Research on Engineering Structures & Materials

www.jresm.org

Research Group

Research Article

Multi modal early warning for mountain bridge structural defects using CBAM enhanced ResNet Mask R CNN

Feng Jiang [*,a], Miao Yu [b]

School of Energy and Water Resources, Shenyang Institute of Technology, Shenyang, China

Article Info

Article History:

Received 27 July 2026

Accepted 30 Aug 2026

Keywords:

ResNet;
Mask R-CNN;
Structural defect early warning;
CBAM;
Mountain bridge

Abstract

To address poor feature extraction adaptability and the gap with practical engineering requirements for mountain bridge structural defect early warning under complex geological backgrounds and variable illumination, this paper proposes a CBAM-ResNet-Mask R-CNN instance-segmentation early warning framework for mountain-bridge disasters. A dual-path residual block is designed for the Stages 3-4 of the ResNet-101 backbone, where dilated convolution is adopted in the main path and 1×1 dimensionality-reduction convolution is introduced in the auxiliary path, so as to suppress complex background noise and enhance weak disaster features under uneven illumination. The CBAM attention module is embedded inside residual blocks of Stages 3-4 to realize joint channel-spatial feature screening. Since Mask R-CNN only outputs two-dimensional pixel-level segmentation masks from RGB images, the visual quantitative features calculated from segmentation masks are further fused with auxiliary structural monitoring data including vibration and strain time-series signals, to realize inversion of three-dimensional structural deformation and disaster severity assessment. The reduction rate of economic loss is estimated by feeding disaster-severity quantification results into a bridge life-cycle cost model. On the independent test set, the proposed model achieves mAP@0.5 of 92.17%, mAP@0.5:0.95 of 76.42% and multi-disaster comprehensive F1-score of 0.93. Its recall rate for small-scale rockfall early warning reaches 93.64%, with an early-warning false-alarm rate of 2.56%. The monitoring accuracy for bearing deformation and concrete spalling proportion both exceeds 90%, and the maximum economic-loss reduction rate reaches 58.46% under test-set conditions. The proposed framework can realize accurate disaster-feature segmentation and multi-source-data joint quantification for mountain-bridge hazards, which provides feasible technical support for mountain-bridge safety monitoring and graded early-warning decision-making.

1. Background

Mountain bridges are susceptible to disasters such as rockfall impact, concrete spalling, and crack propagation due to long-term exposure to complex geological environments and variable climatic conditions [1-2]. Untimely disaster early warning can cause structural damage, traffic disruption, and even casualties. Therefore, constructing accurate and efficient mountain bridge disaster early warning models is of great engineering significance for ensuring bridge operational safety [3-4]. Current early warning methods still have limitations in disaster feature segmentation accuracy and structural stress prediction [5]. Some models tend to mis extract features or miss disasters when dealing with complex backgrounds and weak features in mountain bridge images [6-7]. Residual Network (ResNet) can specifically enhance weak disaster features and suppress redundant background information. Mask Region-based Convolutional Neural Network (Mask R-CNN) can

[*]Corresponding author: ddl10280@163.com

[a]orcid.org/0009-0003-5361-8728; [b]orcid.org/0009-0004-6767-965X

DOI: http://dx.doi.org/10.17515/resm2026-2054im0727rs

restore high-resolution segmentation masks through a four-layer transposed convolution up sampling in the segmentation branch. The Convolutional Block Attention Module (CBAM), due to its flexible modular design, can seamlessly integrate into ResNet backbone networks and Mask R-CNN feature extraction pipelines. Against the background of civil-infrastructure structural health monitoring, vision-based inspection has become an important technical route for bridge condition assessment [8-9]. Nevertheless, mountain-bridge monitoring suffers from complex geological background interference and variable illumination, which easily trigger false detection. Moreover, the logical linkage between image segmentation outputs and disaster quantitative evaluation remains insufficiently addressed in existing research. Accordingly, this paper proposes a mountain-bridge disaster early warning model integrating CBAM, ResNet, and Mask R-CNN. Dual-path residual blocks together with multi-scale feature fusion and CBAM attention modules are adopted to suppress complex mountain background noise and reduce false alarms caused by redundant background features. Pixel-level segmentation masks yielded by the visual model are further combined with vibration and strain time-series sensor data to realize disaster severity quantification, establishing the correlation between image processing results and structural disaster assessment. This work intends to satisfy the requirements of weak-disaster feature extraction and early warning for mountain bridges.

2. Literature Review

Structural health monitoring is essential to guarantee the safe service of civil infrastructures, including contact sensor based and non-contact sensing approaches. Contact sensors can collect one dimensional time series vibration and strain signals; one dimensional convolutional neural network combined with transfer learning are widely used to extract damage related features, while generative adversarial networks are adopted to mitigate the scarcity of real-world damage samples [10-11]. Sensor placement optimization is also a key issue for dynamic structural monitoring of civil infrastructures [12]. Apart from contact monitoring techniques, InSAR time series remote sensing can realize large scale infrastructure deformation monitoring but cannot capture fine grained local damage details of bridges [13]. Non-contact computer vision-based monitoring possesses flexible deployment merits, and vision driven real time structural health monitoring systems have been developed for offshore infrastructures. However, few investigations focus on mountain bridge scenarios disturbed by uneven illumination and complicated geological backgrounds. As an instance segmentation algorithm with multi-task capability and high-precision segmentation, Mask R-CNN can simultaneously output target categories, bounding boxes, and segmentation masks in tasks. Some scholars have conducted explorations based on its characteristics. Setrayana A et al. proposed the application of YOLOv11 and Mask R-CNN models for real-time water level detection to address the problem of slow response in manual systems for urban flood monitoring in Indonesia [14]. Instance segmentation approaches can be divided into two stage and one stage paradigms. Two stage detectors represented by Mask R CNN achieve superior mask boundary precision benefited from RoI Align operation, yet bring non negligible computational overhead. By comparison, one stage segmentation models such as YOLOv8 seg deliver faster inference speed suitable for edge device deployment, at the cost of slight degradation in segmentation boundary accuracy, and have been applied for post seismic structural damage segmentation tasks [15]. Besides conventional ResNet backbones, Swin Transformer and other vision Transformer architectures exhibit advantages in global context modeling and capturing long range dependencies for complex scenes, while suffering higher computational cost. Unlike Mask R-CNN, CBAM can precisely enhance key features and suppress redundant information through dual attention mechanisms of channel and spatial dimensions. Some domestic and international scholars have conducted different studies using this method. Huang Y et al. developed a model that integrated multi-scale convolution with a self-attention mechanism to address seismic event classification tasks, while introducing a voting mechanism to integrate multi-component data. Experiments verified that this method outperformed traditional methods [16]. Wenxuan Z et al. proposed a lightweight neural network method to solve the problems of low accuracy and slow speed in traditional fire detection methods. The results indicated that this approach demonstrated significant superiority in identifying small-scale fire events [17].

In addition, effective disaster early warning can minimize the impact of disasters on life, property, and social systems. Some scholars have proposed different opinions on this issue. Yue Y put forward an analytic hierarchy process optimized by entropy weighting and fuzzy numbers to assess construction risks of highway bridges in plateau mountainous regions. The experimental results showed that this method could effectively identify key risk points such as building materials and provide quantitative basis for precise prevention and control of high risks [18]. Yang P's team proposed a dual-threshold method for the uncertainty problem of critical rainfall thresholds in mountain flood warning. The results indicated that this method had higher reliability compared to traditional single-threshold methods [19]. Rawat D et al. proposed the use of wavelet synchro squeezing transform to analyze seismic signals from station networks for the Chamoli 2021 ice-rock avalanche disaster. The experimental results showed that this method could effectively identify high-frequency collapse signals at 16-18 Hz and low-frequency seismic precursors at 0.01-0.2 Hz [20]. Zhao Y's team innovatively proposed an adversarial interpretive structural modeling method for factors influencing mountain community resilience. The results showed that this method could systematically identify the action paths of key factors such as infrastructure, information access, and emergency planning [21]. Ding Y et al. proposed a comprehensive rockfall scenario simulation method for the difficult problem of rockfall risk simulation in mountain bridges. Experiments showed that this method could effectively analyze rockfall risks under complex coupling effects by constructing virtual risk environments [22].

From the conclusions drawn by domestic and international scholars, although current disaster early warning methods have made certain progress, there still exist problems of insufficient model adaptability and scenario compatibility. Some early warning methods are disconnected from actual engineering needs, and the entropy weight-based analytic hierarchy process and dual-threshold rainfall methods are difficult to directly provide precise technical support for bridge disaster emergency response. Therefore, to improve the adaptability of disaster early warning methods, this study proposed a mountain bridge disaster early warning model combining CBAM, ResNet, and Mask R-CNN. The study expects to provide effective technical support for property loss control and ecological facility protection through this model.

3. Mountain Bridge Disaster Early Warning Model Construction

3.1 Design of Mask R-CNN Disaster Recognition Model Combine Resnet

ResNet introduces residual block structures that allow inputs to directly propagate across layers to outputs. This advantage enables effective network training even when the depth increases. Mask R-CNN provides precise instance segmentation and multi-task capabilities, which can simultaneously identify different disaster types in mountain bridge early warning scenarios. Therefore, this study designs a Mask R-CNN disaster recognition model integrating ResNet, which strengthens weak feature extraction for bridge disasters through ResNet and achieves integrated disaster category recognition, location detection, and pixel-level segmentation with Mask R-CNN. The instance segmentation process is shown in Figure 1.

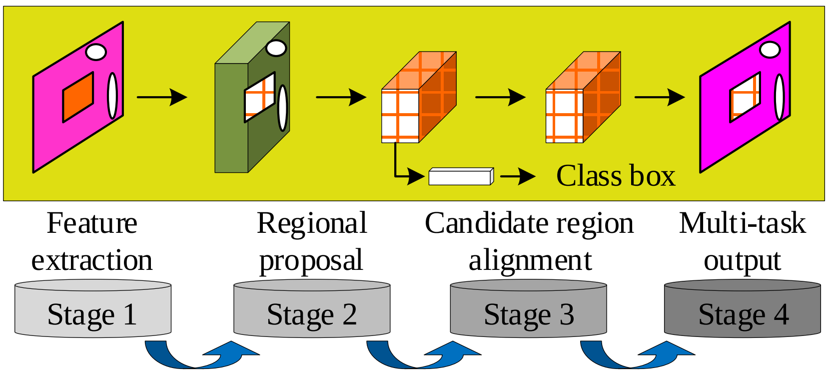


Fig. 1. Instance segmentation process of Mask R-CNN

In Figure 1, Stage 1 first standardizes and augments the original bridge images, then uses the backbone network to gradually reduce image size and increase feature dimensions. Stage 2 presets twelve anchor boxes on the C4 feature map for diverse bridge disaster forms, evaluates whether each anchor box contains a disaster using 1×1 convolution, and outputs confidence scores. Candidate regions are filtered through non-maximum suppression to remove redundant overlapping anchor boxes, producing high-confidence candidate regions. Stage 3 maps candidate rockfall regions from the original image onto the C4 feature map proportionally, divides sampling points in the mapped feature regions, and calculates feature values for each point using bilinear interpolation. Finally, it averages the feature values to generate a fixed-dimension feature vector. The total loss function is shown in Equation (1).

$$L_{total} = L_{cls} + \lambda_1 \cdot L_{reg} + \lambda_2 \cdot L_{mask} \tag{1}$$

Equation (1) shows L_{cls} and L_{reg} representing classification loss and bounding box regression loss, L_{mask} representing Mask segmentation loss, and λ_1 and λ_2 representing loss weights. The classification loss calculation is shown in Equation (2) [23].

$$L_{cls}(p, p*) = -\sum_{k=0}^{K} p_k^* \cdot log(p_k) \tag{2}$$

In Equation (2), $p = [p_0, p_1, \ldots, p_k]$ indicates the predicted class probability distribution, p_k represents the probability that a candidate region belongs to class k, and K is the total number of classes. Stage 4 performs multi-task output. The input feature vector passes through two fully connected layers to compress it to 1×4 dimensions. The Softmax function outputs class probabilities, and another fully connected layer outputs coordinate parameters. The Mask branch enlarges the feature map with four transposed convolution layers, removes low-confidence regions, and overlays the segmentation mask of remaining high-confidence regions onto the original bridge image. The study selects ResNet-101 as the backbone to capture finer-grained disaster features. The ResNet-based feature enhancement module is shown in Figure 2.

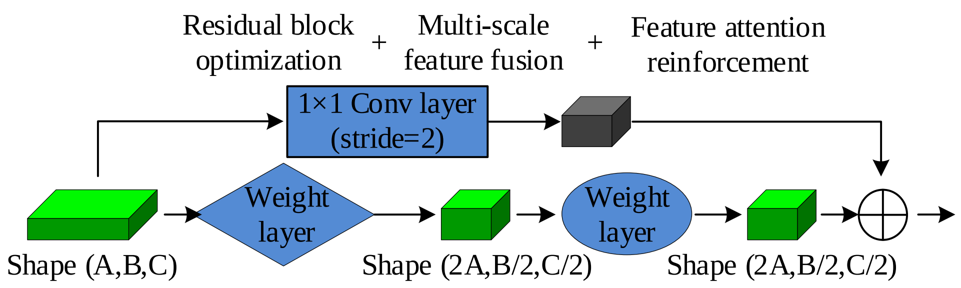


Fig. 2. ResNet-based feature enhancement module

In Figure 2, the model adds an adaptive contrast enhancement module at the ResNet input layer. It uses the Retinex algorithm to separate image illumination and reflection components and applies contrast-limited adaptive histogram equalization to the reflection component. A dual-path residual block structure is designed, where the main path introduces dilated convolution to the original residual block, and the auxiliary path adds a 1×1 convolution for dimensionality reduction and expansion. The two paths are fused through element-wise addition. Different stages of ResNet-101 use different residual block configurations: stages 1–2 use standard residual blocks, while stages 3–4 replace all standard blocks with dual-path residual blocks. Residual block feature propagation is calculated as shown in Equation (3).

$$H(x) = F(x, \{W_i\}) + W_s(x) \tag{3}$$

In Equation (3), $H(x)$ is the output feature map of the residual block, $F(x, \{W_i\})$ is the residual function, $\{W_i\}$ is the convolution kernel weights in the residual function, and $W_s(x)$ represents the shortcut connection. Global feature extraction is calculated as in Equation (4).

$$F_{main}(x) = BN(W_2 \cdot ReLU(BN(W_1 \cdot x)) \tag{4}$$

Equation (4) uses W_1 and W_2 as 3×3 convolution kernels, with BN representing batch normalization. High-frequency weak feature enhancement is calculated in Equation (5).

$$F_{aux}(x) = W_4 \cdot ReLU(W_3 \cdot x) \tag{5}$$

Equation (5) shows W_3 and W_4 as 1×1 convolution kernels, W_3 reduces to one-fourth of input channels, and W_4 expands to the main path output channels. The overall framework of the Mask R-CNN-based disaster recognition model integrated with ResNet is illustrated in Figure 3.

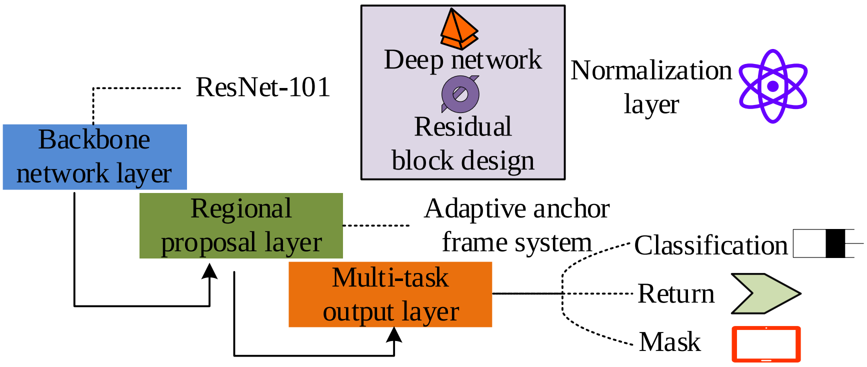


Fig. 3. Mask R-CNN disaster recognition model integrating ResNet

As shown in Figure 3, the backbone network uses ResNet-101 as the core for feature extraction. The 101-layer deep network with residual block design addresses weak feature extraction deficiencies in traditional convolutional neural networks. Batch normalization layers are added in the residual blocks of ResNet stages 3 and 4. The weighted feature fusion is calculated as shown in Equation (6) [24].

$$F_{fusion} = \alpha \cdot C3' + \beta \cdot C4 + \gamma \cdot C5' \tag{6}$$

In Equation (6), F_{fusion} indicates the fused enhanced feature map, α, β, and γ are fusion weights. $C3'$, $C4$, and $C5'$ represent aligned multi-scale feature maps. Channel attention weights are calculated as in Equation (7).

$$M_c(F) = \sigma(W_2{}' \cdot ReLU(W_1{}' \cdot GAP(F))) \tag{7}$$

Equation (7) shows F as the input feature map to the attention module. $W_1{}'$ and $W_1{}'$ are fully connected layer weights, and c represents the number of feature map channels. The region proposal layer designs an adaptive anchor system for diverse bridge disaster forms. The multi-task output layer includes classification, regression, and Mask branches. The model outputs disaster probabilities through activation functions and optimizes bounding box coordinates using Smooth L1 loss. The regression branch outputs pixel-level disaster locations. The Mask branch up samples to the original image size for disaster quantification, using 1×1 convolution and Sigmoid activation to output 28×28 binary segmentation masks.

3.2 Optimization of Disaster Recognition Model with CBAM

Although the Mask R-CNN disaster recognition model combined with ResNet can effectively extract disaster features, it still suffers from background interference that leads to false detections in complex mountain environments. CBAM can accurately enhance key features without significantly increasing the number of model parameters. This advantage makes CBAM well suited for mountain bridge disaster recognition tasks [25]. Therefore, to further improve model performance, this study introduces CBAM to optimize the constructed disaster recognition model. The process of how CBAM adapts to bridge disaster requirements is shown in Figure 4.

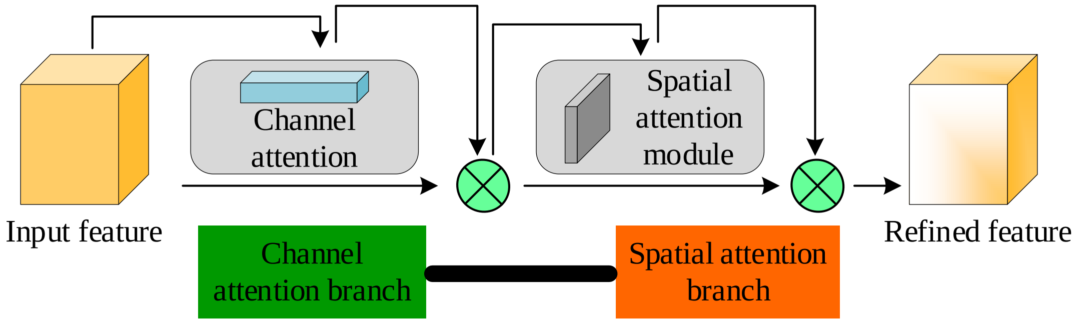


Fig. 4. Schematic diagram of CBAM adapting to bridge disaster requirements

In Figure 4, the channel attention branch in CBAM applies global average pooling to condense the feature map into a 1×1×C vector, which is subsequently processed through two fully connected layers followed by a Sigmoid activation function to produce channel weights. These weights amplify the disaster-related feature channels while attenuating background information. This process enables adaptation to the mixed-channel scenes in bridge disasters. The spatial attention branch addresses the issue that disaster regions occupy a small proportion of the image and are easily covered by backgrounds. It compresses the feature map through channel pooling, and after compression to 2×H×W, applies a 1×1 convolution to generate spatial weights, achieving localization of disaster pixels. The calculation of the final enhanced feature map output by CBAM is shown in Equation (8) [26].

$$F'_{final} = F' \otimes M_c(F') \otimes M_s(F'_c) \tag{8}$$

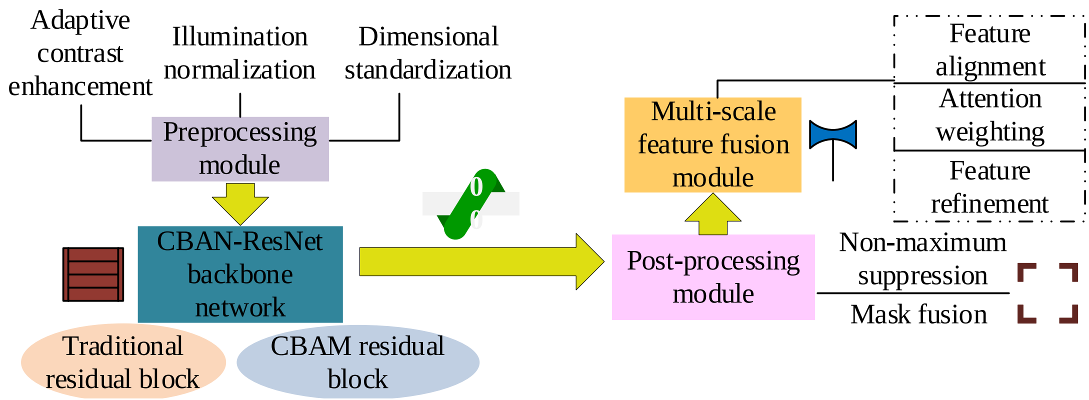


Fig. 5. Architecture of CBAM-ResNet-Mask R-CNN

In Equation (8), F' is the input feature map of CBAM, $M_c(F')$ is the channel attention weight vector, and the two $\otimes$ operations respectively perform channel-level feature selection and spatial-level feature focusing. It should be noted that Mask R-CNN as a two-stage detector imposes higher computational burden compared with one-stage YOLO-seg series models. For mountain-bridge disaster quantitative early-warning tasks, the segmentation boundary accuracy directly governs the calculation reliability of crack width, rockfall volume and subsequent structural-damage evaluation. Boundary errors from segmentation will propagate and trigger incorrect early-warning decisions. Therefore, pixel-level segmentation accuracy is prioritized in this study. Meanwhile, the dual-path residual blocks and embedded CBAM attention module optimize feature extraction and suppress redundant background computation, enabling the model to satisfy the latency requirement for mountain-bridge early warning under given hardware conditions. The study then combines CBAM, ResNet, and Mask R-CNN to construct a new algorithm named CBAM-ResNet-

Mask R-CNN, which serves as the foundation of the disaster early warning model. The architecture is shown in Figure 5.

In Figure 5, the algorithm enhances the local contrast of weak features such as crack edges, while the Retinex decomposition removes uneven lighting effects and scales the image. The CBAM-ResNet backbone serves as the feature extraction core. Stages 1–2 use traditional residual blocks to output C2 and C3 feature maps, while stages 3–4 are replaced with CBAM residual blocks to output C4 and C5 feature maps. Each CBAM residual block includes convolutional feature extraction and CBAM attention selection. The crack width is calculated as shown in Equation (9).

$$W' = \frac{N \times Re\,s}{S} \tag{9}$$

Equation (9) shows N as the pixel length of the crack in the segmentation mask, $Re\,s$ and S as the image resolution and image scaling factor, respectively. Equation (9) is derived under the premise of calibrated near-orthogonal shooting perspective. For practical inclined camera viewpoints, camera intrinsic parameters and planar homography transformation should be introduced to compensate perspective distortion. In this paper, image scaling factors are calibrated for each sampling scene using reference scale markers on bridge surfaces to mitigate measurement bias induced by oblique shooting angles. The rockfall volume is calculated as shown in Equation (10) [27].

$$V = Area \times h \times \rho \tag{10}$$

In Equation (10), $Area$ is the pixel area of the rockfall in the segmentation mask, h and ρ are the average height and shape coefficient of the rockfall. Accurate 3-D volume calculation requires depth information from LiDAR or RGB-D sensors, which are not available in this monocular RGB imaging system. Therefore, the rockfall pixel area calculated from segmentation mask serves as the primary quantitative indicator. The volume calculated in Equation (10) is an estimated reference value derived from empirical average height and shape-coefficient assumptions, rather than a strictly measured true 3-D volume. The average height and shape-coefficient parameters are derived from field statistical data of rockfall events in local mountain-bridge zones, rather than real-time depth measurement from each input image. These empirical parameters are suitable for general-scale rockfall samples in the study area. The estimated rockfall volume carries an inherent relative uncertainty of ±22 % under different actual rockfall morphologies. Volume output serves only as auxiliary reference, and rockfall pixel area from segmentation mask acts as the primary quantitative indicator. In the multi-scale feature fusion module, the region proposal network generates disaster candidate regions based on the fused feature maps. RoI Align performs pixel-level alignment through bilinear interpolation. The classification branch outputs disaster probabilities, the localization branch outputs bounding box coordinates, and the segmentation branch generates a 28×28 binary mask through four transposed convolution layers, which is up sampled to produce pixel-level segmentation results. The post-processing module removes overlapping candidate regions using non-maximum suppression and overlays the segmentation masks onto the original image. The Mask R-CNN only outputs two-dimensional pixel-level segmentation masks of disaster regions from RGB images. Structural indicators such as bearing deformation and structural stress, as well as economic-loss estimation, cannot be directly derived from visual segmentation results alone. In this framework, visual quantitative features computed from masks (including crack length, crack width, spalling area and rockfall area) are fused with multi-source auxiliary structural monitoring data (vibration frequency and strain growth rate collected by sensors, see Equation (11)) for joint calculation, so as to invert three-dimensional structural bearing deformation and stress state. The percentage of economic-loss reduction is evaluated by inputting disaster-severity quantification results into the bridge life-cycle cost analysis model, rather than directly predicted by the instance-segmentation network. The workflow of the mountain bridge disaster early warning model is illustrated in Figure 6.

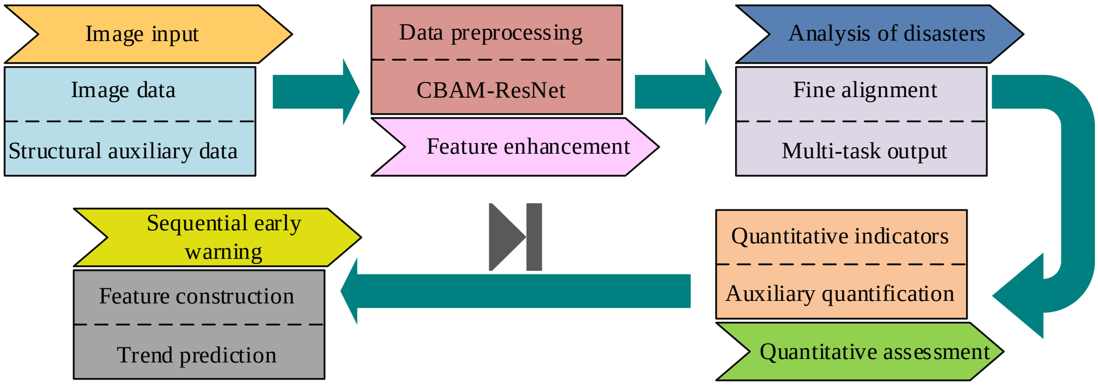


Fig. 6. Workflow of mountain bridge disaster early warning model

Figure 6 shows that the preprocessing and feature enhancement layers mainly solve the problems of complex backgrounds and weak feature extraction in mountain bridge images through data preprocessing and CBAM attention enhancement. This part optimizes images using illumination normalization, contrast enhancement, and size standardization, and filters and normalizes vibration and strain data. The calculation is shown in Equation (11).

$$I_{norm}(i,j) = \frac{I(i,j) - \mu}{\sigma' + \varepsilon} \tag{11}$$

In Equation (11), $I(i,j)$ is the grayscale value of the original image, $I_{norm}(i,j)$ represents the normalized pixel value, μ is the global mean of the image, $\sigma^{'}$ is the global standard deviation. The calculation of mask morphological filtering is shown in Equation (12).

$$m_{filtered} = Close(Open(m,k_1),k_2) \tag{12}$$

In Equation (12), m is the original segmentation mask, $m_{filtered}$ represents the filtered mask, and $Open(m,k_1)$ and $Close(m,k_2)$ are the opening and closing operations, respectively. The calculation of crack length is shown in Equation (13).

$$l = \sum_{i=1}^{N} Dist(s_i, s_{i+1}) \times Re\,s \tag{13}$$

In Equation (13), $Dist(s_i, s_{i+1})$ represents the Euclidean distance between adjacent sampling points, and s_i represents the sampling points on the crack skeleton line. CBAM-ResNet uses ResNet-101 as the backbone network. The disaster instance segmentation layer, based on the enhanced feature map, provides accurate spatial information for subsequent quantitative analysis. In the disaster quantification analysis layer, the model performs auxiliary quantification using structural data, integrating vibration frequency changes, strain growth rate, and image-based quantification results. The time-series trend prediction layer predicts disaster development trends based on multi-period quantitative data. The early warning decision output layer combines the current disaster level and future trend prediction to set multi-level warning thresholds, finally outputting visualized early warning results. Image instance-segmentation outputs visual damage features. The complete early-warning workflow follows sequential steps. Damage geometric parameters derived from segmentation are combined with multi-source sensor time-series data to calculate structural-state indicators. Disaster development trends are predicted using multi-period quantitative results, and multi-level warning thresholds are triggered to generate early-warning outputs. Pixel-level segmentation acts as the core intermediate step of the whole pipeline, not the final early-warning output. The transformation from pixel-level defect geometric features to structural stress and economic indicators adopts a data-driven correlation mapping strategy, without introducing external simulation software or digital-twin models. The defect geometric parameters (crack width, spalling area, rockfall area) output from segmentation masks are correlated with structural stress by linear fitting built upon labeled sensor monitoring datasets. The calculated structural-stress results are further mapped to economic-loss indicators according to empirical maintenance-cost statistics. This mapping assumes a linear correlation between defect geometric size and local structural stress within the service-state range of mountain bridges, and the economic calculation keeps unit repair cost and traffic-interruption loss coefficient unchanged.

4. Performance Verification of Mountain Bridge Disaster Early Warning Model

4.1 Comparison of Model Simulation Performance

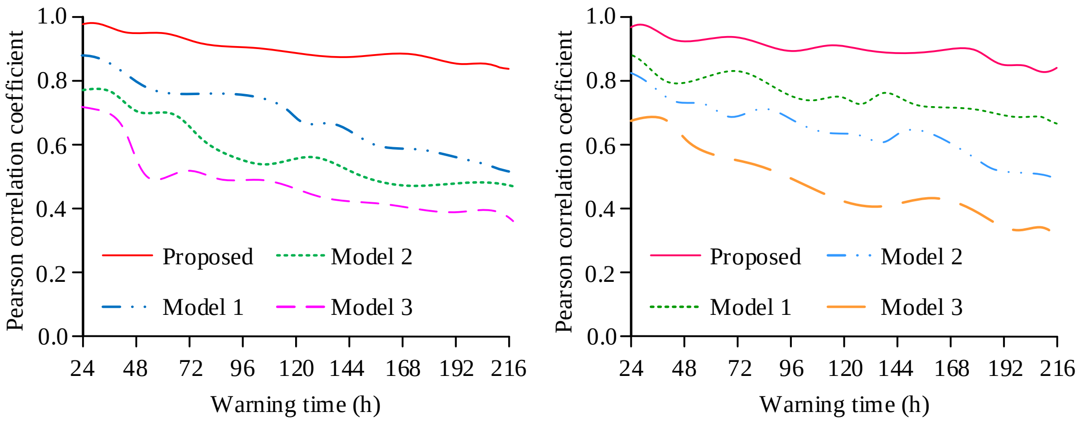
To verify the performance of the proposed mountain bridge disaster early warning model, the study selected the following three early warning models as comparison models, which were early warning models based on ResNet-Mask R-CNN-CNN, ResNet-Mask R-CNN, and Mobile Network Version 3-Mask R-CNN (MobileNetV3-Mask R-CNN). These three comparison models are set as ablation baselines. ResNet-Mask R-CNN serves as the original basic framework without our dual-path residual and CBAM improvement. ResNet-Mask R-CNN-CNN adds extra ordinary convolution branches on the basis of the basic framework to verify the effectiveness of our dual-path residual design. MobileNetV3-Mask R-CNN replaces the ResNet backbone with lightweight MobileNetV3, which is widely adopted for edge-end vision tasks in infrastructure monitoring. The study named them as Model 1, Model 2, and Model 3. Two kinds of datasets are adopted in this work. The input of Mask R-CNN is optical RGB image dataset, which is derived from the public dataset of Mountain Bridge Structure Health Monitoring Laboratory of Chongqing Jiaotong University. This image dataset contains 8264 annotated images, covering 11 mountain-bridge structures, with four damage categories: crack, concrete spalling, rockfall, and bearing deformation-related surface features. The Long-Term Guided-Wave Dataset Under Real-World Conditions contains 7.6 million one-dimensional ultrasonic guided-wave (UGW) time-series signals collected by 14 piezoelectric sensors from simulated bridge metal components. Note that these UGW signals are not fed into the Mask R-CNN network. Instead, they are fused with visual segmentation outputs in the post-processing disaster-quantification module for auxiliary structural-state inversion, and do not provide image-level labels for instance-segmentation training. Ground-truth labels for visual damage categories including crack, concrete spalling, rockfall and surface features of bearing deformation are manually annotated by two professional bridge-inspection engineers with LabelMe software. Ground-truth values of crack width, spalling area and bearing deformation are acquired through total-station and high-precision laser-ruler field measurements at real mountain-bridge sites. Structural-stress ground-truth originates from strain sensor readings of test components. Manual annotation error for image masks is within 3 pixels, and physical measurement uncertainty for crack width reaches ±0.5 μm. The reported 0.68 μm prediction error of crack width is calculated on scenes with high-precision on-site reference measurements. The image dataset is split into training set and test set at a ratio of 7:3 by whole bridge-scenario units instead of random sampling of individual images. All images collected from one specific mountain-bridge scenario are entirely allocated to either the training set or the test set. Images originating from the same bridge-scenario do not appear in both subsets, avoiding data leakage induced by inter-correlated samples from identical bridges, monitoring periods or damage events. The ultrasonic guided-wave dataset follows the identical scenario-based splitting rule. The experimental hardware adopted equipment with CPU of Intel Core i9-13900K (3.0GHz) and GPU of NVIDIA RTX 4090 (24GB video memory), with memory and storage of 64GB and 2TB SSD respectively. In terms of software, it was based on Ubuntu 22.04 system, used Python 3.9 as the development language and employed PyTorch 2.0 deep learning framework, adopted OpenCV 4.8.0 to process image data, and used Matplotlib 3.7.1 for visualization of results. The bounding box regression loss weight and Mask segmentation loss weight were 1.0 and 2.0 respectively, with 100 iterations. The learning rate decayed linearly in the first 50 epochs, and the learning rate for the last 50 epochs was 1e-5. Detailed training hyper-parameter settings are summarized in Table 1 for full experimental reproducibility.

There is no feature fusion inside Mask R-CNN network backbone between optical image and UGW signals; multi-source fusion is implemented only in the post-processing disaster-quantification module. All test images adopt the resolution of 1920 × 1080 pixels. Under the given experimental hardware, the average time consumption of each link is as follows: image pre-processing takes

28 ms, model inference takes 53 ms, post-processing including mask morphological filtering and multi-source data fusion takes 41 ms. In offline laboratory testing without field data-transmission overhead, the total processing latency is 122 ms. Data-transmission delay is not included in this offline test; in practical on-site deployment, additional sensor-to-device transmission latency will be introduced. The reported sub-one-second early-warning response time refers to offline laboratory processing latency excluding remote wireless transmission. The study first compared the Pearson correlation coefficients of each model under different early warning lead times, and the results were shown in Figure 7.

Table 1. Comparison of smooth substrate classes relevant to adhesive debonding

Hyper-parameter	Value
Optimizer	SGD with momentum (momentum=0.9)
Weight decay	0.0001
Total epochs	100
Batch size	2
Initial learning rate	0.001
Learning rate scheduler	Linear decay: decay in first 50 epochs, final lr =1e-5 for epoch 51-100
Anchor box aspect ratios	[0.5, 1, 2]
NMS threshold (proposal)	0.7
NMS threshold (detection output)	0.5
RoIAlign sampling points	4
Classification loss weight	1.0
Bounding-box regression loss weight	1.0
Mask segmentation loss weight	2.0

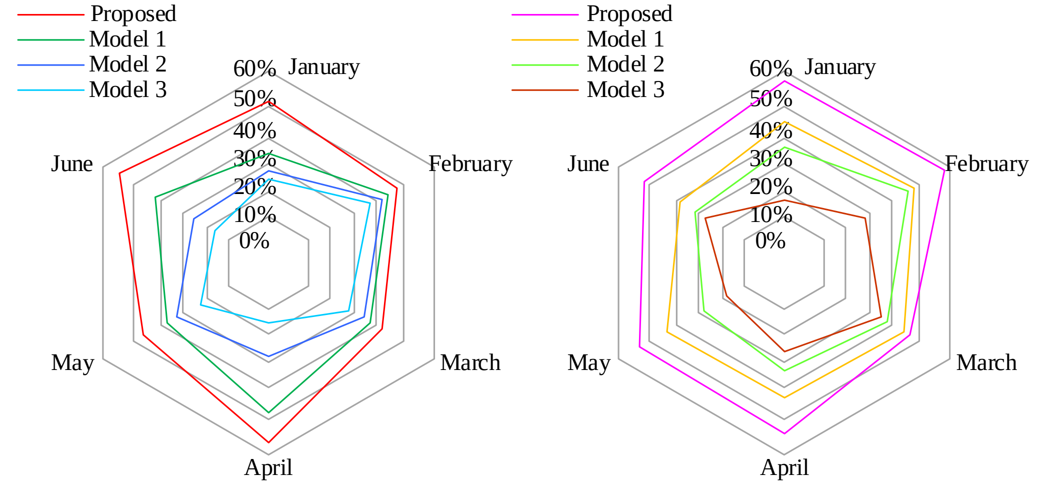


(a) Comparison of correlation coefficients of each model in the training set

(b) Comparison of correlation coefficients of each model in the test set

Fig. 7. Comparison of correlation coefficients

Figure 7(a) showed that in the training set, when the early warning lead time was 144 h, the Pearson correlation coefficient of the proposed model was 0.916. At this time, the Pearson correlation coefficients of Model 1 and Model 2 were 0.682 and 0.547, and the Pearson correlation coefficient of Model 3 was 0.473. In Figure 7(b), the Pearson correlation coefficient of the proposed model remained at a high level under different early warning lead times and was more stable compared to the comparison models. The above results indicated that the proposed model performed best in terms of correlation in disaster early warning and demonstrated stronger stability and precision. Subsequently, the study tested the economic loss reduction of the four models in mountain bridge disaster early warning tasks, and the results were shown in Figure 8. The percentage of economic-loss reduction reflects the relative improvement of expected economic loss under the model-driven early-warning mode compared with the traditional periodic manual-inspection warning mode. The baseline reference corresponds to conventional regular manual inspection for mountain bridges. The comprehensive economic loss covers bridge maintenance expenses, traffic interruption losses and structural repair costs. Relevant economic parameters adopt statistical data of mountain-bridge operation and maintenance in Southwest China. This indicator is calculated under fixed assumptions of disaster occurrence probability and unit repair cost, and it only represents relative improvement effect instead of absolute monetary loss value.

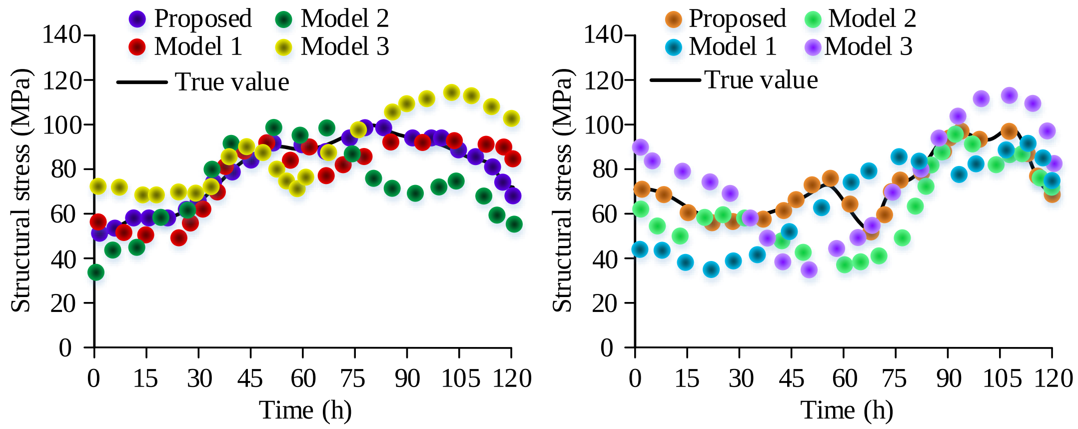


(a) The percentage reduction in economic loss for each model in the training set

(b) The percentage reduction in economic losses of each model in the test set

Fig. 8. Economic loss reduction in training set and test set

In Figure 8(a), in the training set, the economic loss reduction of the proposed model reached 51.23% in January. The economic loss of Model 1 decreased by 34.78% in January, while Model 2 and Model 3 decreased by 28.64% and 24.89% respectively, all lower than the proposed model. This indicated that the proposed model had a positive effect on reducing the economic impact of disasters. As shown in Figure 8(b), the proposed model had the highest economic loss reduction percentage in different months in the test set, especially in February, reaching 58.46%, while the highest economic loss reduction percentage of the comparison models was only 48.21%. The above results showed that the proposed model could more effectively reduce the economic losses caused by mountain bridge disasters. Subsequently, the study compared the prediction results of bridge structural stress for the four models, as shown in Figure 9.



(a) The stress prediction results of bridge structures for each model in the training set

(b) The stress prediction results of bridge structures for each model in the test set

Fig. 9. Bridge structural stress prediction results of each model

From Figure 9, it could be seen that the bridge structural stress prediction values of the proposed model were closer to the true values. In the training set, the proposed model exhibited only a 0.53 MPa deviation between the predicted and actual bridge structural stress at 90 h, whereas Model 2 showed a much larger deviation of 23.47 MPa. In the test set, the prediction points of the proposed model closely fitted the true value curve and performed better in capturing stress patterns. From this, it could be known that the proposed model had more superiority in the stress prediction task of mountain bridge disaster early warning and could provide more reliable technical support for structural safety assessment. Finally, the study compared the mountain bridge disaster early warning effects of the four models, and the results were shown in Table 2.

Table 2. Comparison of mountain bridge disaster early warning effects

Dataset	Index	Proposed	Model 1	Model 2	Model 3
Training	Disaster early warning response time (s)	0.85	1.23	1.07	1.52
	Early warning false alarm rate (%)	3.21	8.57	11.34	15.68
	Rockfall early warning recall rate (%)	92.35	81.62	75.88	68.45
Test	Disaster early warning response time (s)	0.69	1.35	1.18	1.62
	Early warning false alarm rate (%)	2.56	8.98	12.84	16.82
	Rockfall early warning recall rate (%)	93.64	80.54	73.25	65.84

From Table 2, it could be seen that in the training set, the disaster early warning response time of the proposed model was only 0.85 s, and its bearing deformation over-threshold early warning false alarm rate and small-volume rockfall early warning recall rate were 3.21% and 92.35% respectively, all superior to the comparison models. In the test set, all performance indicators of the proposed model were still superior to the comparison models, and its small-volume rockfall early warning recall rate was the highest, reaching 93.64%. Here, the recall rates of Model 1 and Model 2 were 80.54% and 73.25% respectively, and the recall rate of Model 3 was the lowest at 65.84%. The above results indicated that the proposed model could quickly and accurately provide early warning for mountain bridge disasters and had significant efficiency.

4.2 Analysis of Model Practical Application Effects

To further verify the performance effects of the disaster early warning model in practical applications, the study selected the public dataset from the Mountain Bridge Structure Health Monitoring Laboratory of Chongqing Jiaotong University to conduct performance tests on each model. The collection of this dataset strictly matched the actual service scenarios of mountain bridges and was synchronously associated with disaster quantification annotation information of key parts of bridges. The study first compared the crack width prediction results of different models in mountain bridge disaster early warning tasks, and the results were shown in Figure 10.

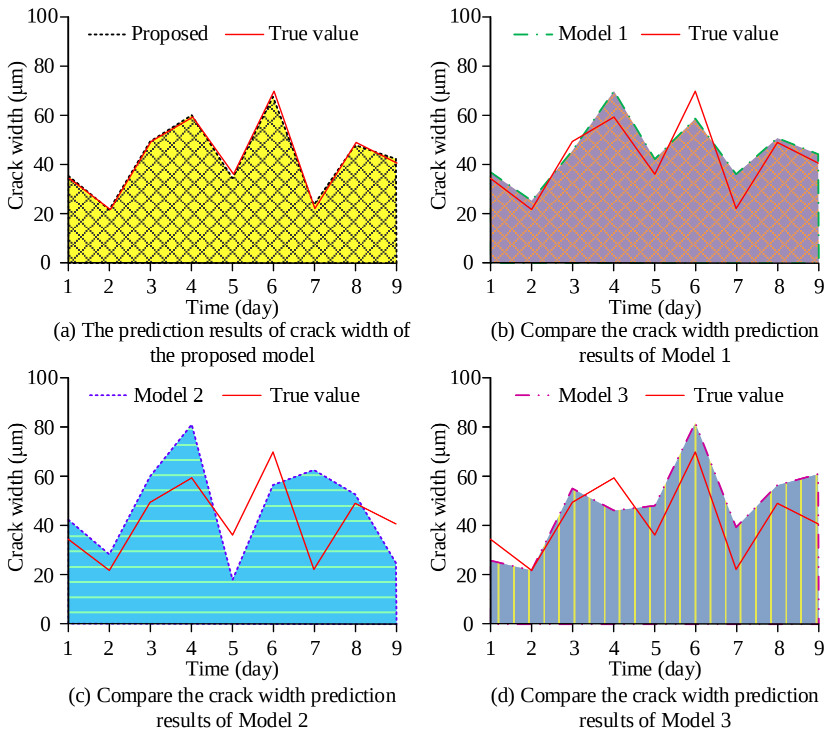


Fig. 10. Comparison of crack width prediction results

In Figure 10, when the time was at the 9th day, the proposed model showed a difference of only 0.68 μm between the predicted and actual crack widths. The prediction range of Model 1 was relatively wide, and the difference reached 1.53 μm. The prediction values of Model 2 differed greatly from the true values in most time periods. The prediction values of Model 3 fluctuated greatly, and when the time was at the 9th day, the difference reached 20.06 μm. The above results indicated that the proposed model could more accurately provide key data on crack development for bridge disaster early warning. Next, the study tested the monitoring accuracy of each model for two indicators: bridge bearing deformation and the proportion of concrete spalling area on the bridge surface, and the results were shown in Figure 11.

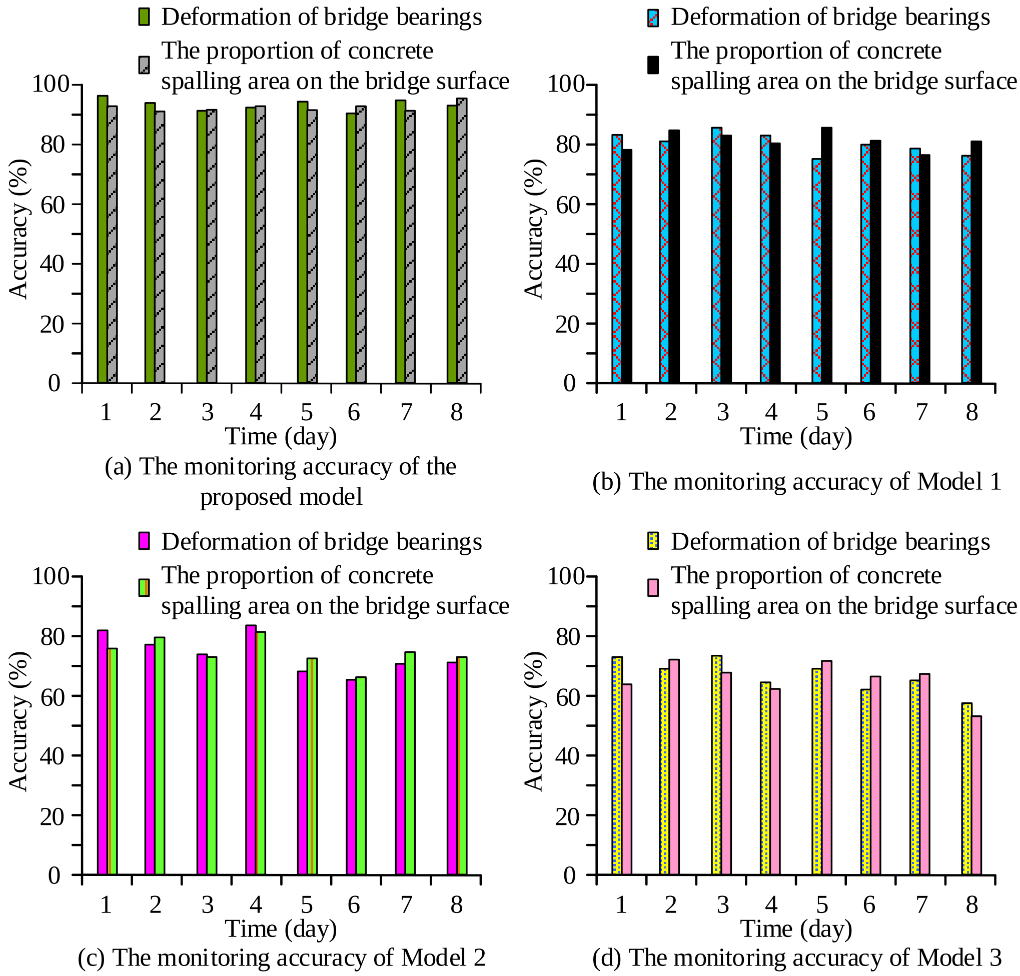


Fig. 11. Comparison of monitoring accuracy for different indicators

In Figure 11, the monitoring accuracy of the proposed model for both bridge bearing deformation and the proportion of concrete spalling area on the bridge surface reached over 90%. When the time was the 5th day, the monitoring accuracy of Model 1 for the proportion of concrete spalling area on the bridge surface was only 85.13%. From the perspective of bridge bearing deformation monitoring accuracy, when the time was at the 6th day, the monitoring accuracy of Model 2 and Model 3 were 64.91% and 62.82% respectively. The above data results indicated that the proposed model could more accurately monitor mountain bridge bearing deformation and concrete spalling on the bridge surface, and performed better in bridge structural damage monitoring tasks. Finally, the study compared the comprehensive performance of each model in mountain bridge disaster early warning tasks, and the results were shown in Table 3.

From Table 3, it could be seen that from the perspective of model inference speed, the proposed model reached 18.72 fps, and its crack identification accuracy in weak light environments was also the highest at 88.56%. Meanwhile, the bridge bearing deformation prediction error of the proposed model was only 0.32mm, indicating that this model could effectively ensure the accuracy of structural safety assessment.

Table 3. Comparison of comprehensive disaster early warning indicator results

Indicator	Proposed model	Model 1	Model 2	Model 3
Inference speed (fps)	18.72	15.36	16.89	12.75
mAP@0.5 (%)	92.17	84.33	80.21	74.58
mAP@0.5:0.95 (%)	76.42	67.25	63.14	58.77
Crack identification accuracy in weak light environments (%)	88.56	76.23	72.91	69.37
Bridge bearing deformation prediction error (mm)	0.32	0.85	1.21	1.57
F1 score for concurrent multi-disaster identification	0.93	0.82	0.75	0.68

The proposed model had the highest F1 score for concurrent multi-disaster identification at 0.93, while the concurrent multi-disaster identification F1 score of Model 3 was only 0.68. In terms of standard instance segmentation evaluation on test set, the proposed model achieves mAP@0.5 of 92.17% and mAP@0.5:0.95 of 76.42%, demonstrating its competitive capability in accurate boundary localization for bridge disaster targets. The above data results indicated that the proposed model had superior comprehensive performance and had good accuracy and efficiency in mountain bridge disaster early warning. Detailed precision, recall and pixel-level F1 for each bridge defect category are listed in Table 4. YOLO11-seg is introduced as a modern one-stage instance-segmentation baseline for comparative evaluation to analyze the trade-off between mask boundary accuracy and inference speed.

Table 4. Instance-segmentation performance comparison on test set

Damage category	Metric	Proposed model	Model 1	Model 2	Model 3	YOLO11-seg	Swin-Transformer-Mask R-CNN
Crack	Precision (%)	90.12	81.77	77.24	70.16	85.41	87.33
	Recall (%)	92.88	79.64	72.17	64.22	86.27	89.15
	Pixel-level F1	0.89	0.79	0.74	0.67	0.83	0.85
Concrete spalling	Precision (%)	91.34	82.45	78.31	71.25	86.52	88.46
	Recall (%)	94.15	81.22	74.36	66.13	87.44	90.31
	Pixel-level F1	0.91	0.80	0.76	0.69	0.84	0.86
Rockfall	Precision (%)	89.66	80.83	76.42	69.47	84.77	86.62
	Recall (%)	93.64	80.54	73.25	65.84	87.52	90.33
	Pixel-level F1	0.88	0.78	0.73	0.66	0.82	0.84
Bearing-deformation-related surface feature	Precision (%)	88.74	79.91	75.58	68.61	83.94	85.77
	Recall (%)	91.26	78.43	71.08	63.17	84.86	87.42
	Pixel-level F1	0.87	0.77	0.72	0.65	0.81	0.83

Five independent repeated training and evaluation runs are conducted with different random initialization. Key segmentation metrics exhibit limited fluctuation: the standard deviation of Mask-AP is 0.72 %, and the standard deviation of pixel-level F1 is 0.013. Such small variability demonstrates the stability of the proposed model under the given dataset and hyper-parameter setting. From Table 4, per-defect-category metrics consistently show that the proposed two-stage

CBAM-ResNet-Mask R-CNN achieves higher pixel-level F1 than YOLO11-seg across all damage categories, which demonstrates its advantages in mask boundary segmentation accuracy benefited from RoI-Align operation. By comparison, YOLO11-seg as a modern one-stage model exhibits faster inference speed but suffers degradation in fine boundary segmentation performance. Swin-Transformer-Mask R-CNN obtains competitive segmentation results, yet it brings heavier computational overhead. For mountain-bridge scenes with small-size and irregular defects, higher-quality mask boundaries are critical for subsequent geometric quantification of crack width and rockfall area.

5. Summary

This study addresses the adaptability and real time engineering requirements for mountain bridge structural defect monitoring. A multi modal early warning framework based on CBAM enhanced ResNet Mask R CNN is proposed for instance segmentation and quantitative evaluation of bridge local defects including cracks, concrete spalling and rockfall impact. Benefiting from dual path residual blocks and CBAM attention enhancement, the proposed method suppresses complex mountain background interference and improves detection performance for weak small-scale defects. Under test set conditions, the structural stress prediction deviation reaches only 0.53 MPa, and the recall rate for small scale rockfall events achieves 93.64%. By fusing visual segmentation outputs and ultrasonic guided wave time series data, the framework realizes the mapping from pixel level defect geometry to structural state indicators and provides technical support for mountain bridge early warning practice.

This work acknowledges that insufficient depth of multi-source data fusion may lead to incomplete disaster-trend prediction. Beyond that, this method exists inherent hardware-driven limitations originating from monocular optical imaging. Without active depth sensors such as LiDAR or RGB-D cameras and stereo-vision systems, accurate absolute three-dimensional rockfall-volume quantification cannot be guaranteed. Volume estimation calculated from 2-D segmentation masks relies on empirical statistical parameters and will produce non-negligible estimation uncertainty. In addition, the optical-vision-dominated framework is vulnerable to harsh mountain environmental conditions. Heavy fog, rainfall and snow cover will degrade image quality and may reduce the detection and segmentation reliability of visual models. Although ultrasonic guided-wave sensor signals are introduced as auxiliary inputs, they cannot fully compensate the performance drop of visual branches under extreme weather.

Future work will further construct correlation models between multi-dimensional monitoring data and structural-defect evolution trends. To tackle the insufficient multi-source-fusion depth, two promising computational frameworks will be considered. The structural-health-monitoring digital-twin (SHMDT) framework can couple visual defect measurements with physical structural simulation, mapping short-term visual damage observations to long-term structural bearing-capacity evolution. Physics-informed neural networks (PINNs) will also be adopted, embedding structural mechanical prior knowledge into data-driven model training, to reduce the dependence of quantitative assessment on purely empirical fitting relations. Further field-site tests under foggy, rainy and snowy weather will also be carried out to improve the robustness of the whole early-warning system.

Fundings

The research is supported by: Basic scientific research project of Liaoning Provincial Department of Education, Application of geological disaster early warning and prevention in the transportation industry, No.: LJ212413201002

References

[1] Han D, Yang G, Lu W, Huang M, Liu S. A multi-level damage assessment model based on change detection technology in remote sensing images. Nat Hazards. 2025;121(6):7367-88. https://doi.org/10.1007/s11069-024-07038-0

[2] Zhang H, Zheng J, Liu C, Dong J, Liu Y, Peng T. Deep learning-based automatic identification of gust fronts from weather radar data. J Meteorol Res. 2024;38(6):1021-31. https://doi.org/10.1007/s13351-024-4067-1

[3] Qiu Q, Xie Z, Zhang D, Ma K, Tao L, Tan Y, Zhang Z, Jiang B. Knowledge graph for identifying geological disasters by integrating computer vision with ontology. J Earth Sci. 2023;34(5):1418-32. https://doi.org/10.1007/s12583-022-1718-4

[4] Preethi P, Mamatha HR. Region-based convolutional neural network for segmenting text in epigraphical images. Artif Intell Appl. 2023;1(2):103-11. https://doi.org/10.47852/bonviewAIA3202905

[5] Cai Z, Jiang R, Lian X, Yang C, Wang Z, Fan Z. Forecasting citywide crowd transition process via convolutional recurrent neural networks. IEEE Trans Mobile Comput. 2024;23(5):5433-45. https://doi.org/10.1109/TMC.2023.3302636

[6] Song S, Zhang S, Dong W, Li G, Pan C. Multi-source information fusion meta-learning network with convolutional block attention module for bearing fault diagnosis under limited dataset. Struct Health Monit. 2024;23(2):818-35. https://doi.org/10.1177/14759217231174987

[7] Yasmin T, Khamis K, Ross A, Sen S, Sharma A, Sen D, Sen S, Buytaert W, Hannah DM. Brief communication: Inclusiveness in designing an early warning system for flood resilience. Nat Hazards Earth Syst Sci. 2023;23(2):667-74. https://doi.org/10.5194/nhess-23-667-2023

[8] Sabato A, Dabetwar S, Kulkarni NN, Fortino G. Noncontact sensing techniques for AI-aided structural health monitoring: A systematic review. IEEE Sens J. 2023;23(5):4672-84. https://doi.org/10.1109/JSEN.2023.3239851

[9] Jiao P, Ye X, Zhang C, Li W, Wang H. Vision-based real-time marine and offshore structural health monitoring system using underwater robots. Comput Aided Civ Inf. 2024;39(2):281-99. https://doi.org/10.1111/mice.13063

[10] Teng S, Chen G, Yan Z, Cheng L, Bassir D. Vibration-based structural damage detection using 1-D convolutional neural network and transfer learning. Struct Health Monit. 2023;22(4):2888-909. https://doi.org/10.1177/14759217221147775

[11] Luleci F, Catbas FN, Avci O. Generative adversarial networks for labeled acceleration data augmentation for structural damage detection. J Civil Struct Health Monit. 2023;13(1):181-98. https://doi.org/10.1007/s13349-022-00630-y

[12] Nicoletti V, Quarchioni S, Amico L, Gara F. Assessment of different optimal sensor placement methods for dynamic monitoring of civil structures and infrastructures. Struct Infrastructure Eng. 2026;22(7):1299-314. https://doi.org/10.1080/15732479.2025.2458428

[13] Macchiarulo V, Milillo P, Blenkinsopp C, Giardina G. Monitoring deformations of infrastructure networks: A fully automated GIS integration and analysis of InSAR time-series. Struct Health Monit. 2022;21(4):1849-78. https://doi.org/10.1177/14759217211044738

[14] Setrayana A, Purwanto E. Comparative analysis of Yolov11 and Mask R-CNN models for river water level detection. J Comput Netw Archit High Perform Comput. 2025;7(3):934-47. https://doi.org/10.47709/cnahpc.v7i3.5936

[15] Yazdanpanah O, Bakhshi EA, Chang M. Post-seismic structural damage segmentation using YOLOv8-seg model. Smart Struct Syst. 2025;36(4):195-202. https://doi.org/10.12989/sss.2025.36.4.195

[16] Huang Y, Xie Y, Liu W, Ma Y, Miao F, Zhang G. Multi-scale convolution networks for seismic event classification with windowed self-attention. J Seismol. 2025;29(1):257-68. https://doi.org/10.1007/s10950-024-10255-y

[17] Wenxuan Z, Yaqin Z, Zhaoxiang Z, Ao L. Lite transformer network with long–short range attention for real-time fire detection. Fire Technol. 2023;59(6):3231-53. https://doi.org/10.1007/s10694-023-01469-8

[18] Yue Y. Construction risk of highway bridges in plateau and mountainous areas based on safety risk assessment. J Comput Method Sci Eng. 2024;24(3):1617-30. https://doi.org/10.3233/JCM-247271

[19] Yang P, Xu Z, Yan X, Wang X. Establishing a rainfall dual-threshold for flash flood early warning considering rainfall patterns in mountainous catchment, China. Nat Hazards. 2024;120(7):6657-84. https://doi.org/10.1007/s11069-024-06489-9

[20] Rawat D, Sharma ML, Varade D, Kumar R, Kanungo DP, Ahmed R, Gupta SC, Singh H, Saxena N. Early warning potential of regional seismic network: Seismic assessment of one of the precursors of Chamoli 2021 disaster. Earth Syst Environ. 2024;8(1):85-104. https://doi.org/10.1007/s41748-023-00334-z

[21] Zhao Y, Cai J, Tang L, Li H, Hu S, Xing H. Hierarchical and networked analysis of resilience factors in mountain communities in Southwest China. Nat Hazards. 2024;120(2):1519-56. https://doi.org/10.1007/s11069-023-06249-1

[22] Ding Y, Wu Y, Zhu Q, Zhang L, Sun Q, Wang W. Virtual geographic environment-based integrated rockfall risk simulation method for canyon bridges. Trans GIS. 2023;27(3):797-820. https://doi.org/10.1111/tgis.13049

[23] Firdausi HM, Utomo SB, Widjonarko W. Smart camera for volcano eruption early warning system based on Faster R-CNN and YOLO. Rekayasa. 2025;18(1):49-63. https://doi.org/10.21107/rekayasa.v18i1.23848
[24] Wang B, Wang J, Xu X, Bao X. Gas mask wearing detection based on Faster R-CNN. J Ambient Intell Smart Environ. 2024;16(1):57-71. https://doi.org/10.3233/AIS-230081
[25] Natha SAS, Ali MARL, Zia SS, Buksh SK, Usmani IA. Deep learning-based predictive analytics for weapon detection: A fusion of FMR-CNN and YOLOv9 Models. VFAST Trans Softw Eng. 2025;13(2):255-72. https://doi.org/10.21015/vtse.v13i2.2132
[26] Khan I, Guo Z, Lim K, Kim J, Kwon YW. Assessment of indoor risk through deep learning-based object recognition in disaster situations. Multimed Tools Appl. 2024;83(12):34669-90. https://doi.org/10.1007/s11042-023-16912-3
[27] Milanez CE, Valadao CT, de Almeida GM, Cuadros MA. Use of mask R-CNN for detection and control of slag scraper wear. Ironmaking Steelmaking. 2023;50(11):1698-706. https://doi.org/10.1080/03019233.2023.2215886