Received: 8 November 2023 | Accepted: 15 April 2024

DOI: 10.1111/mice.13219

INDUSTRIAL APPLICATION

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING

WILEY

Lightweight object detection network for multi-damage recognition of concrete bridges in complex environments

Tianyong Jiang[1] | **Lingyun Li**[1] | **Bijan Samali**[2] | **Yang Yu**[2,3] | **Ke Huang**[1] | **Wanli Yan**[1] | **Lei Wang**[1]

[1]School of Civil Engineering, Changsha University of Science & Technology, Changsha, China

[2]Centre for Infrastructure Engineering, Western Sydney University, Penrith, New South Wales, Australia

[3]Centre for Infrastructure Engineering and Safety, School of Civil and Environmental Engineering, The University of New South Wales, Sydney, New South Wales, Australia

Correspondence
Yang Yu, Centre for Infrastructure Engineering and Safety, School of Civil and Environmental Engineering, The University of New South Wales, Sydney, NSW, Australia.
Email: yang.yu12@unsw.edu.au

Lei Wang, School of Civil Engineering, Changsha University of Science & Technology, Changsha, China.
Email: Leiwang@csust.edu.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2019YFC1511000; National Natural Science Foundation of China, Grant/Award Number: 52378123; Graduate Research Innovation Project of Hunan Province is Hunan Provincial Department of Education, China, Grant/Award Number: CX20220879

Abstract

To solve the challenges of low recognition accuracy, slow speed, and weak generalization ability inherent in traditional methods for multi-damage recognition of concrete bridges, this paper proposed an efficient lightweight damage recognition model, constructed by improving the you only look once v4 (YOLOv4) with MobileNetv3 and fused inverted residual blocks, named YOLOMF. First, a novel lightweight network named MobileNetv3 with fused inverted residual (MobileNetv3-FusedIR) is constructed as the backbone network for YOLOMF. This is achieved by integrating the fused mobile inverted bottleneck convolution (Fused-MBConv) into the shallow layers of MobileNetv3. Second, the standard convolution in YOLOv4 is replaced with the depthwise separable convolution, resulting in a reduction in the number of parameters and complexity of the model. Third, the effects of different activation functions on the damage recognition performance of YOLOMF are thoroughly investigated. Finally, to verify the effectiveness of the proposed method in complex environments, a data enhancement library named Imgaug is used to simulate concrete bridge damage images under challenging conditions such as motion blur, fog, rain, snow, noise, and color variations. The results indicate that the YOLOMF shows excellent multi-damage recognition proficiency for concrete bridges across varying field-of-view sizes as well as complex environmental conditions. The detection speed of YOLOMF reaches 85f/s, facilitating effective real-time multi-damage detection for concrete bridges under complex environments.

1 | INTRODUCTION

Concrete bridges play an essential role in modern transportation systems and road networks. Currently, China boasts a total of 1.032 million highway bridges, including 8816 long-span bridges and 159,600 large bridges (Ministry of Transport of the People's Republic of China, 2022). However, as time goes on, concrete bridges inevitably

experience varying degrees of damage due to factors such as temperature, humidity, loading, and degradation of concrete materials (Pezeshki, Adeli, et al., 2023; Rafiei et al., 2017a, 2017b). Structural damage, such as cracks, concrete spalling, and exposed rebar, can be considered as clear indicators of potential degradation or structural failure (Cha et al., 2018; Koch et al., 2015; C. Zhang et al., 2020). Therefore, the early detection of potential damage of concrete bridges, combined with the implementation of appropriate protective measures, is of paramount significance. These measures not only enhance the structural reliability and service life of the bridges but also reduce maintenance and replacement costs, as well as associated risks.

Damage recognition methods based on vibration information have played a pivotal role within bridge structural health monitoring systems (Amezquita-Sanchez & Adeli, 2019; Perez-Ramirez et al., 2016; Pezeshki, Adeli, et al., 2023). These methods utilize vibration information such as structural frequency, mode, and flexibility to establish damage recognition indicators, enabling qualitative and quantitative evaluation of bridge damage (Amezquita-Sanchez & Adeli, 2016; Altunışık et al., 2019; Ciambella et al., 2019; Pezeshki, Pavlou, et al., 2023). To acquire accurate structural vibration information, it is often necessary to deploy a considerable number of contact sensors. However, challenges arise due to the complex monitoring environment and measurement noise, impeding the acquisition of precise structural vibration data. Meanwhile, non-destructive testing based on piezoelectric sensors (Jiang et al., 2018), ultrasonic detection (Mutlib et al., 2016), fiber optic sensors (Li et al., 2020), and acoustic emission (Verstrynge et al., 2021) have been used to detect the presence of damage such as cracks and voids inside bridge structures. Nonetheless, the deployment of these diagnostic methodologies requires specialized technical expertise and the utilization of dedicated sensors to assess the structural health of bridges. The maintenance of these sensors often poses considerable challenges, thereby circumscribing their range of application. This limitation is especially pronounced in the context of health surveillance for large-scale concrete structures, including extensive span bridges and dams, where significant impediments are frequently encountered (S. Yu et al., 2021).

In recent years, the rapid development of artificial intelligence methods and photogrammetry technology has brought new solutions to the performance evaluation and damage detection of concrete bridge structures (Perez-Ramirez et al., 2019; Rafiei et al., 2016; Wu et al., 2022). Javadinasab et al. (2021) presented the prospects for the development of integrated structural control and health monitoring systems in future smart cities. J. Zhang et al. (2022) used optical cameras and video recording equipment to measure the displacement of long-span bridges. Chou et al. (2022) combined a multi-layer image pyramid with operational modal analysis to obtain the modal characteristics of the model using captured images. Image analysis has become an effective means of efficient structural vibration information measurement and health condition assessment. Consequently, researchers have directed their efforts toward extracting valuable information for structural health monitoring and damage assessment from raw image data (Duque et al., 2018; Seo et al., 2018; Spencer et al., 2019). Computer vision technology has gained significant popularity in the crack detection of bridges, owing to its advantages of being non-contact and low-cost (Dong & Catbas, 2021). Payab et al. (2019) conducted a comprehensive investigation into the recognition of key parameters of cracks on the concrete bridge surface, including the distribution, width, and length. Y. F. Liu et al. (2016) proposed a crack projection recognition method by combining two-dimensional digital image technology with three-dimensional reconstruction technology, addressing the challenging problem of deformation correction in quantitative crack recognition. However, it is worth noting that these methods often necessitate close-range imaging, focusing primarily on small areas of damage. Furthermore, the recognition results are susceptible to background and lighting variations within the imaging environment.

Recently, deep learning (DL) has experienced significant advancements, particularly in the realm of convolutional neural networks (CNNs). These networks have found extensive application in image recognition and classification tasks, showing tremendous potential in the field of damage recognition in civil engineering and infrastructure (Bao & Li, 2021). Cha et al. (2017) compiled a dataset of 40,000 images depicting concrete structural cracks and compared a CNN-based crack detection model with traditional digital image techniques, confirming the reliability of DL in structural damage recognition. Y. Yu, Rashidi, et al. (2022) integrated the revised chicken swarm optimization algorithm into CNN to optimize network structure, thereby improving the detection efficiency of the network. Furthermore, Y. Yu, Samali, et al. (2022) put forth an enhanced dempster shafer (DS) algorithm that integrated the recognition results of 15 pre-trained CNN models to enhance the precision of surface crack detection in concrete structures. H. Zhang et al. (2023) classified bridge images at three levels: structure, component, and surface damage based on the DL algorithm. The above studies have demonstrated the superiority of DL in automatic feature extraction and addressing non-linear classification problems.

Object detection algorithms represent one of the hot topics in DL (Jodas et al., 2022), with wide-ranging applications in various fields such as autonomous driving (Carranza-García et al., 2022), underwater detection

(Foresti et al., 2022), object tracking (Urdiales et al., 2023), and edge detection (Xian et al., 2023). Concurrently, they also demonstrate promising potential in real-time detection of structural surface damage. The you only look once (YOLO; Redmon et al., 2016), integrates target classification and positioning into a single convolutional network, enabling it to predict multiple bounding boxes and class probabilities simultaneously, thus achieving faster detection speeds. C. Zhang et al. (2020) proposed an enhanced YOLOv3 (Redmon & Farhadi, 2018) by introducing a novel transfer learning (TL; Pan & Yang, 2009) method for recognizing surface damage on concrete bridges. Yu et al. (2021) improved YOLOv4 (Bochkovskiy et al., 2020) by using a Focal Loss (Lin et al., 2017) and pruning algorithm and proposed an efficient YOLOV4-FPM, which was subsequently applied to the real-time crack detection of concrete bridges using unmanned aerial vehicle (UAV). Zou et al. (2022) proposed an improved YOLOv4-D for post-earthquake damage identification and reliability assessment of reinforced concrete structures, which can be used to preliminarily determine the degree of structural damage and failure mode. Zhao et al. (2022) used the Swin transformer as the backbone network for YOLOv5 (Jocher, 2020) and introduced coordinate attention modules to propose the YOLOv5s-HSC algorithm, which was applied to damage detection in concrete dams. Recently, Gao et al. (2023) introduced multi-task TL into the Transformer network and proposed a multi-attribute structural damage detection model, which showed superiority in multi-task damage detection. X. Xu and Li (2024) achieved satisfactory detection results by applying YOLOv7 (Wang et al., 2022) for pipeline weld surface defect detection. Sohaib et al., 2024) trained three different sizes of YOLOv8 (Jocher et al., 2023) models using various datasets and employed an ensemble strategy to establish a hybrid YOLOv8 model, enabling efficient detection and segmentation of concrete cracks. Nonetheless, the algorithms deployed in these studies exhibit significant reliance on extensive computational resources. Further, the paucity of research focusing on structural surface damage under complex environmental conditions, such as adverse weather, is conspicuous. Consequently, their effectiveness is curtailed when tasked with pinpointing large-scale damage in real-world scenarios fraught with complexity.

To overcome these challenges, this study adopts YOLOv4, a stable network architecture widely employed in damage identification for concrete structures, as the foundation to propose a novel lightweight object detection network. This network is designed specifically for detecting multi-damages of concrete bridges in complex environments. The main innovations of this paper can be summarized as follows: (1) establishment of a modified YOLOv4, namely, YOLOMF, where MobileNetv3-FusedIR is employed as the lightweight backbone network to enhance the inspection speed and recognition ability for minor damage. (2) Integration of depthwise separable convolution (DSC) in feature fusion networks, replacing standard convolution to decrease the amount of convolution layer parameters and enhance model damage detection accuracy. (3) Selection of optimal activation function combinations through investigating their effects on the accuracy of the damage recognition model, thus improving network feature extraction ability and damage recognition accuracy, and (4) verification of the validity and practicability of the proposed method by identifying the damage of concrete bridges in complex environments, including various field-of-view sizes (small, medium, and large) as well as challenges such as motion blur, fog, rain, snow, and noise disturbance.

2 | DAMAGE RECOGNITION MODEL FOR CONCRETE BRIDGES

2.1 | Overview of YOLOv4

The YOLOv4 model, combined many techniques with YOLOv3, aiming to optimize the recognition process to achieve an overall improvement in recognition accuracy and speed as demonstrated in Figure 1. The backbone network of YOLOv4 employs CSPDarknet53, which builds upon Darknet53 while incorporating the cross-stage partial network (Wang et al., 2020). This integration enhances the diversity and representation capability of image feature extraction. The CBM block, standing for convolution + batch normalization (Ioffe & Szegedy, 2015) + Mish (Misra, 2019), is included among them.

The neck network is utilized for feature fusion across multiple scales in YOLOv4, incorporating both spatial pyramid pooling (SPP; He et al., 2015) and path aggregation network (PANet; Liu et al., 2018). Within the SPP structure, three maximum pooling layers with kernel sizes of 5×5, 9×9, and 13×13 are employed along with a skip connection to significantly expand the receptive field. The PANet structure employs upsample and concat operations to enhance the semantic information and receptive field of features, resulting in a division of the original image into 52×52, 26×26, and 13×13 grids based on the proportion of feature mapping. This enables recognition of small, medium, and large objects. Additionally, the CBL block is a combination of convolution, batch normalization, and Leaky rectified linear units (LeakyReLU) layers (J. Xu et al., 2020). The head network employs an anchor-based regression classifier to calculate the position, category, and confidence information of each detected object, generating object detection results.

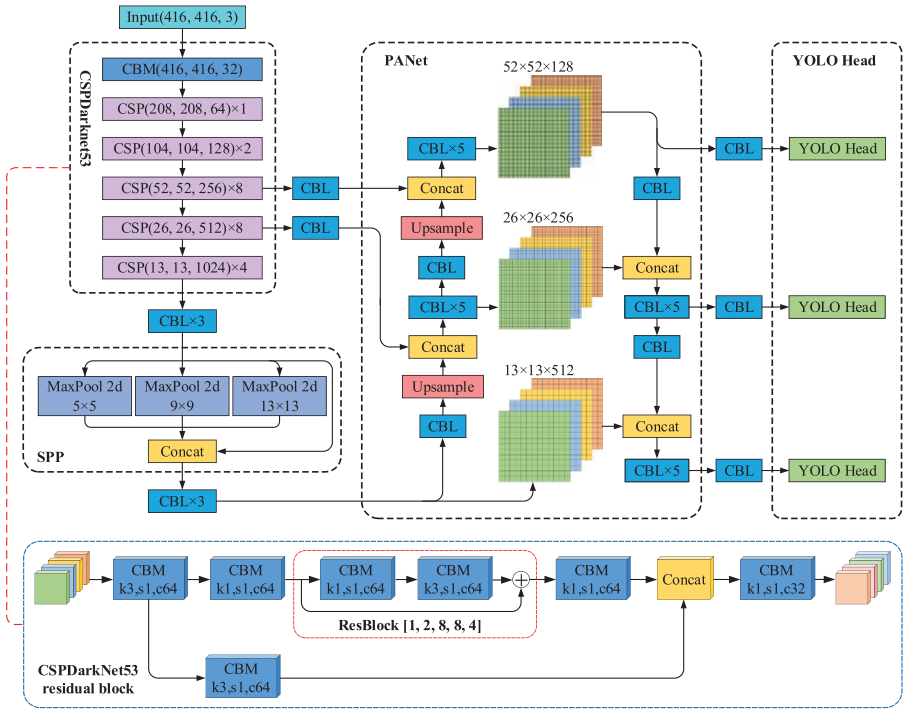

FIGURE 1 Structural diagram of the you only look once v4 (YOLOv4) network.

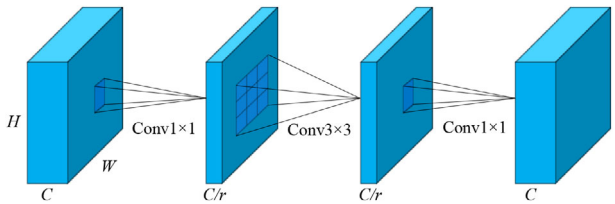

FIGURE 2 Bottleneck structure in the CBL × 3 block. CBL, convolution, batch normalization, and LeakyReLU.

The CBL × n block in YOLOv4 is composed of multiple bottleneck structures (He et al., 2016) as illustrated in Figure 2. The initial 1 × 1 convolution compresses the input feature map ($B \times C \times H \times W$) into ($B \times C/r \times H \times W$), followed by feature extraction through 3 × 3 convolution layers. Finally, a 1 × 1 convolution reshapes the resulting feature map to match the dimensions of the initial feature map. Incorporating a bottleneck structure into the network architecture offers notable advantages in terms of parameter reduction and network depth augmentation.

2.2 | Novel backbone based on MobileNetv3

The standard backbone network of the original YOLOv4 is CSPDarkNet53. Although it can achieve acceptable accuracy in object detection, the large number of model parameters leads to considerable processing time and hinders real-time onsite inspection for actual bridges. To enhance the detection speed and accuracy of YOLOv4 while reducing network complexity and parameter count, the backbone network is replaced with a novel backbone, which is developed by combining the strengths of MobileNetv3 (Howard et al., 2019) and Fused-MBConv (Yang et al., 2018; Zoph & Le, 2016). Additionally, DSC (Howard et al., 2017) is employed in place of standard convolution to reduce computational costs and improve network efficiency.

2.2.1 | DSC

The DSC comprises a depthwise convolution (DWC) layer and a pointwise convolution (PWC) layer as illustrated in Figure 3. First, DWC is conducted on the input features, enabling the acquisition of uncorrelated feature maps for each channel. Subsequently, the relevant feature information from channels that share the same spatial position within the feature maps is effectively integrated through PWC.

Assuming the input feature map with a size of $D_f \times D_f$, convolution kernel with size of $D_k \times D_k$, M input channels, and N output channels. The computation quantity for a standard convolution operation is $D_f \times D_f \times D_k \times D_k \times M \times N$, while the computation quantity of DWC is $D_f \times D_f \times M \times N$, and the computation quantity of PWC is $D_f \times D_f \times D_k \times D_k \times M$. The ratio between the computation quantity of DSC and stan-

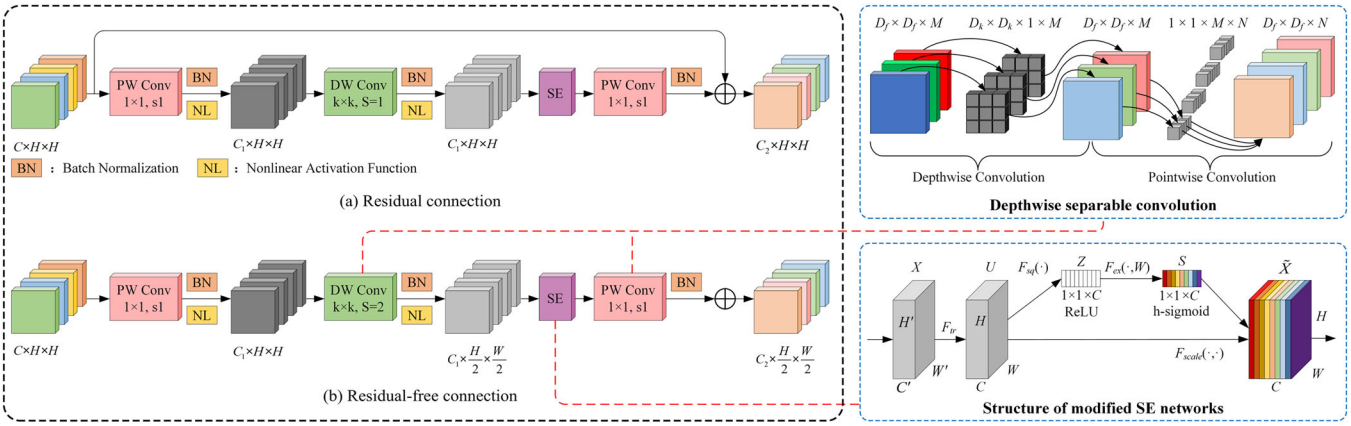

FIGURE 3 Description of the inverted residual blocks (IRBlock) architecture.

dard convolution can be calculated using the following Equation (1):

$$\frac{D_f \times D_f \times D_k \times D_k \times M + D_f \times D_f \times M \times N}{D_f \times D_f \times D_k \times D_k \times M \times N} = \frac{1}{D_k^2} + \frac{1}{N} \quad (1)$$

As shown in Equation (1), the main advantage of DSC is the reduction in model parameter count and computational complexity, leading to improved lightweight performance of the model. While maintaining a certain level of accuracy, it significantly reduces the model size and computational requirements. This characteristic renders the DSC particularly well-suited for application scenarios with limited computing resources, such as real-time detection using mobile devices like UAV.

2.2.2 | MobileNetv3

MobileNetv3 is a lightweight CNN proposed by Google, which optimizes and improves the network structure through network architecture search (Zoph & Le, 2016) and NetAdapt algorithm (Yang et al., 2018), based on MobileNetv1 (Howard et al., 2017) and MobileNetv2 (Sandler et al., 2018). It is primarily utilized for tasks like image classification and object detection on mobile and embedded devices.

The MobileNetv3 architecture is composed of a sequence of inverted residual blocks (IRBlock) featuring a linear bottleneck as shown in Figure 3. The bottleneck structure of the IRBlock is distinct from the traditional bottleneck structure depicted in Figure 2. First, channel C of the feature map is expanded to C_1 using a 1×1 convolution, followed by DWC with kernel size $k \times k$ (where MobileNetv3 adopts 3×3 and 5×5) to extract feature information. Finally, the quantity of channels is reduced from C_1 to C_2 using 1×1 convolution, where $C_1 > C \geq C_2$. The IRBlock has two forms, using a residual connection when the stride of the DWC is set to 1. The linear bottleneck is implemented by utilizing the linear activation function in the final 1× 1 convolution layer of IRBlock, which can avoid information loss due to nonlinear activation functions acting on low-dimensional features

To enhance the extraction capability of effective features, MobileNetv3 incorporates a modified squeeze and excitation network (SENet; Hu et al., 2018) after DWC processing as described in Figure 3.

As a channel attention mechanism (AM), the modified SENet primarily comprises the squeeze and excitation operations, which enable learning of importance weights for each feature channel, followed by reweighting of the feature channel. For the input feature layer X, the standard convolution operator F_{tr} is used to derive the feature layer U. The spatial dimension ($H \times W$) of each feature channel is compressed through global average pooling, condensing the information of each feature channel into a weight vector $\boldsymbol{Z}$.

To comprehensively capture the correlation between feature channels, the weight vector $\boldsymbol{Z}$ executes a dimension reduction layer with parameter W_1, followed by a ReLU activation layer and a dimension increase layer with parameter W_2. Finally, an H-Sigmoid activation function is applied to achieve the interrelated importance weight vector $\boldsymbol{S}$.

To update the feature layer, the weight value s_c of each feature channel in the acquired weight vector $\boldsymbol{S}$ is multiplied by its corresponding element u_c in the feature layer U. This multiplication operation ensures that each feature channel is appropriately adjusted based on its associated weight value, leading to the updated feature layer $\tilde{X}$.

TABLE 1 Parameter setting of the MobileNetv3-FusedIR.

Stage	Type	Kernel size	Stride	Channels	AM	NL	Layers
0	Conv2d	3 × 3	2	16	–	H-Swish	1
1	Fused-MBConv 1,	3 × 3	1	16	SE0.25	ReLU	1
2	Fused-MBConv 4, 3	3 × 3	2, 1	24	SE0.25	ReLU	2
3	Fused-MBConv 3, 3, 3	3 × 3	2, 1, 1	40	SE0.25	ReLU	3
4	IRBlock 6, 2.5, 2.3, 2.3	3 × 3	2, 1, 1, 1	80	SE0.25	H-Swish	4
5	IRBlock 6, 6	3 × 3	1, 1	112	SE0.25	H-Swish	2
6	IRBlock 6, 6, 6	5 × 5	2, 1, 1	160	SE0.25	H-Swish	3
7	Conv2d	1 × 1	1	960	–	H-Swish	1
8	Pool	7 × 7	1	–	–	–	1
9	Conv2d, NBN	1 × 1	1	1280	–	H-Swish	1
10	Conv2d, NBN	1 × 1	1	k	–	–	1

Note: Conv2d denotes convolution. Fused-MBConv 4, 3 means that the first 3 × 3 convolution layer of the Fused-MBConvBlock increases the number of channels for the input feature matrix by four and three times, respectively.
Abbreviations: AM, attention mechanism; IRBlock, inverted residual blocks; MobileNetv3-FusedIR, MobileNetv3 with fused inverted residual; NBN, no batch normalization; NL, nonlinear activation function.

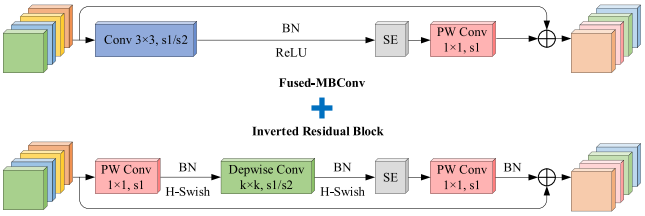

FIGURE 4 Structure of the MobileNetv3-FusedIR.

2.2.3 | Innovative lightweight backbone network

Although the DSC has demonstrated exceptional performance in lightweight CNNs, it often falls short of fully leveraging the capabilities of modern accelerator parallel computing, particularly in shallow network architectures (Tan & Le et al., 2021). Therefore, while MobileNetv3 exhibits satisfactory performance in object detection, its training speed and accuracy do not meet the desired level. To efficiently extract features from shallow network architectures, the Fused-MBConv (Gupta & Tan, 2019) has been proposed as a replacement for DSC in shallow networks of lightweight CNNs.

This study used a novel backbone network named MobileNetv3-FusedIR, which replaces the IRBlock in the shallow network of MobileNetv3 with Fused-MBConv, and by integrating the strengths of both IRBlock and Fuse-MBConv, the feature extraction process of the MobileNetv3 backbone network is optimized. The Fused-MBConv is obtained by replacing the 1 × 1 PWC and 3 × 3 DWC in IRBlock with the 3 × 3 standard convolution as depicted in Figure 4.

Table 1 presents the specific parameters of MobileNetv3-FusedIR, where Stages 1–3 are set to Fused-MBConv and Stages 4–6 are set to IRBlock. In addition, SE0.25 is used in Stages 1–6, representing the count of nodes in the first fully connected layer as 0.25 times that of the input feature matrix channels of the SENet, that is, the number of elements in weight vector $\boldsymbol{Z}$.

2.3 | Activation function in CNN

Activation functions play a crucial role in neural networks by introducing nonlinearity to the linear transformation within each layer. They are implemented as point-wise functions, operating on individual elements, to enable the network to capture complex relationships and nonlinear patterns in the data. In early literature, Sigmoid and TanH were extensively used; however, they subsequently proved ineffective in deep neural networks. Compared to Sigmoid and TanH, ReLU offers advantages such as simple computation, nonlinear and gradient stability, better generalization ability, and faster convergence. However, it also comes with disadvantages, including the presence of dead neurons, non-zero-centered output, and non-differentiability. Over the years, many activation functions have been proposed to address the drawbacks of ReLU. They include ReLU6 (Sandler et al., 2018), LeakyReLU (J. Xu et al., 2020), sigmoid linear unit (SiLU) (Ramachandran et al., 2017), and H-Swish (Howard et al., 2019) as shown in Figure 5, and these activation functions can be defined as Equations (2) to (6), respectively.

$$\mathrm{ReLU}(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \le 0 \end{cases} \tag{2}$$

$$\mathrm{ReLU6}(x) = \begin{cases} 6, & \text{if } x \ge 6 \\ x, & \text{if } 0 < x < 6 \\ 0, & \text{if } x \le 0 \end{cases} \tag{3}$$

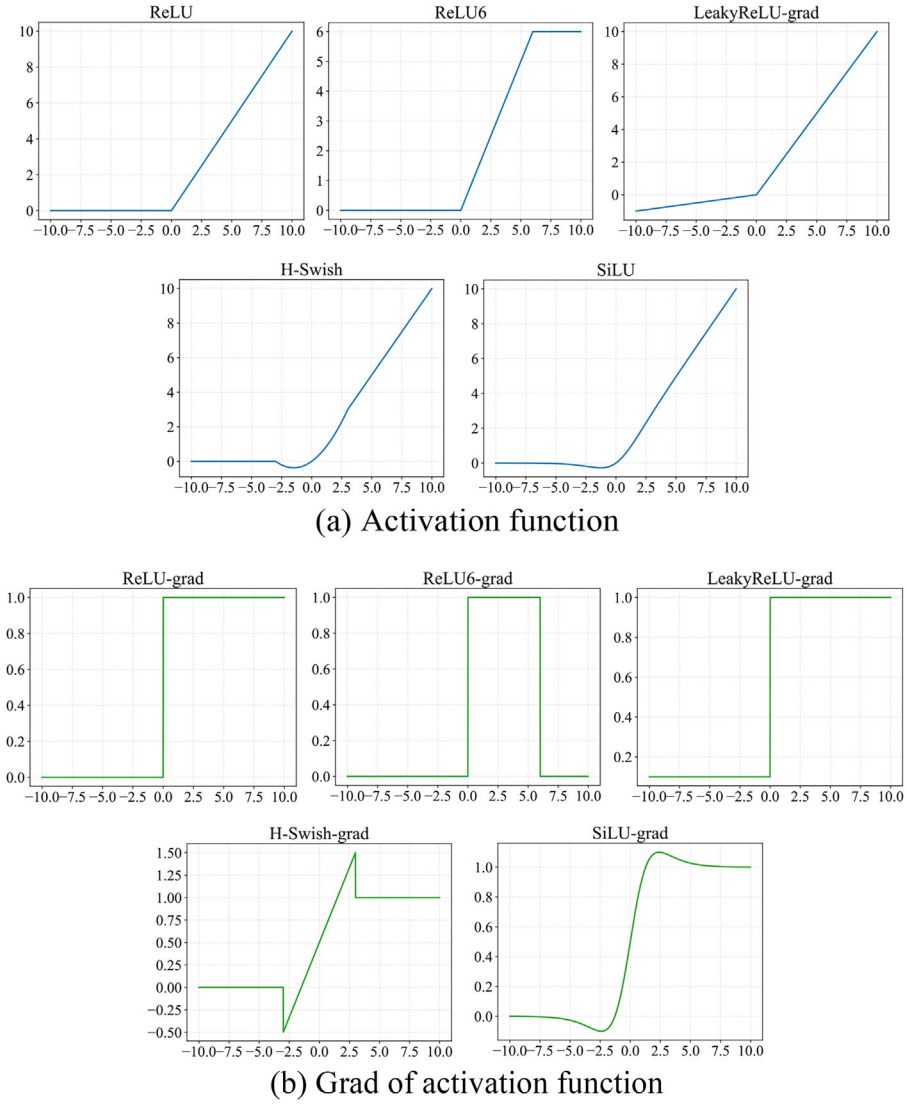

FIGURE 5 Different activation functions and their grad.

$$\text{LeakyReLU}(x) = \begin{cases} x, \text{ if } x > 0 \\ ax, \text{ if } x \le 0 \end{cases} \tag{4}$$

$$\text{H} - \text{Swish}(x) = x\frac{\text{ReLU6}(x+3)}{6} \tag{5}$$

$$\text{SiLU}(x) = x \cdot \text{Sigmoid}(x) = \frac{x}{1+e^{-x}} \tag{6}$$

The H-Swish solves the problems of gradient vanishing, sparsity, and discontinuity that are present in the ReLU. By improving model performance and efficiency, the expression and generalization capabilities of neural networks are enhanced. Moreover, it is especially well-suited for lightweight CNNs, which can effectively reduce the computational cost of mobile device detection.

This study discussed the influence of different activation functions on YOLOMF. The settings of the activation functions are shown in Table 2, where CB-NL × 3/CB-NL ×5 represents the bottleneck structure, and CB-NL represents the standard convolution block. This article replaces the 3 × 3 standard convolution in Figure 1 with DSC, which contributes to a more lightweight model.

TABLE 2 Activation functions setting in you only look once v4 (YOLOv4) with MobileNetv3 and fused inverted residual blocks (YOLOMF).

Models	CB-NL × 3/CB-NL × 5	CB-NL
YOLOMF1	ReLU6	ReLU6
YOLOMF2	Hard-Swish	LeakyReLU
YOLOMF3	Hard-Swish	SiLU

Abbreviation: CB-NL, standard convolution block.

2.4 | YOLOMF for concrete bridge damage recognition

This study aims to provide a comprehensive comparative analysis of three distinct backbone networks: DarkNet53,

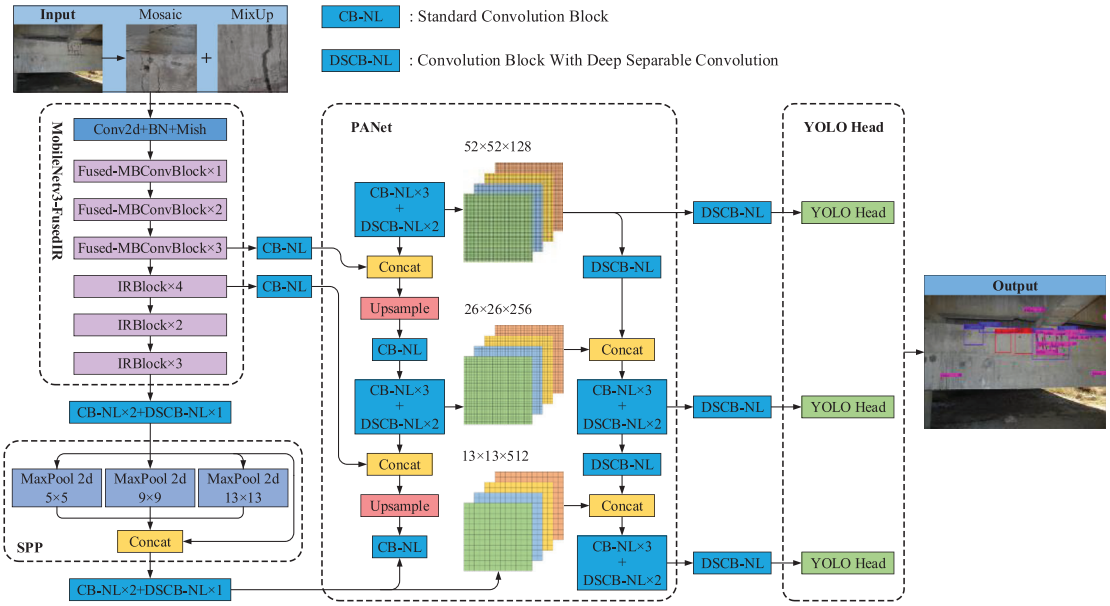

FIGURE 6 Structural diagram of the YOLOMF network.

MobileNetv3, and MobileNetv3-FusedIR. These backbone networks are closely associated with their respective main networks, namely, YOLOv4, YOLOM, and YOLOMF. Through a thorough examination and comparison of these network combinations, the study seeks to contribute valuable insights to the field of computer vision and object detection.

The main process of damage recognition for concrete bridges using YOLOMF is shown in Figure 6. In terms of YOLOv4, the YOLOMF surpasses it in three aspects. First, it employs a new backbone network, MobileNetv3-FusedIR, replacing CSPDarknet53 as the backbone of YOLOMF. In MobileNetv3-FusedIR, Fused-MBConv is introduced as a feature extraction block for the backbone shallow network due to the poor performance of DSC in this context. Second, to further minimize the model parameters, the 3×3 standard convolution was replaced with DSC. Third, by studying the effect of different activation functions on network performance, the optimal combination of activation functions is selected to enhance both the efficiency and precision of damage detection.

3 | DATASETS OF CONCRETE BRIDGE DAMAGE

3.1 | Data sources

It is well known that the generalization capacity of CNNs principally relies on a sufficient volume of training data, particularly labeled samples. In circumstances where the training data are insufficient, the risk of overfitting becomes pronounced. The task of constructing a dataset for concrete bridge damage is a challenging endeavor, demanding substantial time and human resources. Current research predominantly focuses on main damage types such as crack, rebar exposure, and spalling. However, in practical engineering scenarios, the forms of damage to concrete bridge structures are manifold and diverse. Relying on a single type of damage detection cannot accurately reflect the true state of concrete bridge deterioration. Consequently, researchers are required to further explore a broader spectrum of damage types and collect more comprehensive and diversified data, to enhance the precision and reliability of damage detection in concrete bridge structures.

In this study, the dataset used for concrete bridge damage recognition adopts the format of the PASCAL VOC dataset. During the process of collecting image data, the project team members used a UAV cloud control platform and smartphones to conduct on-site data collection in Changsha, Hunan. The images obtained through the UAV and smartphones have resolutions of 5472×3078 and 4096×3072, respectively. During the fieldwork, more than 90% of the images with damaged surfaces were collected, while a minority were sourced through Internet searches. A concrete bridge damage dataset containing 2235 original images was established. The dataset includes seven common types of damage, such as crack, spalling, rebar exposure, separation, corrosion, voids pits, and hole. During model training, the damage dataset is partitioned into a training set, validation set, and test set. The proportion between the training set and the validation set is 9:1,

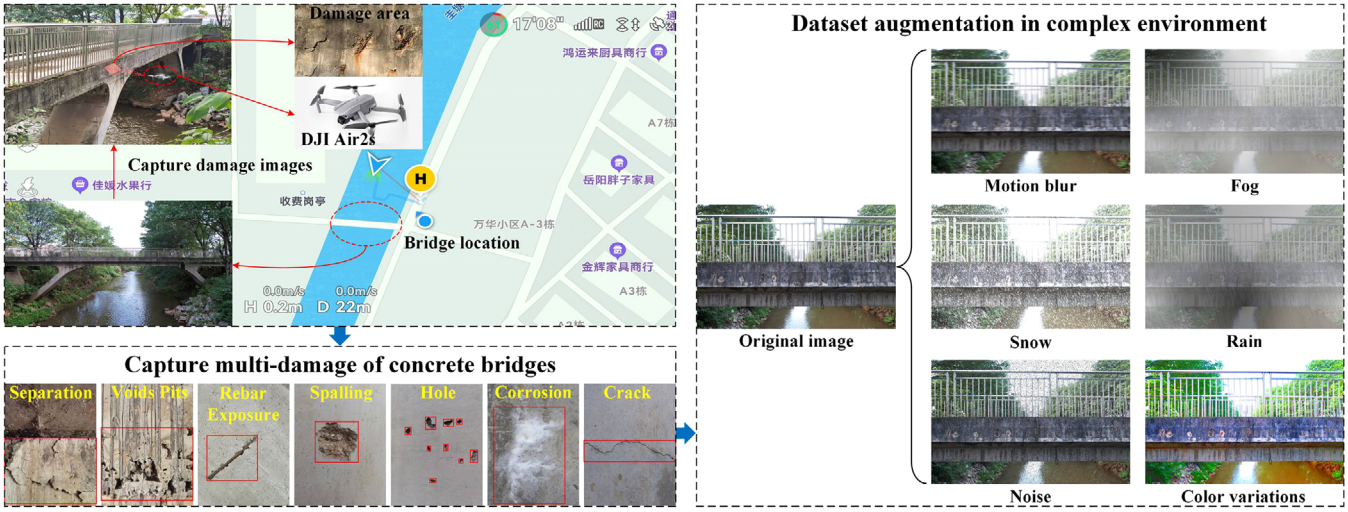

FIGURE 7 Damage dataset in complex environments.

while the proportion between the combined training and validation set and the test set is also 9:1.

The impact of model performance is widely acknowledged to depend on the amount of valid information in the training database. Nevertheless, in practical scenarios, training models with deeper network architectures using abundant image samples encompassing multiple damage types becomes unrealistic. To mitigate the potential risk of model overfitting and enhance its generalization capability, the Python image augment library Imgaug is used to augment the established database, simulating complex environments under the disturbances of severe weather, noise, and lighting conditions. The creation process of the damage dataset of concrete bridges in complex environments is shown in Figure 7.

The complex environment augments method is an offline data augmentation technique that involves the following steps: first, randomly selecting two to three transformations from a set of five, including motion blur, fog, rain, snow, and noise. Then, a single transformation is randomly selected from three options: halving or doubling contrast, increasing brightness, and enhancing color saturation, to create the enhancement sequence. Finally, the original dataset is enhanced for a set number of cycles. In this experiment, 8940 images of concrete bridge damage in complex environments were obtained by three augments to the original dataset.

3.2 | Data annotation

The LabelImg software was utilized to annotate images of concrete bridge damage. In accordance with various damage categories and the annotation format of the PASCAL VOC dataset, a total of 32,488 damage targets have been annotated in the original dataset. Specific statistics regarding the damage information are presented in Table 3. In the dataset, the quantities of labels for crack, spalling, rebar exposure, and separation are relatively balanced, while corrosion and voids pits have fewer labels. The label quantity for holes is the highest due to their concentrated and abundant presence in the damaged images. To address the issue of imbalanced distribution among different damage samples, this study adopts Focal Loss to enhance the recognition and classification ability of the model toward samples with fewer instances of damage.

TABLE 3 Statistical of concrete bridge damage datasets.

Damage type	Labeling number	Proportion (%)
Crack	3845	11.83
Spalling	4187	12.89
Rebar exposure	6193	19.06
Separation	2529	7.87
Corrosion	1188	3.66
Voids pits	787	2.42
Hole	13,759	42.35
Total	32,488	–

4 | EXPERIMENTS ON MULTI-DAMAGE RECOGNITION OF CONCRETE BRIDGES

To evaluate the damage recognition capability and application prospect of YOLOMF, the model was trained using data obtained from a complex environment data augmentation program. Its damage identification accuracy was subsequently assessed on the test data. The experiment

comprises two primary stages: training the model and testing the model. First, the YOLOMF model undergoes training using the designated training set. Simultaneously, the validation set is utilized to assess the convergence of the network throughout the training procedure. Second, three distinctive test experiments are established to holistically assess the efficacy of YOLOMF in multi-damage recognition for concrete bridges, including (1) comparing the damage recognition capabilities of YOLOv4, YOLOM, and YOLOMF; (2) comparing the damage recognition capabilities of YOLOMF with other excellent CNN models and (3) analyzing the effect of complex environment on damage recognition accuracy of YOLOMF.

4.1 | Experimental environment

In this experiment, the damage recognition algorithm is trained and tested on a personal computer running the Ubuntu 20.04.5 LTS operating system. The computer is equipped with a central processing unit (CPU) of Intel i9-13900KF with 24-core 32-threaded, and two graphics processing units (GPUs) of NVIDIA RTX4090 with 24GB graphics memory. For streamlined computation, CUDA 11.3 and cuDNN 8.2.0 are used. The programming framework is supported by Python 3.7 and PyTorch 1.11.

4.2 | Model performance assessment metrics

To evaluate the efficiency and precision of the damage recognized based on the object detection algorithm, a set of metrics has been chosen as evaluation indicators. These metrics include average precision (*AP*), mean AP (*mAP*, the intersection over union threshold is set to 0.5.), *F*1 score, frames per second (*FPS*), parameter count, and computation workload (floating-point operations, *FLOPs*). The calculation equations for each of these metrics are provided below:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$f1 = \frac{2P \cdot R}{P + R} \tag{9}$$

$$F1 = \frac{\sum f1}{n} \tag{10}$$

$$AP = \int_0^1 P(R)dR \tag{11}$$

$$mAP = \frac{\sum AP}{n} \tag{12}$$

$$FPS = \frac{N_{image}}{T_{totol}} \tag{13}$$

where P represents the precision rate, which is defined as the proportion of correctly predicted positives out of all predicted positives. R represents recall rate, which is defined as the proportion of correctly predicted positives out of all actual positives. The *PR* curve plots a curve by calculating precision and recall at different classification thresholds, with the horizontal axis representing recall and the vertical axis representing precision. *AP* represents the area under the *PR* curve and is used to measure the overall performance of the model. The count of accurately identified targets is referred to as true positive (*TP*), while false positive (*FP*) pertains to the count of mistakenly identified non-targets as targets, and false negative (*FN*) reflects the count of undetected targets. n represents the count of damage classes, T_{total} represents the overall duration for image identification, and N_{image} corresponds to the total count of images successfully detected.

4.3 | Model training

In the area of bridge damage recognition, it is a challenge for researchers to collect satisfactory training data. In this context, the TL approach stands out as a potent tool, which not only minimizes the reliance on training data quantity but also maximizes the exploitation of existing resources. The TL partitions the dataset into a source domain and a target domain. The fundamental concept is to apply the knowledge acquired from the source domain to the target domain, thereby improving the learning efficiency and performance of the target task. In situations where two datasets are associated, certain basic features can be shared within the CNN models, whereas advanced features can be fine-tuned through TL techniques.

In this study, both YOLOM and YOLOMF execute TL on the PASCAL VOC dataset to enhance the generalization capability and recognition accuracy of models. During the TL process for YOLOM, the model weights of YOLOv4 are used as pre-training weights, thereby endowing YOLOM with prior knowledge from the PASCAL VOC dataset. Similarly, YOLOMF is pre-trained using model weights of YOLOM on the PASCAL VOC dataset, followed by fine-tuning using the dataset of concrete bridge damage, ultimately achieving a model capable of multi-damage recognition of concrete bridges. Through the above across backbone network TL strategy, the prior knowledge learned by YOLOv4 and YOLOM from the

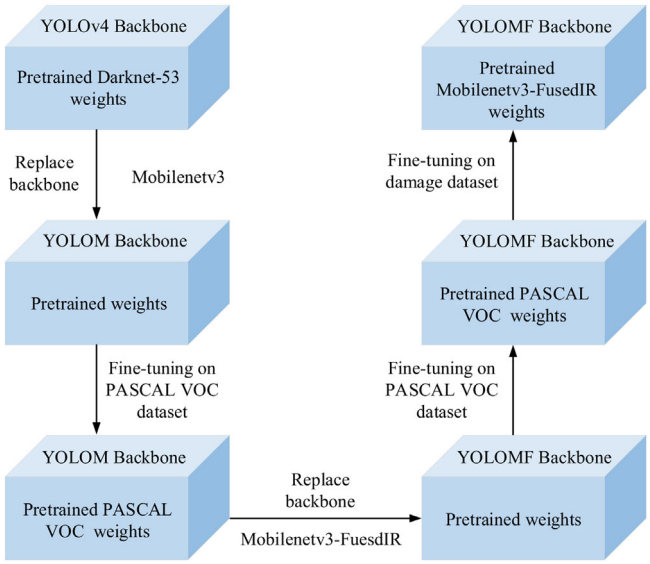

FIGURE 8 Transfer learning method.

PASCAL VOC dataset can be smoothly transferred into the YOLOMF model, thereby significantly reducing training time and improving the generalization capabilities of YOLOMF. The TL process of the model is depicted in Figure 8.

In the process of TL, the hyperparameter settings for fine-tuning YOLOMF using PASCAL VOC dataset and bridge damage dataset are as follows: The momentum parameter is set to 0.937 using the Adam optimizer. The initial learning rate is 0.001, and a cosine annealing strategy is used to gradually reduce the learning rate, reaching its minimum of 0.00001. This experiment used a freeze training strategy for 300 epochs. In the freezing phase, the model backbone network is frozen, the network is only fine-tuned without changing the weights of the feature extraction network. This strategy can prevent potential damage to backbone weights, thereby improving training efficiency. This phase involves 50 epochs of training with a batch size of 16. Subsequently, unfreezing training is performed to unfreeze the backbone of the model, resulting in a change in the feature extraction network, thus updating all network parameters. During this stage, training spans 250 epochs with a batch size of 8.

The loss curve and mAP curve of YOLOMF is depicted in Figure 9. During the freeze training stage in the initial 50 epochs, the loss of the model rapidly decreases, with the rate of decrease gradually slowing as the number of training iterations increases. After the 50th epoch, the backbone network is unfrozen, and the entire network model engages in the training. A slight increase in the loss of the model can be observed during the training at the 51st epoch as evidenced by the figure. After 210 epochs, the loss of YOLOMF gradually stabilizes, indicating that the model is gradually converging. The final training loss

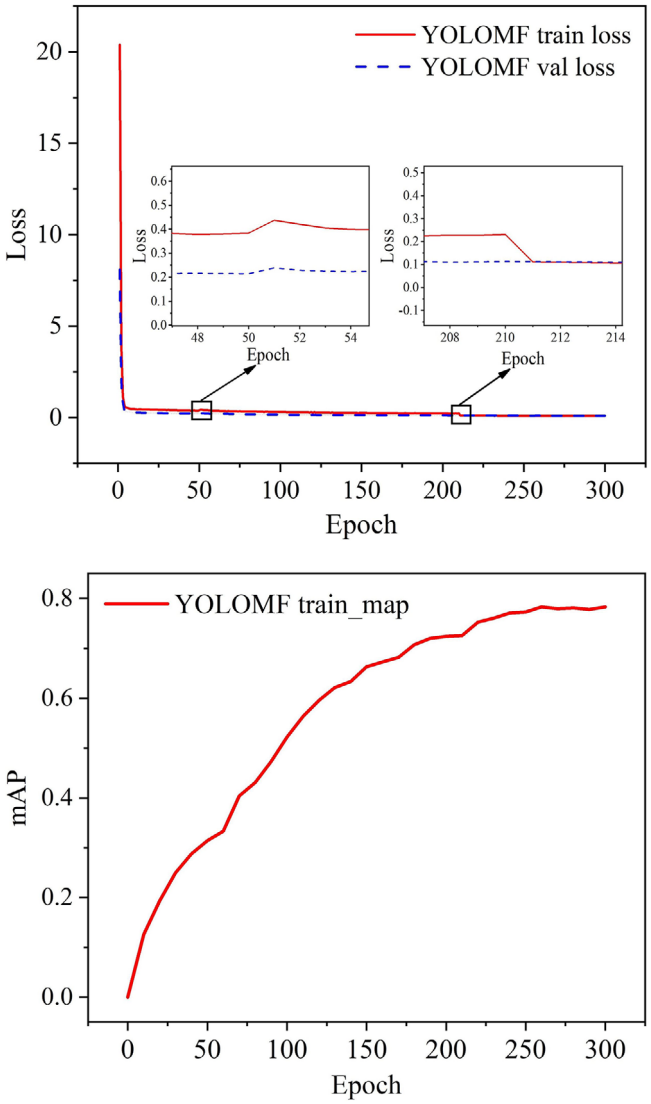

FIGURE 9 The curves of the YOLOMF model training.

and validation loss values are 0.094 and 0.099, respectively. The mAP curve for YOLOMF shows a clear upward trend after the 70th epoch following the unfreezing of the backbone network and after the 220th epoch when the model converges, implying the robust generalization ability of the YOLOMF model. Ultimately, it reaches a peak of 78.34% after the 260th epoch.

4.4 | Damage recognition results and analysis

4.4.1 | Comparison of damage recognition results with YOLOv4

To validate the damage identification ability of the modified approach used in this study, the performance of YOLOv4, YOLOv4-D (Yu et al., 2021), YOLOv4-FPM (Zou

TABLE 4 Comparison of performance evaluation indicators between YOLOv4 and other models.

Network	*Params* (MB)	Floating-point operations (*FLOPs;* GB)	*Precision* (%)	*Recall* (%)	*F*1 (%)	Mean average precision (*mAP;* %)
YOLOv4	63.97	142.00	89.21	49.53	63.36	70.76
YOLOv4-D	42.51	108.55	89.79	49.03	63.00	65.37
YOLOv4-FPM	11.64	20.21	89.54	49.48	63.25	70.47
YOLOM	11.43	16.98	89.80	47.37	61.46	68.67
YOLOMF1	11.53	18.75	90.37	52.95	66.50	72.84
YOLOMF2	11.53	18.75	90.56	59.25	71.19	76.97
YOLOMF3	11.53	18.75	90.66	60.32	71.95	77.32

et al., 2022), YOLOM, and YOLOMF in multi-damage recognition for concrete bridges is compared.

Compared with YOLOv4, YOLOV4-D uses DSC to replace the 3 × 3 standard convolution in the feature pyramid structure and adopts the H-Swish activation function to enhance the feature extraction capabilities of the model. This approach somewhat reduces the parameters and complexity of the model. However, its detection accuracy is 5.39% lower than that of YOLOv4. YOLOv4-FPM, on the other hand, uses a pruning algorithm to eliminate redundant feature channels in the YOLOv4, resulting in a reduction of the parameters and complexity of the model by 81.80% and 85.77%, respectively. The mAP declines by only 0.29%, compared to YOLOv4, indicating that the pruning algorithm is an efficient means of model lightening. YOLOM and YOLOMF use lightweight backbone networks, MobileNetv3 and MobileNetv3-FusedIR, contributing to the reductions of numbers of network parameters by 82.13% and 81.97%, respectively, as well as the decrease in *FLOPs* by 88.04% and 86.80%, respectively, as shown in Table 4. This significantly alleviates the computational burden of real-time multi-damage inspection of concrete bridges using mobile terminals. It is worth noting that although YOLOM has a lightweight network architecture, its *precision* is only 0.59% higher than YOLOv4, while the *recall*, *F*1, and *mAP* have decreased by 2.16%, 1.09%, and 2.09%, respectively. This decrease is attributed to the low efficiency of DSC in shallow networks and the inability to extract rich damage features.

By contrast, the YOLOMF has demonstrated significant improvements in *precision*, *recall*, *F*1, and *mAP*, while incurring minimal increases in computation. As can be seen, the MobileNetv3-FusedIR backbone has effectively improved the deficiencies of MobileNetv3. The analysis shows that using the combined activation function of Hard-Swish and SiLU has better damage recognition performance. In comparison with YOLOv4, YOLOMF3 boasts improvements in *precision*, *recall*, *F*1 score, and *mAP* by 1.45%, 10.79%, 8.59%, and 6.56%, respectively. On the other hand, the damage recognition performance of YOLOMF2 with the combination activation function of Hard-Swish and Leaky ReLU was reduced. The YOLOMF1, which adopts the ReLU6 activation function, exhibits the worst performance. The main reason contributing to this phenomenon is that ReLU6 introduces information loss in high-dimensional inputs. It sets all negative values to zero and clips positive values greater than 6, causing important information loss. Consequently, YOLOMF3 was selected as the model for multi-damage detection of concrete bridges.

The *AP* and *F*1 scores of YOLOv4 and YOLOMF for multi-damage recognition of concrete bridges are shown in Figure 10. The results demonstrate that YOLOMF outperforms YOLOv4 in recognition accuracy. Notably, the YOLOMF displays substantial improvements in identifying multi-damage of concrete bridges, including corrosion, crack, rebar exposure, and voids pits. Among these, the most significant enhancement has been observed in the recognition of voids pits, with an increase of 12.19% in the *AP* value and a remarkable 20% increase in the *F*1 score. Both models exhibit good recognition capabilities for spalling, mainly because of their prominent features, especially in terms of shape and color. The YOLOMF has a relatively small improvement in hole and rebar exposure recognition accuracy. The main reason is that the hole is a small target, and the rebar exposure is a specific target with an extreme aspect ratio. In addition, these two types of damage often occur in concentrated patterns, posing challenges for feature extraction and making multi-target recognition demanding.

4.4.2 | Comparison of damage recognition results with classical object detection models

This section compares YOLOMF with classic object detection models such as Faster R-CNN (Ren et al., 2015), single shot detection (SSD) (Liu et al., 2016), YOLOv4, YOLOv5-s, YOLOv7-tiny, and YOLOv8-s using the same dataset and computer. Table 5 shows the number of parameters, *FLOPs*, damage recognition precision (*mAP*), performance

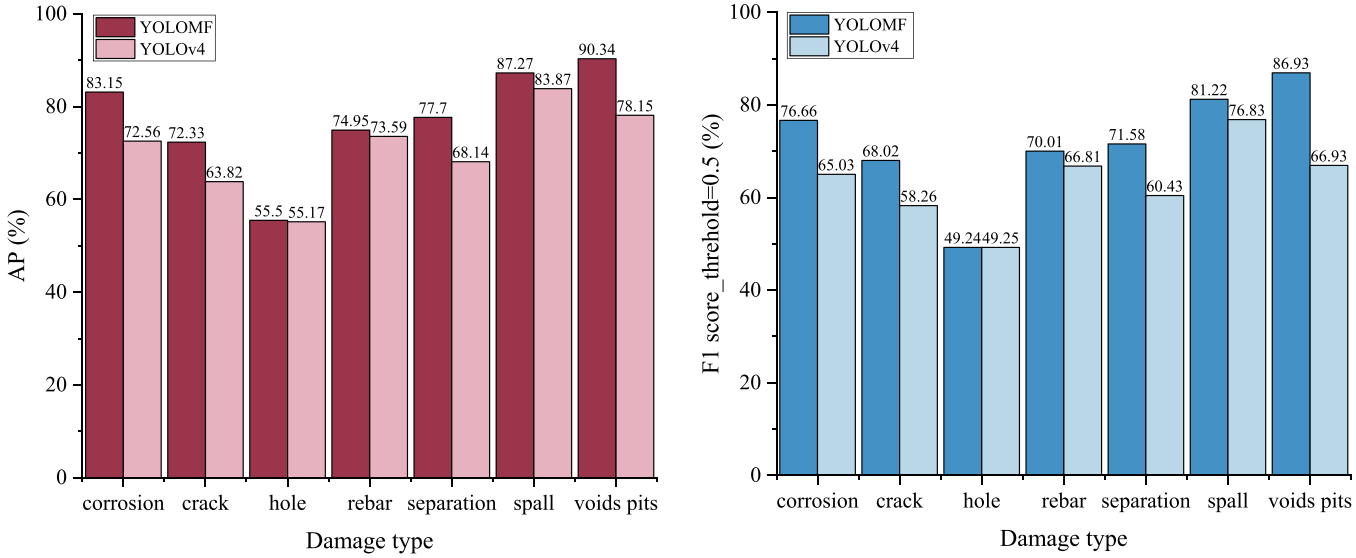

FIGURE 10 Comparison of damage recognition performance between YOLOv4 and YOLOMF.

TABLE 5 Comparison of performance evaluation indicators between YOLOMF and other models.

Network	*Params* (MB)	*FLOPs* (GB)	*mAP* 0.5(%)	*F*1(%)	*FPS* (f/s)
Faster R-CNN	136.81	401.84	38.03	39.39	50
SSD	26.29	62.75	27.04	16.64	120
YOLOv4	63.97	142.00	70.76	63.36	66
YOLOv5-s	7.08	16.53	72.56	68.88	150
YOLOv7-tiny	**6.03**	**13.23**	62.19	57.19	**225**
YOLOv8-s	11.14	28.66	73.69	69.97	151
YOLOMF	11.53	18.75	**77.32**	**71.95**	85

Abbreviation: *FPS*, frames per second.
The optimal value is shown in bold.

(*F*1), and speed (*FPS*) for each model. It can be observed that the two-stage detection algorithm Faster R-CNN has a high number of parameters and complexity, while the structures of SSD and YOLOv4 are relatively cumbersome. In contrast, YOLOv5-s and YOLOv7-tiny possess relatively fewer computational parameters. The number of parameters YOLOv8-s is slightly lower than that of YOLOF. However, YOLOv8-s exhibits a higher *FLOPs* value and a more complex model structure.

Contrastingly, Faster R-CNN, despite its wealth of model parameters, exhibits subpar performance in handling complex concrete damage. Models such as YOLOv4, YOLOv5-s, and others, although they demonstrate commendable recognition accuracy, fall short in their feature extraction capabilities for identifying multi-damage in concrete bridges, particularly when compared to YOLOMF, which achieves an *mAP* of 77.32%. SSD has some advantages in inspection speed, but it has the poorest damage identification accuracy. The YOLOv7-tiny, with its lightweight network, has the fastest recognition speed, reaching 225f/s. YOLOv5-s and YOLOv8-s also exhibit relatively fast recognition speeds. YOLOMF, despite some compromise in identification speed, attains a rate of 85f/s, marking a 28.79% enhancement in damage identification speed over YOLO4. This competence can satisfy the demands of real-time damage detection in concrete bridges.

The recognition results of multi-damage in concrete bridges about each model are presented in Tables 6 and 7. The YOLOMF exhibits an absolute advantage in the recognition accuracy for corrosion, hole, rebar exposure, separation, and spalling. However, its accuracy in identifying crack and voids pits is not as high as that of YOLOv8-s. Overall, when considering the multi-damage recognition performance, YOLOMF is on par with YOLOv8-s, while surpassing it in terms of multi-damage recognition accuracy.

The multi-damage identification results of YOLO series networks in small, medium, and large field of view of

TABLE 6 Comparison of damage recognition results between YOLOMF and other models.

Network	Corrosion		Crack		Hole		Rebar exposure	
	AP (%)	*F*1 (%)	*AP* (%)	*F*1 (%)	*AP* (%)	*F*1 (%)	*AP* (%)	*F*1 (%)
Faster R-CNN	43.30	45.27	37.60	38.95	2.70	8.55	36.37	40.30
SSD	21.22	9.15	21.24	12.33	17.52	7.60	25.35	11.78
YOLOv4	72.56	65.03	63.82	58.26	55.17	49.25	73.59	66.81
YOLOv5-s	77.00	72.99	66.22	63.91	54.56	**52.35**	67.88	63.13
YOLOv7-tiny	65.55	64.80	51.32	49.32	49.51	37.20	62.44	57.29
YOLOv8-s	81.99	**79.46**	**74.54**	**71.43**	34.16	25.76	69.44	66.82
YOLOMF	**83.15**	76.66	72.33	68.02	**55.50**	49.34	**74.95**	**70.01**

TABLE 7 Comparison of damage recognition results between YOLOMF and other models.

Network	Separation		Spalling		Voids pits	
	AP (%)	*F*1 (%)	*AP* (%)	*F*1 (%)	*AP* (%)	*F*1 (%)
Faster R-CNN	41.07	41.32	59.68	55.68	45.51	44.66
SSD	29.21	21.26	50.44	46.05	24.05	8.28
YOLOv4	68.14	60.43	83.87	76.83	78.15	66.93
YOLOv5-s	75.03	73.58	85.23	80.83	82.01	74.39
YOLOv7-tiny	58.81	57.85	81.08	76.30	66.62	57.85
YOLOv8-s	77.12	**75.44**	86.28	**84.29**	**91.59**	86.24
YOLOMF	**77.70**	71.58	**87.27**	81.22	90.34	**86.93**

TABLE 8 Statistics of damage recognition results between YOLOMF and other models in different field of view.

Damage type	The number of damage detected by object detection models in different field of view														
	YOLOv4			YOLOv5-s			YOLOv7-tiny			YOLOv8-s			YOLOMF		
	S	M	L	S	M	L	S	M	L	S	M	L	S	M	L
Corrosion	0/0	0/0	0/0	0/0	0/1	0/0	0/0	0/0	0/0	0/0	0/1	0/0	0/0	0/1	0/0
Crack	2/0	0/0	0/0	2/1	0/0	0/0	2/3	0/0	0/0	**2/2**	0/0	0/0	**2/2**	0/0	0/0
Hole	36/28	0/0	0/0	**36/31**	0/0	0/0	36/27	0/0	0/0	36/21	0/0	0/0	36/28	0/0	0/0
Rebar exposure	0/0	23/9	11/1	0/0	**23/13**	**11/4**	0/0	23/11	11/2	0/0	23/8	11/0	0/0	23/9	11/3
Separation	0/0	6/2	13/7	0/0	**6/5**	13/7	0/0	6/4	13/7	0/0	6/4	13/6	0/0	6/4	13/4
Spalling	0/0	22/17	**10/8**	0/0	**22/20**	10/6	0/1	22/19	10/6	0/0	22/19	10/7	0/0	22/17	10/7
Voids pits	0/0	2/0	0/0	0/0	2/0	0/0	0/0	2/0	0/0	0/0	**2/1**	0/0	0/0	**2/1**	0/0

Note: x/y form of data representation: the actual damage number in the image to be detected/the damage number detected by model. S, M, and L represent small, medium, and large field of view, respectively. The linear underlining indicates that there are false positives in the damage recognition.

concrete bridges are shown in Figure 11. At the same time, Table 8 shows the statistics of the multi-damage identification results of YOLOMF and other models in different fields of view. In the small field of view, although the holes are small in shape and seem difficult to identify, they are densely distributed and numerous in the damage images, accounting for 42.35% of the damage dataset labels. Therefore, all models can identify most of the holes. Among them, YOLOv8-s detects the least number of holes, identifying 21 out of 36 holes, with an identification rate of 58.3%. The YOLOMF and YOLOv8-s show commendable identification ability for both holes and cracks, although occasional missed detections still occur. YOLOv4 and YOLOv5-s are not sensitive enough to crack, YOLOv7-tiny has a good identification effect on crack, but there are false positive problems.

Within the medium field of view, all models demonstrated good recognition capabilities for separation, spalling, and rebar exposure, but there were false detections and missed detections in recognition of corrosion

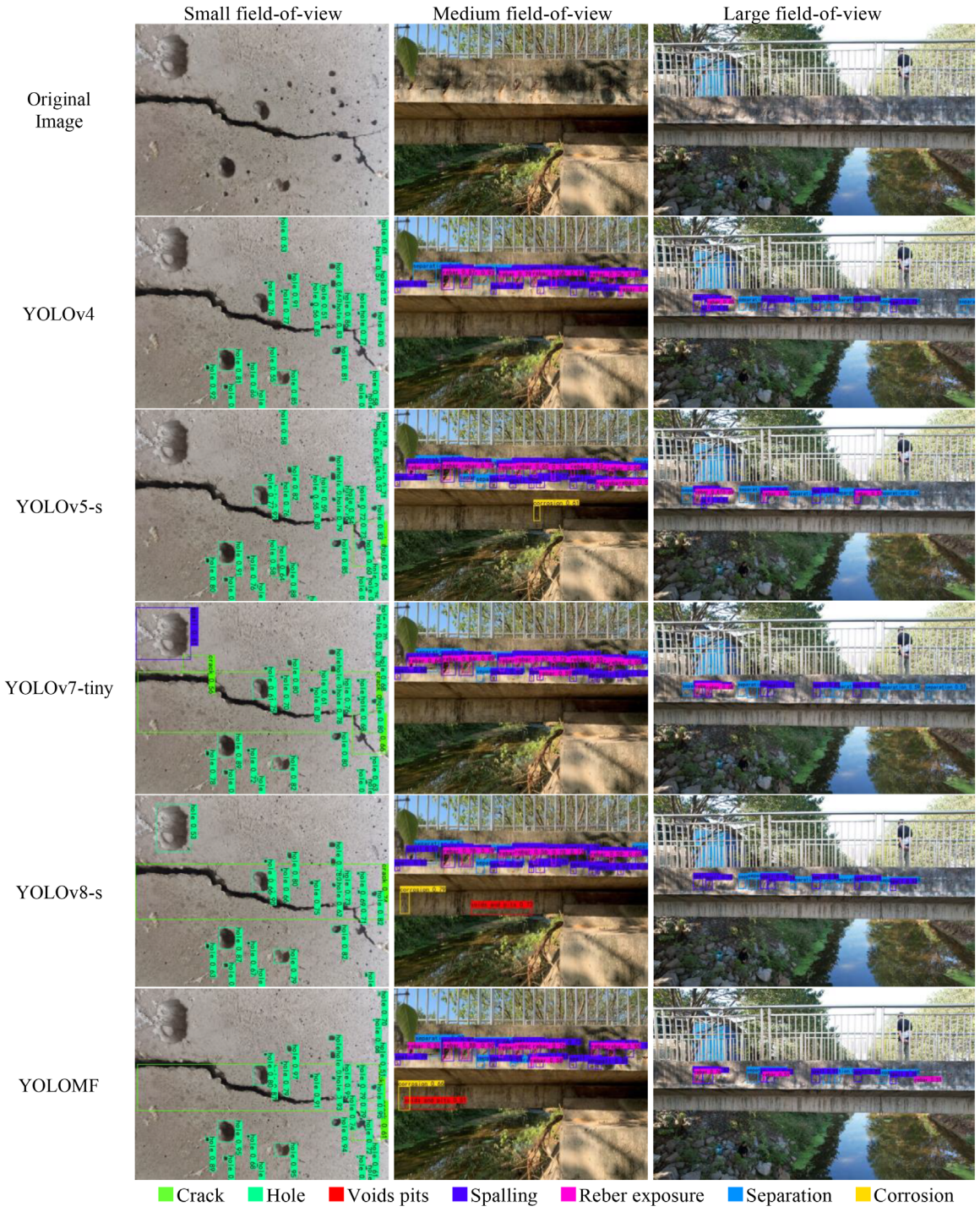

FIGURE 11 Damage recognition results using YOLOMF and other models in different field of view.

and voids pits. For example, YOLOv4, YOLOv5-s, and YOLOv7-tiny failed to recognize voids pits, while YOLOv5-s, YOLOv8-s, and YOLOMF misidentified water stains as corrosion. This may be because these two types of damage were less frequent in the damage dataset, and the models could not effectively extract damage features. It can be seen from Table 8 that YOLOv5-s had a slight advantage in the number of damage identifications. However, in terms of overall damage recognition performance, YOLOv8-s and YOLOMF performed better.

In the large field of view, the detection of damaged objects was hindered by the challenges posed by the small size of the damaged area, compared to the entire image. The results showed that the damage recognition

TABLE 9 Statistics of damage recognition results between YOLOMF and other models in complex environments.

		The number of damage recognition by models in complex environments											
		Motion blur				Fog				Rain			
Damage	Number	v4	v5-s	v8-s	MF	v4	v5-s	v8-s	MF	v4	v5-s	v8-s	MF
Corrosion	1	1	1	1	0	0	0	1	0	0	0	1	0
Crack	0	0	0	0	0	0	0	0	0	0	0	1	0
Hole	0	0	0	0	0	0	0	0	0	0	0	0	0
Rebar exposure	30	11	10	2	**14**	7	5	2	**11**	10	11	3	**13**
Separation	0	0	0	0	0	0	0	0	0	0	0	0	0
Spalling	9	1	2	**7**	6	1	4	6	**7**	2	2	3	**5**
Voids pits	2	0	1	**2**	**2**	1	**2**	**2**	**2**	0	**2**	**2**	**2**

performance of each model was not satisfactory. They could only partially identify separation and spalling, and the recognition effect of rebar exposure was the worst. Rebar exposure was a smaller form of damage, and YOLOv8-s could not even recognize it.

4.4.3 | Impact of complex environments on damage recognition accuracy of the YOLOMF

To verify the multi-damage recognition performance of concrete bridges using YOLOMF in complex environments, the collected damage dataset was subjected to specific image augmentation operations, including motion blur, fog, rain, snow, noise conditions, and color variations. The multi-damage recognition results of YOLOv4, YOLOv5-s, YOLOv8-s, and YOLOMF in complex environments are depicted in Figure 12.

The results indicate that under the adverse conditions of image augmentation simulation, YOLOv4 can only effectively capture a portion of the rebar exposure and spalling. Notably, it exhibits a substantial number of missed detections when there is noise interference and color variations. Additionally, its ability to extract voids pits features is inadequate due to the resemblance of the shape and color characteristics of voids pits to the concrete background. In comparison to YOLOv4, both YOLOv5-s and YOLOv8-s can recognize a broader range of damage types, particularly demonstrating superiority in identifying spalling and voids pits. However, they fail to significantly identify elongated types of damage, such as rebar exposure. Specifically, YOLOv8-s identifies only 6.7% of the rebar exposure in the damaged images under conditions of motion blur, foggy weather, and color variations as shown in Tables 9 and 10. Similarly, noise strongly interferes with the damage recognition of both YOLOv5-s and YOLOv8-s, with YOLOv8-s even erroneously identifying large areas of noise background as voids pits.

Compared to YOLOv4, YOLOv5-s, and YOLOv8-s, YOLOMF demonstrates superior potential for damage recognition under various challenging conditions. YOLOMF exhibits satisfactory performance in damage identification under conditions involving motion blur, fog, and rain, accurately identifying most surface damages. This suggests that the effective collaboration between the MobileNetv3-FusedIR backbone network, and the optimally chosen activation functions enhance the ability of the model to recognize surface damages with challenging feature extraction.

In situations of snow and color variations, although YOLOMF showed false negatives in the detection of spalling and smaller rebar exposure, and its inability to identify voids pits, it still retains the capacity to recognize major damage areas. This surpasses the damage recognition abilities of YOLOv4, YOLOv5-s, and YOLOv8-s. However, it is noteworthy that noise interference continues to significantly impair the performance of YOLOMF in damage recognition, with only a portion of the damage correctly identified and some noise backgrounds being erroneously detected as voids pits. Consequently, it is imperative to enhance the capability of YOLOMF to resist noise interference and recognize small damage.

Furthermore, for corrosion and voids pits damage, the models employed in the experiment showed subpar performance. This could be attributed to their relatively scarce representation in the damage dataset, accounting for only 3.66% and 2.42% of the total labels. Their shape and color attributes are easily confused with the background of the concrete surface generated by image enhancement, resulting in false detection. This issue is also manifested in damage recognition under different fields of view. Clearly, the diversity, balance, and authenticity of the damage dataset are crucial for improving the ability of the model to recognize damage. Therefore, it is necessary to enhance and refine the damage dataset of concrete bridges in complex environments.

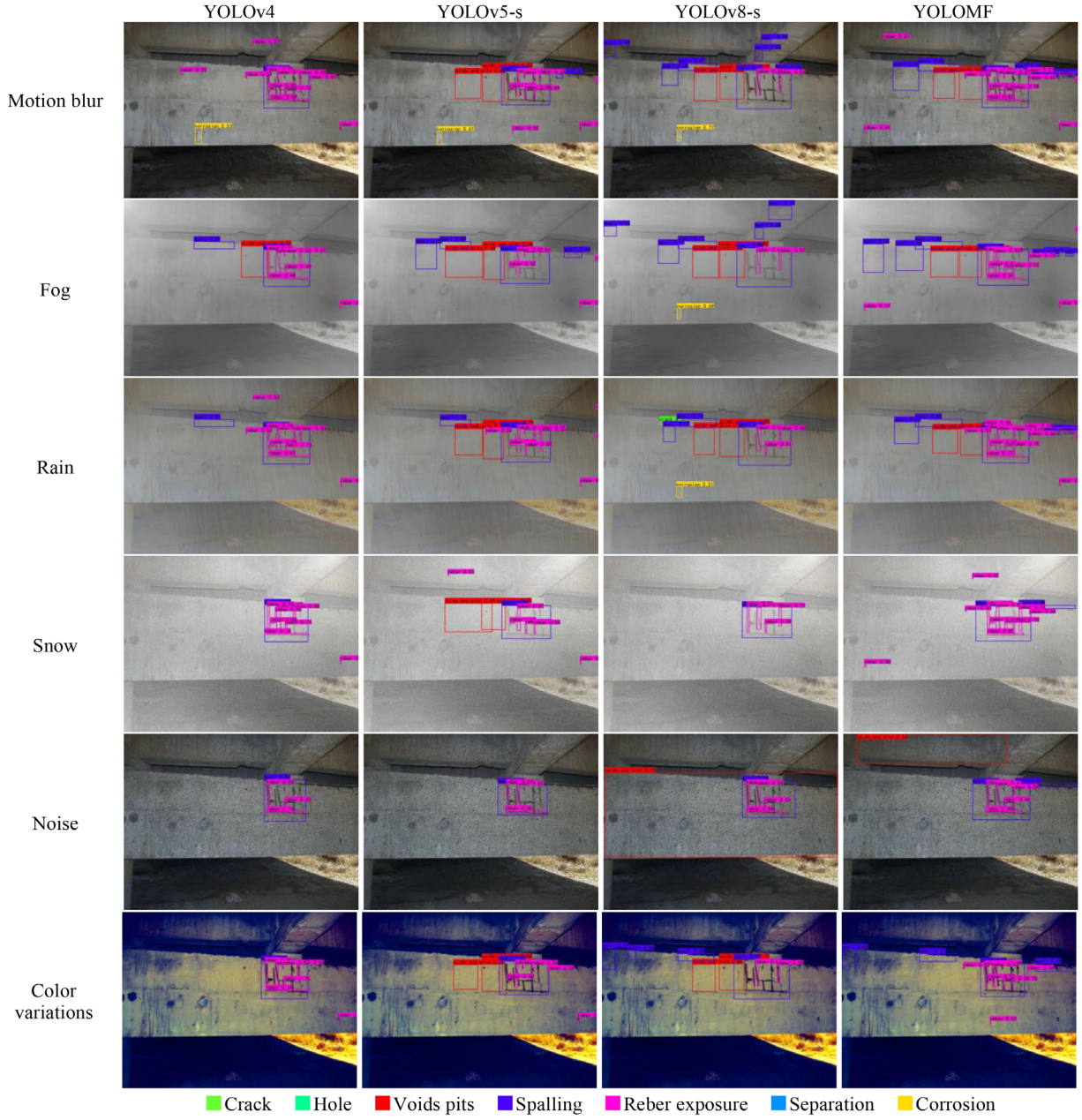

FIGURE 12 Damage recognition results using YOLOv4, YOLOv8-s, and YOLOMF in complex environments.

TABLE 10 Statistics of damage recognition results between YOLOMF and other models in complex environments.

		The number of damage recognition by models in complex environments											
		Snow				Noise				Color variation			
Damage	**Number**	**v4**	**v5-s**	**v8-s**	**MF**	**v4**	**v5-s**	**v8-s**	**MF**	**v4**	**v5-s**	**v8-s**	**MF**
Corrosion	1	0	0	0	0	0	0	0	0	0	0	0	0
Crack	0	0	0	0	0	0	0	0	0	0	0	0	0
Hole	0	0	0	0	0	0	0	0	0	0	0	0	0
Rebar exposure	30	8	6	4	**11**	3	3	4	**5**	7	6	2	**9**
Separation	0	0	0	0	0	0	0	0	0	0	0	0	0
Spalling	9	1	1	1	**2**	1	1	1	**2**	1	1	4	4
Voids pits	2	0	0	0	0	0	0	1	1	0	**2**	0	0

5 | CONCLUSION

In this paper, a novel multi-damage identification model for concrete bridges called YOLOMF is proposed based on YOLOv4. This model achieves high-precision multi-damage real-time detection of concrete bridges in complex environments. The primary findings obtained from this research can be summarized as follows.

1. YOLOM employs MobileNetv3 as its backbone network and manages to significantly reduce network parameters and model complexity through the use of DSC. The model *params* to merely 11.43 MB, and the *FLOPs* are 16.98GB, which, respectively, represent a reduction of 82.13% and 88.04%, compared to YOLOv4. However, due to the inability of DSC to fully capture effective spatial features in shallow networks, the multi-damage detection precision of YOLOM is only 0.59% higher than that of YOLOv4, while the *recall*, *F*1 score, and *mAP* have decreased by 2.16%, 1.09%, and 2.09%, respectively.
2. YOLOMF uses MobileNetv3-FusedIR as its backbone network, effectively mitigating the spatial feature loss caused by the insufficient performance of DSC by introducing the Fused-MBConv module into the shallow network of MobileNetv3, thereby enhancing the precision of the model in multi-damage recognition. The efficient combination of MobileNetv3-FusedIR with Hard-Swish and SiLU activation functions has enabled YOLOMF to achieve *mAP* of 77.32% and *F*1 score of 71.95%, which are 6.56% and 8.59% higher than the original YOLOv4, respectively. Furthermore, the damage recognition performance of YOLOMF surpasses that of advanced object detection models such as YOLOv8-s. Although the detection speed of YOLOMF has somewhat decreased, it still reaches 85 f/s, representing a 28.79% improvement over the original YOLOv4. This makes it suitable for real-time detection of multi-damages in concrete bridges.
3. YOLOMF shows excellent performance in multi-damage recognition of concrete bridges in various field-of-view scenarios of small, medium, and large. By employing image enhancement techniques to simulate complex real-world scenarios such as motion blur, fog, rain, or snow conditions; noise interference; and color variations, the applicability of the YOLOMF model for multi-damage identification under intricate environments was validated. The damage recognition performance of YOLOMF is better than that of existing classical networks such as Faster R-CNN, SSD, YOLOv5-s, and YOLOv8-s. However, the damage identification accuracy of YOLOMF for small damage and noise interference scenes needs to be further improved.

More details and code are available at https://github.com/llyun1995/YOLOMF. In future work, the authors will further understand the role of different activation functions in the target detection network, optimize the feature fusion strategy, and integrate super-resolution technology and more advanced AMs into the damage feature extraction block to improve the multi-damage recognition accuracy and generalization ability of the model.

ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (No. 2019YFC1511000), the National Natural Science Foundation of China (No. 52378123), and the Graduate Research Innovation Project of Hunan Province is Hunan Provincial Department of Education, China (No. CX20220879).

Open access publishing facilitated by University of New South Wales, as part of the Wiley - University of New South Wales agreement via the Council of Australian University Librarians.

REFERENCES

Altunışık, A. C., Okur, F. Y., Karaca, S., & Kahya, V. (2019). Vibration-based damage detection in beam structures with multiple cracks: Modal curvature vs. modal flexibility methods. *Nondestructive Testing and Evaluation*, *34*(1), 33–53.

Amezquita-Sanchez, J. P., & Adeli, H. (2016). Signal processing techniques for vibration-based health monitoring of smart structures. *Archives of Computational Methods in Engineering*, *23*, 1–15.

Amezquita-Sanchez, J. P., & Adeli, H. (2019). Nonlinear measurements for feature extraction in structural health monitoring. *Scientia Iranica*, *26*(6), 3051–3059.

Bao, Y., & Li, H. (2021). Machine learning paradigm for structural health monitoring. *Structural Health Monitoring*, *20*(4), 1353–1372.

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). *YOLOv4: Optimal speed and accuracy of object recognition*. arXiv preprint. arXiv:2004.10934. https://arxiv.org/abs/2004.10934

Carranza-García, M., Galán-Sales, F. J., Luna-Romera, J. M., & Riquelme, J. C. (2022). Object detection using depth completion and camera-LiDAR fusion for autonomous driving. *Integrated Computer-Aided Engineering*, *29*(3), 241–258.

Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack damage recognition using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, *32*(5), 361–378.

Cha, Y. J., Choi, W., Suh, G., Mahmoudkhani, S., & Büyüköztürk, O. (2018). Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, *33*(9), 731–747.

Ciambella, J., Pau, A., & Vestroni, F. (2019). Modal curvature-based damage localization in weakly damaged continuous beams. *Mechanical Systems and Signal Processing*, *121*, 171–182.

Chou, J. Y., Chang, C. M., & Spencer, B. F., Jr. (2022). Out-of-plane modal property extraction based on multi-level image pyramid

reconstruction using stereophotogrammetry. *Mechanical Systems and Signal Processing*, *169*, 108786.

Dong, C. Z., & Catbas, F. N. (2021). A review of computer vision-based structural health monitoring at local and global levels. *Structural Health Monitoring*, *20*(2), 692–743.

Duque, L., Seo, J., & Wacker, J. (2018). Bridge deterioration quantification protocol using UAV. *Journal of Bridge Engineering*, *23*(10), 04018080.

Foresti, G. L., & Scagnetto, I. (2022). An integrated low-cost system for object detection in underwater environments. *Integrated Computer-Aided Engineering*, *29*(2), 123–139.

Gao, Y., Yang, J., Qian, H., & Mosalam, K. M. (2023). Multiattribute multitask transformer framework for vision-based structural health monitoring. *Computer-Aided Civil and Infrastructure Engineering*, *38*(17), 2358–2377.

Gupta, S., & Tan, M. (2019). EfficientNet-EdgeTPU: Creating accelerator-optimized neural networks with AutoML. Google AI Blog.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV (pp. 770–778).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint. arXiv:1704.04861. https://arxiv.org/abs/1704.04861

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., & Adam, H. (2019). Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea (pp. 1314–1324).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT (pp. 7132–7141).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning, Lille, France (pp. 448–456).

Javadinasab Hormozabad, S., Gutierrez Soto, M., & Adeli, H. (2021). Integrating structural control, health monitoring, and energy harvesting for smart cities. *Expert Systems*, *38*(8), e12845.

Jiang, T., Wu, Q., Wang, L., Huo, L., & Song, G. (2018). Monitoring of bolt looseness-induced damage in steel truss arch structure using piezoceramic transducers. *IEEE Sensors Journal*, *18*(16), 6677–6685.

Jocher, G. (2020). YOLOv5 by Ultralytics. https://github.com/ultralytics/yolov5

Jocher, G., Chaurasia, A., & Qiu, J. (2023). YOLO by Ultralytics. https://github.com/ultralytics/ultralytics

Jodas, D. S., Yojo, T., Brazolin, S., Velasco, G. D. N., & Papa, J. P. (2022). Detection of trees on street-view images using a convolutional neural network. *International Journal of Neural Systems*, *32*(1), 2150042.

Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., & Fieguth, P. (2015). A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. *Advanced Engineering Informatics*, *29*(2), 196–210.

Li, C., Sun, L., Xu, Z., Wu, X., Liang, T., & Shi, W. (2020). Experimental investigation and error analysis of high precision FBG displacement sensor for structural health monitoring. *International Journal of Structural Stability and Dynamics*, *20*(6), 2040011.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy (pp. 2980–2988).

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT (pp. 8759–8768).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. *European Conference on Computer Vision*, Amsterdam, the Netherlands (pp. 21–37).

Liu, Y. F., Cho, S., Spencer, B. F., Jr.,, & Fan, J. S. (2016). Concrete crack assessment using digital image processing and 3D scene reconstruction. *Journal of Computing in Civil Engineering*, *30*(1), 04014124.

Ministry of Transport of the People's Republic of China. (2022). Statistical bulletin on development of transport industry. https://xxgk.mot.gov.cn/2020/jigou/zhghs/202306/t20230615_3847023.html

Misra, D. (2019). Mish: A self regularized non-monotonic activation function. arXiv preprint. arXiv:1908.08681. https://arxiv.org/abs/1908.08681

Mutlib, N. K., Baharom, S. B., El-Shafie, A., & Nuawi, M. Z. (2016). Ultrasonic health monitoring in structural engineering: Buildings and bridges. *Structural Control and Health Monitoring*, *23*(3), 409–422.

Pezeshki, H., Adeli, H., Pavlou, D., & Siriwardane, S. C. (2023). State of the art in structural health monitoring of offshore and marine structures. *Proceedings of the Institution of Civil Engineers-Maritime Engineering*, *176*(2), 89–108.

Pezeshki, H., Pavlou, D., Adeli, H., & Siriwardane, S. C. (2023). Modal analysis of offshore monopile wind turbine: An analytical solution. *Journal of Offshore Mechanics and Arctic Engineering*, *145*(1), 010907.

Perez-Ramirez, C. A., Amezquita-Sanchez, J. P., Adeli, H., Valtierra-Rodriguez, M., Camarena-Martinez, D., & Romero-Troncoso, R. J. (2016). New methodology for modal parameters identification of smart civil structures using ambient vibrations and synchrosqueezed wavelet transform. *Engineering Applications of Artificial Intelligence*, *48*, 1–12.

Perez-Ramirez, C. A., Amezquita-Sanchez, J. P., Valtierra-Rodriguez, M., Adeli, H., Dominguez-Gonzalez, A., & Romero-Troncoso, R. J. (2019). Recurrent neural network model with Bayesian training and mutual information for response prediction of large buildings. *Engineering Structures*, *178*, 603–615.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Payab, M., Abbasina, R., & Khanzadi, M. (2019). A brief review and a new graph-based image analysis for concrete crack quantification. *Archives of Computational Methods in Engineering*, *26*, 347–365.

Rafiei, M. H., Khushefati, W. H., Demirboga, R., & Adeli, H. (2016). Neural network, machine learning, and evolutionary approaches for concrete material characterization. *ACI Materials Journal*, *113*(6), 781–789.

Rafiei, M. H., Khushefati, W. H., Demirboga, R., & Adeli, H. (2017a). Novel approach for concrete mixture design using neural dynamics model and virtual lab concept. *ACI Materials Journal*, *114*(1), 117–127.

Rafiei, M. H., Khushefati, W. H., Demirboga, R., & Adeli, H. (2017b). Supervised deep restricted Boltzmann machine for estimation of concrete. *ACI Materials Journal*, *114*(2), 237–244.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. arXiv preprint. arXiv:1710.05941. https://arxiv.org/abs/1710.05941

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV (pp. 779–788).

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An incremental improvement*. arXiv preprint. arXiv:1804.02767. https://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, vol. *28*, Montreal Canada.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT (pp. 4510–4520).

Seo, J., Duque, L., & Wacker, J. (2018). Drone-enabled bridge inspection methodology and application. *Automation in Construction*, *94*, 112–126.

Sohaib, M., Jamil, S., & Kim, J. M. (2024). An ensemble approach for robust automated crack detection and segmentation in concrete structures. *Sensors*, *24*(1), 257.

Spencer, Jr., B. F., Hoskere, V., & Narazaki, Y. (2019). Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering*, *5*(2), 199–222.

Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller models and faster training. *International Conference on Machine Learning*, Virtual (pp. 10096–10106).

Urdiales, J., Martín, D., & Armingol, J. M. (2023). An improved deep learning architecture for multi-object tracking systems. *Integrated Computer-Aided Engineering*, *30*(2), 121–134.

Verstrynge, E., Lacidogna, G., Accornero, F., & Tomor, A. (2021). A review on acoustic emission monitoring for damage recognition in masonry structures. *Construction and Building Materials*, *268*, 121089.

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA (pp. 390–391).

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7464–7475).

Wu, T., Tang, L., Du, P., Liu, N., Zhou, Z., & Qi, X. (2022). Non-contact measurement method of beam vibration with laser stripe tracking based on tilt photography. *Measurement*, *187*, 110314.

Xian, R., Lugu, R., Peng, H., Yang, Q., Luo, X., & Wang, J. (2023). Edge detection method based on nonlinear spiking neural systems. *International Journal of Neural Systems*, *33*(1), 2250060.

Xu, J., Li, Z., Du, B., Zhang, M., & Liu, J. (2020). Reluplex made more practical: Leaky ReLU. 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France (pp. 1–7).

Xu, X., & Li, X. (2024). Research on surface defect detection algorithm of pipeline weld based on YOLOv7. *Scientific Reports*, *14*(1), 1881.

Yang, T. J., Howard, A., Chen, B., Zhang, X., Go, A., Sandler, M., & Adam, H. (2018). NetAdapt: Platform-aware neural network adaptation for mobile applications. Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany (pp. 285–300).

Yu, S., Xu, Z., Su, Z., & Zhang, J. (2021). Two flexible vision-based methods for remote deflection monitoring of a long-span bridge. *Measurement*, *181*, 109658.

Yu, Y., Rashidi, M., Samali, B., Mohammadi, M., Nguyen, T. N., & Zhou, X. (2022). Crack recognition of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm. *Structural Health Monitoring*, *21*(5), 2244–2263.

Yu, Y., Samali, B., Rashidi, M., Mohammadi, M., Nguyen, T. N., & Zhang, G. (2022). Vision-based concrete crack recognition using a hybrid framework considering noise effect. *Journal of Building Engineering*, *61*, 105246.

Yu, Z., Shen, Y., & Shen, C. (2021). A real-time recognition approach for bridge cracks based on YOLOv4-FPM. *Automation in Construction*, *122*, 103514.

Zhang, C., Chang, C. C., & Jamshidi, M. (2020). Concrete bridge surface damage detection using a single-stage detector. *Computer-Aided Civil and Infrastructure Engineering*, *35*(4), 389–409.

Zhang, H., Shen, Z., Lin, Z., Quan, L., & Sun, L. (2023). Deep learning-based automatic classification of three-level surface information in bridge inspection. Computer-Aided Civil and Infrastructure Engineering. Advance online publication. https://doi.org/10.1111/mice13117

Zhang, J., Zhou, L., Tian, Y., Yu, S., Zhao, W., & Cheng, Y. (2022). Vortex-induced vibration measurement of a long-span suspension bridge through noncontact sensing strategies. *Computer-Aided Civil and Infrastructure Engineering*, *37*(12), 1617–1633.

Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. arXiv preprint. arXiv:1611.01578. https://arxiv.org/abs/1611.01578

Zou, D., Zhang, M., Bai, Z., Liu, T., Zhou, A., Wang, X., & Zhang, S. (2022). Multicategory damage detection and safety assessment of post-earthquake reinforced concrete structures using deep learning. *Computer-Aided Civil and Infrastructure Engineering*, *37*(9), 1188–1204.

How to cite this article: Jiang, T., Li, L., Samali, B., Yu, Y., Huang, K., Yan, W., & Wang, L. (2024). Lightweight object detection network for multi-damage recognition of concrete bridges in complex environments. *Computer-Aided Civil and Infrastructure Engineering*, *39*, 3646–3665. https://doi.org/10.1111/mice.13219