Embedded Selforganizing Systems

Intelligent Detection of Road Cracks Based on Improved YOLOv5

Zhiyan Zhou
Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application
Harbin University of Science and Technology
Harbin, China
zhouzhiyan526@163.com

Xiaoyu Yu
College of Electron and Information,
University of Electronic Science and Technology of China，
Zhongshan Institute,
Zhongshan, China
yuxy@zsc.edu.cn

Yuji Iwahori
Department of Computer Science,
Chubu University,
Aichi, Japan
iwahori@isc.chubu.ac.jp

Qing Wu
Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application
Harbin University of Science and Technology
Harbin, China
wuqing@hrbust.edu.cn

Haibin Wu*
Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application
Harbin University of Science and Technology
Harbin, China
woo@hrbust.edu.cn

Aili Wang*
Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application
Harbin University of Science and Technology
Harbin, China
aili925@hrbust.edu.cn

***Abstract*—With the gradual increase of highway coverage, the frequency of road cracks also increases, which brings a series of security risks. It is necessary to detect road cracks, but the traditional detection method is inefficient and unsafe. In this paper, deep learning is used to detect road cracks, and an improved model BiTrans-YOLOv5 is proposed. We add Swin Transformer to YOLOv5s to replace the original C3 module, and explore the performance of Transformer in the field of road crack detection. We also change the original PANet of YOLOv5s into a bidirectional feature pyramid network (BIFPN), which can detect small targets more accurately. Experiments on the data set Road Damage show that BiTrans-YOLOv5 has improved in Precision, Recall, F1 score and mAP@0.5 compared with YOLOv5s, among which mAP@0.5 has improved by 5.4%. It is proved that BiTrans-YOLOv5 has better performance in road detection projects.**

Keywords—Road Cracks, Deep Learning, YOLOv5, Swin Transformer, BiFPN

I. Introduction

With the urbanization of China, the coverage rate of expressways has gradually increased. According to the Report on Sustainable Traffic Development in China, by the end of 2020, the total mileage of highways in China has reached 5.198 million kilometers, of which the coverage rate of expressways in cities with a population of more than 200,000 exceeds 95%, with a total mileage of 161,000 kilometers[1]. The rapid development of highway construction has brought great convenience to people's travel, but due to the increase of service time, the influence of natural environment and improper construction, cracks in roads are inevitable, which brings a series of security risks[2]. Various kinds of cracks are one of the early diseases of road bearing capacity decline and pavement damage, which are usually used as an important index to evaluate the state of road damage. By detecting road cracks, road collapse and irreparable road damage can be effectively avoided[3]. The National Bureau of Statistics reported that there were 347 road collapse accidents in 2021, and the average growth rate of collapse accidents in 2019-2020 was 80.93%. In addition, road cracks will also lead to car accidents, traffic jams and other problems, which will bring great threats to people's life safety and property safety.

The traditional road crack detection is through manual field detection, which is a time-consuming, laborious and dangerous work. With the development of digital technology, the current mainstream detection method is manual discrimination based on images. Although this method improves the security, it is still an inefficient repetitive work[4]. In recent years, the development of computer vision technology has opened up a new direction for road crack detection. Using computers to evaluate road conditions can save a lot of manpower costs and improve the detection speed and accuracy. Intelligent detection methods based on deep learning have achieved high performance results in many fields[5].

At present, YOLO series deep learning algorithm is one of the most frequently used methods in object detection, which has strong generalization ability. Among them, YOLOv5 has achieved smaller model size and lower computing resource consumption by using modern deep learning technology and structural optimization. It shows high speed and precision in real-time target detection tasks, and is very friendly to practical engineering applications[6]. Therefore, based on YOLOv5, this paper proposes a model combining Swin Transformer and BiFPN model, which shows excellent algorithm robustness and generalization in experiments.

II. PROPOSED METHODOLOGY

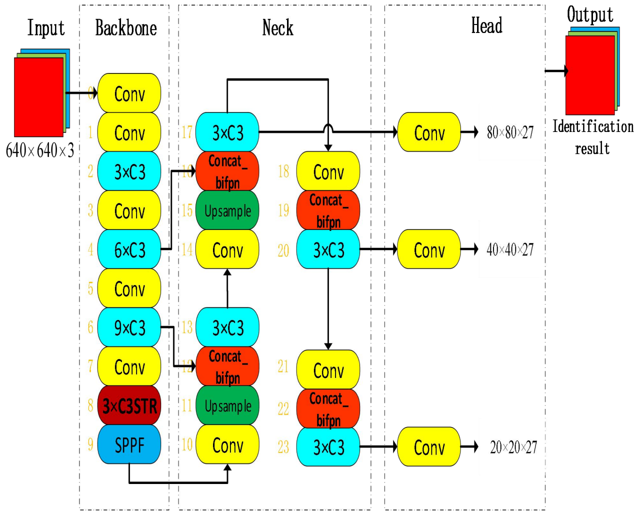


Fig. 1. The architecture of BiTrans-YOLOv5

A. YOLOv5

YOLOv5 is composed of input, backbone, neck, head and output. The backbone of the latest version of YOLOv5 is mainly composed of Conv module, C3 module and SPPF module, which is used to extract image features. Neck uses PANet for multi-scale fusion; head has three groups of output for detection. In addition, YOLOv5 uses mosaic data augmentation, adaptive anchor box computing and image scaling at the input, nms non-maximum suppression at the output and CIOU_Loss as the loss function of Bounding box[7].

YOLOv5 is divided into four types: Yolo V5s, Yolo V5 m, Yolov5l and Yolo V5x, and the model size and complexity increase one by one. The depth of each CSP structure in the four network architectures and the number of convolution cores in different stages is different. YOLOv5s performs best on devices with limited computing resources and has the fastest detection speed, so it is suitable for mobile devices or edge devices[8]. Considering that road crack detection needs to be integrated into mobile unmanned aerial vehicles before it can be used in engineering, we choose YOLOv5s as the Baseline model.

B. BiTrans-YOLOv5

Based on YOLOv5 network architecture, we use a Swin-Transformer in backbone, and use BiFPN instead of PANet to strengthen higher-level feature fusion. The pictures at the input end are uniformly filled or scaled to 640*640. There are three output layers, which correspond to the feature layers of different scales, and their output size is related to the number of classifications (nc). They are 80×80×(3×(nc+5)), 40×40×(3×(nc+5)) and 20×20×(3×(nc+5)) respectively, and nc = 4 in this data set. The architecture of BiTrans-YOLOv5 is shown in Fig. 1.

1) Swin-Transformer

Recently, Transformer has received a lot of attention in the field of computer vision, and achieved outstanding results in a variety of visual tasks[9]. According to the paper[10], integrating the Transformer Prediction Heads into YOLOv5 can effectively improve the performance of the model, which proves the feasibility of combining transformer and YOLOv5.

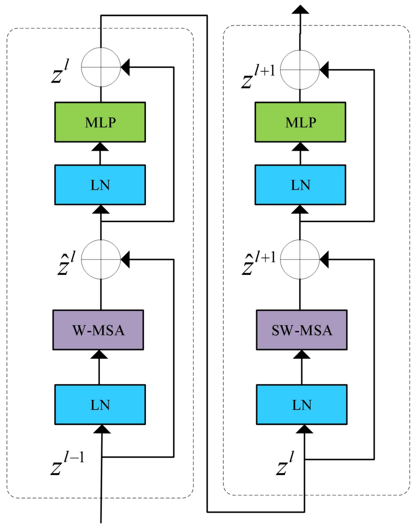


Fig. 2. The architecture of Swin Transformer Layers

Swin Transformer aims to solve the problem of high calculation and memory cost of traditional transformer when processing large-scale images. It introduces a strategy called Shifted Window, which divides the image into uniform local windows and obtains the global context information through the cross attention between windows. Swin Transformer downsamples and upsamples image features with multiple resolutions through hierarchical structure, so that the model can handle global and local information at the same time[11]. The core component of this hierarchical structure is patch merging and patch partitioning strategy based on Shifted Window. The architecture diagram of two consecutive Swin Transformer Layers is shown in Fig. 2. Each Swin Transformer layer consists of Multi-Head Self-Attention (MSA) and Multi-Layer Perceptron (MLP). Layer Normalization (LN) is used to normalize the input of each layer in the neural network.

We only use Swin Transformer in backbone, and integrate it into C3 module to form C3STR module.

2) Bidirectional Feature Pyramid Network (BiFPN)

Inspired by the paper[12], we use BiFPN instead of PANet in neck. BiFPN is a feature pyramid network for target detection, which aims to solve the problems of feature fusion and information transmission in multi-scale target detection[13]. BiFPN makes the feature pyramid network more expressive by introducing bidirectional paths, which is suitable for detecting road cracks of tiny pixels. Its architecture diagram is shown in Fig. 3.

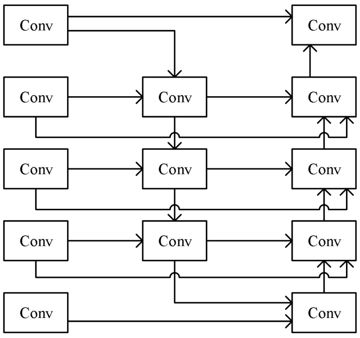


Fig. 3. The architecture of BiFPN

III. EXPERIMENTS AND RESULTS ANALYSIS

A. *Experimental preparation*

1) Datasets

We use Road Damage data set to evaluate whether the model improvement is effective. There are four types of cracks in Road Damage, namely Pothole, Alligator Cracking, Lateral Cracking and Longitudinal Cracking, with a total of 3321 pictures. A picture may contain many kinds of cracks, and the statistical diagram of labels is shown in Fig. 4.

It can be found that the data set has the problem of sample imbalance, which can be improved to some extent by image weighting reassignment and data enhancement such as cropping, rotation and flipping. Use the ratio of 6:2:2 to divide the data set to training set, verification set and test set.

2) Experimental Setting

We implement BiTrans-YOLOv5 on Pytorch 1.8.1, Python 3.8. The computer configuration used for model training and testing includes: CPU is Xeon-4210, running memory is 32GB, GPU is 2*RTX3090, and the system is Ubuntu 18.04, 64-bit operating system.

3) Evaluation metrics

Accuracy is the most commonly used index in classification problems, and it is the ratio of correctly classified predictions to total predictions. However, for unbalanced multi-classification scenarios, Accuracy is deceptive and highly sensitive to data changes, so it is difficult to judge the performance of the model[14]. Therefore, we choose Precision, Recall, F1-Score, AP and mAP@0.5 as evaluation metrics.

Through the confusion matrix, we can know the true positive (TP), true negative (TN), false positive (FP) and false negative (FN), and then we can calculate the precision, recall and F1-score as defined in (1)–(3).

On the smoothed PR curve, take the value of Precision of 10 bisectors (including 11 breakpoints) on the horizontal axis 0-1, and calculate its average value as the final AP, which is defined in (4). When IOU=0.5, the mean Average Precision (mAP) of all crack types is defined in (5).

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{F1}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \tag{3}$$

$$\text{AP}=\frac{1}{11}\sum\nolimits_{0,0.1\cdots1.0}\text{P}_{\text{smooth}}\left(\text{i}\right) \tag{4}$$

$$\text{mAP}=\frac{\sum_{\text{i=1}}^{4}\text{AP}_{\text{i}}}{4} \tag{5}$$

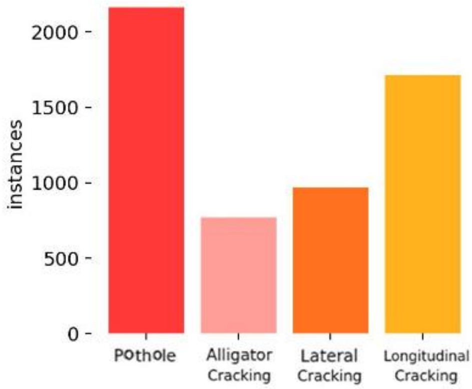


Fig. 4. Label statistics for each category

B. *Model Training*

For BiTrans-YOLOv5 model, three kinds of loss curves and three kinds of metrics curves of the training set are drawn as shown in Fig. 5. Cls_loss calculates whether the anchor frame and the corresponding calibration classification are correct. Box_loss is the error between the prediction frame and the calibration frame. Obj_loss calculates the confidence of the network. The smaller the loss, the better the model performance.

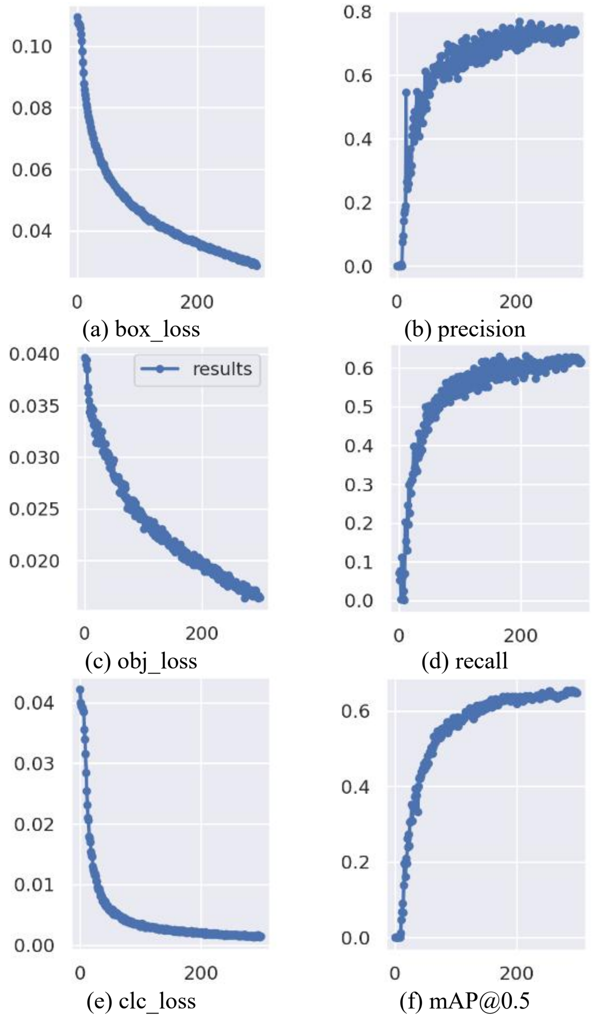
As shown in Fig. 5, each loss tends to be stable and finally converges to a very small value, and the metrics curve finally tends to be stable, which proves that the hyperparameters set by the model is reasonable.

TABLE I. ABLATION STUDY ON TEST-SET OF ROAD DAMAGE

Methods	**Precision (%)**	**Recall (%)**	**F1 (%)**	**mAP@0.5 (%)**
YOLOv5	73.6	61.5	61.0	60.0
BiTrans-YOLOv5	76.3 (↑2.7)	62.8 (↑1.3)	67.0 (↑6.0)	65.4 (↑5.4)

TABLE II. COMPARISON OF AP FOR EACH CATEGORY

Methods	**Pothole (%)**	**Alligator Cracking (%)**	**Lateral Cracking (%)**	**Longitudinal Cracking (%)**
YOLOv5	71.7	70.3	35.5	62.6
BiTrans-YOLOv5	75.5 (↑ 3.8)	73.4 (↑ 3.1)	53.2 (↑ 17.7)	59.7 (↓ 2.9)



(a) box_loss (b) precision

(c) obj_loss (d) recall

(e) clc_loss (f) mAP@0.5

Fig. 5. Loss and metrics

C. *Ablation Study*

We set the epoch to 300. YOLOv5 and BiTrans-YOLOv5 were trained respectively. We use the test set independent of the training set and the verification set to evaluate the model performance with IoU=0.5. The obtained model test results are shown in TABLE I.

Because we are discussing the multi-classification problem, we also list the AP value of a specific type of crack, as shown in TABLE II.

According to TABLE I, BiTrans-YOLOv5 has improved 2.7% in Precision, 1.3% in Recall, 6.0% in F1 score and 5.4% in mAP@0.5 compared with YOLOv5. This proves that our model improvement is effective.

According to TABLE II, we find that the AP of Lateral Cracking is particularly small. On the one hand, because the samples are not balanced, and on the other hand, because the pixels of Lateral Cracking are too small compared with other kinds, even if we make image weighting reassignment, Lateral Cracking is still more difficult to detect. This also explains that although Alligator Cracking is the least in the statistics of labels in Fig. 4, the AP is very high. Because both the number of samples and the size of the target have an impact on the AP, the size of Alligator Cracking is larger, and it contributes more to the whole picture. It means that even if the number of samples is small, the detection result of large targets may still be good[15]. Anyway, by adding Swin Transformer and BiFPN models, we successfully improved the AP of Lateral Cracking by 17.7%.

We choose some pictures of BiTrans-YOLOv5 detected in the test set as the actual results. As shown in Fig. 6, all four types of cracks can be effectively detected.

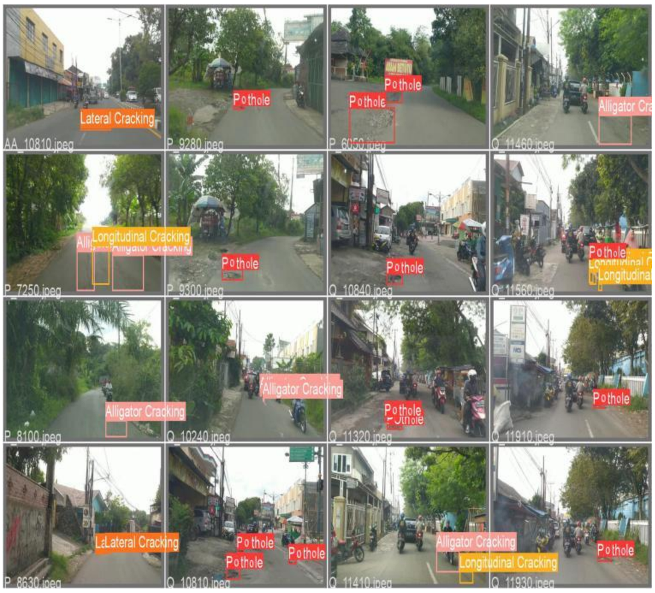


Fig. 6. Visualization results from BiTrans-YOLOv5 on test-set

IV. CONCLUSIONS

In this paper， we propose an improved model BiTrans-YOLOv5 for Road Damage dataset. First of all, we choose YOLOv5s architecture suitable for embedded development as the baseline, and choose the data set Road Damage for analysis. BiTrans-YOLOv5 is mainly improved in feature extraction layer and neck layer, and the performance of the model can be effectively improved by using Swin Transformer in backbone and BiFPN module in neck. All evaluation metrics selected in this paper have all been improved. In addition, we tested the test set and found that BiTrans-YOLOv5 can still detect four kinds of road cracks well in the noisy background.

ACKNOWLEDGMENT

Qing Wu acknowledges the National Natural Science Foundation of China (Grant No. 62205091), the China Postdoctoral Science Foundation Funded Project (Grant No. 2022M710983) and HeiLongJiang Postdoctoral Foundation (Grant No. LBH-Z22201). Supported by the Fundamental Research Fundation for Universities of Heilongjiang Province: 2022-KYYWF-0121.

REFERENCES

[1] DING J X and LI W B. Promote the construction of smart highways. [J]. China Investment, 2023, No.551(Z2): 80-81.

[2] ZHAO D L. On the harm of road cracks and its prevention measures [J]. Technology & economy in areas of communications, 2006, (05): 68+70.

[3] XIAO M J. Research on infrared image fusion and recognition extraction method for road crack detection [D]; the Beijing University of Civil Engineering and Architecture,2022 DOI:10.26943/d.cnki.gbjzc.2022.000284.

[4] NI C S, LI L, LUO W T, QIN Y, YANG Z, et al. Improved YOLOv7 asphalt pavement disease detection [J]. Computer Engineering and Application: 1-16.

[5] Mandal V, Uong L, Adu-Gyamfi Y. Automated road crack detection using deep convolutional neural networks[C]//2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018: 5212-5215.

[6] CHEN X L, LIAN Q W, CHEN X L and SHANG J. Surface crack detection method for coal rock based on improved YOLOv5 [J]. Applied Sciences-Basel, 2022, 12(19) :9695.

[7] QI J, LIU X, LIU K, XU F, GUO H, et al. An improved YOLOv5 model based on visual attention mechanism: application to recognition of tomato virus disease [J]. Computers and Electronics in Agriculture, 2022, 194 : 106780.

[8] XU R J, LIN H F, LU K J, CAO L and LIU Y F. A forest fire detection system based on ensemble learning [J]. Forests, 2021, 12(2) : 217.

[9] MA J Y, TANG L F, FAN F, HUANG J, MEI X G, et al. SwinFusion: cross-domain long-range learning for general image fusion via Swin Transformer [J]. Ieee-Caa Journal of Automatica Sinica, 2022, 9(7):1200-1217.

[10] ZHAO Q, LIU B H, LYU S, WANG C L and ZHANG H. TPH-YOLOv5++: boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer [J]. Remote Sensing, 2023, 15(6) : 1687.

[11] LIU Z, LIN Y T, CAO Y, HU H, WEI Y X, et al. Swin Transformer: hierarchical vision transformer using shifted windows [J/OL] 2021, arXiv:2103. 14030[https://ui.adsabs.harvard.edu/abs/2021arXiv210314030L. 10.48550/arXiv.2103.14030.

[12] DU F J, JIAO S J. Improvement of lightweight convolutional neural network model based on YOLO algorithm and its research in pavement defect detection [J]. Sensors, 2022, 22(9) : 3537.

[13] GUO Y, CHEN S Q, ZHAN R H, WANG W and YANG J. SAR ship detection based on YOLOv5 using CBAM and BIFPN; proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, MALAYSIA, F Jul 17-22, 2022 [C]. 2022 : 2147-2150.

[14] HE H B, GARCIA E A. Learning from imbalanced data [J]. Ieee Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[15] Ren S Q, He K M, Girshick R and SUN J. Faster r-cnn: towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.