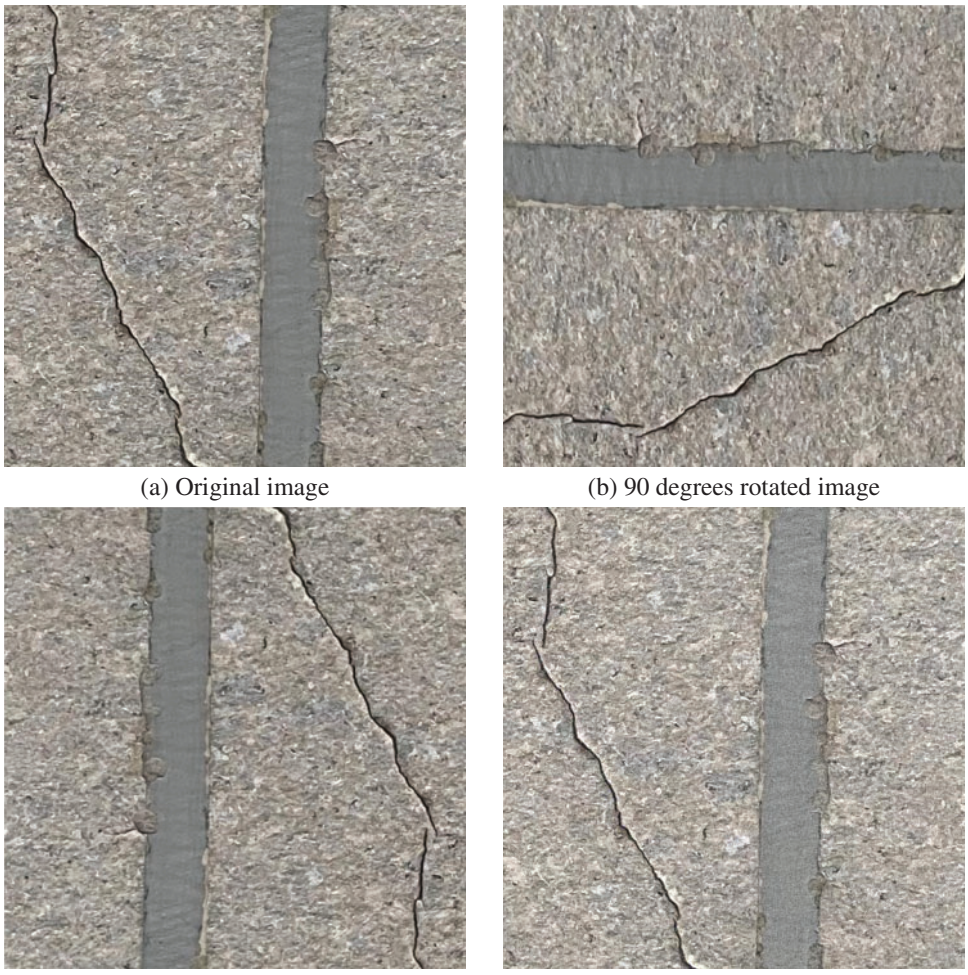

(a) Single Crack Labeling Image (b) Multi-Crack Labeling Image

**Figure 5:** Crack labeling images

Fig. 5 shows that the negative sample area of the crack image is much larger than the positive sample area. A total of 1920 images were annotated in this experiment, of which 80% were training sets. Images in the dataset can be of any size. Before training, the image size can be adjusted to $512 \times 512$ pixels without distortion through the program.

### *3.1.3 Data Enhancement*

The neural network is trained by fitting the commonality of all data in the data set. The deep neural network has a large number of parameters. When the training set is small, the deep neural network will fit the characteristics of all data in the training set, making the network only suitable for the training set and poor generalization. Due to the high cost of cracks collection and annotation, this paper adopts the methods of rotating 90 degrees, rotating 180 degrees, and adding Gaussian noise to enhance the data set. The value of mean and variance in adding Gaussian noise are zero and one respectively. Fig. 6 shows the enhanced results, after data enhancement, a total of 7680 images were obtained for training and validating.



(a) Original image (b) 90 degrees rotated image

(c) 180 degrees rotated image (d) Adding Gaussian noise

**Figure 6:** Data enhancement

### *3.2 Model Improvement*

#### *3.2.1 DenseNet*

Cracks belong to topological targets, which take too small proportion of the image and are long and thin. Therefore, the network is easy to learn background features as crack features. When the image size is compressed, the features of cracks are easy to be lost. Thus, in the present paper, ResNet, the backbone network in Mask R-CNN, is replaced by DenseNet proposed by Huang et al. [27]. It is worth mentioning that DenseNet, similar to ResNet, short-circuits the connections without network degradation. Also, DenseNet retains a large number of underlying features through dense connections between convolution blocks. However, residual joins of DenseNet and ResNet are stacking channels and accumulating channels, respectively. As a result, DenseNet preserves crack characteristics for repeated learning more effectively. DenseNet is composed of Dense Blocks and Transition Blocks. Dense Block is a unique module of DenseNet, as shown in Fig. 7. All convolution layers are densely connected in this module, and the size of feature layer is not changed while the number of channels is changed. Transition Block is a partition of Dense Block that integrates stacked features and pools

them to reduce their sizes. In addition, DenseNet retains more information about cracks than ResNet. Therefore, DenseNet is adjusted to change the number of connections in each Dense Block to make it suitable for Mask R-CNN.

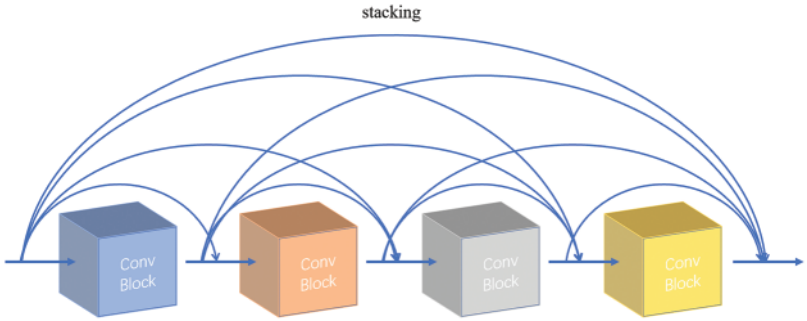



**Figure 7:** Dense block

### *3.2.2 Batch Normalization*

Note that Dropout is used in ResNet to prevent gradient disappearance [28]. However, Batch Normalization (BN) [29] that is used in DenseNet not only prevents gradient disappearance but also enhances the generalization ability of the model to some extent. Due to the limited GPU computing capacity and the increase of training sets number, the neural network training is often unable to be completed in a group. Therefore, the neural network generally adopts the batch gradient descent method to optimize parameters. By setting batch size, data is divided into several batches to reduce the amount of calculation each time. As part of the Batch Normalization process, the mean value ($\mu_B$) and variance ($\sigma_B^2$) of the Mini-batch obtained by Eqs. (1) and (2) are substituted into Eq. (3) to normalize each element of the Batch.

$$\mu_B = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

where $n$ is the batch size, $x_i$ is a sample in the batch, and $\mu_B$ is the mean value.

$$\sigma_B^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_B)^2 \tag{2}$$

$$x_i' = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{3}$$

where $x_i'$ is the result of sample normalization, $\varepsilon$ is the default minimum to prevent the denominator from being zero.

Batch Normalization introduces learnable parameters $\gamma$ and $\beta$ for feature reconstruction to prevent Normalization from destroying the feature distribution learned by the previous layer. As shown in Eq. (4), the network can learn the distribution characteristics by itself and optimize the network without reducing the network accuracy.

$$y_i = \gamma x_i' + \beta \tag{4}$$

where $y_i$ is the sample's output result.

The BN layer can optimize the network, enhance network generalization ability, and make training faster. Meanwhile, the output of samples is determined by all samples in Mini-batch, which is composed of different samples with different outputs. It should also be noted that the present paper compares the detection effects of Dropout and Batch Normalization.

## 4 Training Results and Analysis

The test platform was GPU NVIDIA GeForce RTX 2080Ti, and the test environment was Python 3.7, Tensorflow-GPU 2.0.0, Keras 2.3.1, Compute Unified Device Architecture 10.0, and cudNN 7.6. The image size was $512 \times 512$ pixels. The data in the experiment were obtained after training 100 epochs for each model.

### *4.1 Evaluation Indicators*

In object detection, the Intersection-over-Union (IoU) between the prediction boxes and ground truth boxes is used to determine whether the network can detect the object properly. When IoU is greater than the threshold, it is recorded as True Positive (TP), while when IoU is less than or equal to the threshold, it is recorded as False Positive (FP). Also, when IoU is not detected, it is recorded as False Negative (FN). Note that Precision and Recall can be obtained by Eqs. (5) and (6).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

Since Precision and Recall cannot comprehensively evaluate the model performance, AP is introduced to evaluate the accuracy of classification. As shown in Fig. 8, a Precision-Recall curve can be drawn with Recall and the corresponding Precision of the test set of Mask-R-CNN. Note that the area under the curve is the AP value of this classification.

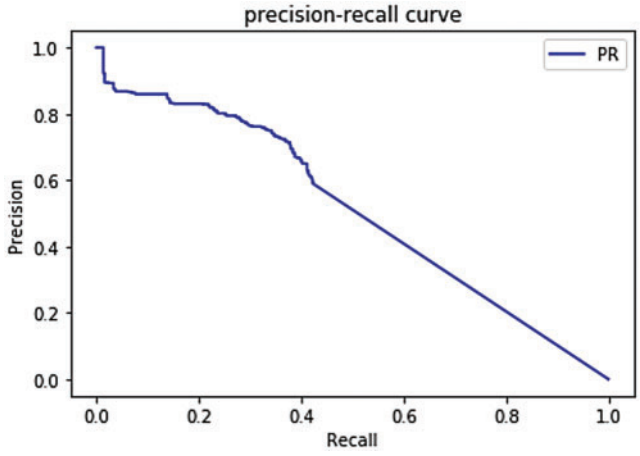



**Figure 8:** Precision-recall curve of mask R-CNN

### *4.2 Analysis of Training Results*

#### *4.2.1 Model Performance Comparison*

Compared with the mainstream object detection networks, the original Mask R-CNN has a better crack detection effect. However, serious missing phenomena still exist in detecting the original Mask R-CNN on the dataset collected by UAV. The main reason for missing detection is that the cracks have few pixels, and the shallow features of the cracks are not effectively utilized, which leads to the failure of the model to locate the cracks. Therefore, the present paper improves the model based on crack characteristics by replacing ResNet with DenseNet and compares the detection performance of DenseNet with different crack depths. Also, the utilization rate of crack features is improved by using the shallow features repeatedly.

The test results, provided in Tab. 4, show that DenseNet has the best performance when the growth rate is 32. Also, when the growth rate is 16, the network depth is too deep, and the gradient disappears, resulting in a low AP value. In addition, when the growth rate is 64, the stack times of DenseNet are greatly reduced, but the model performance is also greatly reduced. Therefore, in this paper, the growth rate of the improved Mask R-CNN was set to 32. In addition, when the BN layer in DenseNet is replaced by the Dropout layer, the AP value of the model decreases significantly, proving that the BN layer has a better effect on avoiding model overfitting. It should also be noted that the detection effect of the original Mask R-CNN network on cracks was not ideal, and the false detection rate was high, resulting in a low AP value. This may be because the ResNet does not effectively retain the characteristic information of cracks. Note that when the cracks are thin, Mask R-CNN cannot identify the cracks. Also, when DenseNet was used as the backbone network, the number of parameters increases by 84.4% due to the deepening of network depth, but the network performance was greatly improved. Also, Precision increases by 24.5%, Recall increases by 89.6%, and AP increases by 163.93%. This proves that the improved DenseNet can effectively retain the characteristics of cracks, has an excellent detection performance of cracks, and its accuracy can meet the requirements.

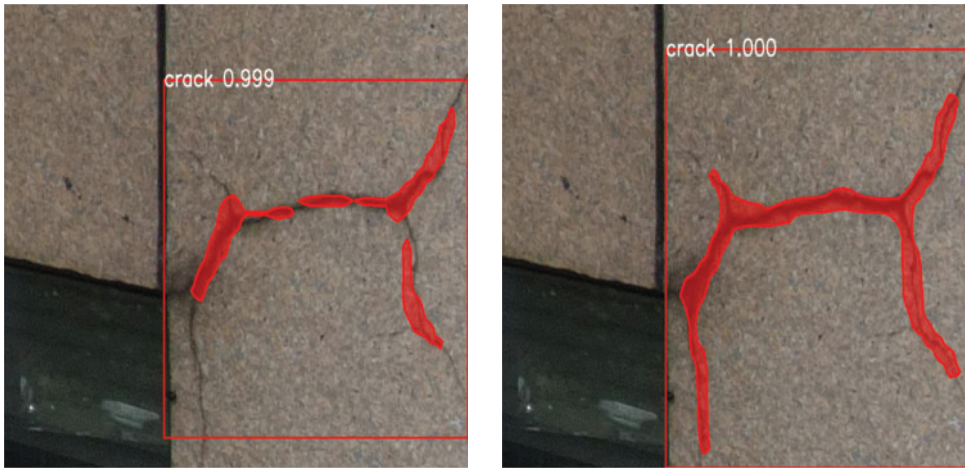
**Table 4:** Comparison of object detection networks performances

| Backbone | Preventing overfit | Growth rate | Depth | Precision | Recall | AP |
|---|---|---|---|---|---|---|
| Resnet | Dropout | None | Middle | 0.770 | 0.259 | 0.341 |
| DenseNet | BN | 16 | Deep | 0.647 | 0.152 | 0.164 |
| DenseNet | BN | 32 | Middle | 0.959 | 0.491 | 0.900 |
| DenseNet | BN | 64 | Shallow | 0.569 | 0.167 | 0.169 |
| DenseNet | Dropout | 32 | Middle | 0.616 | 0.441 | 0.465 |

#### *4.2.2 Comparative Analysis of Crack Detection Results*

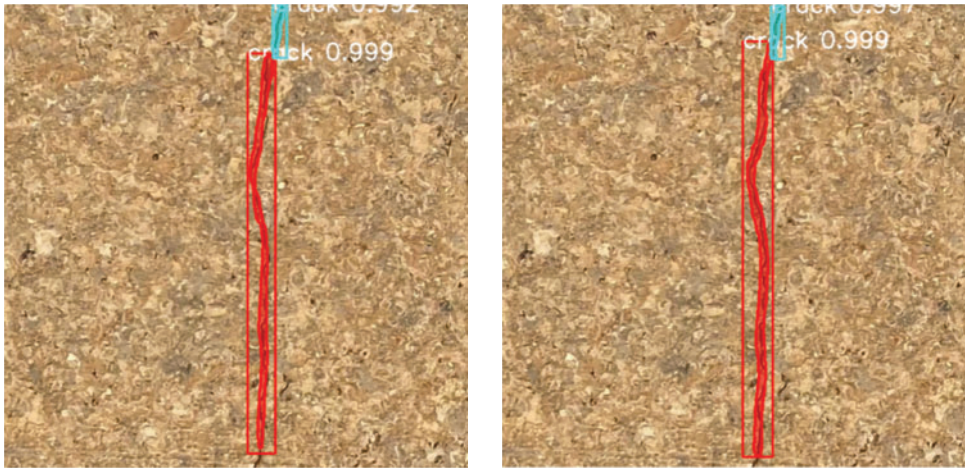
The improved Mask R-CNN effectively increases the crack detection accuracy of the model, removes the defect of the original Mask R-CNN, and improves the quality of Mask fitting. Fig. 9a shows that the original Mask R-CNN model has many missed detection phenomena in window corner cracks detection. However, the positioning of the Prediction box is relatively accurate, even though some cracks are still outside the target frame, and the Mask fitting quality is poor. The improved Mask R-CNN not only improves the positioning accuracy of the Prediction box, but also greatly improves the quality of Mask fitting, and the Mask covers most of the cracks. Fig. 10 compares the crack detection

effects under intense light. As shown, the original Mask R-CNN has a good detection effect under intense light, but masks do not cover some cracks in the target box. However, the improved Mask R-CNN increases the fitting quality of masks. Fig. 11 shows the detection effect of cracks truncated by shadows under intense light. Also, Fig. 11a shows that, near the shadow, the original Mask R-CNN has a poor fitting of cracks and a high rate of missed detection. However, the improved Mask R-CNN can accurately detect the cracks around the shadow. The image of Fig. 12 is a shot in a dark environment, resulting in cracks with a similar color to the background. In addition, cracks across dark brickwork joints are difficult to detect. The original Mask R-CNN had a serious problem of missing detection in the environment with similar background color, while the improved Mask R-CNN completely detected cracks, and the detection results were not cut by brickwork joint.



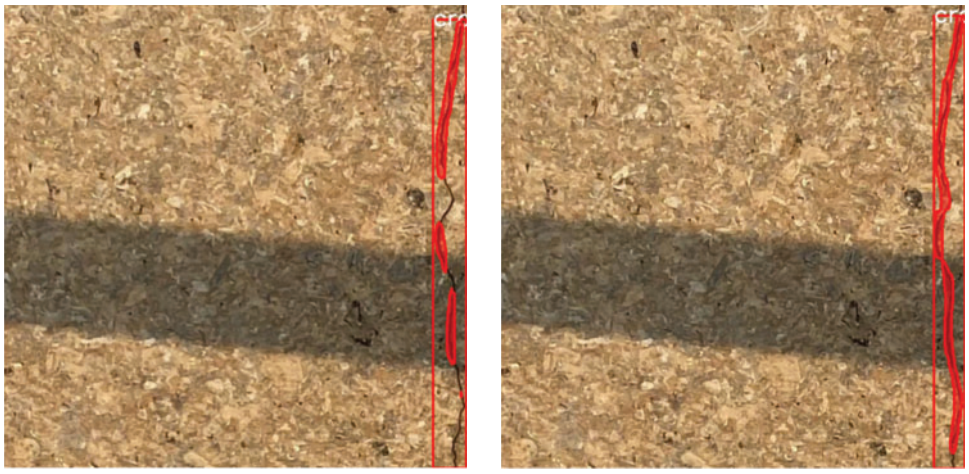

(a) Original Mask R-CNN (b) Improved Mask R-CNN

**Figure 9:** Detection of window corner crack




(a) Original Mask R-CNN (b) Improved Mask R-CNN

**Figure 10:** Detection of crack under intense light



(a) Original Mask R-CNN (b) Improved Mask R-CNN

**Figure 11:** Detection of crack under shadow

(a) Original Mask R-CNN (b) Improved Mask R-CNN

**Figure 12:** Detection of crack across brick joint

The results show that the improved Mask R-CNN model has strong robustness and improves the detection accuracy and Mask quality compared with the original model. Also, it can effectively detect the crack images collected by UAV in different environments. In addition, UAV can collect high-quality crack images in different environments, and can detect cracks with high detection efficiency with no preprocessing.

## 5 Conclusions

Manual detection methods for building exterior wall cracks have high risk and low efficiency. Thus, this paper applied the UAV technology and computer vision technology to propose an intelligent recognition method of building exterior cracks based on UAV and improved Mask R-CNN. For this purpose, a crack dataset of 1920 images was established, and the images of a residential building

exterior wall under different lighting conditions were collected. Also, the average crack detection precisions of different methods including the SSD, YOLOv3, YOLOv4, Faster R-CNN, and Mask R-CNN methods were compared, and the Mask R-CNN method with the best performance and average precision of 0.34 was selected. The following conclusions can be drawn from the results.

(1) The quality of crack image acquisition shows that UAV can collect high-resolution images of cracks in dark, intense light, and other environments, and can perform the collection task of crack detection properly. The crack database can be established by using UAV to collect crack images under different building backgrounds and to provide supporting data for the development of crack detection technology.
(2) By improving the utilization rate of shallow crack features, the improved Mask R-CNN greatly increases the crack detection accuracy and Mask quality of the model. Note that the AP value increases from 0.341 to 0.900. The experimental results show that the improved Mask R-CNN has strong robustness and can accurately identify the cracks in images under different environments, meeting the requirements of crack detection accuracy of building external walls.
(3) In the training stage, the model proposed in this paper can learn parameters by itself without setting parameters and thresholds manually. In the use stage, the crack image can be used for crack detection without preprocessing, and the detection process does not need manual intervention, which is highly intelligent. It was found that the proposed method greatly reduces the cost and risk of manual inspection of building exterior wall cracks and realizes the efficient identification and labeling statistics of building exterior wall cracks.

**Acknowledgement:** The authors acknowledge the support of Changsha University of Science and Technology and the support of National Natural Science Fund of China.

**Funding Statement:** This work was supported in part by the National Natural Science Foundation of China under Grant 51408063, author W. C, http://www.nsfc.gov.cn/; in part by the Outstanding Youth Scholars of the Department of Hunan Provincial under Grant 20B031, author W. C, http://kxjsc.gov.hnedu.cn/.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] Y. Q. Bao and H. Li, "Artificial intelligence for civil engineering," *China Civil Engineering Journal*, vol. 52, no. 5, pp. 1–11, 2019. https://doi.org/10.15951/j.tmgcxb.2019.05.001.
[2] Y. Zhou and T. Liu, "Computer vision-based crack detection and measurement on concrete structure," *Journal of Tongji University(Natural Science)*, vol. 47, no. 9, pp. 1277–1285, 2019.
[3] Y. D. Wang, L. Q. Zhu, H. M. Shi, E. Q. Fang and Z. L. Yang, "Vision detection of tunnel cracks based on local image texture calculation," *Journal of the China Railway Society*, vol. 40, no. 2, pp. 82–90, 2018.
[4] R. Wang and T. Y. Qi, "Study on crack characteristics based on machine vision detection," *China Civil Engineering Journal*, vol. 49, no. 7, pp. 123–128, 2016. https://doi.org/10.15951/j.tmgcxb.2016.07.012.
[5] A. M. A. Talab, Z. C. Huang, F. Xi and H. M. Liu, "Detection crack in image using Otsu method and multiple filtering in image processing techniques," *Optik*, vol. 127, no. 3, pp. 1030–1033, 2016. https://doi.org/10.1016/j.ijleo.2015.09.147.
[6] F. C. Pereira and C. E. Pereira, "Embedded image processing systems for automatic recognition of cracks using UAVs," *IFAC PapersOnLine*, vol. 48, no. 10, pp. 16–21, 2015. https://doi.org/10.1016/j.ifacol.2015.08.101.

[7] N. Yang, C. Zhang and T. H. Li, "Design of crack monitoring system for Chinese ancient wooden buildings based om UAV and CV," *Engineering Mechanics*, vol. 38, no. 03, pp. 27–39, 2021.
[8] X. J. Han and Z. C. Zhao, "Structural surface crack detection method based on computer vision technology," *Journal of Building Structures*, vol. 39, no. S1, pp. 418–427, 2018. https://doi.org/10.14006/j.jzjgxb.2018.S1.055.
[9] Y. Xue, B. Onzo, R. F. Mansour and S. Su, "Deep convolutional neural network approach for covid-19 detection," *Computer Systems Science and Engineering*, vol. 42, no. 1, pp. 201–211, 2022.
[10] R. Rajakumari and L. Kalaivani, "Breast cancer detection and classification using deep cnn techniques," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1089–1107, 2022.
[11] T. Jeslin and J. A. Linsely, "Agwo-cnn classification for computer-assisted diagnosis of brain tumors," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 171–182, 2022.
[12] S. I. Saleem and A. M. Abdulazeez, "Hybrid trainable system for writer identification of arabic handwriting," *Computers, Materials & Continua*, vol. 68, no. 3, pp. 3353–3372, 2021.
[13] Y. Xue, Y. Tong, Z. Yuan, S. Su, A. Slowik *et al.,* "Handwritten character recognition based on improved convolutional neural network," *Intelligent Automation & Soft Computing*, vol. 29, no. 2, pp. 497–509, 2021.
[14] W. Liu, A. Dragomir, E. Dumitru, S. Christian, R. Scott *et al.,* "SSD: Single shot multibox detector," ArXiv preprint, 2018. https://arxiv.org/abs/1512.02325.
[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," ArXiv preprint, 2018. http://arxiv.org/abs/1804.02767.
[16] A. Bochkovskiy, C. -Y. Wang and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," ArXiv preprint, 2020. http://arxiv.org/abs/2004.10934.
[17] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," ArXiv preprint, 2014. http://arxiv.org/abs/1311.2524.
[18] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," ArXiv preprint, 2016. http://arxiv.org/abs/1506.01497.
[19] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," ArXiv preprint, 2018. http://arxiv.org/abs/1703.06870.
[20] L. Attard, C. J. Debono, G. Valentino, M. Di Castro, A. Masi *et al.,* "Automatic crack detection using mask R-CNN," in *2019 11th International Symposium on Image and Signal Processing and Analysis*, Xiamen, China, pp. 152–157, 2019.
[21] J. X. Dong, J. H. Liu, N. N. Wang, H. Y. Fang, J. P. Zhang *et al.,* "Intelligent segmentation and measurement model for asphalt road cracks based on modified mask R-CNN algorithm," *Computer Modeling in Engineering & Sciences*, vol. 128, no. 2, pp. 541–564, 2021. https://doi.org/10.32604/cmes.2021.015875.
[22] Y. F. Zhang, J. F. Wang, B. Chen, T. Feng and Z. Y. Chen, "Pavement crack detection algorithm based on improved mask R-CNN," *Journal of Computer Applications*, vol. 40, no. S2, pp. 162–165, 2020.
[23] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," ArXiv preprint, 2015. http:/arxiv.org/abs/1512.03385.
[24] S. Zagoruyko and N. Komodakis, "Wide residual networks," ArXiv preprint, 2017. http://arxiv.org/abs/1605.07146.
[25] O. Delalleau and Y. Bengio, "Shallow *vs.* deep sum-product networks," *Advances in Neural Information Processing Systems*, vol. 24, pp. 666–674, 2011.
[26] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," ArXiv preprint, 2016. http://arxiv.org/abs/1512.03965.
[27] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "Densely connected convolutional networks," ArXiv preprint, 2018. http://arxiv.org/abs/1608.06993.
[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov *et al.,* "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," ArXiv preprint, 2015. http://arxiv.org/abs/1502.03167.