IJRCS

International Journal of Robotics and Control Systems
Vol. 3, No. 2, 2023, pp. 259-269
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

ASCEE

Bridge Crack Detection Based on Attention Mechanism

Geng Chuang [a, d, 1, *], Cao Li-jia [b, c, d, 2]

[a] Sichuan University of Science & Engineering, School of Automation & Information Engineering, Yibin 644000, China
[b] School of Computer Science & Engineering, Sichuan University of Science & Engineering, Yibin 644000, China
[c] Artificial Intelligence Key Laboratory of Sichuan Province, Yibin 644000, China
[d] Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing, Yibin, 644000, China
[1] geng_c2222@163.com; [2] caolj@suse.edu.cn
* Corresponding Author

ARTICLE INFO

Article history
Received February 22, 2023
Revised April 08, 2023
Accepted April 12, 2023

Keywords
Object detection;
Bridge Crack Detection;
Attention Mechanism;
Robot

ABSTRACT

With the strong support of the country for bridge construction and the increase in supervision of the safety of old bridges, the visual-based bridge crack target detection has a problem of incomplete target framing due to the characteristics of the bridge crack target, reflecting the current algorithm model's poor ability to accurately identify targets. In this paper, YOLO V5 algorithm was used to address the issue of poor accuracy in bridge crack target detection, and a relevant bridge crack detection dataset was created. Three attention mechanisms, SENet, ECALayer, and CBAM, were respectively fused to improve the model's feature fusion part, and comparative experiments were conducted. The experimental results show that the improved algorithm has increased from 80.5% to 87% in mAP50-95 indicators compared to the original algorithm.

This is an open-access article under the CC–BY-SA license.

1. Introduction

In recent years, with the increase in budget for bridge construction and the strengthening of supervision of the safety of old bridges by the government, the issue of bridge crack detection has become a hot topic. Currently, the detection of bridge cracks mainly relies on bridge inspection vehicles, which use folding or telescopic arms attached to the vehicle or a hybrid aerial ladder to move a platform carrying inspection personnel under the bridge for crack detection while driving along the edge of the bridge. However, this method wastes manpower, materials, and financial resources, and there are significant limitations on the viewing angle of the inspection personnel, and the detection results depend on their level of expertise, making it difficult to carry out large-scale inspections. Therefore, visual-based bridge inspection projects have emerged, using cameras mounted on drones to capture relevant bridge images and utilizing bridge crack target detection algorithms to detect cracks on bridges.

In the past two decades, traditional visual-based methods for detecting bridge cracks have mainly used graphical analysis [1], pattern recognition [2], edge detectors [3], [4], line detectors [5], [6], and threshold segmentation [7]. These methods can achieve good detection accuracy for continuous cracks with high contrast, demonstrating the feasibility of automated detection based on vision. However, during the actual process of collecting bridge crack images, factors such as collection equipment, shooting angle, external lighting, and vibration often affect the model's ability to achieve good

detection results. In response to this situation, simple detection algorithms based on traditional image processing can no longer meet the growing production demands in terms of robustness.

With the development of the machine learning field, it has become possible to detect bridge cracks based on vision in complex environments. Henrique et al. [8] proposed a machine learning-based method that detects crack blocks by statistical processing of the mean and standard deviation of the gray values within image blocks. Edrardo Zalama et al. [9] proposed an instrumented vehicle for detecting cracks based on an imaging system, two inertial profilers, differential global positioning systems, and network cameras. They designed a method based on Gabor filters to identify horizontal and vertical cracks, improved single classifier results using the Adaboost algorithm [10], validated the feasibility of the solution and method with a large database, and obtained good results through rigorous testing. Prateek Prasanna et al. [11] proposed a new automatic crack detection algorithm called STRUM (Spatially Tuned Robust Multi-Feature), which was tested on real bridge data using an advanced robot bridge scanning system. This algorithm mainly avoids manually adjusting threshold parameters through machine learning classification algorithms. It fits potential cracks in space and calculates visual features in that area, and the entire algorithm is built using reasonable crack information representation and a classifier trained multiple times. Through scientific experiments, the improved algorithm achieved the highest accuracy of up to 95%. Cord and Chambon [12] described the texture features of cracks in images using a model designed with the AdaBoost algorithm. Shi et al. [13] proposed a model based on random forests to extract image features and detect cracks in the CrackForest road crack dataset. These traditional object detection methods based on expert features have deviations in robustness in practical applications due to their dependence on expert features.

With the development of deep learning, CNN (Convolutional Neural Networks) based on deep learning has gradually become a hot research direction in object detection. Region-based methods such as R-CNN [14], Fast R-CNN [15], and Faster R-CNN [16] achieve high detection accuracy but have slow detection speeds, which cannot meet practical needs. Regression-based algorithms such as YOLO [17] and SSD [18] have good performance in both detection accuracy and speed and have become popular model architectures in the field of object detection.

Convolutional neural network algorithms based on deep learning can also be applied to bridge crack detection problems. Zhang et al. [19] used a convolutional neural network to achieve single-pixel classification, which can predict whether a single pixel belongs to a crack. However, this method did not utilize the semantic information of crack targets very well and required manually designed feature extractors for image preprocessing, which lacks universality. Zou et al. [20] proposed DeepCrack, which is the first detector to use multiscale convolutional features to detect cracks, opening up a new path for pixel-level crack detection in bridges based on deep learning. Wang et al. [21] used HDCBs to learn spatial features by adding them to the neural network, enlarging the receptive field of the convolutional kernel, avoiding the loss of a large amount of semantic information due to the grid effect, and maintaining the continuity of pixel-level cracks. Shuai Teng et al. [22] tested different object detection algorithms for the detection of bridge surface defects by introducing Gaussian white noise. Through experiments, it was found that using transfer learning and data augmentation methods to improve the YOLO V3 [23] network can effectively improve the bridge defect detection capability but does not address the most important bridge crack recognition problem. Jinsong Zhu et al. [24] improved the VGG-16 network classifier and collected real bridge surface defect images labeled into seven categories. Through comparison with multiple detection and classification algorithms, it was found that the improved algorithm is superior to other algorithms, providing a feasible solution for bridge surface defect detection and classification. Philipp Hüthwohl et al. [25] collected defect data from many different bridges to create a relevant classification dataset and proposed a three-level concrete defect classifier for bridge defect detection, achieving a detection accuracy of 85% through experiments. Sizeng Zhao et al. [26] combined the YOLO V5 algorithm with 3D photogrammetric reconstruction methods to propose a defect detection method for concrete dams. An improved algorithm was proposed for the problems of complex backgrounds and blurred boundaries, which improved accuracy by 3.8% compared to the original algorithm, especially for small object detection. Gang Li et al. [27] used a fully convolutional network to extract bridge crack

features and then used a naive Bayes data fusion model to segment the cracks. Compared with traditional visual feature detection methods, there was a significant improvement in accuracy and detection time.

However, none of the above-mentioned studies have explored the issue of high-precision localization in bridge crack detection. Insufficient high-precision localization ability can lead to successful recognition of the target, but the label box cannot fully enclose the target, which affects the subsequent risk assessment of bridge cracks. Therefore, this paper will conduct relevant research on the high-precision localization of bridge cracks.

2. Related Work

The attention mechanism is a method of allocating limited computing resources to important local information. This method is consistent with the cognitive rules of the human brain and eyes and is a bionic neural network-assisted algorithm. In recent years, it has been widely used in the field of computer vision and has been proven to be beneficial in improving model performance. The essence of the attention mechanism is to locate the information that is of interest and beneficial to the recognition results, suppress irrelevant information, and output the results in the form of a probability map or probability feature vector.

2.1. Channel Attention Mechanism

If we divide them by dimension, convolutional neural networks in the field of image processing are two-dimensional. One dimension contains information about the spatial scale of the image, namely its width and height. The other dimension contains information about the image's channels. There are two commonly used channel attention mechanisms: SENet (Squeeze and Excitation Net) [28] and ECA modules (Efficient Channel Attention) [29].

SENet is a channel-based attention mechanism model that models the importance of each feature channel in an image and enhances or suppresses different channel information for different recognition tasks. The principle diagram of the SENet module is shown in Fig. 1.

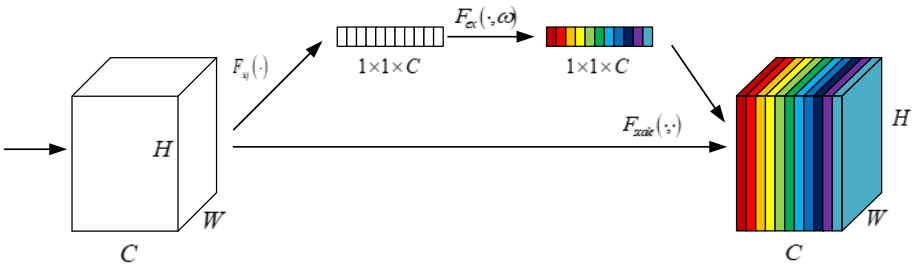

Fig. 1. SENet Module

In Fig. 1, H and W represent the height and width of the feature map in the spatial dimension, while C represents the number of channels in the feature map.

First, the feature map is compressed along the spatial dimensions using a Squeeze operation $F_{sq}(\cdot)$, similar to global average pooling. After this operation, the number of feature channels remains the same. For each feature map channel, a weight value ω is generated using a function $F_{ex}(\cdot,\omega)$, and then the weights are normalized and multiplied with the original feature map channel-wise using a function $F_{scale}(\cdot,\cdot)$ to complete the channel attention operation. The feature weights are learned using a fully connected network based on the result of the loss function, avoiding feature weight obtained solely based on the numerical values of feature channels. This ensures that the weights of effective feature channels are larger, resulting in higher learning efficiency.

However, the dimensionality reduction used in SENet can affect the predictive performance of channel attention and result in low efficiency in capturing channel dependencies in images. Therefore, the ECA module was developed to reduce the dimensionality reduction and improve cross-channel

interaction capabilities, leading to improved model performance. The principle diagram of the ECA module is shown in Fig. 2.

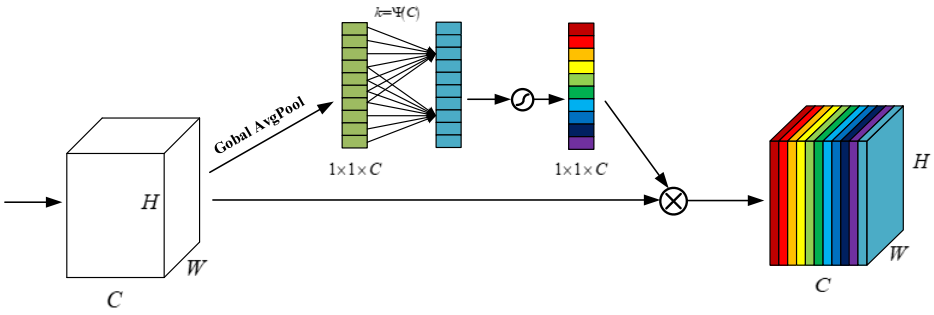

Fig. 2. ECA Module

First, the input feature map is compressed in spatial dimension by using global average pooling. Then, the inter-channel dependencies of the compressed feature map are learned by applying a 1×1 convolution. Next, the learned channel attention information, which contains the weight information, is multiplied by the input feature map channel-wise.

SENet uses fully connected layers (FC) to globally learn the input channel features, while the ECA module uses a 1×1 convolution to locally learn the channel correlation information. By using a dynamic convolutional kernel size, the ECA module can learn the correlation between different channels. When the number of channels is large, a larger kernel size is used to perform 1×1 convolution to achieve cross-channel interaction with more channel information. When the number of channels is small, a smaller kernel size is used to perform 1×1 convolution to achieve cross-channel interaction with less channel information.

The adaptive function of a dynamic convolution kernel is

$$k = \psi(C) = \left| \frac{log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{1}$$

where k is the convolution kernel size, C is the number of channels, which $|\ \ |_{odd}$ means an odd number for the result, γ and b generally set to 2 and 1, which is used to change the ratio between the number of channels C and the convolution kernel size k.

2.2. Mixed Attention Mechanism

CBAM (Convolutional Block Attention Module) [30] is one of the representative methods in the hybrid attention mechanism, which combines channel attention and spatial attention mechanisms. The structure diagram of the CBAM module is shown in Fig. 3.

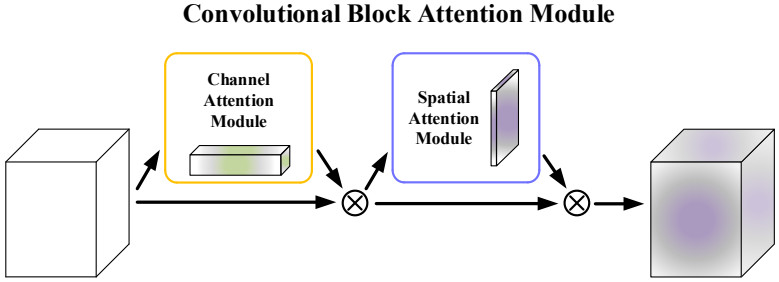

Fig. 3. CBAM Module

CBAM is an improvement based on the SENet method, which models the importance of channel features using channel attention and the degree of attention to spatial positions using spatial attention. CBAM learns the channel and spatial features of the feature map separately, which allows it to improve model performance in most cases and also has a wider range of applications.

The channel attention in CBAM is similar to SENet, and its block diagram is shown in Fig. 4.

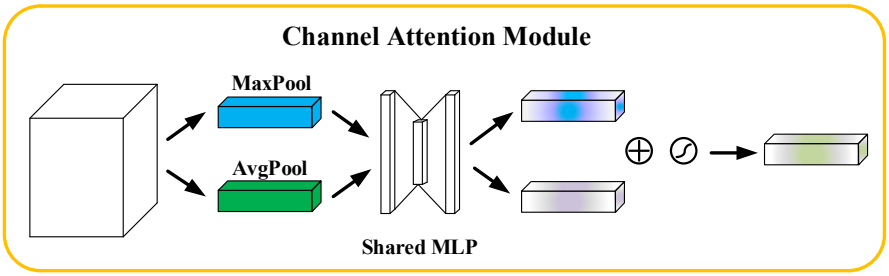

Fig. 4. Channel attention in CBAM

In Fig. 4, both max pooling and average pooling algorithms are used to compress the input feature map dimension and then passed through several Shared MLP (Multilayer Perceptron) layers. The Shared MLP layers use a 1×1 convolution to extract channel feature information. Finally, the attention weights of channel features are obtained through a sigmoid activation function after fusing on two channels. The obtained attention weights of channel features are then fused with the original feature map and sent as input to the spatial attention module.

The principle underlying the spatial attention mechanism is that different regions of an image contribute differently to the recognition task, and improving the model's performance only requires focusing on the regions that have a higher contribution to the task, which can enhance the model's performance and reduce computation. Essentially, the spatial attention mechanism locates the target and performs some transformations to obtain corresponding weights during the learning process. In the mixed-domain attention mechanism, the spatial attention mechanism is shown in Fig. 5.

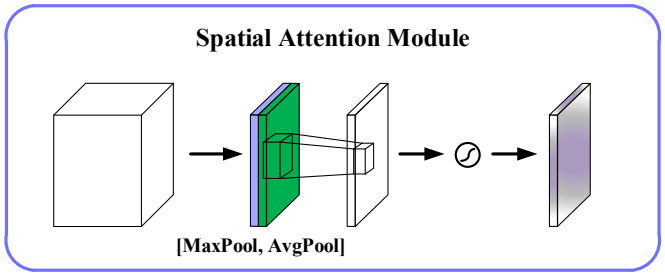

Fig. 5. Spatial attention in CBAM

As shown in Fig. 5, the spatial attention mechanism in CBAM first reduces the dimensionality of the channels by applying both max-pooling and average-pooling operations. Then, the results are concatenated into a feature map, which is further processed by a convolutional layer to learn spatial features. Finally, a sigmoid activation function is applied to obtain the attention weights for the spatial features.

In order to improve the performance of the CBAM module for bridge crack recognition tasks and enhance the learning efficiency of the module for features, this paper added three convolutional layers on top of the CBAM module. The enhanced CBAM module with the additional three convolutional layers is referred to as CBAMC3, which has the advantages of high lightweight, strong applicability, and strong performance improvement.

2.3. Fusion with YOLO V5

YOLO V5 algorithm is a one-stage object detection algorithm based on regression. In the data preprocessing process, the same mosaic image online enhancement method as YOLO V4 algorithm is used to expand the number of small targets in a single batch, which improves the network's ability to recognize small target objects and increases the data information of a single batch. In the backbone network, FPN feature pyramid structure is used to extract and fuse feature information from the bottom up. In the neck structure, the PAN (Path Aggregation Network) network structure is used to fuse the

top-down PAN, shortening the path between the bottom-level features and the prediction layer. The CSP (Cross Stage Partial Network) layer is used instead of the residual structure connection layer, which enhances the model's learning ability, lightweight the model, maintains the model's accuracy performance and reduces the computational bottleneck.

YOLO V5 algorithm improves the problem of class imbalance that existed in previous YOLO algorithms. In previous YOLO algorithms, positive samples were defined based on the IOU value between the anchor box and the true target box. When the IOU value was greater than the threshold, the anchor box was set as a positive sample. However, due to the one-to-one correspondence between anchor boxes and true target boxes, there could only be as many positive samples as true target boxes, resulting in class imbalance. YOLO V5 algorithm defines positive samples based on the aspect ratio between anchor boxes and true target boxes. When the aspect ratio is less than a threshold, it is defined as a positive sample. Additionally, YOLO V5 predicts the same target in nearby grids simultaneously to increase the number of positive samples, effectively solving the problem of class imbalance.

Regarding the problem of bridge crack recognition, because crack targets are elongated, discontinuous, and have large-scale changes, using the YOLO V5 algorithm can, to some extent, avoid the problem of imbalance between positive and negative samples and further improve model performance. However, the accuracy of YOLO V5 algorithm in recognizing some bridge crack targets cannot meet the requirements. Therefore, a method of fusing attention mechanism is used to further improve the performance of YOLO V5 algorithm. The improved YOLO V5 algorithm network structure is shown in Fig. 6.

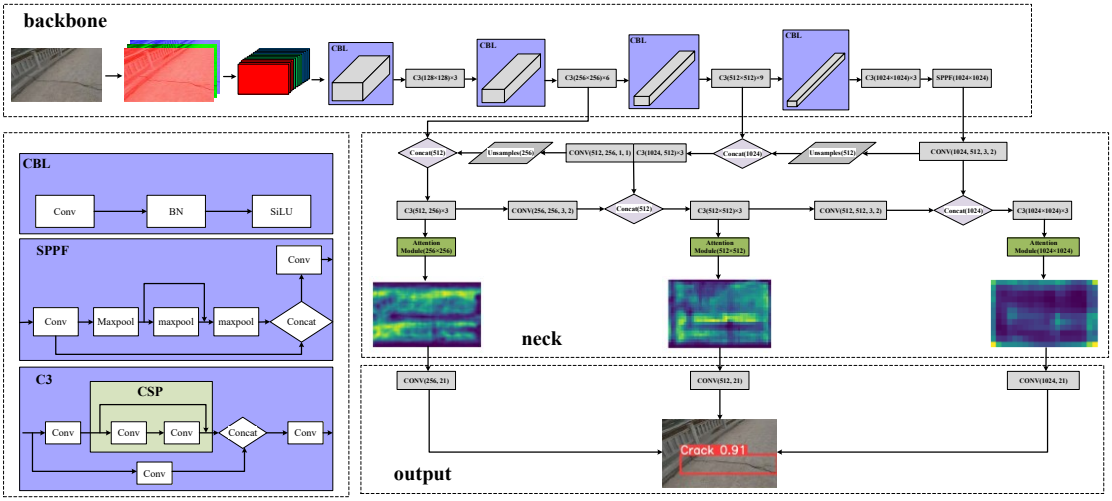

Fig. 6. Improved YOLO V5 Module

Incorporating attention mechanisms before the convolutional layers in the YOLO V5 network's prediction layer can increase the influence of the learned attention weights on the final performance of the model. By validating the effectiveness of different attention mechanisms in addressing the bridge crack recognition problem, it is possible to better solve real-world bridge crack recognition problems.

3. Experiment

3.1. Experiment Environment and Evaluation Indicators

The experimental environment is a high-performance server dedicated to deep learning object recognition, with two high-performance RTX 8000 graphics cards running the stable version of Ubuntu 20.04.3 LTS. The object recognition framework is PyTorch, version 1.12, and the basic YOLO V3 algorithm is the PyTorch version from Ultralytics. The experimental dataset is a self-made dataset, which includes images collected from the bridge crack detection dataset and real-world bridge crack images. The dataset has been manually annotated and verified multiple times for accuracy and

practicality in addressing bridge crack problems. The number of samples in the dataset is shown in Table 1.

Table 1. Number of samples in the dataset

Dataset	Number of images	Bridge crack samples	Bridge damage samples
Train	1225	2568	547
Val	306	642	137
Test	306	642	137
Total	1531	3210	684

The evaluation metric uses the concept of "average precision" referenced in the current mainstream VOC 2007. Precision P represents the proportion of correct predictions made by the model, while recall R represents the coverage of the target category in the recognition results. In the object recognition task, there are two types of samples: positive and negative, and two types of detection results: correct and incorrect. Positive samples predicted correctly are defined as TP, positive samples predicted incorrectly as FP, negative samples predicted correctly as TN, and negative samples predicted incorrectly as FN. Precision and recall can be calculated as (2)-(3).

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + TN} \tag{3}$$

According to precision, the average precision (AP) for each class can be calculated, $AP = \int_0^1 p(r)\, dr$ and then the mean average precision (mAP) for all classes can be computed, $mAP = \frac{1}{m} \int AP$ where m is the number of classes, the commonly used mAP has different versions depending on the IOU threshold used, such as mAP50 and mAP50-95. The latter means that the AP values are calculated with IOU thresholds ranging from 0.5 to 0.95 with a step size of 0.05, and then the average value is taken. Compared to mAP50, mAP50-95 can better reflect the performance of the algorithm and also demonstrates the algorithm's ability to recognize targets with high accuracy and confidence.

3.2. Comparative Experiments

To verify the performance improvement brought by the fusion of three different attention mechanisms with YOLO V5, a comparative experiment method was adopted to validate. Meanwhile, the performance of the improved YOLO V5 algorithm was compared with the original YOLO V5 algorithm and YOLO V3 algorithm in the bridge crack detection task to evaluate the performance parameters. The relevant experiments have carried out the re-clustering of anchor boxes in advance. The clustering algorithm used was K-Means algorithm, and the same experimental conditions and dataset were used for re-clustering to eliminate external factors that may interfere with the performance comparison of the models. The performance indicators of the five algorithm models are shown in Table 2.

Table 2. Performance comparison experiment

Method	Precision (%)	Recall (%)	mAP50 (%)	mAP50-95 (%)
YOLO V3	95.3	80.4	94.5	81.4
YOLO V5	97.4	89.3	97.1	80.5
YOLO V5 + SENet	97.2	91	95.3	82.8
YOLO V5 + ECALayer	97.3	90.8	95	84.2
YOLO V5 + CBAMC3	**99.3**	**92.6**	**97.6**	**87**

According to Table 2, the YOLO V5 algorithm with CBAMC3 module fusion shows the best performance, surpassing other algorithms in precision, recall, mAP50, and mAP50-95. The two improved algorithms with fusion channel attention mechanisms did not surpass the original algorithm

in mAP50 but did so in mAP50-95, indicating that the improved fusion channel attention mechanisms have good performance in addressing the high-precision issue of bridge crack target detection tasks. The improved algorithm with fused hybrid domain channel attention mechanism, CBAMC3, achieved a 0.5% increase in mAP50 and a 6.57% increase in mAP50-95 compared to the original algorithm, showing good improvement for high-precision localization issues.

3.3. Experiment Results

The F1 curve and PR curve can effectively demonstrate the convergence process and performance of the model. The F1 curve and PR curve for YOLO V5 and the three improved YOLO V5 algorithms with fusion attention mechanisms are shown in Fig. 7.

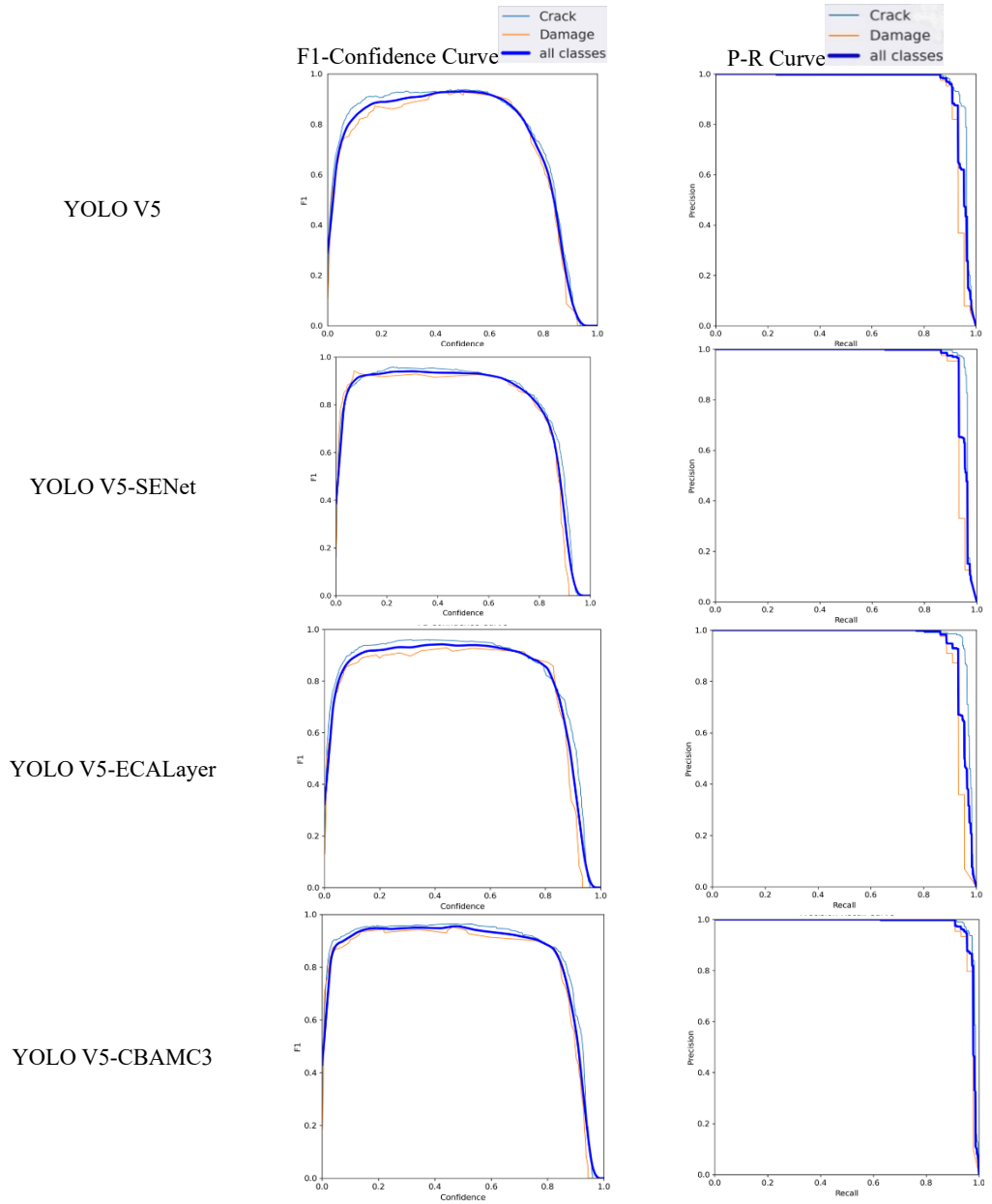

Fig. 7. F1 and PR curve

As shown in Fig. 7, it can be observed that the improved YOLO V5 algorithm with fused hybrid domain attention mechanism, CBAMC3, has the largest area under the F1 curve and PR curve, indicating that the YOLO V5 algorithm with CBAMC3 module fusion has the best convergence effect compared to the other three algorithms.

4. Conclusion

This article addresses the issue of incomplete box selection caused by the lack of high-precision localization capabilities in bridge crack target detection tasks and selects the YOLO V5 algorithm as the backbone network. In response to the YOLO V5 algorithm's inability to achieve expected performance in high-precision localization, fusion attention mechanism is used to optimize the YOLO V5 algorithm's high-precision localization problem. Two-channel attention mechanisms, SENet and ECALayer modules, and one hybrid domain attention mechanism, CBAMC3, were selected for fusion, and relevant experiments were conducted. The results show that the fusion attention mechanism method can effectively improve the YOLO V5 algorithm's high-precision localization performance. Among them, the CBAMC3 module fused with the YOLO V5 algorithm has the best effect, and the mAP50-95 is improved by 6.5% for high-precision localization issues. In the future, we will conduct research on bridge crack problems in more complex scenarios. We will expand the bridge crack dataset by collecting more bridge crack images and conducting algorithm improvement research on the object detection problem of bridge crack images under high resolution. High-resolution images require object detection algorithms to have a larger receptive field range than commonly used algorithms, but simply increasing the receptive field may not significantly improve the model's performance. Therefore, a better attention mechanism that utilizes information more fully is needed to extract feature information from large receptive fields. This will be a feasible method for object recognition tasks in high-resolution image detection.

Author Contribution: All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding: This work was Supported by the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (No.2021QZJ01), Sichuan University of Science & Engineering Graduate Innovation Fund (No. y2021064).

Acknowledgment: This work was Supported by the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (No.2021QZJ01), Sichuan University of Science & Engineering Graduate Innovation Fund (No. y2021064).

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] A. Ammouche, D. Breysse, H. Hornain, O. Didry, and J. Marchand, "A New Image Analysis Technique for the Quantitative Assessment of Microcracks in Cement-Based Materials," *Cement and Concrete Research*, vol. 30, pp. 25-35, 2000, https://doi.org/10.1016/S0008-8846(99)00212-4.

[2] H. D. Cheng, J. Chen, C. Glazier, and Y. G. Hu, "Novel Approach to Pavement Cracking Detection Based on Fuzzy Set Theory," *Journal of Computing in Civil Engineering,* vol. 13, pp. 270-280, 1999, https://doi.org/10.1061/(ASCE)0887-3801(1999)13:4(270).

[3] M. Win, A. R. Bushroa, M. A. Hassan, N. M. Hilman, and A. Ide-Ektessabi, "A Contrast Adjustment Thresholding Method for Surface Defect Detection Based on Mesoscopy," *IEEE Transactions on Industrial Informatics*, vol. 11, pp. 642-649, 2015, https://doi.org/10.1109/TII.2015.2417676.

[4] H. Zhao, G. Qin and X. Wang, "Improvement of Canny Algorithm Based on Pavement Edge Detection," in *2010 3rd International Congress on Image and Signal Processing*. vol. 2, pp. 964-967, 2010, https://doi.org/10.1109/CISP.2010.5646923.

[5] Z. Zhang, F. Xing, X. Shi, and L. Yang, "SemiContour: A Semi-Supervised Learning Approach for Contour Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 251-259, 2016, https://doi.org/10.1109/CVPR.2016.34.

[6] A. Sironi, E. Türetken, V. Lepetit, and P. Fua, "Multiscale Centerline Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1327-1341, 2016, https://doi.org/10.1109/TPAMI.2015.2462363.

[7] M. Kamaliardakani, L. Sun and M. K. Ardakani, "Sealed-Crack Detection Algorithm Using Heuristic Thresholding Approach," Journal of Computing in Civil Engineering, vol. 30, pp. 1943-5487, 2016, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000447.

[8] O. Henrique and C. P. Lobato, "Automatic Road Crack Detection and Characterization," I*EEE Transactions on Intelligent Transportation Systems*, vol. 14, pp. 155-168, 2013, https://doi.org/10.1109/TITS.2012.2208630.

[9] E. Zalama, J. G. García-Bermejo, R. Medina, and J. L. Fernández, "Road Crack Detection Using Visual Features Extracted by Gabor Filters," *Computer Aided Civil and Infrastructure Engineering*, vol. 29, pp. 342-358, 2014, https://doi.org/10.1111/mice.12042.

[10] R. Lienhart and J. Maydt, "An Extended Set of Haar-Like Features for Rapid Object Detection," in *Proceedings. International Conference on Image Processing.* vol. 1, pp. I-I, 2002, https://doi.org/10.1109/ICIP.2002.1038171.

[11] P. Prasanna, K. J. Dana, N. Gucunski, B. B. Basily, H. M. La, R. S. Lim, and H. Parvardeh, "Automated Crack Detection on Concrete Bridges," *IEEE Transactions on Automation Science and Engineering*, vol. 13, pp. 591-599, 2016, https://doi.org/10.1109/TASE.2014.2354314.

[12] A. Cord and S. Chambon, "Automatic Road Defect Detection by Textural Pattern Recognition Based on AdaBoost," *Computer-Aided Civil and Infrastructure Engineering*, vol. 27, pp. 244-259, 2012, https://doi.org/10.1111/j.1467-8667.2011.00736.x.

[13] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic Road Crack Detection Using Random Structured Forests," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, pp. 3434-3445, 2016, https://doi.org/10.1109/TITS.2016.2552248.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014, https://doi.org/10.1109/CVPR.2014.81,

[15] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440-1448, 2015, https://doi.org/10.1109/ICCV.2015.169.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137-1149, 2017, https://doi.org/10.1109/TPAMI.2016.2577031.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016, https://doi.org/10.1109/CVPR.2016.91.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," *In Computer Vision–ECCV 2016: 14th European Conference*, pp. 21-37, 2016, https://doi.org/10.1007/978-3-319-46448-0_2.

[19] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road Crack Detection Using Deep Convolutional Neural Network," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3708-3712, 2016, https://doi.org/10.1109/ICIP.2016.7533052.

[20] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "DeepCrack: Learning Hierarchical Convolutional Features for Crack Detection," *IEEE Transactions on Image Processing,* vol. 28, pp. 1498-1512, 2019, https://doi.org/10.1109/TIP.2018.2878966.

[21] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding Convolution for Semantic Segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1451-1460, 2018. https://doi.org/10.1109/WACV.2018.00163.

[22] S. Teng, Z. Liu and X. Li, "Improved YOLOv3-Based Bridge Surface Defect Detection by Combining High- and Low-Resolution Feature Images," *Buildings*, vol. 12, 2022, https://doi.org/10.3390/buildings12081225.

[23] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv e-prints*, pp., 2018, https://doi.org/10.48550/arXiv.1804.02767.

[24] J. S. Zhu and J. B. Song, "An Intelligent Classification Model for Surface Defects on Cement Concrete Bridges," *Applied Sciences-Basel*, vol. 10, no. 3, p. 972, 2020, https://doi.org/10.3390/app10030972.

[25] P. Huthwohl, R. Lu and I. Brilakis, "Multi-Classifier for Reinforced Concrete Bridge Defects," *Automation In Construction*, vol. 105, 2019, https://doi.org/10.1016/j.autcon.2019.04.019.

[26] S. Zhao, F. Kang and J. Li, "Concrete Dam Damage Detection and Localisation Based on YOLOv5s-HSC and Photogrammetric 3D Reconstruction," *Automation in Construction*, vol. 143, 2022, https://doi.org/10.1016/j.autcon.2022.104555.

[27] G. Li, Q. W. Liu, S. M. Zhao, W. T. Qiao, and X. L. Ren, "Automatic Crack Recognition for Concrete Bridges Using a Fully Convolutional Neural Network and Naive Bayes Data Fusion Based on a Visual Detection System," *Measurement Science and Technology*, vol. 31, 2020, https://doi.org/10.1088/1361-6501/ab79c8.

[28] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* Salt, pp. 7132-7141, 2018, https://doi.org/10.1109/CVPR.2018.00745.

[29] W. Qilong, W. Banggu, Z. Pengfei, L. Peihua, Z. Wangmeng, and H. Qinghua, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11531-11539, 2020, https://openaccess.thecvf.com.

[30] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *In Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3-19. 2018, https://doi.org/10.1007/978-3-030-01234-2_1.