Hindawi
Structural Control and Health Monitoring
Volume 2024, Article ID 7857012, 15 pages
https://doi.org/10.1155/2024/7857012

Research Article

Autonomous Identification of Bridge Concrete Cracks Using Unmanned Aircraft Images and Improved Lightweight Deep Convolutional Networks

Fei Song, Ying Sun, and Guixia Yuan

School of Information Technology, Jiangsu Open University, Nanjing 210017, China

Correspondence should be addressed to Guixia Yuan; yuangx@jsou.edu.cn

Received 4 September 2023; Revised 19 December 2023; Accepted 25 January 2024; Published 5 February 2024

Academic Editor: Jian Zhang

The identification of the development of structural defects is an important part of bridge structure damage diagnosis, and cracks are considered the most typical and highly dangerous structural disease. However, existing deep learning-based methods are mostly aimed at the scene of concrete cracks, while they rarely focus on designing network architectures to improve the vision-based model performance from the perspective of unmanned aircraft system (UAS) inspection, which leads to a lack of specificity. Because of this, this study proposes a novel lightweight deep convolutional neural network-based crack pixel-level segmentation network for UAS-based inspection scenes. Firstly, the classical encoder-decoder architecture UNET is utilized as the base model for bridge structural crack identification, and the hourglass-shaped depthwise separable convolution is introduced to replace the traditional convolutional operation in the UNET model to reduce model parameters. Then, a kind of lightweight and efficient channel attention module is used to improve model feature fuzzy ability and segmentation accuracy. We conducted a series of experiments on bridge structural crack detection tasks by utilizing a long-span bridge as the research item. The experimental results show that the constructed method achieves an effective balance between reasoning accuracy and efficiency with the value of 97.62% precision, 97.23% recall, 97.42% accuracy, and 93.25% IOU on the bridge concrete crack datasets, which are significantly higher than those of other state-of-the-art baseline methods. It can be inferred that the application of hourglass-shaped depth-separable volumes can actively reduce basic model parameters. Moreover, the lightweight and efficient attention modules can achieve local cross-channel interaction without dimensionality reduction and improve the network segmentation performance.

1. Introduction

There are nearly 1 million road bridges and 22,000 kilometers of high-speed railways in China, which means that the cumulative length of high-speed railway bridges exceeds 50% of the total length of the line [1, 2]. However, at present, 40% of bridges in China have entered the old age of use for more than 25 years, and the aging of bridge materials and degradation of service performance have become key issues worthy of attention.

Figure 1 shows the commonly used artificial inspection methods in bridge safety management. It can be inferred that the traditional bridge inspection mainly uses the bridge bottom inspection channel and the bridge inspection vehicle (BIV) for inspection [3]. The former is constructed at the same time as the bridge, and due to the aging of components during operation, it usually loses its function within the design service life. On the other hand, the BIV technique is widely used in current bridge inspection activities, but its shortcomings are also very prominent. Firstly, the BIV technique will inevitably occupy the road, which seriously affects the traffic operation efficiency of bridges. Secondly, manual defect detection and identification are still used in BIV inspection tasks. Manual detection is time-consuming and laborious, and the comprehensiveness and objectivity of the results are difficult to guarantee. In addition, when using the above two methods for inspection, the inspectors are usually at a height of tens of meters, which poses high safety hazards.

Figure 1: Common artificial bridge inspection methods.

Traditional nondestructive testing instruments and manual testing methods have certain advantages in detecting internal structural defects of bridges [4–7]. However, there are still certain shortcomings in realizing large-scale defect detection of bridges, which are limited by inspection efficiency and accuracy [8]. With the continuous development of aerial photography technology and remote sensing technology, the application of the unmanned aircraft system (UAS) technique has penetrated more into all walks of life, including the maintenance of civil infrastructure [9–11]. Specifically, the bridge intelligent detection scheme with UAS as the carrier and artificial intelligence as the core is gradually applied to bridge engineering inspection [12]. UAS detection technology has significant advantages, which can realize fixed-point hovering observation, real-time transmission of pictures, and self-control flight. Even for dangerous places such as bridge piers, bases, and bellies, there is no need to build a frame or a hanging basket to cooperate with personnel detection, which greatly improves safety. For some unreachable bridge parts (like belly and cable), UAS can get close to observe and acquire more details [11].

With the development of image recognition technology, automatic identification of structural defects based on computer vision has become a research hotspot [13–18]. With the development of deep neural networks, semantic segmentation has achieved tremendous progress [15, 19, 20]. For this purpose, the encoder-decoder structures are widely used in many actual inspection scenarios [21]. For example, Ren et al. [22] proposed an improved deep fully convolution neural network, called CrackSegNet, to realize the dense pixelwise crack segmentation for tunnel concrete structures. Teng et al. [23] developed a concrete crack detection model using the well-known feature extractor model and the YOLO v2 Network. Xu et al. [24] proposed a lightweight semantic segmentation model for bridge structural damage under complicated backgrounds. Li and Zhao [25] designed an image-based crack detection method using the deep convolutional neural network. Zhang et al. [26] developed an improved UNET-based concrete crack detection algorithm using deep learning technology. Based on the abovementioned literature, it can be inferred that these networks are effective in semantic segmentation, but they suffer from a degree of inadequacy and lack of performance due to the challenging and specialized nature of the task [9, 27]. The specific deficiencies are manifested in the following aspects. (a) Due to the small width of bridge cracks (low pixel ratio), the deep learning-based semantic segmentation model will perform pooling operations multiple times, and the reduced resolution image will cover up the feature information of small cracks and their existence. (b) Cracks in actual engineering are continuously closed, but vision-based crack detection models are prone to fractures and discontinuities, which affect the recognition effect. (c) Inference efficiency is also a concern, so it is important to reduce model parameters and improve computational efficiency. Moreover, most of these studies focus on the cracks in concrete materials. However, compared with half of the concrete materials, bridge concrete cracks are more subtle and complex in characteristics, making their identification more difficult. In addition, the uniqueness of the UAS detection scene puts forward higher requirements for the real-time performance and inference efficiency of the visual defect detection algorithm [28, 29]. The above factors make it urgent to propose a high-efficiency and applicable visual identification method for bridge concrete defects.

To solve the above problems, this study first utilizes UAS detection technology to develop a dataset of bridge concrete structure cracks. Then, the UNET-based encoder-decoder network is used as the base model for training, and the hourglass-like deep separable convolution is inserted into the UNET model to replace the conventional convolutional operation for improving model calculation efficiency. Then, the lightweight efficient channel attention (ECA) is introduced to improve network performance. In the second step, the cheap operation module is utilized in the feature graph simple mapping to further compress the network and the structure reparameterization is utilized to decouple the training and the inference models, further improving the feature fusion ability and inference efficiency.

The main contributions of this study can be attributed as follows:

(1) The application of hourglass-shaped depth-separable volumes can actively reduce basic model parameters, and the lightweight and efficient attention modules can achieve local cross-channel interaction without dimensionality reduction and improve the network segmentation performance.

(2) Cheap operation is used to generate ghost feature maps by reducing redundant feature maps, and parameter reconstruction is utilized to reduce the size of model parameters.

(3) The experimental results show that the constructed method achieves an effective balance between reasoning accuracy and efficiency with the value of 97.62% precision, 97.23% recall, 97.42% accuracy, and 93.25% IOU on the bridge concrete crack dataset, which are significantly higher than those of other state-of-the-art baseline methods.

The remainder of this research is described as follows. Firstly, the basic theory about the lightweight convolution network, parameter reconstruction, and model compression is shown in Section 2. Then, a steel truss girder suspension bridge with two towers and two spans is utilized as the case study, and the details of the UAS-based bridge inspection process are described in Section 3. The indicators and performance parameters of the model in the training, verification, and testing phases are elaborated in Section 4. Lastly, the conclusions and limitations of this research are provided in the final part of this paper.

2. Methodology

According to the calculation requirements of the bridge UAS detection scene, a lightweight and efficient deep convolutional neural network framework for bridge structural crack detection is proposed in this paper. The main work of building a lightweight network model is to ensure that the network performance is basically unchanged or slightly degraded while reducing the number of parameters. Figure 2 shows the architecture of the proposed lightweight and efficient bridge structural defect identification network. It can be inferred that firstly the hourglass-like deep separable convolution is inserted in the UNET-based encoder-decoder network to replace the conventional convolutional operation for model lightweight, and then the lightweight ECA module is introduced to improve network performance. Then, in the second step, the cheap operation module is utilized in the feature graph simple mapping to further compress the network, and the structure reparameterization is utilized to decouple the training and the inference models, further improving the feature fusion ability and the inference efficiency.

2.1. Lightweight UNET-Based Network for Bridge Defect Detection. In semantic segmentation, the goal is to classify each pixel in an image into a specific class. Among them, UNET is a popular architecture used for semantic segmentation tasks, which utilizes an encoder-decoder architecture that allows it to capture both low-level features and high-level context information. Figure 3 shows the general schematic diagram of the UNET network. It can be inferred from this figure that the encoder part of the UNET architecture consists of a series of convolutional and pooling layers that downsample the input image, while the decoder part consists of upconvolutional and concatenation layers that upsample the feature maps to the original image size. This allows the model to capture both local and global context information, which is important for accurate semantic segmentation. Overall, UNET captures both low-level features and high-level context information making it a powerful architecture for semantic segmentation tasks.

A lightweight encoder-decoder network is a type of neural network architecture that is designed to have a smaller number of parameters and computational complexity compared to traditional encoder-decoder networks. It is typically used in scenarios where computational resources are limited or where faster inference is required, like UAS inspection scenes. The lightweight encoder-decoder network achieves its efficiency by employing techniques such as parameter sharing, layer reduction, or using lightweight building blocks like depthwise separable convolutions. These techniques help reduce the number of parameters and operations while still maintaining reasonable performance.

In this study, the classical encoder-decoder architecture UNET is utilized as the base model for bridge structural defect identification, and the hourglass-shaped depthwise separable convolution is introduced to replace the traditional convolutional operation in the UNET model to reduce model parameters. Then, a kind of lightweight and efficient channel attention module is used to improve model feature fuzzy ability and segmentation accuracy.

2.2. The Hourglass-Shaped Depthwise Separable Convolution. Hourglass-shaped deep separable convolution refers to a type of convolutional neural network architecture that uses separable convolutions in an hourglass shape. Figure 4 shows the overall diagram of the hourglass-shaped deep separable convolution. It can be seen from this figure that separable convolutions are used to reduce the number of parameters and improve efficiency while maintaining accuracy. The hourglass shape allows for multiscale feature extraction and improves the network's ability to capture both local and global features. Regarding the hourglass convolution as the basic module of the UNET model, the calculation process for replacing the traditional general-purpose convolution is as follows:

Step 1: 3×3 layer-by-layer convolution (depthwise conv) is used to extract features in the depth direction of the input feature map. At this time, the input X_i is not compressed in dimension, and the extracted spatial features are more expressive.

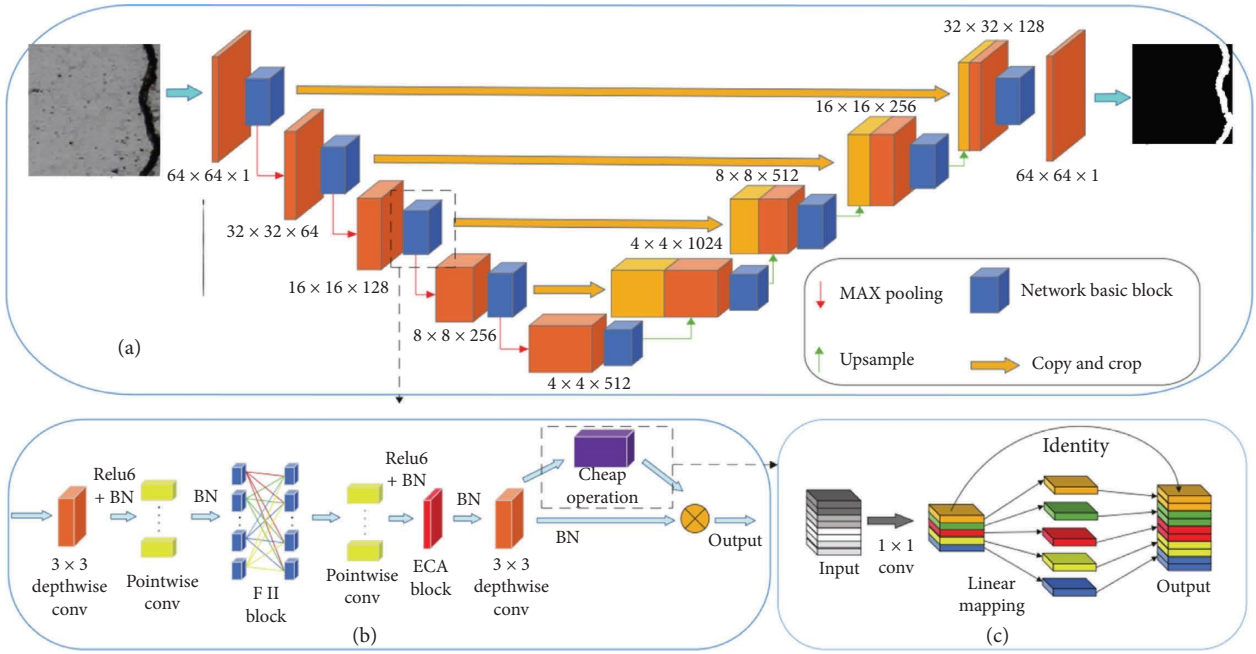


FIGURE 2: Architecture diagram of the proposed lightweight vision-based bridge defect segmentation method: (a) network backbone; (b) network basic block; (c) cheap operation.

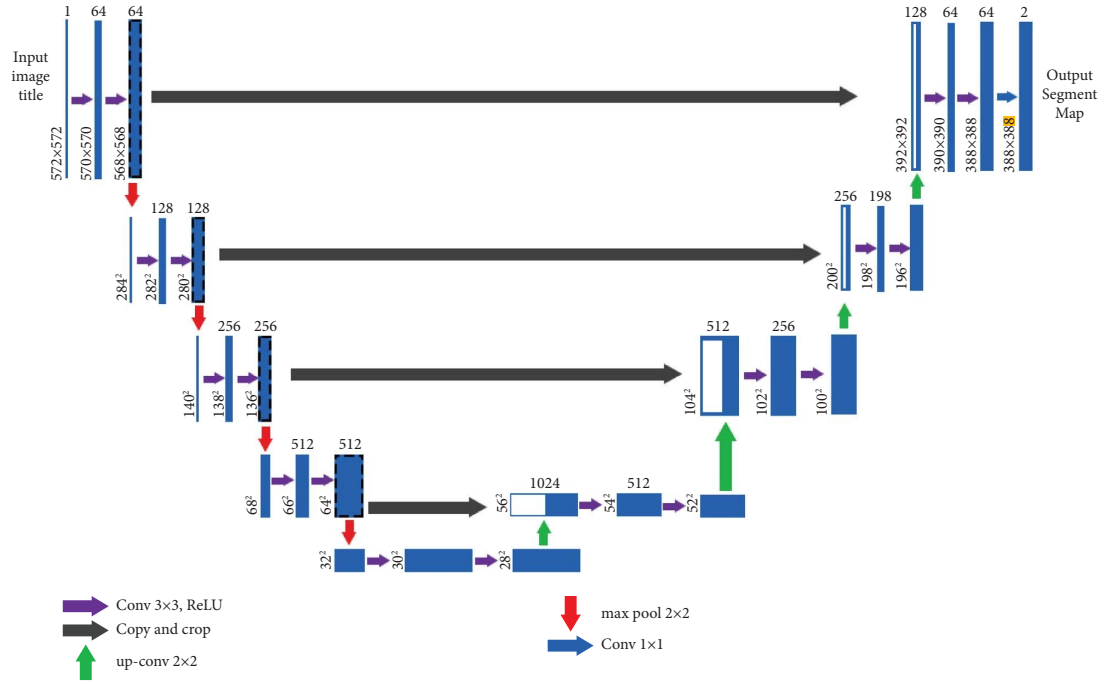


FIGURE 3: The UNET-based semantic segmentation architecture.

Step 2: the feature map $X' \in R^{H\times W\times C}$ after the first layer of layer-by-layer convolution is used as the input of the hourglass-shaped pointwise convolution layer (1×1 pointwise conv).

Step 3: to make up for the loss of these features, a layer of 3×3 layer-by-layer convolutional layers is added at the end to supplement the spatial feature information and make up for the lost part of the spatial information.

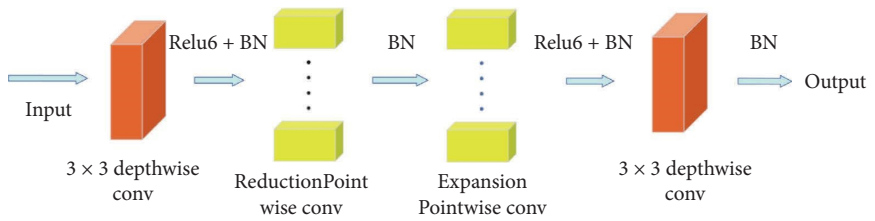


Figure 4: The overall diagram of the hourglass-shaped deep separable convolution.

2.3. The Lightweight Efficient Channel Attention Module. The squeeze-and-excitation (SE) module is a technique used in convolutional neural networks (CNNs) to enhance their representational power [30]. The fundamental function is to introduce a mechanism that allows the network to adaptively recalibrate the feature responses according to their importance. In the SE module, the "squeeze" operation reduces the spatial dimensions of the feature maps, typically using global average pooling. This reduces the number of parameters and computational complexity. Then, the "excitation" operation applies a set of fully connected layers to model the interdependencies between channels. These layers learn channelwise weights, which are then used to rescale the feature maps. The relevant mathematical formulas are expressed as follows:

$$w = \sigma\left(f_{\{W_1,W_2\}}(g(x))\right), \tag{1}$$

where $g(x) = 1/\mathrm{WH}\sum_{i=1,j=1}^{W,H} x_{ij}$, x_{ij} denotes the channel global average pooling, and $\sigma(\cdot)$ represents the sigmoid function. To avoid high model complexity, the sizes of W are set to $C \times (c/r)$.

By incorporating SE blocks into CNN architectures, models can selectively emphasize or suppress certain channels based on their relevance to the task at hand. This helps improve the discriminative power of the network and leads to better performance in various computer vision tasks, such as image classification and object detection. Practical application has proven that the utilization of the SE module can improve network performance, but the addition of most attention modules increases the performance of the network while adding a large amount of computing burden. Therefore, the focus of this study is whether effective channel attention can be learned more efficiently.

To address the abovementioned limitation, this study develops an improved ECA module, which can enhance model performance by introducing several key parameters. ECA is a technique used in computer vision to selectively attend to informative channels in a convolutional neural network. It helps to reduce the computational cost of a network while maintaining or improving its accuracy. ECA achieves this by adaptively weighting the feature maps of each channel based on their importance. This method has been proven to be effective in various computer vision tasks, such as image classification, object detection, and semantic segmentation. Figure 5 demonstrates the overall view of the ECA network. The mathematical expression for this ECA mechanism is as follows:

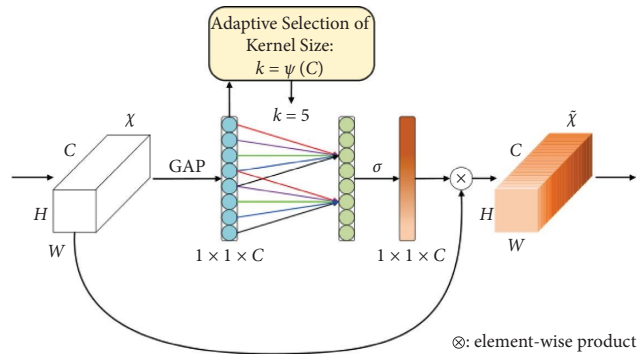


Figure 5: The overall view of the ECA module.

$$g(\varphi) = \frac{1}{\mathrm{WH}}\sum_{i=1,j=1}^{W,H}\varphi_{ij},$$
$$\varphi_o = g(\varphi_s). \tag{2}$$

Then, the channel weights through one-dimensional convolution with extremely low computation are generated, which are as follows:

$$W_j = \varepsilon\left(\sum_{i=1}^{k}\gamma^i y_j^i\right), \tag{3}$$

where y_j^i represents the set of k adjacent channels in φ and ε denotes the sigmoid function.

Finally, the output feature map is obtained by the product of the channel weight and the input feature map, which can make the increase of the operation negligible while improving the defect recognition effect of semantic segmentation network performance.

2.4. Structure Reparameterization and Model Compression for Redundant Feature Map. Structure reparameterization is a technique used in deep learning to modify the structure of a neural network during training. It involves changing the number of neurons or layers in the network to improve its performance. This can be done by adding or removing neurons or by changing the connections between them. The goal of structure reparameterization is to find the optimal network architecture for a given task, which can improve the accuracy and efficiency of the model.

Structural reparameterization is a technique used in machine learning to simplify and improve the training of deep neural networks. It involves modifying the network's

architecture by introducing new parameters that are learned during training, which can help to reduce the number of parameters and improve the overall performance of the network. Figure 6 shows the application of the structural reparameterization for the building blocks of the network. It can be inferred from this figure that a 3×3 convolution filtering, 1×1 convolution filtering, and two bias term transformations are obtained. Then, the 1×1 convolution filter is padded with 0 to form a $3 * 3$ convolution filter, and the convolution filter and offset term distribution are added to the final inference convolution filter W and bias term B.

Feature map-based model compression technology can reduce the use of computing and storage resources while maintaining model accuracy, so it has broad application prospects in resource-constrained environments such as mobile devices and embedded systems. From the perspective of feature map redundancy, one of two very similar feature maps can be regarded as the "shadow" of the other feature map. This type of feature map can be obtained through cheap operation, thereby reducing a large number of 1×1 convolutional operations.

Figure 7 shows the integration of the cheap operation module into the UNET-based encoder-decoder network. The actual compression parameter ratio of cheap operation can be formulated as follows:

$$\frac{C_0 \times H_i \times W_i \times C_i}{(K-1) \times C_o/K + C_o \times H_i \times W_i \times C_i/K} = \frac{K \times H_i \times W_i \times C_i}{(K-1) \times H_i \times W_i \times C_i}. \tag{4}$$

The number of input channels in a deep network is often hundreds or thousands, that is, $C_i \gg K$. Thus, equation (4) can be approximated as follows:

$$\frac{K \times H_i \times W_i \times C_i}{(K-1) + H_i \times W_i \times C_i} \approx K. \tag{5}$$

2.5. Loss Function and Evaluation Metrics. The network is optimized using the sum of the binary cross-entropy loss function and the Dice loss function as the total loss function. The higher the accuracy of the prediction, the lower the loss value, which is a good measure of the difference between the two probability distributions. The binary classification cross-entropy loss function is formulated using the following equals:

$$\text{BCELoss}\,(x_n, y_n) = -w_n\left[y_n \times \lg x_n + (1-y_n) \times \lg(1-x_n)\right], \tag{6}$$

where x_n represents the predicted image (the range is between 0 and 1); y_n represents the real label (the value is 0 or 1); and w_n represents the scaling factor of the loss value, which is used to adjust the weight between samples, and this paper takes 1.

The Dice coefficient is a set similarity measure function, usually used to represent the similarity of two samples:

$$\text{Dice\_Coefficient} = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{7}$$

where $|X \cap Y|$ represents the intersection of two sets X and Y and $|X|$ and $|Y|$ represent the number of elements.

$$\text{Dice\_Loss} = 1 - \frac{2|X \cap Y| + \text{smooth}}{|X| + |Y| + \text{smooth}}. \tag{8}$$

In this study, several evaluation indicators about the inference efficiency are selected to calculate the calculation efficiency of the vision-based defect detection model. The specific mathematical expressions are as follows:

$$\begin{aligned} \text{MFLOPs} &= 2\text{HW}\left(K^2 \cdot C_{l-1} + 1\right) \cdot C_l \cdot 10^{-6}, \\ \text{MParams} &= \sum_{l=1}^{L} K_l^2 \cdot C_{l-1} \cdot C_l \cdot 10^{-6}, \end{aligned} \tag{9}$$

where H and W represent the height and the width of input images, K denotes the kernel size, and C_{l-1} are C_l the input and output channels. It is worth noting that a smaller value of MFLOPs and MParams represents a smaller parameter size of the model, that is, a higher inference efficiency.

In addition to the detection efficiency, the detection accuracy is also an important indicator worthy of attention.

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1 &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \end{aligned} \tag{10}$$

where TN, TP, FP, and FN denote the number of test bridge defect images for true negatives, true positives, false positives, and false negatives, respectively. It is worth noting that a larger value of these indexes represents a more accurate and comprehensive defect identification effect.

3. Case Study

3.1. Project Description. The case object for UAS-based inspection is a steel truss girder suspension bridge with two towers and two spans. The total span of this bridge is 1480 m, and its architectural design and real shots are shown in Figure 8. It is currently the world's largest span plate-truss combined stiffened girder suspension bridge, ranking second in the world and first in China. Since it is near the key control project of Hangrui Expressway, the construction

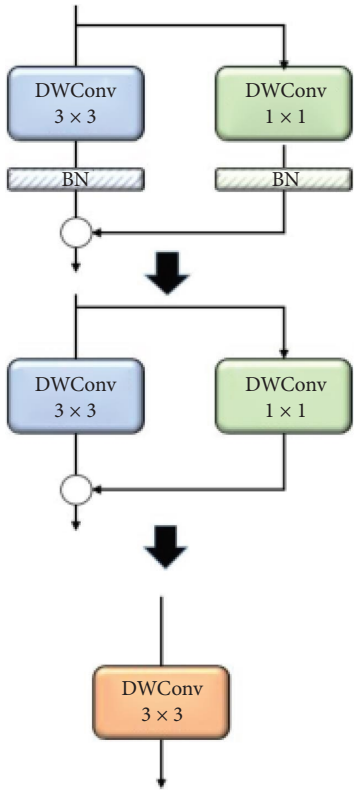


Figure 6: Structural reparameterization of building blocks of network.

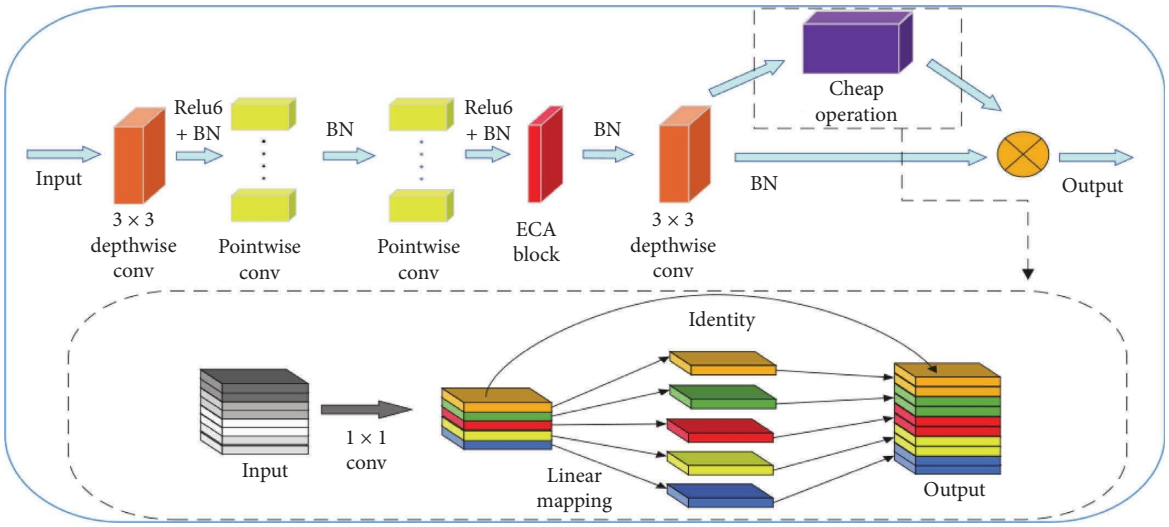


Figure 7: Basic network architecture integrating cheap operation.

Figure 8: The steel truss girder suspension bridge in this study.

conditions of the bridge are complicated and the scale is grand.

3.2. The UAS-Based Inspection Process for Bridge Appearance Detection. The height of the cable-stayed or suspension bridge is high, and due to the obstruction of the cables, it is impossible to use the bridge inspection vehicle to inspect the bottom plate. As a result, the effect of traditional visual inspection is poor, and it is difficult to inspect the outer surface of the cable tower.

To solve the above problems, this study introduces the UAS-based inspection technology equipped with zoom high-resolution cameras for bridge structural appearance detection. Figure 9 shows the UAS-based inspection process for bridge appearance detection. Table 1 shows the operating performance parameters of the UAS equipment used in this bridge inspection case. Table 2 shows the basic parameters of the visual inspection system carried by the UAS. It can be seen from the figure that during the aerial photography flight of the UAS, different safety distances are controlled according to the differences of the detected objects. Bridge piers and tower columns are generally controlled at about 1.5–3 m, while complex parts such as cables and steel components are generally controlled at about 3–5 m. During flight operations, only one side is inspected and photographed at a time, and the upper and lower parts of the bridge are inspected separately. The inspection sequence is from left to right, first up and then down. During the inspection of the lower part, the upper camera is used to operate on the same side as the pilot, and when the upper part is inspected, the lower camera is used to operate on the same side as the pilot.

3.3. Dataset Preparation and Augmentation for Deep Learning. About 100 images of cracks in bridge concrete structures collected by UAS were used for model training and verification evaluation. Figure 10 shows the images of bridge concrete cracks and the manual labeling results. It can be inferred from the figure that the resolution of the image of bridge concrete cracks is high, but the image-pixel ratio of cracks is relatively low because its accurate identification is a challenging task.

After labeling is complete, the defect images collected by the UAS are divided into multiple small images to input into the network for training. To further improve data diversity, data enhancement technology is introduced for data expansion. Data augmentation for images is a technique used to increase the size of a dataset by applying various transformations to the original images. This can help improve the accuracy of DL-based models by exposing them to a wider range of variations in the data. Figure 11 demonstrates the application of data enhancement technology in the bridge crack dataset. It can be inferred from the figure that a series of data enhancement techniques are applied in this study, including flipping, rotation, scaling, saturation, contrast map, and brightness adjustment, to enrich the diversity of the image. A total of approximately 6,000 images with an image resolution of $200 * 200$ after data enhancement were used for defect recognition model training and verification evaluation. In this study, the bridge concrete structural defect dataset was first divided into the training, validation, and test sets according to the ratio of 7 : 2 : 1. There are 4200 defect images used for model training, 1,200 defect images used for model validation, and 600 defect images used for model test. The validation metric was calculated by inference on the validation dataset at the end of each epoch to evaluate the training effect. It should be noted that the indicator to evaluate the defect segmentation ability of the model on the validation set is intersection over union (IOU).

4. Experimental Result and Discussion

4.1. The Model Training Process. This experiment is performed on the Windows 10 operating system. The hardware environment used in this study is $2 \times$ Xeon (R) Gold 5118, 256 GB memory, $1 \times$ Tesla T4, and 1 TB SSD. The software environment is Python 3.7, Cuda12, and Cudnn7.7. The method used in this study is based on the deep learning framework TensorFlow and the software platform Vscode for coding and implementation. Specifically, the number of iterations is set to 200 and the number of batch processing is set to 8. Cosine annealing learning rate is a type of learning rate schedule commonly used in training deep learning models. This study introduces the cosine annealing learning rate change method to adjust the learning rate. Specifically, the initial learning rate of the model is set to 0.01, and in the remaining iterations, the model adjusts the learning rate according to the cosine function curve.

Figure 12 demonstrates the loss functions and metric changes for the 200-epoch process of the developed model. It can be inferred from the figure that the loss function of the built model in the training set gradually decreases steadily and finally tends to converge, indicating that the model has learned enough effective information from sufficient crack data. Moreover, it can be also seen that the BCE loss function has a faster convergence rate and smaller numerical changes, while the Dice loss function has a larger numerical value, greater fluctuations, and a slower convergence rate. Correspondingly, the segmentation performance evaluation index of the model in the validation set gradually increases and eventually tends to converge, indicating that the model has a good crack segmentation ability.

To avoid the model from overfitting or falling into a local optimal solution, in this study, during the model iteration process, the weight coefficients were saved for each iteration, and the optimal model was selected based on the strategy with the highest evaluation index (IOU in this study) on the validation set. In this study, the highest iteration number of the validation set evaluation index occurs in the 162nd iteration, and its maximum IOU value is 0.8306.

4.2. Ablation Study. To further verify the effectiveness of the proposed method, a series of ablation experiments were introduced in this study. Specifically, different mechanisms are introduced to change the architectural composition of the neural network.

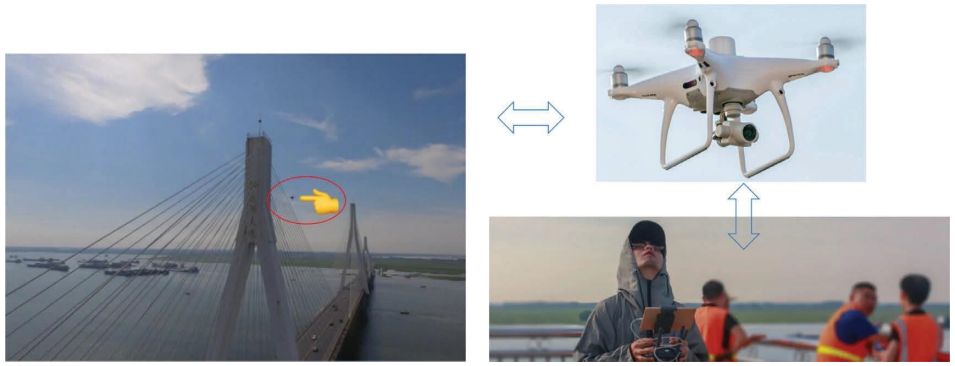


Figure 9: Bridge inspection process display using UAS detection.

Table 1: Main flight performance parameters of UAS.

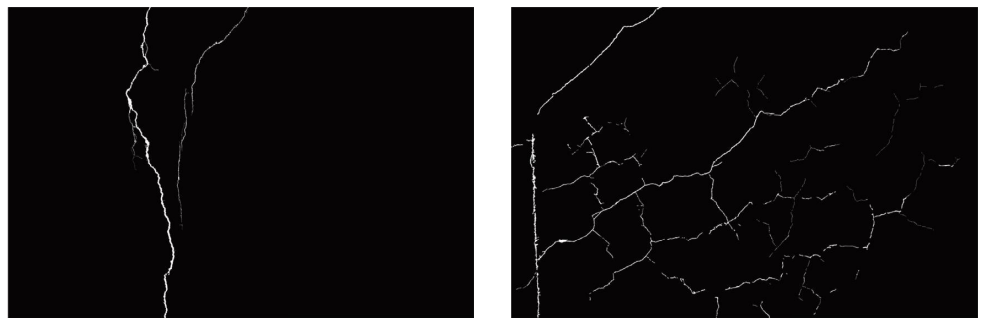
Main indicator	Parameter value
Weight	1375 g
Wheelbase	350 mm
Flight time	30 min
Maximum ascent speed	6 m/s
Maximum descent speed	4 m/s
Maximum horizontal flight speed	72 km/h
Maximum flight altitude	6000 m
Maximum wind resistance rating	10 m/s
Maximum transmission distance	7.5 km
Working temperature	0–40 degrees

Table 2: Main parameters of the visual inspection system of UAS.

Main indicator	Parameter value
Visual system	Forward vision system Downward-looking visual system
Speed measurement range	Flight speed ≤10 m/s (height 2 m, sufficient light)
Height measurement range	0–10 m
Precision hover range	0–10 m
Maximum video stream	60 mbps
Camera resolution	4000 ∗ 3000 pixel

(a)

Figure 10: Continued.



(b)

Figure 10: The bridge concrete crack images and corresponding labeling results. (a) Original images. (b) Manual label annotations.

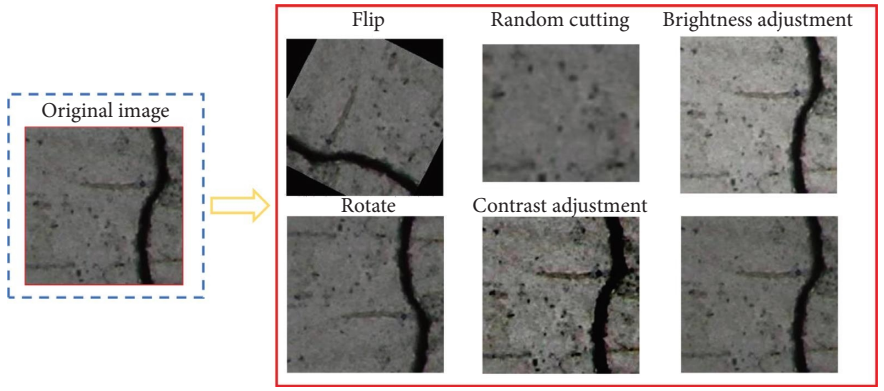


Figure 11: The application of data augmentation in bridge crack images.

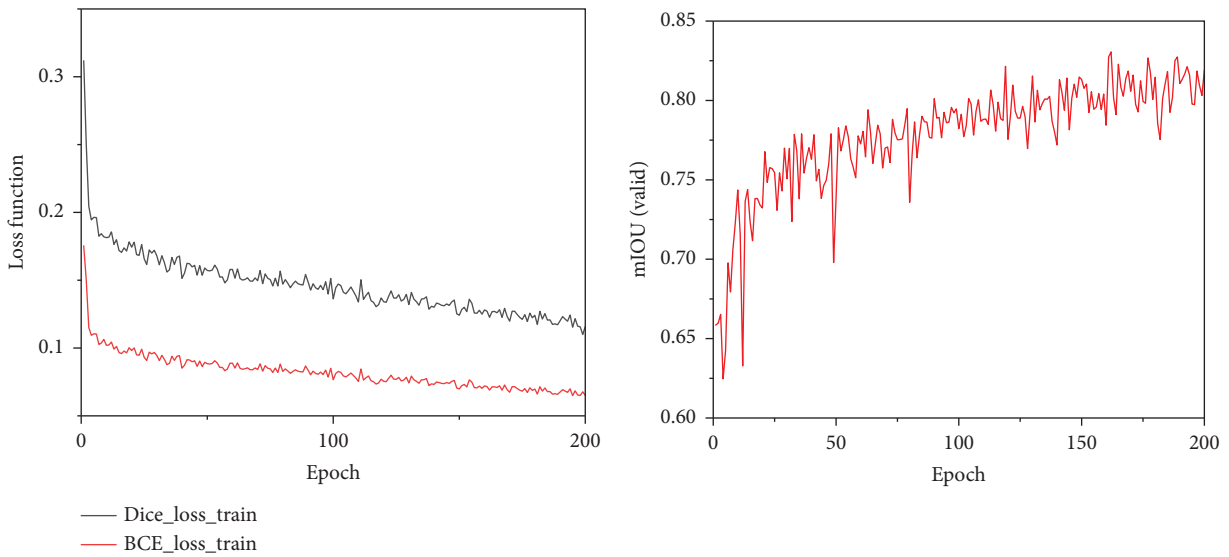


Figure 12: Process lines for loss functions and evaluation metrics: (a) training set; (b) model validation.

(i) Model I: the UNET model with the conventional convolutional network

(ii) Model II: the UNET model with the hourglass-shaped depthwise convolutional network

(iii) Model III: the UNET model with the lightweight ECA attention mechanism

(iv) Model IV: the UNET model with the SE mechanism

(v) Model V: the UNET model with the hourglass-shaped depthwise convolutional network, ECA, and the structural reparameterization

Table 3 shows the detection performance comparison of the ablation experiment results. It can be observed from Table 3 that the introduction of the hourglass convolution module will not significantly reduce the accuracy and performance of the model in defect identification, but at the same time, it can significantly reduce the size of the model parameters and improve the inference efficiency. In addition, the introduction of two types of attention mechanisms (e.g., SE and ECA mechanisms) can improve the feature fusion and defect identification capabilities of neural networks to a certain extent. In addition, the introduction of two types of attention mechanisms can improve the feature fusion and defect identification capabilities of neural networks to a certain extent. However, the lightweight ECA mechanism module brings less increase in the number of parameters, which can improve the efficiency of model reasoning to a certain extent. In addition, the combination of the model structural parameter reconstruction and compression technology can significantly reduce model parameters and improve the real-time reasoning ability of the model. Moreover, it can be also seen from Table 3 about the efficiency comparison of the proposed method and other benchmark methods. It can be inferred that the introduction of the hourglass convolution module can effectively improve the reasoning efficiency of the model, while the introduction of the ECA lightweight attention mechanism will not significantly reduce the reasoning efficiency. In addition, the joint application of model parameter reconstruction and feature compression technology can significantly accelerate the efficiency of model inference.

4.3. Comparison with Other Algorithms. To illustrate the effectiveness of our proposed defect detection model, we compare the proposed model with some popular lightweight vision-based inspection models. The setting of the model parameters and the proposed method are consistent in terms of the training data of the algorithm. For the proposed method and the comparison method, this study uses the same hyperparameters for model training and inference, that is, the number of iterations, the number of batch processing, and the optimal model selection method. Specifically, the number of iterations is set as 200, the number of batch processing is set to 8, and the optimal model was determined according to the strategy with the highest evaluation index (IOU in this study) on the validation set. (i.e., the highest accuracy segmentation index). Finally, the model weight with the highest evaluation index in the validation set is saved and used for model recommendation and performance testing.

These models are described as follows:

(i) UNET: It was proposed by [31] in 2015. It uses an encoder-decoder architecture that allows for a precise segmentation of objects in images. The UNET architecture has become popular in image segmentation tasks due to its ability to effectively capture contextual information and its ability to handle limited training data.

(ii) DEEPLABV3: It was proposed by Liu et al. [32] in 2018. DeepLab V3 uses atrous convolution (also known as dilated convolution) to capture multiscale context information from the input image. It also employs a spatial pyramid pooling module to capture object context at multiple scales. DeepLab V3 has achieved state-of-the-art performance on several benchmark datasets for semantic image segmentation.

(iii) Pyramid Scene Parsing Network (PSPNET): It was proposed by Zhao et al. [33] in 2017. PsPNet utilizes a pyramid pooling module to capture contextual information at multiple scales, allowing for a more comprehensive understanding of the scene. This architecture has achieved excellent performance in various scene parsing benchmarks and has been widely used in computer vision applications.

(iv) Fully Convolutional Network (FCN): It was proposed by Yan et al. in 2015 [34]. FCN replaces the fully connected layers of a traditional CNN with convolutional layers, allowing the network to accept input images of any size and produce output feature maps that are also of variable size. FCN has been widely used in various applications, including medical image segmentation, autonomous driving, and object detection.

Table 4 demonstrates the evaluation of the constructed method and other benchmark advanced algorithms on the test set of the bridge concrete crack dataset. It can be seen from the table that even considering the model detection accuracy, precision, and recall rate at the same time, the constructed method performs better than the other methods, indicating that the method has a strong crack identification performance.

In addition, model inference efficiency is an important evaluation index to measure the applicability of the model in the actual bridge crack inspection process. Table 4 shows the speed comparison between the constructed method and the baseline method in inferring a bridge concrete crack image with a resolution of $600 * 400$ pixels. It can be inferred from the table that the constructed model has a real-time detection effect, and its efficiency is significantly higher than that of other baseline methods. This is mainly due to the comprehensive use of the lightweight convolutional segmentation network PsPNet and the ResNet backbone feature network, which reduces the parameters required for model calculation and the number of network layers.

Table 3: Ablation experiment results.

Models	MFLOPs	MParams	Precision (%)	Accuracy (%)	Recall (%)	Speed (FPS)
Model I	11.26	0.045	96.25	95.21	95.15	18.34
Model II	9.32	0.042	96.30	95.18	95.16	24.05
Model III	9.36	0.043	97.23	96.28	96.83	23.56
Model IV	12.35	0.052	96.82	95.95	95.62	16.36
Model V	9.11	0.033	97.62	97.23	97.42	28.71

Table 4: Model identification performance comparison of different algorithms.

Models	Precision (%)	Recall (%)	Accuracy (%)	IOU (%)
The developed method	97.62	97.23	97.42	93.25
UNET	94.23	94.35	94.29	93.12
FCN	93.89	92.95	93.21	92.86
PSPNET	92.78	92.57	92.82	91.38
DEEPLABV3	93.85	93.74	93.78	92.82

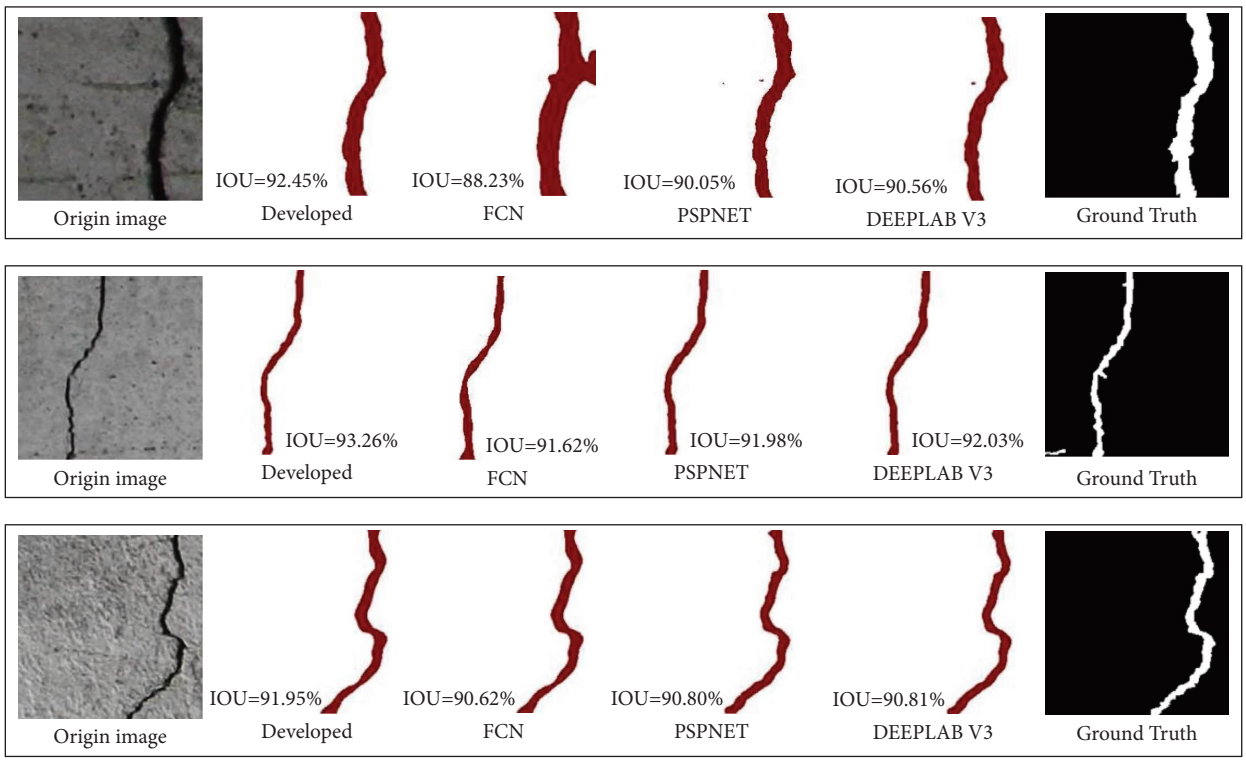


Figure 13: Comparison of the constructed method and other state-of-the-art methods.

Figure 13 shows the bridge structural defect identification comparison of the constructed method and other benchmark models. It can be inferred from the figure that compared with the existing mainstream semantic segmentation algorithms, the crack details obtained by the proposed algorithm are richer and closer to the results of manual observation.

4.4. Real Inspection Scene Validation. To further test the practical application effect of the proposed method in bridge detection, some complex scenes were selected for model evaluation. Among them, the proposed method refers to the model based on the UNET network and lightweight ECA mechanism. Figure 14 demonstrates the effect of applying the constructed defect recognition method to actual bridge structure image recognition. It can be seen from the figure that the proposed method shows better segmentation capabilities in defect images with different backgrounds, including rough concrete background surfaces, holes, red paint distractors, and pockmarked scenes. Even for those scenes with extremely dark lighting conditions and severely insufficient light, the proposed method can still accurately identify and segment the geometry of concrete cracks, which shows that the method has strong generalization and adaptability. This shows that the machine vision-based crack

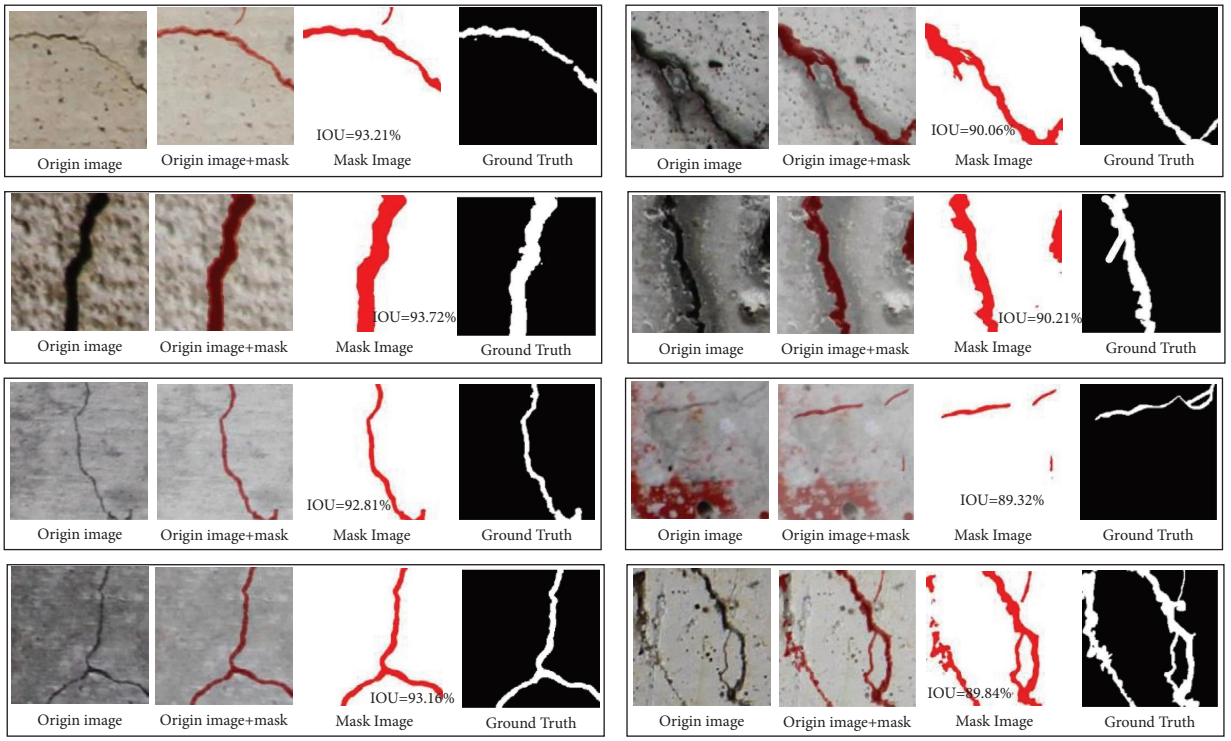


Figure 14: Evaluation of model detection results in complex real inspection scene.

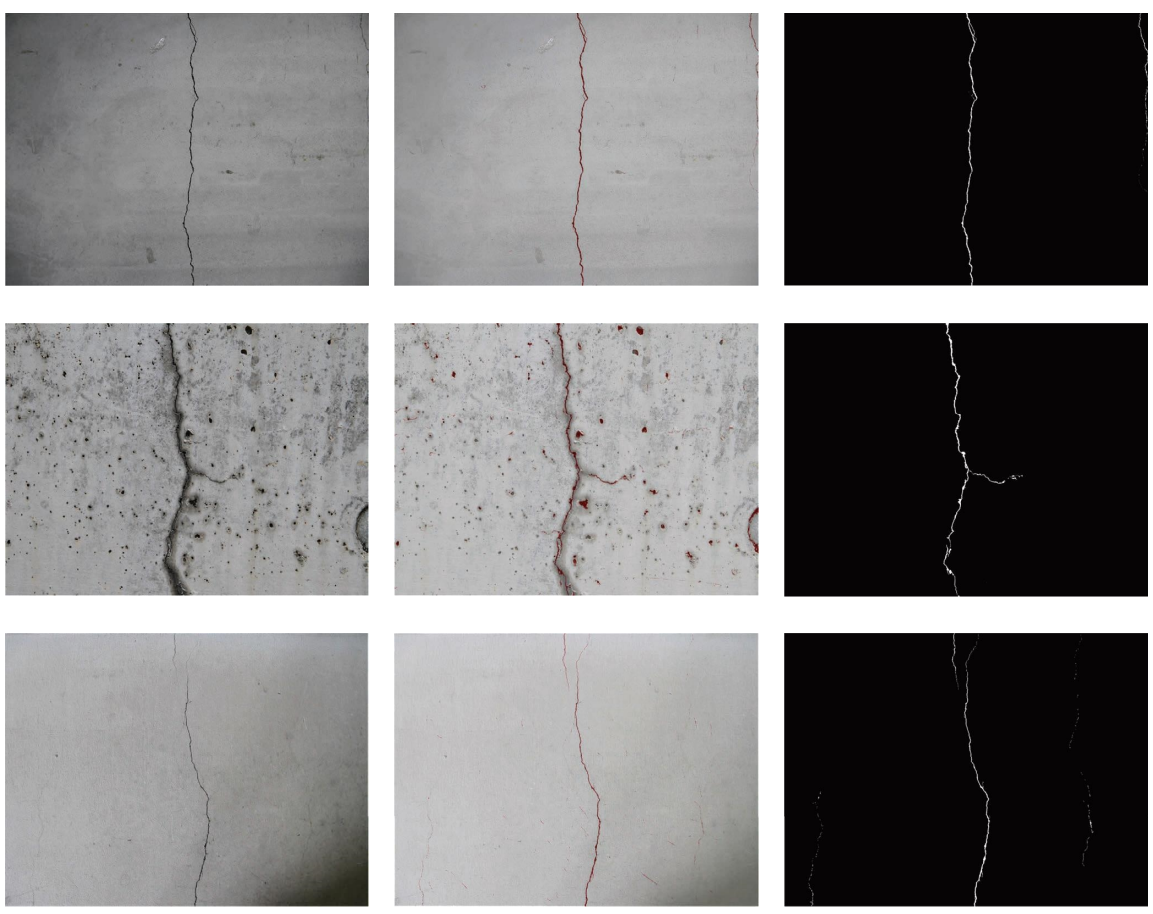


Figure 15: Pixel-level segmentation results of bridge defects using the proposed method. (a) Original image. (b) Defect segmentation result. (c) Ground truth.

identification method has better environmental adaptability than manual detection. Moreover, it can also be seen from the figure that the evaluation indicators of the proposed method's bridge defect segmentation results with different complex backgrounds are higher than those of the comparative method, indicating its better recognition performance.

4.5. Qualitative Evaluation on Large-Scale Images. To achieve high-resolution bridge information collection, UAS equipment is usually equipped with high-definition cameras, so the recognition of high-definition full-scale images is an important part of evaluating the practical application ability of visual inspection models. In this study, using the proposed defect recognition method as a tool, three different images of bridge structural defects are used to verify the effectiveness of the method. It should be noted that the application of the developed method on the high-resolution large-scale structural defect images is a test scenario independent of the training and validation data. Specifically, the model construction in this study is mainly based on small-scale bridge concrete structure defect images for model training and verification evaluation, while further testing of the model robustness is based on high-resolution defect images newly collected by UAS. Figure 15 shows the identification effect of the developed model applied to full-scale bridge structural defect images obtained by UAS-based inspection technology. It can be seen from the figure that the constructed bridge structural defect segmentation model has a good defect-shape segmentation ability, and the defect geometry can be segmented from large-scale images. Even for the cracks at the edge of the bridge structure images, the morphological features of small defects can still be accurately identified and segmented.

5. Conclusions and Discussion

5.1. Conclusions. In this study, a machine vision detection system suitable for bridge UAS detection scenes is proposed. The system can realize high-resolution image acquisition of bridge concrete structural defects through UAS-based photography technology. Firstly, the classical encoder-decoder architecture UNET is utilized as the base model for bridge structural defect identification, and the hourglass-shaped depthwise separable convolution is introduced to replace the traditional convolutional operation in the UNET model to reduce model parameters. Then, a kind of lightweight and efficient channel attention module is used to improve model feature fuzzy ability and segmentation accuracy. On this basis, an end-to-end lightweight efficient inference neural network is proposed to achieve pixel-level bridge structural defect segmentation in different noise scenarios.

The specific contributions of this study are as follows:

(1) Cheap operation is used to generate ghost feature maps by reducing redundant feature maps, and parameter reconstruction is utilized to reduce the size of model parameters.

(2) The experimental results show that the constructed method achieves an effective balance between reasoning accuracy and efficiency with the value of 97.62% precision, 97.23% recall, 97.42% accuracy, and 93.25% IOU on the bridge concrete defect dataset, which are significantly higher than those of other state-of-the-art baseline methods.

5.2. Limitations and Future Discussion. However, some limitations need to be further addressed. Firstly, this study mainly focuses on typical defects such as concrete cracks. In subsequent research, we will study detection and identification methods for multicategory structural defects like calcium precipitation, aggregate exposure, and holes, to further expand the application scope of machine vision inspection technology. Moreover, this study mainly focuses on fast and efficient defect identification and detection, but there is still a lack of research on the follow-up bridge digital twin scene construction and defect mapping. Based on visual defect detection, the construction of the digital twin and bridge information model techniques is an important research content in the future. Realizing the three-dimensional mapping interaction between machine vision inspection models and digital twin scenes is important research content for future bridge safety management. In future research, the proposed method can be further extended to the operation and maintenance of various civil infrastructures such as housing construction and transportation to improve automation and intelligence level.

Data Availability

The data presented in this study are available from the corresponding authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank Liu Bo for his contribution to the code of this study and scientific research assistants for their assistance. This work was supported in part by the National Natural Science Foundation of China (grant no. 62102184) and in part by the Natural Science Foundation of Jiangsu Province (grant no. BK20200784).

References

[1] R. Li, J. Yu, F. Li, R. Yang, Y. Wang, and Z. Peng, "Automatic bridge crack detection using Unmanned aerial vehicle and Faster R-CNN," *Construction and Building Materials*, vol. 362, Article ID 129659, 2023.

[2] S. Sony, S. Gamage, A. Sadhu, and J. Samarabandu, "Multiclass damage identification in a full-scale bridge using optimally tuned one-dimensional convolutional neural network," *Journal of Computing in Civil Engineering*, vol. 36, no. 2, pp. 1–14, 2022.

[3] H. Xu, X. Su, Y. Wang, H. Cai, K. Cui, and X. Chen, "Automatic bridge crack detection using a convolutional neural network," *Applied Sciences*, vol. 9, no. 14, p. 2867, 2019.

[4] A. M. Alani, M. Aboutalebi, and G. Kilic, "Applications of ground penetrating radar (GPR) in bridge deck monitoring and assessment," *Journal of Applied Geophysics*, vol. 97, pp. 45–54, 2013.

[5] E. M. Abdelkader, T. Zayed, and N. Faris, "Synthesized evaluation of reinforced concrete bridge defects, their non-destructive inspection and analysis methods: a systematic review and bibliometric analysis of the past three decades," *Buildings*, vol. 13, no. 3, p. 800, 2023.

[6] Y. Lin, Z. Nie, and H. Ma, "Dynamics-based cross-domain structural damage detection through deep transfer learning," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 1, pp. 24–54, 2022.

[7] L. Chen, W. Chen, L. Wang et al., "Convolutional neural networks (CNNs)-based multi-category damage detection and recognition of high-speed rail (HSR) reinforced concrete (RC) bridges using test images," *Engineering Structures*, vol. 276, Article ID 115306, 2023.

[8] S. Perez Jimeno, J. Capa Salinas, J. A. Perez Caicedo, and M. A. Rojas Manzano, "An integrated framework for non-destructive evaluation of bridges using UAS: a case study," *Journal of Building Pathology and Rehabilitation*, vol. 8, no. 2, p. 80, 2023.

[9] M. Mohammadi, M. Rashidi, V. Mousavi, A. Karami, Y. Yu, and B. Samali, "Quality evaluation of digital twins generated based on uav photogrammetry and tls: bridge case study," *Remote Sensing*, vol. 13, no. 17, pp. 3499–3522, 2021.

[10] J. Chen and D. Liu, "Bottom-up image detection of water channel slope damages based on superpixel segmentation and support vector machine," *Advanced Engineering Informatics*, vol. 47, Article ID 101205, 2021.

[11] A. Galdelli, M. D'imperio, G. Marchello et al., "A novel remote visual inspection system for bridge predictive maintenance," *Remote Sensing*, vol. 14, no. 9, pp. 2248–2317, 2022.

[12] D. Jana, S. Nagarajaiah, and Y. Yang, "Computer vision-based real-time cable tension estimation algorithm using complexity pursuit from video and its application in Fred-Hartman cable-stayed bridge," *Structural Control and Health Monitoring*, vol. 29, no. 9, pp. 1–21, 2022.

[13] Q. Mei, M. Gül, and M. R. Azim, "Densely connected deep neural network considering connectivity of pixels for automatic crack detection," *Automation in Construction*, vol. 110, Article ID 103018, 2020.

[14] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "Uav-yolo: small object detection on unmanned aerial vehicle perspective," *Sensors (Switzerland)*, vol. 20, no. 8, pp. 2238–2312, 2020.

[15] H. Zoubir, M. Rguig, M. El Aroussi, A. Chehri, R. Saadane, and G. Jeon, "Concrete bridge defects identification and localization based on classification deep convolutional neural networks and transfer learning," *Remote Sensing*, vol. 14, no. 19, p. 4882, 2022.

[16] Y. Xu, Y. Fan, Y. Bao, and H. Li, "Task-aware meta-learning paradigm for universal structural damage segmentation using limited images," *Engineering Structures*, vol. 284, Article ID 115917, 2023.

[17] Y. Xu, S. Li, D. Zhang et al., "Identification framework for cracks on a steel structure surface by a restricted Boltzmann machines algorithm based on consumer-grade camera images," *Structural Control and Health Monitoring*, vol. 25, no. 2, pp. 20755–e2120, 2018.

[18] J. Zhao, F. Hu, W. Qiao et al., "A modified U-net for crack segmentation by Self-Attention-Self-Adaption neuron and random elastic deformation," *Smart Structures and Systems*, vol. 29, pp. 1–16, 2022.

[19] Y. Su, J. Cheng, H. Bai, H. Liu, and C. He, "Semantic segmentation of very-high-resolution remote sensing images via deep multi-feature learning," *Remote Sensing*, vol. 14, no. 3, pp. 533–625, 2022.

[20] S. Teng, X. Chen, G. Chen, L. Cheng, and D. Bassir, "Structural damage detection based on convolutional neural networks and population of bridges," *Measurement*, vol. 202, Article ID 111747, 2022.

[21] J. Chen, D. Zhang, H. Huang, M. Shadabfar, M. Zhou, and T. Yang, "Image-based segmentation and quantification of weak interlayers in rock tunnel face via deep learning," *Automation in Construction*, vol. 120, Article ID 103371, 2020.

[22] Y. Ren, J. Huang, Z. Hong et al., "Image-based concrete crack detection in tunnels using deep fully convolutional networks," *Construction and Building Materials*, vol. 234, Article ID 117367, 2020.

[23] S. Teng, Z. Liu, G. Chen, and L. Cheng, "Concrete crack detection based on well-known feature extractor model and the YOLO_v2 network," *Applied Sciences*, vol. 11, no. 2, pp. 1–13, 2021.

[24] Y. Xu, Y. Fan, and H. Li, "Lightweight semantic segmentation of complex structural damage recognition for actual bridges," *Structural Health Monitoring*, vol. 22, no. 5, pp. 3250–3269, 2023.

[25] S. Li and X. Zhao, "Image-based concrete crack detection using convolutional neural network and exhaustive search technique," *Advances in Civil Engineering*, vol. 2019, Article ID 6520620, 12 pages, 2019.

[26] L. Zhang, J. Shen, and B. Zhu, "A research on an improved Unet-based concrete crack detection algorithm," *Structural Health Monitoring*, vol. 20, no. 4, pp. 1864–1879, 2021.

[27] H. Kim, J. Yoon, and S. H. Sim, "Automated bridge component recognition from point clouds using deep learning," *Structural Control and Health Monitoring*, vol. 27, no. 9, pp. 1–13, 2020.

[28] Y. Xu, Y. Bao, J. Chen, W. Zuo, and H. Li, "Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images," *Structural Health Monitoring*, vol. 18, no. 3, pp. 653–674, 2019.

[29] Y. Xu, W. Qian, N. Li, and H. Li, "Typical advances of artificial intelligence in civil engineering," *Advances in Structural Engineering*, vol. 25, no. 16, pp. 3405–3424, 2022.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June, 2018.

[31] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 568–576, 2020.

[32] C. Liu, L.-C. Chen, F. Schroff et al., "Auto-deeplab: hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 82–92, Long Beach, CA, USA, June, 2019.

[33] J. Zhou, M. Hao, D. Zhang, P. Zou, and W. Zhang, "Fusion PSPnet image segmentation based method for multi-focus image fusion," *IEEE Photonics Journal*, vol. 11, no. 6, pp. 1–12, 2019.

[34] L. Yan, D. Liu, Q. Xiang et al., "PSP net-based automatic segmentation network model for prostate magnetic resonance imaging," *Computer Methods and Programs in Biomedicine*, vol. 207, Article ID 106211, 2021.