The recognition and detection of cracks in historical landscape buildings through deep learning technology

S. Wang [1], D.H. Zhu [1]
[1] College of Urban Design, Shanghai Art & Design Academy, Shanghai, 201808, China

Abstract

This paper added the spatial attention fusion module with an attention mechanism to the You Only Look Once version 5 algorithm to improve its building crack detection performance. Simulation experiments were conducted. Ablation experiments were conducted on the improved model to verify the role of the spatial attention fusion module module. It was then compared with three target detection algorithms: Faster-regional-based convolutional neural network, You Only Look Once version 4, and You Only Look Once version 6. It was found that after the introduction of the self-attentive feature map module, the precision of the improved You Only Look Once version 5 on the test set increased from 0.852 (the traditional You Only Look Once version 5) to 0.989, the recall rate increased from 0.843 to 0.988, the F1 value increased from 0.847 to 0.988, the mean average precision increased from 0.789 to 0.987, the intersection over union increased from 0.841 to 0.986, and the increase in each indicator was approximately 14.0 – 19.8 %. In comparisons with other algorithms, the mean average precision of the improved You Only Look Once version 5 was 0.253, 0.189, and 0.158 higher than Faster-regional-based convolutional neural network (0.734), You Only Look Once version 4 (0.798) and You Only Look Once version 6 (0.829) respectively. Its intersection over union was 0.245, 0.174, and 0.138 higher, respectively. These improvements mean that the model is more reliable in both crack identification and location: on the one hand, higher precision and recall rates can reduce missed detections and false detections, and lower the costs of manual rechecks and secondary inspections; on the other hand, a higher intersection over union indicates a more accurate fit of the crack location and range, which is conducive to quantitative analysis of parameters such as crack length and width, providing a more reliable basis for the preventive protection and scientific restoration of historical landscape buildings, and thus bringing practical benefits to heritage protection, tourism safety, and public life and property safety.

Keywords: building crack, deep learning, crack identification, You Only Look Once version 5.

Citation: Wang S, Zhu DH. The recognition and detection of cracks in historical landscape buildings through deep learning technology. Computer Optics. 2026; 50(5): 1864. doi: 10.18287/COJ1864.

Introduction

Historical landscape architecture is an important material carrier of human civilization, carrying rich cultural memory and artistic value [1]. These buildings not only reflect the technological level and aesthetic style of a specific historical period but also play an irreplaceable role in urban cultural identity, tourism development, and heritage protection [2]. However, over time, under the influence of factors such as natural environment erosion, climate change, human destruction, and material aging, historical landscape buildings suffer from varying degrees of structural damage, among which cracks are the most common and representative type of disease [3]. The appearance of building cracks not only affects the appearance of buildings but also indicates a decline in structural stability. In severe cases, it can threaten the safety of buildings and the lives and property of the public. Timely and accurate detection of cracks in historic landscape buildings is a key link in achieving health monitoring, preventive protection, and scientific restoration. Traditional manual inspection and visual identification have problems such as low efficiency, strong subjectivity, and difficulty in quantification, which are difficult to meet the needs of the protection of historical landscape buildings [4]. The development of computer vision, image processing technology, and artificial intelligence has demonstrated powerful capabilities in tasks such as image recognition and target detection, providing new ways for the identification of cracks in historical landscape buildings. Zhang et al. [5] introduced an optimized version of the You Only Look Once version 8 (YOLOv8) model, called MBE-YOLOv8, for detecting building cracks. The experiments found that the ability to detect building cracks was significantly enhanced compared to the previous YOLOv8 model. Sharma et al. [6] proposed a new convolutional and composite attention transformation network model and found through experiments that this method could accurately, efficiently, and reliably detect cracks in building structures. Lou et al. [7] presented a deep learning-based three-dimensional crack detection method that uses light detection and ranging point cloud data to detect building cracks. Golding et al. [8] proposed an autonomous crack detection method based on a convolutional neural network (CNN) technology and discovered that the deep learning-based crack detection did not rely on color. Krishnan et al. [9] filled the research gap and addressed the data shortage in the detection of brick wall cracks by developing a novel deep learning image processing technique. They also trained and validated multiple deep learning models to classify brick wall images into cracked or normal states. Katsigiannis et al. [10] proposed a deep learning method for crack detection on the

exterior walls of brick buildings using limited labeled data. The results showed that this method was very effective in crack detection. After end-to-end training of the neural network, its accuracy and F1 value reached up to 100 %. Ali et al. [11] developed a customized CNN for crack detection in concrete structures. This method was compared with four existing deep learning-based methods, and the results showed that this model performed better in crack detection and localization of concrete structures. The above-mentioned relevant studies have all analyzed the detection of building cracks. Some of them focus on the algorithms capable of identifying crack images, while others direct their research towards the data collection methods for building cracks. This paper focuses on the algorithms used to identify cracks in building images. This study employs YOLOv5 for the detection of building crack images. The selection of YOLOv5 is primarily due to its enhanced architecture compared to earlier versions (YOLOv1-YOLOv4), which significantly improves both detection speed and accuracy. Although YOLOv6-YOLOv8 achieve higher overall precision via structural refinements, these advancements come at the cost of increased training time, hindering efficient model iteration and optimization. In applications requiring rapid iteration and deployment, such delays can become a constraint. Moreover, detecting building cracks falls under the category of small-object detection, where moderate accuracy is often sufficient, and pursuing marginal gains in precision is unnecessary. In such cases, the more complex architectures of YOLOv6–YOLOv8, with improved accuracy, may introduce excessive computational overhead without providing significant benefits. YOLOv5 achieves a favorable balance between detection accuracy and inference speed, and its traditional architectural design has good potential for enhancement. Thus, this paper finally improves YOLOv5 by introducing the SAFM module with an attention mechanism. Then, a simulation experiment was carried out. The novelty of this paper lies in the introduction of the SAFM module with an attention mechanism into YOLOv5 for building crack recognition, enabling to reduce the computational load while maintaining the recognition performance.

1. Building crack recognition based on deep learning

The technology for detecting cracks in historical landscape buildings must have a high enough recognition accuracy, be able to automate processing to reduce human intervention, and be able to quantitatively analyze parameters such as crack length and width. Traditional manual inspection relying on the experience of technicians is labor-intensive, and the inspection results are subject to subjective influence [12]. Infrared thermal imaging technology uses the differences in infrared images caused by temperature differences to identify targets but has limited detection effect on cracks because cracks do not have a significant impact on the heat of buildings. Deep learning technology can extract features and identify crack types from building crack images with its feature extraction capabilities and nonlinear modeling capabilities [13].

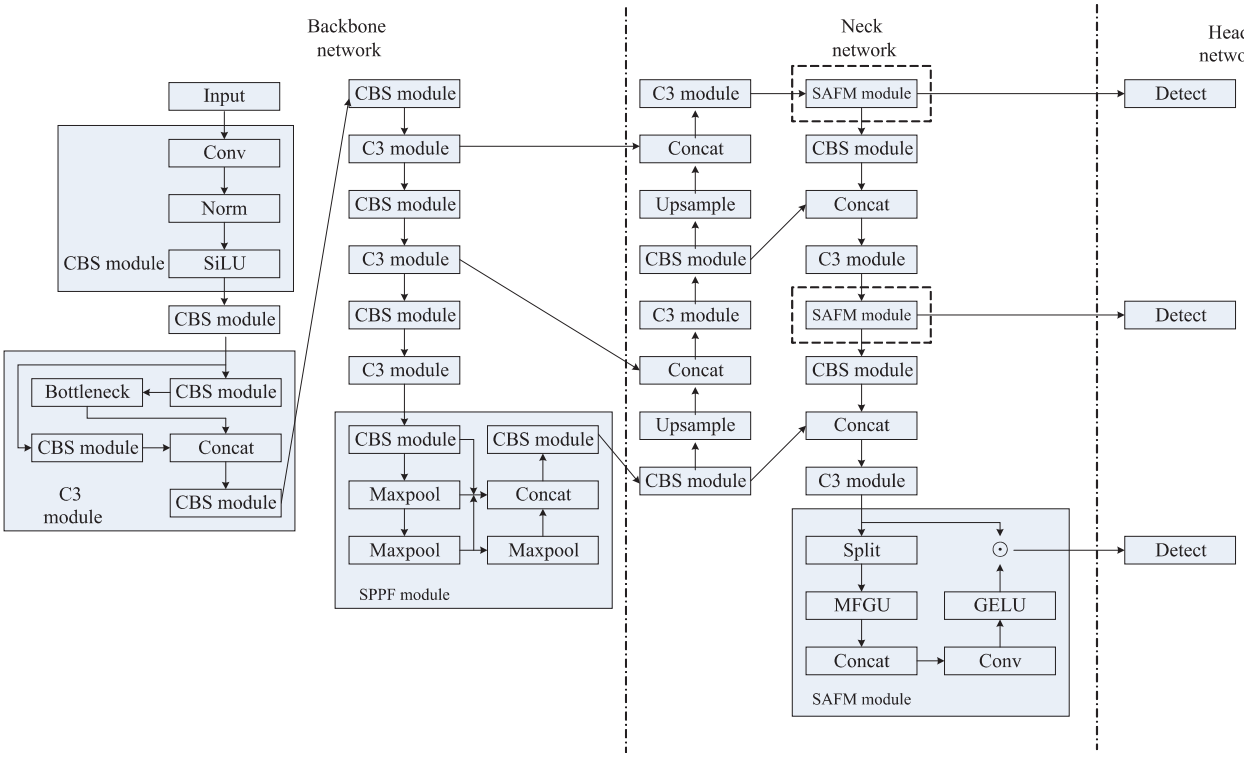

Fig. 1. The architecture of the improved YOLOv5 model

Crack detection on building surfaces is an image target detection task. In this paper, the YOLOv5 model is chosen for the detection and recognition of building cracks. To improve its recognition performance, an attention mechanism is introduced into it. Fig. 1 shows the architecture of the improved YOLOv5 model, where the SAFM module in the virtual box is the introduced attention mechanism module [14]. The attention mechanism of this module can add weights to the

feature channels of the feature images extracted by the backbone network. In this way, during training and usage, it can focus on important feature information and ignore secondary feature information, so as to jointly maintain the training and recognition effects and reduce the computational complexity. The process of detecting and identifying building cracks with the improved YOLOv5 model is as follows.

① A camera is used to take pictures of the building surface, and the captured images are preprocessed, including denoising and enhancement, before being input into the improved YOLOv5. In this paper, the bilateral filtering method [15] is used to denoise the image, and the formula for enhancing the image is:

$$f(x,y) = (G(x,y) - m_g)\frac{cs_f}{cs_g + (1-c)cs_f} + bm_f + (1-b)m_g, \tag{1}$$

where $G(x,y)$ is the grayscale value of pixel (x,y) in the original image, $f(x,y)$ is the grayscale value of (x,y) after the enhancement operation, m_g is the mean grayscale value of pixels in the original image, $\mathrm{m_f}$ is the mean grayscale value of image pixels after the enhancement operation, s_g is the standard deviation of the grayscale value of image pixels in the original image, s_f is the standard deviation of the grayscale value of image pixels after the enhancement operation, c is the contrast spread constant, and b indicates the luminance coefficient.

② The preprocessed building surface image is input into the backbone network of the improved YOLOv5 for layer-by-layer calculation according to the structure shown in Fig. 1 [16].

③ The backbone network extracts image features of different scales, and they are fused in the neck network [17]. The SAFM module consists of segmentation layers, multi-feature gate unit (MFGU) layers, concatenation layers, convolutional layers, Gaussian error linear unit (GELU) layers, and element product layers. The calculation formula for the entire module is:

$$\begin{cases} [B_0, B_1, B_2, B_3] = Split(B) \\ \hat{B}_0 = Conv_{3\times3}(B_0) \\ \hat{B}_i = \uparrow_p (Conv_{3\times3}(\downarrow_{\frac{p}{2^i}} (B_i)) \\ \hat{B} = Conv_{1\times1}(Concat(\hat{B}_0, \hat{B}_1, \hat{B}_2, \hat{B}_3)) \\ \bar{B} = GELU(\hat{B}) \otimes B \end{cases}, \tag{2}$$

where B is the input image features, $Split()$ is the channel segmentation function, B_0 , B_1, B_2, and B_3 are four image features after segmentation, $\hat{B}_0$ is the convolution-processed B_0, $Conv_{3\times3}( )$ is a convolution operation that indicates a 3×3 convolution kernel, $\hat{B}_i$ is the pooled, convolved, and upsampled B_i, $\downarrow_{\frac{p}{2^i}}( )$ indicates pooling features to a size of $\frac{p}{2^i}$, $\uparrow_p( )$ indicates upsampling features to a size of p , *Concat*( ) indicates a concatenation operation, $Conv_{1\times1}( )$ indicates a convolution operation of a 1 × 1 convolution kernel, $\hat{B}$ is the feature obtained after concatenating and convolving $\hat{B}_0$ and $\hat{B}_i$, $GELU()$ is the activation function, $\otimes$ represents the element product operation, and $\bar{B}$ is an output feature of the SAFM module.

④ After the neck network fuses the extracted image features, the head network detects targets in the fused image features. Candidate boxes of different sizes are used for target detection of the image features output by the three SAFM modules of the neck network, identify the content in the candidate boxes respectively, determine whether there are cracks in the candidate boxes and the types of cracks, and finally perform regression calculation on the grid points of the candidate boxes to determine their coordinates in the original image [18].

2. Simulation experiment

2.1. Experimental data

The data used in the simulation experiments were partly from the building structure crack annotation dataset (https://opendatalab.com/jiange1236/StructuralCracksDataset), and the remaining were from self-taken building surface images. There were totally 14,000 images of building surface cracks. All of them included brick and concrete structures. When marking the sample images, the target boxes fit the cracks as closely as possible. The types of building cracks in the sample set included fine cracks, adjacent edge-through fracture cracks, opposite edge-through fracture cracks, diagonal through cracks, defect cracks, perforated cracks, and no cracks (Fig. 2). Each type included 2,000 images, among which 1,000 images were set as the training set, 500 as the validation set, and 500 as the test set.

2.2. Experimental setup

The structure settings for the improved YOLOv5 model are shown in Figure 1. Some of the structure settings have been described in the previous process, and the relevant parameter settings for the remaining parts are shown in Table 1. The improvement of the YOLOv5 model lies in the introduction of the attention mechanism, i.e., the addition of the SAFM module reflected in the model structure. Therefore, to verify the effectiveness of the introduced mechanism, ablation experiments were conducted on the improved YOLOv5 model, and the same training and testing were carried out on the algorithm after the SAFM module was removed.

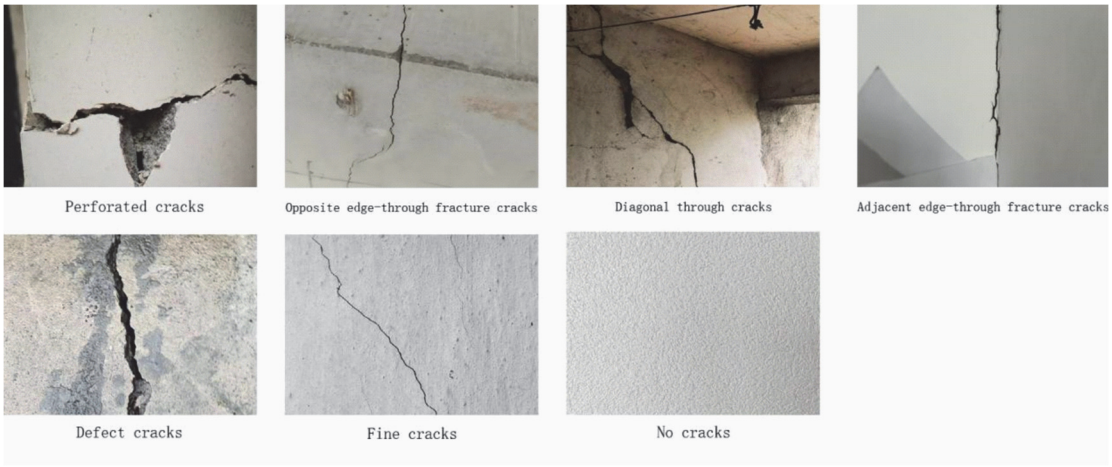

Fig. 2. Crack examples of different types

Table 1. Relevant parameters for the improved YOLOv5 model

Parameter	Setting	Parameter	Setting
Input layer	4,000×3,600	CBS module	A convolutional layer with 32 convolution kernels of a 3×3 specification, a moving step size of 2, and a SiLU activation function
SPPF module	A 3× pooling layer, max-pooling, a moving step size of 2	Number of training sessions	200
Batch	8	Optimizer	Adam [19]
Learning rate	0.01		

Moreover, to further verify the target detection performance of the improved YOLOv5 model, it was compared with other algorithms, namely Faster-regional-based convolutional neural network (Faster-R-CNN), YOLOv4, and YOLOv6.

2.3. Evaluation indicators

Common indicators for evaluating target detection algorithms include precision, recall rate, and F1 value:

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN}, \\ F = \frac{2PR}{P+R} \end{cases} \tag{3}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. The precision-recall rate curve of the target detection model can be plotted by precision and recall rate, and the area below the curve is the average precision (AP) of the detection model. If it is a multi-class detection model, the corresponding precision-recall rate curve can be plotted for each corresponding class to obtain the AP of the corresponding class. The mean average precision (mAP) value obtained by summing and averaging the AP of all classes is the detection accuracy of the multi-class detection model, which is between 0 and 1.

In addition to identifying the type of target, the target detection algorithm also needs to locate the target through the target box, so its evaluation indicator also includes the positioning accuracy of the target box:

$$IOU = \frac{DR \cap GT}{DR \cup GT}, \tag{4}$$

where IOU represents the degree of overlap of the target box, DR represents the target box predicted by the algorithm, and GT represents the actual target box in the image.

2.4. Experimental results

To test the effect of the introduced attention mechanism, ablation experiments were conducted on the improved YOLOv5 model with the SAFM module eliminated. After removing the SAFM module, the improved YOLOv5 model was equivalent to the traditional YOLOv5 model. It can be seen from Table 2 that the performance of the improved model in identifying and detecting building cracks and and the performance of crack positioning was reduced after the SAFM module was removed, indicating that the SAFM module can effectively enhance the target detection performance of the YOLOv5 model.

After verifying the effectiveness of the introduced attention mechanism, three target detection algorithms, Faster-R-CNN, YOLOv4, and YOLOv6, were used to compare with the improved YOLOv5 algorithm. The training and validation sets were

used to check whether the four algorithms were overfitting, and the test set was employed to verify the performance comparison among the four algorithms. The results are displayed in Table 3, and the partial detection results for building crack detection are presented in Fig. 3. It can be seen from the table that the recognition effects of the four algorithms on the cracks in the training set were all excellent, and the differences among them were not significant. However, for the validation set, except for the improved YOLOv5 algorithm, the recognition effects of the other three algorithms declined. The algorithms were trained with the samples in the training set, and they were adjusted according to the recognition errors during the training process, so they performed well for the training set. However, the validation set contained new samples for the algorithms, and the algorithms used the parameters obtained after training to make predictions. Except for the improved YOLOv5 algorithm, the other three algorithms suffered from over-fitting, which made them perform excellently when dealing with training samples but led to a significant decrease in performance when facing unknown new samples. When the test set was used, the improved YOLOv5 algorithm had the best detection and localization performance for building cracks, followed by the YOLOv4 and YOLOv6 algorithms, and the Faster-R-CNN algorithm had the worst performance. As can be seen from Fig. 2, the improved algorithm could effectively identify crack types and fit crack locations effectively. The YOLOv6 algorithm also identified crack types but was less precise in locating crack locations than the improved YOLOv5 model. The YOLOv4 and Faster-R-CNN algorithms misidentified crack types and were rough in crack location.

Table 2. Results of ablation experiments

Evaluation indicator	Traditional YOLOv5	Improved YOLOv5
Precision	0.852	0.989
Recall rate	0.843	0.988
F1	0.847	0.988
mAP	0.789	0.987
Intersection over union (IOU)	0.841	0.986

Table 3. Performance comparison of different building crack target detection algorithms

Sample set	Evaluation indicator	Faster-R-CNN	YOLOv4	YOLOv6	Improved YOLOv5
Training set	Precision	0.989	0.991	0.992	0.990
	Recall rate	0.987	0.990	0.991	0.989
	F1	0.988	0.990	0.990	0.989
	mAP	0.989	0.989	0.987	0.988
	IOU	0.988	0.988	0.986	0.987
Validation set	Precision	0.782	0.806	0.854	0.990
	Recall rate	0.781	0.815	0.848	0.989
	F1	0.781	0.805	0.845	0.989
	mAP	0.733	0.799	0.828	0.989
	IOU	0.742	0.813	0.847	0.988
Test set	Precision	0.784	0.801	0.853	0.989
	Recall rate	0.783	0.811	0.849	0.988
	F1	0.783	0.806	0.847	0.988
	mAP	0.734	0.798	0.829	0.987
	IOU	0.741	0.812	0.848	0.986

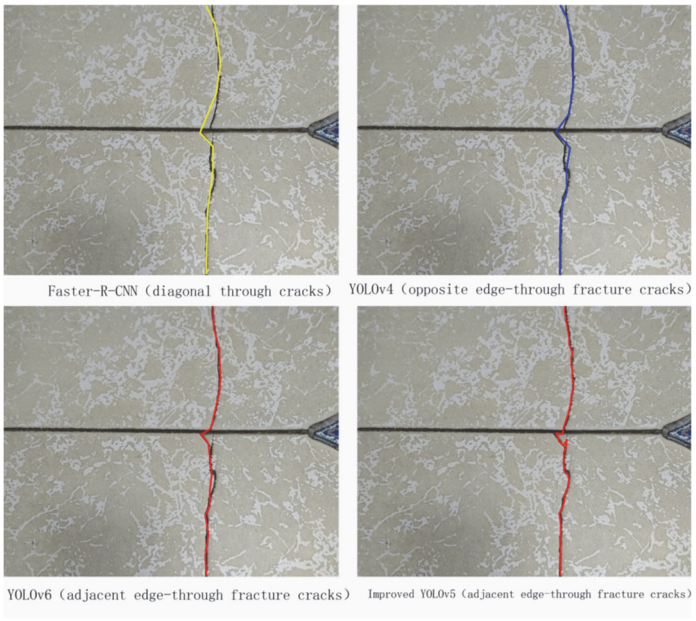

Fig. 3. Partial detection results of different building crack target detection algorithms

Conclusions

This paper introduced the SAFM module with an attention mechanism to the YOLOv5 model to improve its detection performance. Simulation experiments were conducted. Ablation experiments were first conducted on the improved YOLOv5 model to verify the effectiveness of the SAFM module, The improved algorithm was also compared with the other three algorithms. It was found that the attention mechanism of the SAFM module could effectively improve the target detection performance of the YOLOv5 model. The improved YOLOv5 algorithm had the best detection and location performance for building cracks, followed by the YOLOv4 and YOLOv6 algorithms, and the Faster-R-CNN algorithm performed worst. In terms of the detection results of building crack images, the improved YOLOv5 model could identify crack types and effectively fit crack locations.

References

[1] Luo Q, Ge B, Tian Q. A fast adaptive crack detection algorithm based on a double-edge extraction operator of FSM. Construction and Building Materials 2019; 204: 244-254. DOI: 10.1016/j.conbuildmat.2019.01.151.
[2] Hadiani R, Hidayah Q. Characteristic of fiber optic on concrete outer structure for crack detection system. AIP Conference Proceedings 2024; 2920(1): 070001. DOI: 10.1063/5.0190090.
[3] Ye W, Deng S, Ren J, Xu X, Zhang K, Du W. Deep learning-based fast detection of apparent concrete crack in slab tracks with dilated convolution. Construction and Building Materials 2022; 329: 127114. DOI: 10.1016/j.conbuildmat.2022.127114.
[4] Kirthiga R, Elavenil S. A survey on crack detection in concrete surface using image processing and machine learning. Journal of Building Pathology and Rehabilitation 2024; 9(1): 25. DOI: 10.1007/s41024-024-00444-0.
[5] Zhang Z, Hu Z, Chen K, Zhou Q, Zhang H. MBE-YOLOv8: Enhancing building crack detection with an advanced YOLOv8 framework. Measurement Science and Technology 2025; 36(2): 026005. DOI: 10.1088/1361-6501/ad9e16.
[6] Sharma B, Yadav DP, Dhaou IB, Chauhan S. Bridging convolutional neural networks and transformers for efficient crack detection in concrete building structures. Sensors 2024; 24(13): 4294. DOI: 10.3390/s24134294.
[7] Lou Y, Meng S, Zhou Y. Deep learning-based three-dimensional crack damage detection method using point clouds without color information. Structural Health Monitoring 2025; 24(2): 657-675. DOI: 10.1177/14759217241259646.
[8] Golding VP, Gharineiat Z, Munawar HS, Ullah F. Crack detection in concrete structures using deep learning. Sustainability 2022; 14(13): 8117. DOI: 10.3390/su14138117.
[9] Krishnan SSR, Karuppan MKN, Khadidos AO, Selvarajan S, Tandon S, Balusamy B. Comparative analysis of deep learning models for crack detection in buildings. Scientific Reports 2025; 15: 2125. DOI: 10.1038/s41598-025-85983-3.
[10] Katsigiannis S, Seyedzadeh S, Agapiou A, Ramzan N. Deep learning for crack detection on masonry façades using limited data and transfer learning. Journal of Building Engineering 2023; 76: 107105. DOI: 10.1016/j.jobe.2023.107105.
[11] Ali L, Alnajjar F, Jassmi HA, Gocho M, Khan W, Serhani MA. Performance evaluation of deep CNN-based crack detection and localization techniques for concrete structures. Sensors 2021; 21(5): 1688. DOI: 10.3390/s21051688.
[12] Wang Z, Leng Z, Zhang Z. A weakly supervised transformer-based hybrid network with multi-attention for pavement crack detection. Construction and Building Materials 2024; 411: 134673. DOI: 10.1016/j.conbuildmat.2023.134673.
[13] Liu C, Li Q. Building crack detection based on digital image processing technology and multiscale feature analysis automatic detection algorithm. Advances in Multimedia 2022; 2022: 8170469. DOI: 10.1155/2022/8170469.
[14] Deng L, Song J, Zhu XH. When infrared small target detection meets tensor ring decomposition: A multiscale morphological framework. IEEE Transactions on Aerospace and Electronic Systems 2022; 58(4): 3162-3176. DOI: 10.1109/TAES.2022.3152767.
[15] Chen D, Sun S, Lei Z, Shao H, Wang Y. Ship target detection algorithm based on improved YOLOv3 for maritime image. Journal of Advanced Transportation 2021; 2021: 6684243. DOI: 10.1155/2021/6684243.
[16] Jiang X, Chen WX, Nie HT, Hao ZC. Real-time ship target detection based on aerial remote sensing images. Opto-Electronic Engineering 2020; 28(10): 2360-2369. DOI: 10.37188/OPE.20202810.2360.
[17] Fei G, Teng H, Sun J, Wang J, Hussain A, Yang E. A new algorithm of SAR image target recognition based on improved deep convolutional neural network. Cognitive Computation 2018; 10(5): 1-16. DOI: 10.1007/s12559-018-9561-6.
[18] Kim S. Infrared variation reduction by simultaneous background suppression and target contrast enhancement for deep convolutional neural network-based automatic target recognition. Optical Engineering 2017; 56(6): 063108. DOI: 10.1117/1.OE.56.6.063108.
[19] Pei J, Huang Y, Sun Z, Zhang Y, Yang J, Yeo TS. Multiview synthetic aperture radar automatic target recognition optimization: Modeling and implementation. IEEE Transactions on Geoscience and Remote Sensing 2018; 56(11): 6425-6439. DOI: 10.1109/TGRS.2018.2844856.

About authors

Shuai Wang (b. September 1990) graduated from Shanghai University with a doctoral degree in September 2021. He is working at Shanghai Art & Design Academy as a lecturer. He is interested in design science.

Dahuang Zhu (b. February 1976) graduated from Shanghai Jiaotong University with a master's degree in March 2017. He is working at Shanghai Art & Design Academy as an associate professor. She is interested in environmental art design. E-mail: *zdahuang@outlook.com*

Received October 20, 2025. *The final version – March 23, 2026.*