Journal of Artificial Intelligence Practice (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2025.080401
ISSN 2371-8412 Vol. 8 Num. 4

Hybrid Detection Method for Concrete Cracks Based on Maskr-CNN and Swin Transformer

Hongbo Luo[1,2,3,4,a,*], Xiangyuan Ma[1,2,3,4,b], Fan Yu[5,c]

[1]Yangtze River Spatial Information Technology Engineering Co., Ltd. (Wuhan), Wuhan, China
[2]Changjiang Survey, Planning, Design and Research Co., Ltd, Wuhan, China
[3]Engineering Technology Research Center of Water Conservancy Information Perception and Big Data, Hubei Province Hubei, Wuhan, China
[4]Key Laboratory of Watershed Water Security in Hubei Province, Wuhan, China
[5]Beijing University of Civil Engineering and Architecture (BUCEA), Beijing, China
[a]hongbo_1983a@126.com, [b]maxiangyuan@whu.edu.cn, [c]yufan@bucea.edu.cn
[*]Corresponding author

Keywords: Mask R-CNN, Swin Transformer, Concrete Cracks, Precision

Abstract: This study proposes a hybrid detection model based on Mask R-CNN and Swin Transformer for the detection and segmentation of concrete cracks. This method fully utilizes the advantages of Mask R-CNN in precise object localization and pixel level segmentation, while introducing Swin Transformer to compensate for the shortcomings of traditional convolutional neural networks in capturing global contextual information, thereby improving the detection accuracy of the model in complex backgrounds and diverse crack shapes. The experimental results show that the proposed model outperforms traditional models such as U-Net, TransUNet, and Mask R-CNN in terms of Dice coefficient, accuracy, recall, and F1 Score evaluation metrics, especially in the detection and localization of complex crack images, demonstrating high robustness.

1. Introduction

The excellent ability of combining computer vision and deep learning (DL) methods has attracted increasing attention to the health monitoring of civilian infrastructure [1,2]. Previous statistical data on bridge collapses indicate that 46% of them were caused by pre-existing defects [3]. The proliferation of minor structural defects poses significant risks to infrastructure integrity, yet these vulnerabilities can be effectively mitigated through cost-efficient early detection strategies. This urgency has catalyzed the evolution of Structural Health Monitoring and Inspection (SHM&I)—a multidisciplinary field dedicated to prolonging the service life of critical assets while safeguarding human lives and economic stability [4, 5]. Tracing its origins to the 19th century, SHM&I emerged from rudimentary practices such as engineers relying on auditory cues from hammer strikes to assess railroad wheel integrity [6]. Over time, technological innovations have transformed these rudimentary methods into sophisticated systems integrating manual inspections, vibration analysis, and vision-based diagnostics. Despite this progress, conventional approaches face persistent challenges: 1) Manual assessments remain labor-intensive, prone to human error, and hazardous in unstable environments [7-9]. 2) Vibration-based monitoring, while precise, often involves prohibitive installation and operational costs [10]. So researchers have been searching for high-performance digital technologies and machine learning (ML) techniques that can achieve SHM&I process automation [11,12]. Concrete and asphalt are two commonly used materials in structures and roads, and cracks in them can reveal damage and its severity. Traditionally, research has utilized image processing techniques aimed at automatically detecting cracks to replace manual inspections of civil engineering infrastructure such as bridges and roads. Traditional algorithms typically use a set of manually formulated rules to process digital data to identify crack areas, including but not limited to Gabor filtering, Otsu methods, and morphological methods [13].

While models based on Fully Convolutional Networks (FCN) and U-Net have demonstrated considerable efficacy in road damage detection, they exhibit certain limitations when applied to asphalt pavement crack detection and quantification. As noted by Ji et al. [14], DeepLabv3+ presents a viable alternative within the DeepLab series, which employs an encoder-decoder architecture integrating Spatial Pyramid Pooling (SPP) and dilated convolution. The SPP module facilitates the assimilation of multi-scale contextual information from input features and progressively reconstructs spatial information, enabling more precise boundary delineation of larger objects. Concurrently, dilated convolution operations serve to expand the receptive field, thereby enhancing the model's capacity to capture and aggregate multi-scale characteristics. Notwithstanding the high accuracy and precision of conventional models like U-Net and Faster R-CNN in identifying road damage, their inherent dependence on standard convolutional operations imposes constraints. These include difficulties in resolving fine-grained crack details and an limited ability to model long-range dependencies within image data. Moreover, the performance of these models can be susceptible to degradation under variable lighting conditions and complex road surface textures.

To address these challenges, we propose a hybrid architecture that synergistically combines Mask R-CNN with the Swin Transformer. This integration harnesses the precise instance segmentation capabilities of Mask R-CNN alongside the powerful global context modeling afforded by the Swin Transformer's self-attention mechanism. Consequently, the proposed model demonstrates enhanced proficiency in detecting subtle and complex crack patterns across diverse environmental conditions, yielding improvements in accuracy, robustness, and overall comprehensiveness for road damage detection tasks.

2. Algorithm Introduction and Analysis

2.1 Mask R-CNN

Mask R-CNN represents a significant advancement in deep learning frameworks for computer vision, designed to address both object detection and instance segmentation tasks. As an extension of Faster R-CNN, it incorporates a parallel branch dedicated to predicting pixel-level segmentation masks alongside the existing classification and bounding box regression branches. This allows the model not only to identify and localize objects but also to generate precise binary masks delineating each instance's shape and boundaries. Such a capability makes it especially suitable for applications demanding high spatial accuracy, such as road crack detection, where understanding the exact morphology of anomalies is critical. The architecture of Mask R-CNN is composed of several interconnected components that work in sequence (as shown in Figure 1). The process begins with a backbone network, typically a deep convolutional neural network like ResNet or ResNeXt, often enhanced with a Feature Pyramid Network (FPN). This backbone is responsible for extracting multi-scale feature maps from the input image, forming a foundational representation for all subsequent stages. Following this, a Region Proposal Network (RPN) scans these feature maps to generate candidate object bounding boxes, each associated with an "objectness" score indicating the likelihood of containing an object. Key innovation that differentiates Mask R-CNN from its predecessor is the RoIAlign layer. This layer addresses a critical limitation of the RoIPooling technique used in Faster R-CNN, which involved quantization steps (rounding operations) that could lead to misalignments between the input features and the extracted region features. Such misalignments are detrimental to pixel-accurate mask prediction. RoIAlign eliminates this issue by using bilinear interpolation to compute the exact values of the input features at regularly spaced locations within each Region of Interest (RoI), thereby preserving spatial fidelity and ensuring precise alignment crucial for high-quality segmentation.

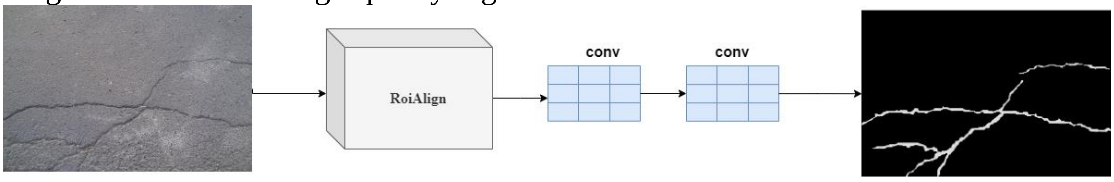

Figure 1 Mask R-CNN framework for pavement concrete crack segmentation

2.2 Swin Transformer

Unlike traditional CNN, ViT captures global context by processing images into fixed size patch sequences and utilizing self attention. Each layer of ViT has a downsampling rate of 16 times, which is not suitable for predicting dense tasks. At the same time, its grasp of multi-scale features will weaken, and for detection and segmentation tasks, multi-scale features are very important, and its self attention is always carried out on the entire image, which is a global modeling, and its computational complexity is quadratic with the image size. Therefore, Swin Transformer draws on many design concepts and prior knowledge of CNN: small window self attention (assuming that the same object will appear in adjacent places, so small window self attention is actually sufficient, while global self attention actually consumes some resources). The reason why CNN can capture multi-scale features is because of the pooling operation (which can increase the receptive field of each convolution kernel). Therefore, Swin Transformer also proposed a pooling like operation, which synthesizes adjacent small patches into a large patch.

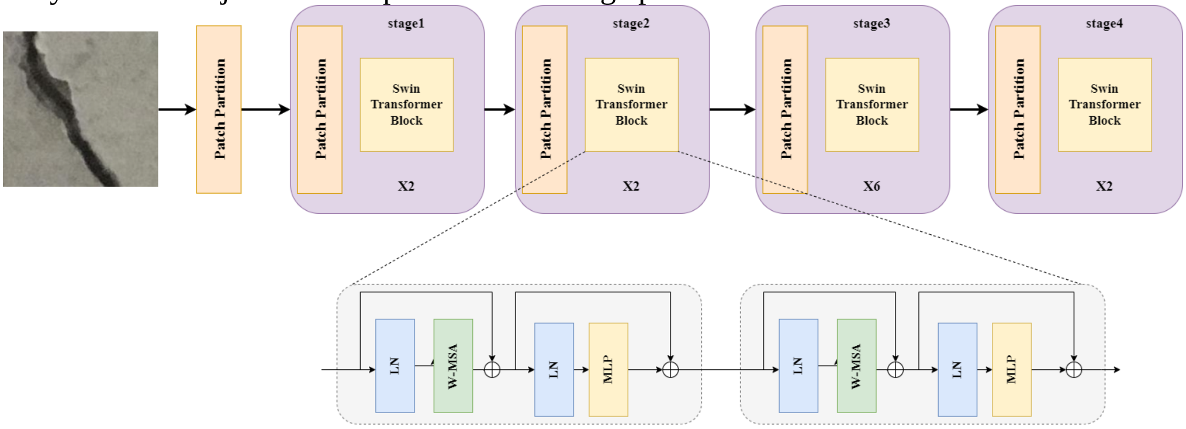

Figure 2 Swin Transformer Architecture

This study uses Swin Transformer (a multi-level hierarchical visual transformer) as the backbone of Mask RCNN network. Figure 2 depicts a Swin Transformer design consisting of four stages: (1) slice segmentation, (2) slice merging, (3) linear embedding, and (4) Swin Transformer blocks. Initially, the process begins with the patch segmentation module dividing the input image into small non overlapping patches (i.e., 4x4), and each patch is considered a token. The linear embedding layer projects the original value features onto any dimension (C). Then, feed the embedded feature vectors into multiple Swin Transformer blocks. The Swin Transformer block includes Window Multi Head Self Attention (W-MSA), Shift Window Multi Head Self Attention (W-MSA), and Multi Layer Perceptron (MLP). To ensure stability and promote smoother training dynamics, each instance of MSA module and MLP in these blocks is preceded by a LayerNorm (LN) layer. Swin Transformer introduces an effective method to manage self attention by restricting them within non overlapping local windows through a moving window approach. We need to use small windows to calculate attention, and move the window along the spatial dimension to the model boundary and global features. The shift of window partitioning between subsequent self attention layers allows for cross window interaction.

2.3 Mixture model

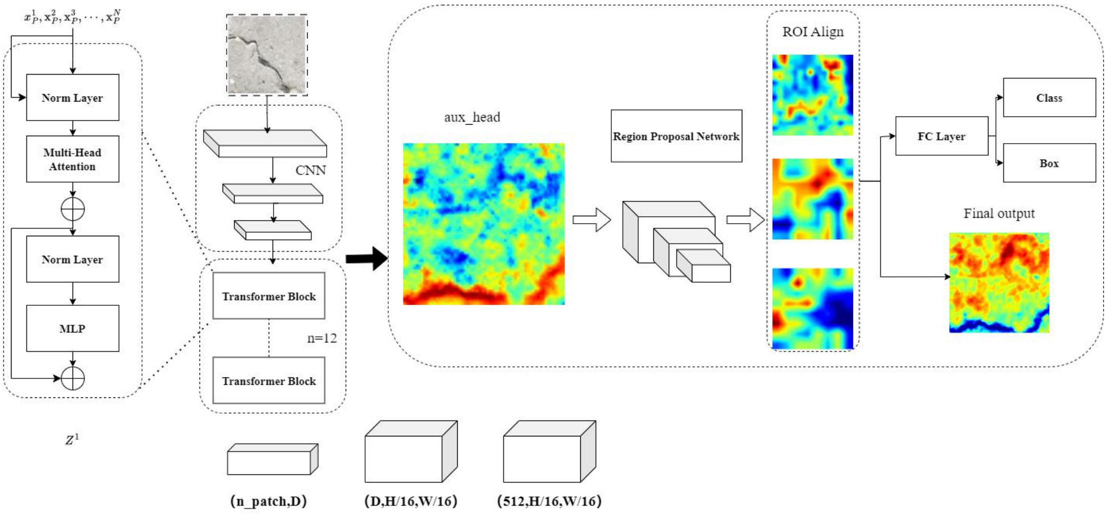

Figure 3 Hybrid Model Architecture of MaskR CNN and Swin Transformer

Region-based two-stage networks are recognized for their high accuracy in object detection, though they typically operate at slower speeds compared to single-stage detectors. A prominent example is Mask R-CNN, developed by He et al., which extends the Faster R-CNN framework by incorporating a parallel branch for predicting instance segmentation masks. Both architectures operate in two phases: the first stage involves a Region Proposal Network (RPN) that scans the image to generate candidate object proposals with an associated "objectness" score. The second stage, a region-based convolutional network, performs classification, bounding-box regression, and—uniquely in Mask R-CNN—pixel-level mask prediction for each Region of Interest (RoI). A key enhancement in Mask R-CNN over its predecessor is the integration of advanced backbone architectures, such as Residual Networks (ResNet), often combined with a Feature Pyramid Network (FPN), instead of the VGG16 network used in Faster R-CNN. This, coupled with the introduction of the RoIAlign layer which preserves precise spatial localization by replacing the quantizing RoIPool operation, significantly improves the accuracy of mask prediction, especially for small objects. Although Mask R-CNN's primary task is instance segmentation, its capability for end-to-end learning of both detection (bounding boxes) and segmentation tasks has allowed it to achieve top-tier detection performance among two-stage algorithms.In the present study, we further augment this architecture by replacing the conventional ResNet-101 backbone with a Swin Transformer for feature extraction on the Concrete Crack Images for Classification dataset. The Swin Transformer is a state-of-the-art hierarchical vision transformer that utilizes a self-attention mechanism computed within shifted windows. This design enables it to efficiently model long-range dependencies while maintaining a linear computational complexity relative to input image size, leading to features that often yield higher accuracy and generalization capability compared to ResNet-101. Owing to these exceptional properties, the Swin Transformer has been widely adopted as a powerful backbone for numerous vision-based algorithms. The overall network structure of our adapted Mask R-CNN model, illustrating the integration of the Swin Transformer architecture, is depicted in Figure 3.

3. Experimental data

3.1 Dataset description

This study uses the Concrete Crack Images for Classification dataset, which includes 40000 concrete surface images labeled as "positive" and "negative" (20000 each). All images have a resolution of 227 × 227 pixels, derived from standardized cropping of high-resolution raw images, without applying data augmentation to preserve the original features (Figure 4). The data is divided into training set, validation set, and testing set in a ratio of 7:2:1, and the training data is enhanced by random flipping and translation. In order to enable the labels to be processed correctly by the neural network, we performed encoding operations on the labels. Specifically, we convert category labels into one hot encoding format. For example, for binary classification problems, label 0 without cracks is encoded as [1,0], and label 1 with cracks is encoded as [0,1].

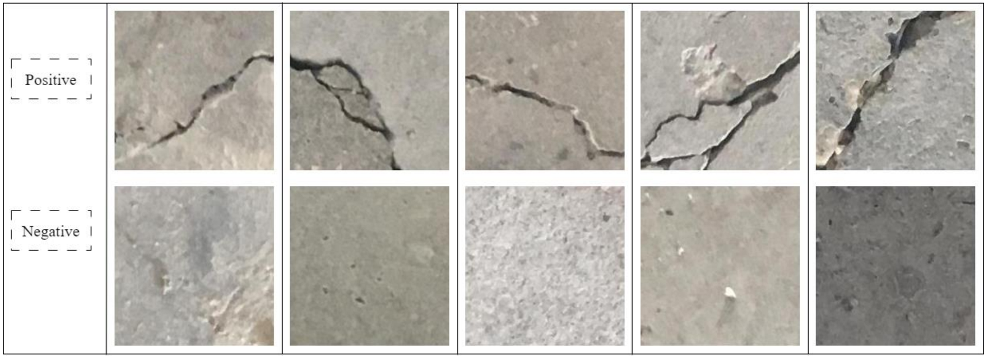

Figure 4 Concrete Crack Images for Classification Dataset

3.2 Model parameter configuration

The model parameter configuration is shown in the following Table 1:

Table 1 The model parameter configuration

Parameters	Concrete Crack Images for Classification
Objective function	DiceLoss+WBCE
2D Sliding window inference batch size	4 with 0.25 overlap
Optimizer	AdamW
Augmentation	True
Input Size	(256, 256)
#of feature map in base layer	16
#of attention head	12
MLP dimension	768
#of epochs	300
Initial learning rate	0.00001
Warmup epoch	10
Learning rate scheduler	OneCycleLR
Max_lr	0.001

A comparative analysis was conducted to evaluate the performance of several deep learning models for road crack detection using a standardized benchmark dataset. The assessed models included U-Net, TranU-Net, Mask R-CNN, nnUNet, and our proposed MaskR-Swin Transformer. The evaluation employed multiple metrics—DSC, Accuracy, Recall, and F1 Score—to provide a holistic assessment of their segmentation and classification capabilities. The benchmark dataset ensured a consistent and fair comparison across all models. Quantitative results indicated varying strengths among the architectures. For instance, nnUNet demonstrated high segmentation accuracy, while our proposed MaskR-Swin Transformer showed a balanced and competitive performance across all four metrics, suggesting an effective integration of instance segmentation and global context modeling. This analysis underscores the distinct advantages of different model families. The results are critical for guiding the selection of appropriate architectures based on specific application requirements, such as the need for precise pixel-level segmentation or efficient instance-level crack identification in automated road inspection systems.

In this study, our proposed MaskR Swin Transformer model demonstrated excellent performance on the Concrete Crack Images for Classification dataset, with a DSC of 80.04%, outperforming other models such as U-Net (66.07%) and TransUNet (68.96%). This significant improvement emphasizes the effectiveness of Masker Transformer in achieving precise segmentation in Table 2. In terms of recall rate, Mask R-CNN leads with an accuracy of 79.954%, followed closely by our MaskR SwinTransformer with an accuracy of 78.892%, indicating its strong ability to accurately identify relevant pixels for crack detection.

Table 2 Model Comparison Analysis Evaluation Indicators

Model	Concrete Crack Images for Classification				
	DSC	Precision	Recall	F1-Score	mIOU
U-Net	66.07	64.298	68.947	66.692	66.692
TransUNet	68.96	61.442	69.63	65.565	65.565
Mask R-CNN	76.44	70.411	79.954	72.59	75.59
nnUNet	74.22	67.917	79.145	72.767	72.767
MaskR-Swin Transfomer	80.04	72.31	78.892	76.698	75.851

3.3 Comparative analysis of performance of pavement concrete crack models

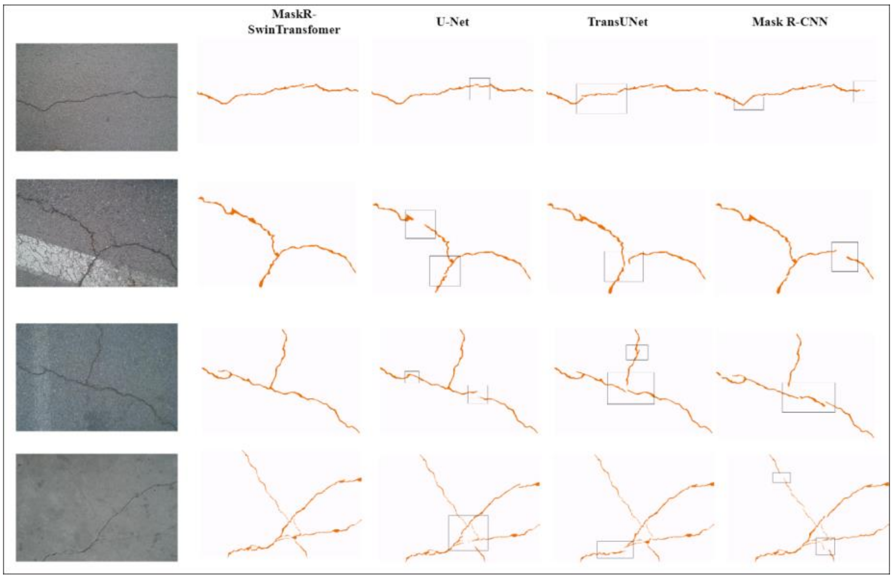

Figure 5 Visual comparison and analysis of results from different models

Figure 5 shows the comparative performance of various deep learning models for road crack detection on the Crack500 dataset. The first column displays the original image of the road surface, followed by the ground truth annotation in the second column, highlighting the actual cracks. The following columns display outputs from different models: U-Net, TransUNet, Mask R-CNN, and our model (MaskR-SwinTransformer). Compared to the real situation on the ground, the effectiveness of each model is visualized through their ability to detect and accurately outline cracks. It is worth noting that our model (rightmost column) shows higher accuracy in depicting cracks, closely matching the real ground conditions, especially in scenes with complex crack patterns (highlighted in black boxes).

4. Conclusion

The crack detection model based on the hybrid architecture of Mask R-CNN and Swin Transformer proposed in this study has achieved significant performance improvement in concrete crack detection tasks. From the experimental results, the model outperforms traditional U-Net, TransUNet, and other common deep learning segmentation networks in key evaluation metrics such as DSC, accuracy, recall, and F1 Score. This is mainly due to the model fully utilizing the advantages of Mask R-CNN in object localization and instance segmentation, while introducing Swin Transformer to compensate for the shortcomings of traditional convolutional networks in capturing global information. Through the complementary fusion of the two, this model can effectively capture remote dependencies and global contextual information while extracting local details in the image, thereby greatly improving the accuracy and robustness of crack segmentation.

In the context of concrete crack detection, cracks often have diverse features such as width, length, direction, and texture. Traditional convolutional network-based detection methods are prone to false or missed detections when dealing with subtle cracks and complex backgrounds. And this model utilizes Mask R-CNN to accurately segment crack edges, combined with the powerful global feature modeling ability of Swin Transformer, which significantly improves the detection results in terms of detail expression and overall consistency. The experimental data shows that the proposed model achieves 80.04% in DSC index, which is significantly improved compared to traditional models. This indirectly proves the improvement of data accuracy and provides more accurate basis for subsequent crack quantification analysis and engineering decision-making.

The reasonable application of data preprocessing and enhancement techniques in this model is also an important factor in improving accuracy. By normalizing, randomly cropping, and flipping the data, the diversity of training samples is effectively expanded, the risk of overfitting is reduced, and the model can maintain high robustness under different lighting conditions, background complexity, and changes in crack morphology. This combination of deep learning and data augmentation provides a practical and feasible path for automated detection in complex scenarios.

Combining the advantages of big data and cloud computing platforms, future research can also attempt to establish a cross regional and multi-scale structural health monitoring platform. By integrating multi-source data with advanced deep learning models, dynamic monitoring and predictive maintenance of the entire lifecycle of infrastructure can be achieved. This can not only significantly reduce the cost of manual inspection, but also provide early warning of potential risks and scientific and timely decision support for engineering management departments, thus playing an important role in ensuring public safety and extending the service life of facilities.

In summary, the hybrid detection model proposed in this article has significant advantages in improving data accuracy, providing solid technical support for accurate detection and subsequent quantitative analysis of concrete cracks; At the same time, research on multimodal data fusion, model lightweighting, and intelligent monitoring platform construction in the future has shown broad development prospects. I believe that with the continuous evolution of related technologies, this field will usher in new breakthroughs in intelligent detection, automated monitoring, and structural health management, contributing greater strength to the safety management of civil infrastructure.

Acknowledgments

This work was supported by Open Research Fund of Key Laboratory of Digital Mapping and Land Information Application, Ministry of Natural Resources: Visualization Method for Complex Urban Underground Space Structures Using Field Models (Fund No. ZRZYBWD202301), the Independent Innovation Project of Hubei Provincial Water Conservancy Information Perception and Big Data Engineering Technology Research Center: Research on Realistic 3D Reconstruction Method of Complex Urban Underground Space Structures (Fund No. CX2023Z07-2) and the Independent Science and Technology Innovation Foundation of Project from Key Laboratory of Watershed Water Security in Hubei Province（No. CX2023Z04-5）

References

[1] Azimi M, Pekcan G. Structural health monitoring using extremely compressed data through deep learning[J]. Computer-Aided Civil and Infrastructure Engineering, 2020, 35(6): 597-614.
[2] M. Naser, Autonomous fire resistance evaluation, Journal of Structural Engineering, 146 (2020), Article 04020103, doi: 10.1061/(ASCE)ST.1943-541X.0002641.
[3] Dung C V. Autonomous concrete crack detection using deep fully convolutional neural network[J]. Automation in Construction, 2019, 99: 52-58.
[4] Naser M Z. Enabling cognitive and autonomous infrastructure in extreme events through computer vision[J]. Innovative Infrastructure Solutions, 2020, 5(3): 99.
[5] Ewald V, Groves R, Benedictus R. Integrative approach for transducer positioning optimization for ultrasonic structural health monitoring for the detection of deterministic and probabilistic damage location[J]. Structural Health Monitoring, 2021, 20(3): 1117-1144.
[6] Taheri F, Shadlou S, Esmaeel R A. Computational Modelling of Delamination and Disbond in Adhesively Bonded Joints and the Relevant Damage Detection Approaches[J]. Reviews of Adhesion and Adhesives, 2013, 1(4): 413-458.
[7] Rao A S, Nguyen T, Palaniswami M, et al. Vision-based automated crack detection using convolutional neural networks for condition assessment of infrastructure[J]. Structural Health Monitoring, 2021, 20(4): 2124-2142.
[8] Andrushia A D, Anand N, Arulraj G P. Evaluation of thermal cracks on fire exposed concrete structures using Ripplet transform[J]. Mathematics and Computers in Simulation, 2021, 180: 93-113.
[9] Dais D, Bal I E, Smyrou E, et al. Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning[J]. Automation in Construction, 2021, 125: 103606.
[10] Dong C Z, Catbas F N. A review of computer vision–based structural health monitoring at local and global levels[J]. Structural Health Monitoring, 2021, 20(2): 692-743.
[11] Bal I E, Dais D, Smyrou E, et al. Novel invisible markers for monitoring cracks on masonry structures[J]. Construction and Building Materials, 2021, 300: 124013.
[12] Ali R, Kang D, Suh G, et al. Real-time multiple damage mapping using autonomous UAV and deep faster region-based neural networks for GPS-denied structures[J]. Automation in Construction, 2021, 130: 103831.
[13] Mohan A, Poobal S. Crack detection using image processing: A critical review and analysis[J]. alexandria engineering journal, 2018, 57(2): 787-798.
[14] Ji A, Xue X, Wang Y, et al. An integrated approach to automatic pixel-level crack detection and quantification of asphalt pavement[J]. Automation in Construction, 2020, 114: 103176.