Proceedings of the 43rd Chinese Control Conference
July 28-31, 2024, Kunming, China

Road Damage Detection with Attention Mechanism and Multiscale Feature Fusion

Kang Shao[1], Qi Lei[1], Le Huang[2], Qianqian Zhu[2], Bin Xie[1]
1. School of Automation, Central South University, Changsha 410083, P. R. China
E-mail: xiebin@csu.edu.cn
2. Zhejiang Zhejue Technology Co., Ltd, Hangzhou 310000, P. R. China
E-mail: 18873173679@168.com

Abstract: Currently, due to the problem of irregular size and shape of road damage and the exceptionally complex background information in road pictures, it is difficult for existing detection algorithms to accurately recognize the damage. To solve the above problems, we propose a new algorithm YOLOv5s-DCA for road damage detection based on YOLOv5s, combining multi-scale feature fusion and attention mechanism. Firstly, the decoupled detection head is introduced to enhance the classification and localization ability of the model. Meanwhile, a new detection layer is added for large targets such as long cracks. Second, a new multi-scale feature fusion structure is constructed using a skip connection, which enhances the feature fusion in the neck network. Besides, the coordinate attention module is introduced to enhance the model's key feature extraction ability and to inhibit the interference of the background information. Finally, the C2f module is used to replace the C3 module in the backbone to improve the detection accuracy. Combined with the practical application background, we verify the effectiveness of the proposed algorithm based on two public datasets, RDD2022 and Baidu datasets, using three damage categories: crack, pothole, and patch. The experimental results indicate that the enhanced algorithm YOLOv5s-DCA, in comparison to the YOLOv5s algorithm, has shown improvements of 2.3% and 1.9% in mAP@0.5 and mAP@0.5:0.95.

Key Words: road damage detection, YOLOv5s, decoupled detection head, skip connection, coordinate attention

1 Introduction

The road network is the foundation of the economy, supporting traffic and transportation systems for people, commerce, and industry. Over time, roads can develop distresses, like cracks and potholes, due to factors such as oxidation, repeated stresses, rain erosion, and other environmental factors. Untreated pavement distress can escalate, causing collapses and escalating repair costs, posing a threat to economic and personal safety. Therefore, timely detection of pavement distress is crucial for preventing road damage and ensuring traffic safety. In the early days, road disease detection relied primarily on manual patrols and inspections, demanding significant manpower and material resources and resulting in relatively low inspection efficiency. As shown in Fig. 1, there are three common types of roadway diseases, namely patch, pothole, and crack.

(a) pothole (b) crack (c) patch

Fig. 1: Three common types of road damage(the rectangular box is where the damage exists). It shows that road damages are usually small area defects. (a)-(c) are figures of the three damages, in order, pothole, crack, and path.

Recently, road disease detection technology has been categorized into traditional methods and target detection algorithms based on deep learning. The majority of traditional road disease detection methods rely on image processing technology and machine learning, involving manual design or feature extraction, followed by classification using a designed classifier. For instance, Talab et al. proposed a method to detect major cracks in concrete structures using the Otsu method[1]. The process involves utilizing a Sobel filter to detect cracks, obtaining the area of the cracked region through threshold segmentation, and ultimately identifying major cracks using the Otsu method. Shi et al. proposed a novel framework for detecting roadway cracks using the random forest classifier "CrackForest"[2]. This classifier creates a high-performance crack detector by introducing a random structure forest capable of recognizing arbitrarily complex cracks. However, these methods rely on manual feature extraction and are susceptible to environmental factors, leading to poor detection performance in complex environments.

In recent years, the rapid development of deep learning technology has led to remarkable achievements in target detection. Target detection technology is primarily categorized into single-stage algorithms and two-stage algorithms. Widely employed two-stage algorithms encompass R-CNN[3], Fast R-CNN[4], Faster R-CNN[5], and others. Similarly, prevalent single-stage algorithms comprise YOLO[6], SSD[7], and others.

Fang et al[8]. pointed out the challenge of applying target detection algorithms directly to lesion detection due to the characteristics of a low signal-to-noise ratio and non-fixed shape of pavement lesions. They addressed this challenge by combining a deep learning model and Bayesian analysis. They trained the Faster RCNN with the Bayesian Integral Algorithm to detect crack blocks with an appropriate signal-to-noise ratio, achieving a high true detection rate and a low false detection rate for cracks. Liu et al. proposed a two-step pavement crack detection and segmentation method based on a convolutional neural network[9]. In the first step, they

This work is supported by the Central South University Research Programme of Advanced Interdisciplinary Studies (Project No.2023QYJC009). Corresponding author: Bin Xie.

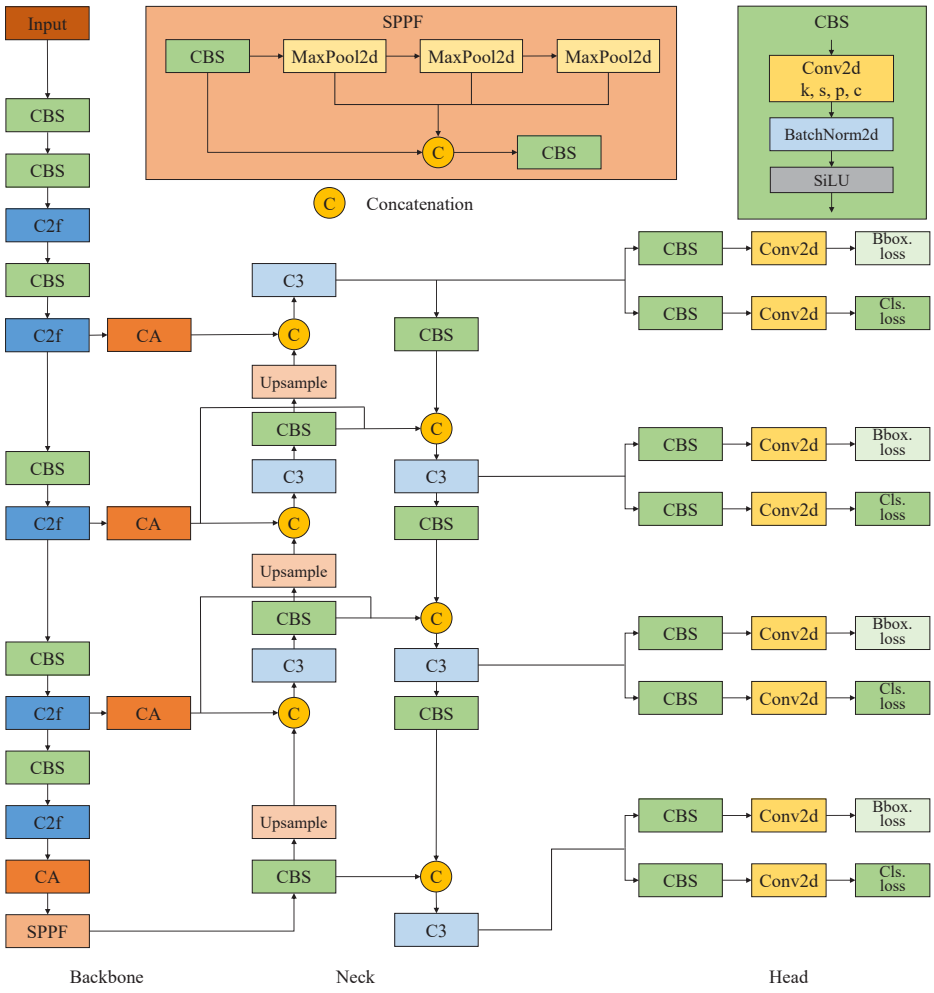

Fig. 2: YOLOv5s-DCA network structure

used the improved YOLOv3 algorithm to detect pavement cracks. The second step involved using the improved U-Net algorithm to segment the pavement cracks detected in the first step. Experimental results demonstrated that the accuracy of the two-step pavement detection method is higher than that of the one-stage pavement crack detection and segmentation algorithm. Liu et al. proposed a method that combines YOLOv5 with 3D ground-penetrating radar (GPR) images to identify internal defects in asphalt pavements[10]. They compared the effectiveness of traditional inspection with GPR inspection by evaluating the maintenance benefits, ultimately demonstrating the effectiveness of 3D ground-penetrating radar in pavement inspection and maintenance.

Firstly, the single-stage target detection algorithm YOLO has attracted considerable attention from scholars since its introduction. Continuous optimization and updates have positioned it as one of the most popular target detection algorithms. The YOLOv5 algorithm, introduced by the Ultralytics team in 2020, has undergone continuous updates, with the latest version being 7.0 in 2022. The YOLOv5 model is classified into yolov5n, yolov5s, yolov5m, yolov5l, and yolov5x based on the gradual increase in network width and depth. YOLOv5s has been identified as meeting real-time performance requirements while also offering higher detection accuracy and a lighter network structure. This makes it suitable for road damage detection. Moreover, there are challenges in using the target detection algorithm directly for road damage detection: (1) The irregular size and shape of cracks make accurate detection difficult. (2) Road images include complex background objects like various materials on the road surface, vehicles, pedestrians, and buildings. These elements may interfere with the model's disease detection, causing it to misjudge the background as road damage.

To solve the above problems, this paper proposes a new improved algorithm, YOLOv5s-DCA, based on YOLOv5, which can effectively improve detection accuracy. The main contributions of this paper can be summarized as follows.

- The task space is decoupled using the decoupled head structure. Furthermore, a large target detection head is added to enhance the prediction performance of the classification and regression tasks, thereby improving detection accuracy.
- Based on skip connection, the paper introduces a new feature layer fusion structure. This involves adding feature layers from the backbone network to the neck network for fusion, further optimizing the effectiveness of model feature fusion.
- Lastly, the paper optimizes the backbone network by replacing the C3 module with the C2f module from yolov8[11]. Furthermore, the Coordinate-Attention(CA)[12] module is added to enhance the model's ability to extract key feature information.

2 Method

YOLOv5 was introduced in 2020 and has undergone continuous updates and iterations. In this paper, we utilize the latest updated version, YOLOv5s 7.0, in 2022. YOLOv5s is mainly composed of three parts: backbone network (backbone), neck network (neck), and output detection head (head). The backbone comprises the CBS module, C3 module, and SPPF module. The neck network utilizes the FPN and PAN structure to fuse feature maps of different scales, and the output detection head incorporates the Detect module, corresponding to three different sizes of detection feature layers. In this paper, we propose an enhanced network based on the YOLOv5s model structure. Firstly, we employ the Decoupled head detection head to separate the task space and introduce the P6 detection layer. Next, we introduce a novel feature layer fusion structure to bolster feature fusion. Subsequently, we incorporate the CA attention Module to augment the model's feature expression ability. Finally, the C3 module in the backbone is replaced by the C2f module with a more abundant gradient flow. By integrating these four key improvements, the enhanced network is denoted as YOLOv5s-DCA, and the modified model structure is illustrated in Fig. 2.

2.1 Decoupled Detection Head

Through an examination of the road disease dataset characteristics, we identified numerous long cracks, along with densely distributed cracks and crack patches. These elements occupy a substantial pixel area in the image and are not effectively detected by yolov5s' three feature layers alone. To address this, we suggest incorporating a dedicated large target detection layer to enhance the model's ability to detect these significant targets. Meanwhile, we introduce a detection head specifically designed for large target detection. This additional detection head, integrated into the model, forms a four-detection head structure alongside the existing three detection heads. While this inclusion increases computational cost, it substantially enhances the model's performance in detecting large targets.

Furthermore, the yolov5s model employs a shared coupled detector head for both classification and localization tasks. This detector head is tasked with simultaneously handling target classification and bounding box regression. Song et al[13] observed that the regions of interest for classification and localization differ, with classification focusing more on saliency region information and localization emphasizing edge region information. YOLOX[14] highlighted the drawbacks of coupled detector heads on detection performance and introduced decoupled detector heads into the YOLO series. In this study, we integrate the decoupled head from YOLOv8 into the YOLOv5s model, employing two parallel branches to independently manage the classification and localization tasks. The specific structure of the decoupled detection head is depicted as the head in Fig. 2.

2.2 New Feature Layer Fusion Architecture

YOLOv5 employs a PAN structure in the neck network for fusing shallow graphical features from the backbone with deep semantic features. As depicted in Fig. 2, the neck initially up-samples the deep feature map and expands the scale through interpolation to enable the fusion of shallow feature maps from the backbone. After completing the upward fusion, the neck continues downsampling to acquire different scales of feature maps for detecting various sizes of target objects. Concurrently, it further fuses shallow graphical features with deep semantic features. Upon analysis, we observed that the neck only fuses the up-fused feature maps in the second fusion and does not re-fuse the feature maps in the backbone. The series of up-sampling and down-sampling operations in the neck can result in some information loss, while the feature maps in the backbone preserve the original image information. Therefore, fusing the feature maps in the backbone during the down-sampling stage of the neck proves advantageous for transferring image information. Given the addition of a large target detection layer, we use the skip connection twice in the right path of the neck, as illustrated in Fig. 2, to fuse the feature maps from the backbone and the left path, respectively, thereby enhancing the feature fusion effect.

2.3 Coordinate Attention Module

While we have successfully fused shallow graphical features and deep semantic features through the neck, the shallow feature maps often include significant background interference information. Addressing this issue, we incorporate the CA attention mechanism into the backbone to aid the network in concentrating on the vital and sensitive regions in the image, thereby minimizing the transmission of redundant background information. Consequently, the enhanced method proposed in this paper integrates the CA attention mechanism alongside skip connection to augment the feature fusion effect of the neck network.

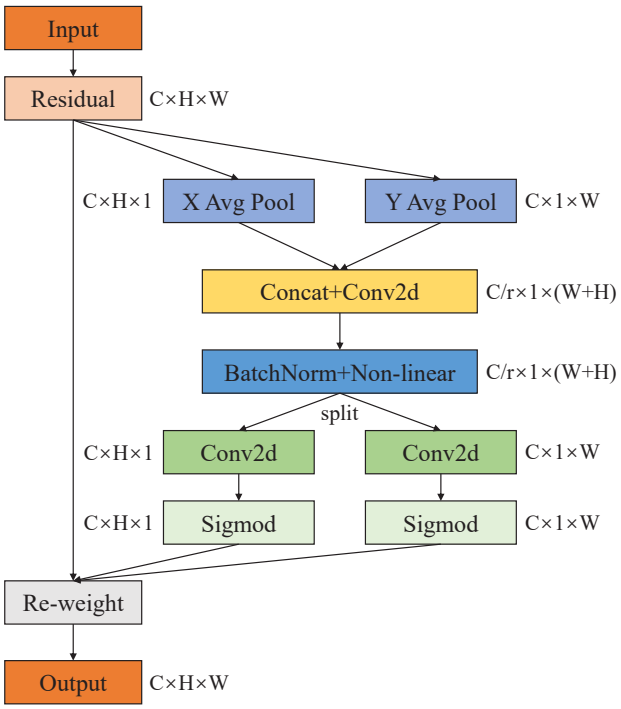

Fig. 3: Coordinate Attention structure

The structure of the CA module, as illustrated in Fig. 3, initiates with the decomposition of global average pooling. The feature maps, with dimensions $C \times H \times W$, undergo

pooling in the width and height directions, resulting in feature maps with dimensions $C \times H \times 1$ and $C \times 1 \times W$. Subsequently, the features in the width and height directions are spliced and undergo dimensionality reduction, producing a feature map with dimensions $C/r \times 1 \times (H + W)$. Finally, through a split into two parallel phases and dimensionality enhancement processing, the attention situation in the width and height dimensions is obtained. This result is then multiplied by the original feature map to constitute the CA attention mechanism. The CA attention mechanism can effectively consider both channel and positional information via feature decomposition in the width and height dimensions, aiding the model in more accurately localizing and identifying the target of interest.

2.4 Introduction of the C2f module

The structure of C3 and C2f is depicted in Fig. 4. C2f is adapted from the latest target detection model, YOLOv8, incorporating more jump connections and additional Split operations on the foundation of the C3 module. This modification allows YOLOv8 to acquire more comprehensive gradient flow information while maintaining a lightweight design. Consequently, in this study, the C2f module replaces the C3 module in the backbone, augmenting the feature extraction capability of the backbone while ensuring the model's lightweight nature.

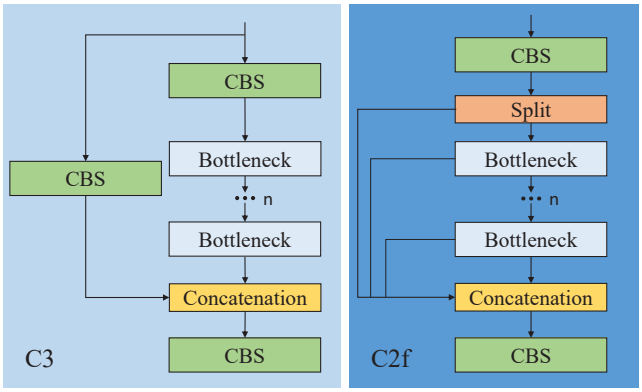

Fig. 4: C3 and C2f structure

3 Experiments and analysis of results

3.1 Introduction to the dataset

The RDD2022 dataset[15] comprises 47,420 road images from six countries: Japan, India, Czech Republic, Norway, USA, and China. The Baidu Flying Paddle dataset comprises 6,000 images distributed across eight categories: Crack, Manhole, Net, Pothole, Patch-Crack, Patch-Net, Patch-Pothole, and others.

This paper consolidates three categories by merging data from the RDD2022 dataset and the Baidu Flying Paddle dataset: Crack, Pothole, and Patch (representing patched cracks and potholes). The dataset is derived from the original Baidu dataset. Using the open-source software Labelimg, we labeled additional bounding boxes for the selected three categories. Ultimately, we have compiled a dataset of 14,691 images for this paper, encompassing three categories: Crack, Pothole, and Patch. The dataset is partitioned into training, validation, and test sets following a 7:2:1 ratio. The training set comprises 10,313 images, the validation set includes 2,968 images, and the test set consists of 1,410 images. In addition, we counted the number of instances of each disease species in the training set, validation set, and test set, as shown in Table 1.

Table 1: Number of road damage instances

datasets	Patch	Crack	Pothole
train	3883	16898	4553
val	1198	5141	1306
test	629	2340	681

3.2 Experimental equipment and evaluation indicators

The experimental environment employs the Ubuntu 20.04 operating system, utilizing the open-source deep learning framework PyTorch as the network framework and CUDA 11.4 for accelerated training. For hardware testing, the CPU selected is Intel(R) Xeon(R) W-2255 CPU @ 3.70GHz, and the GPU chosen is NVIDIA's RTX 3090 with 24GB of video memory. In training, the experiment employs the data enhancement algorithm from the original YOLOv5. The input image size is set to 640, and SGD serves as the optimization function for model training. The model undergoes training for 300 epochs with a batch size of 32, and the initial learning rate is set to 0.01.

This paper employs various evaluation metrics, comprising F1 score, mean average precision (mAP), number of parameters (Params), and billion floating-point operations per second (GFLOPs). The base metrics in this context are checking accuracy and checking completeness. The final comprehensive evaluation metrics of the model are derived from the F1 score and mAP calculated based on checking accuracy and completeness. Params and GFLOPs typically gauge the size and computational complexity of the model. A smaller value for Params and GFLOPs indicates lower computational power requirements and less demand for hardware performance. A model with smaller Params and GFLOPs demands less computational power and places lower requirements on hardware performance.

As shown in Equation 1, the F1 score is the harmonized average of Precision and Recall scores.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

For the object detection task, mAP@0.5 and mAP@0.5:0.95 are computed by assessing the area under the Precision-Recall curve, with a focus on distinct IoU threshold ranges.

$$\text{mAP@0.5} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_{\text{iou=0.5}} \tag{2}$$

$$\text{mAP@0.5 : 0.95} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_{\text{iou=0.5:0.95}} \tag{3}$$

In Equation 2 and Equation 3, N represents the number of categories, and $AP_{iou} = 0.5$ signifies the average accuracy per category at an IoU threshold of 0.5. $AP_{iou} = 0.5 : 0.95$ represents the average precision for each category within the IoU threshold range from 0.5 to 0.95.

3.3 Ablation experiments

To validate the effectiveness of the algorithm improvements proposed in this paper, we conducted ablation experiments using the original YOLOv5s model as the baseline.Evaluation metrics including F1, mAP@0.5, and mAP@0.5:0.95 were employed.The experimental results are summarized in Table 2. Where "DH" stands for decoupled detection head, "skip" stands for skip connection. Firstly, after incorporating the decoupled detection head, F1, mAP@0.5, and mAP@0.5:0.95 increased by 0.3%, 0.6%, and 0.9%, respectively. Following the addition of the large-object detection layer, F1, mAP@0.5, and mAP@0.5:0.95 increased by 0.7%, 0.2%, and 0.3%, respectively. Subsequently, with the adoption of the new feature fusion structure in the neck network, mAP@0.5 improved by 0.8%, while F1 decreased by 0.2%. This demonstrates that the three-layer feature fusion structure proposed in this paper effectively enhances the model's feature fusion capabilities. After integrating the coordinate attention (CA) mechanism into the backbone, F1, mAP@0.5, and mAP@0.5:0.95 increased by 0.5%, 0.3%, and 0.5%, respectively, indicating that the CA attention mechanism helps the model focus on crucial features. Furthermore, replacing the C3 module with C2f in the backbone allowed the model to capture richer gradient flow information, further enhancing its feature extraction capabilities. F1, mAP@0.5, and mAP@0.5:0.95 increased by 0.3%, 0.4%, and 0.3%, respectively.

The experimental results in Table 2 demonstrate that the improved YOLOv5s-DCA model proposed in this paper outperforms the original YOLOv5s model, with increases of 1.6%, 2.3%in terms of F1, mAP@0.5, respectively. This confirms the effectiveness of the proposed algorithm improvements in this paper.

Table 2: Results of ablation experiments

DH	P6	skip	CA	C2f	F1	mAP@0.5
×	×	×	×	×	0.548	0.515
✓	×	×	×	×	0.551	0.521
✓	✓	×	×	×	0.558	0.523
✓	✓	✓	×	×	0.556	0.531
✓	✓	✓	✓	×	0.561	0.534
✓	✓	✓	✓	✓	**0.564**	**0.538**

3.4 Comparison of different algorithms

To further validate the superiority of the proposed YOLOv5s-DCA algorithm, we compared its performance with eight other algorithms under the same conditions, using mAP@0.5, mAP@0.5:0.95, Params, and GFLOPs as evaluation metrics. The comparative experimental results are presented in Table 3. Firstly, YOLOv5s outperformed Faster-RCNN, RetinaNet[16], and YOLOv8s in mAP@0.5, while having smaller Params and GFLOPs. Although YOLOv6s[17] exhibited higher mAP@0.5 and mAP@0.5:0.95 than YOLOv5s, its Params and GFLOPs were nearly three times that of YOLOv5s. Additionally, YOLOv7-tiny[18] had slightly smaller Params and GFLOPs than YOLOv5s, but its mAP@0.5 and mAP@0.5:0.95 were significantly lower than YOLOv5s.The improved YOLOv5s-DCA algorithm, despite a slight increase in Params and GFLOPs compared to the original YOLOv5s, exhibited substantial improvements in mAP@0.5 and mAP@0.5:0.95, with increases of 2.3% and 1.9%, respectively. Furthermore, YOLOv5s-DCA outperformed both YOLOv5m and YOLOv8m in mAP@0.5, while having smaller Params and GFLOPs. However, its mAP@0.5:0.95 was slightly lower than YOLOv8m. This discrepancy may be attributed to the fact that the YOLOv5s-DCA improvement algorithm might not have been fine-tuned further, resulting in a lower sensitivity to different IoU thresholds.

3.5 Comparison of detection results for two algorithms

To further verify the detection performance of the model on different categories of road diseases, Table 4 presents the detection accuracy of YOLOv5s and the improved model YOLOv5s-DCA for three types of road diseases. From Table 4, it is evident that YOLOv5s-DCA outperforms YOLOv5s in the detection accuracy for all three categories: Patch, Crack, and Pothole. The improvement is particularly notable for Patch, with an increase of 2.7% and 2.4% in mAP@0.5 and mAP@0.5:0.95, respectively. Additionally, among the three road diseases, Pothole exhibits the lowest detection accuracy. This is attributed to the small size of Pothole targets, making them prone to being misclassified as background. Moreover, the limited number of training samples for Pothole makes it challenging for the model to learn effective features.

We further visualize the detection results of YOLOv5s and the improved model YOLOv5s-DCA on the test set for comparison, as shown in Fig. 5. This visual comparison provides additional evidence of the effectiveness of the algorithm improvements proposed in this paper.

Fig. 5: Comparison of YOLOv5s (top) and YOLOv5s-DCA (bottom) detection results

4 Conclusion

In this paper, based on two public datasets, the RDD2022 datasets and the Baidu datasets, three categories: crack, pothole, and patch are used for our detection work. Subsequently, we propose an efficient and lightweight YOLOv5s-DCA model for road damage detection. The improved YOLOv5s-DCA model demonstrates a 2.3% and 1.9% enhancement in mAP@0.5 and mAP@0.5:0.95, respectively, compared to the original YOLOv5s model. In the future, we will explore the lightweight issues of the model, aiming to

Table 3: Comparison results of different algorithms

Algorithms	mAP@0.5	mAP@0.5:0.95	Params/M	GFLOPs
Faster R-CNN	0.421	0.184	41.36	134.5
RetinaNet	0.416	0.182	32.24	127.8
YOLOv6s	0.525	0.255	18.5	45.3
YOLOv7-tiny	0.501	0.224	6.01	13.0
YOLOv8s	0.509	0.248	11.13	28.4
YOLOv8m	0.530	**0.262**	25.84	78.7
YOLOv5s	0.515	0.235	7.02	15.8
YOLOv5m	0.536	0.253	20.86	47.9
YOLOv5s-DCA	**0.538**	0.254	16.38	24.7

Table 4: Comparison of detection results by disease type

Algorithms	Type	mAP@0.5	mAP@0.5:0.95
YOLOv5s	All	0.515	0.235
	Patch	0.551	0.270
	Crack	0.491	0.234
	Pothole	0.503	0.201
YOLOv5s-DCA	All	0.538 (+2.3%)	0.254 (+1.9%)
	Patch	0.578 (+2.7%)	0.294 (+2.4%)
	Crack	0.518 (+2.7%)	0.252 (+1.8%)
	Pothole	0.518 (+1.5%)	0.216 (+1.5%)

achieve a balance between high detection accuracy and faster detection speed, facilitating the deployment of the model on mobile devices.

References

[1] Ahmed Mahgoub Ahmed Talab, Zhangcan Huang, Fan Xi, and Liu HaiMing. Detection crack in image using otsu method and multiple filtering in image processing techniques. *Optik*, 127(3):1030–1033, 2016.

[2] Yong Shi, Limeng Cui, Zhiquan Qi, Fan Meng, and Zhensong Chen. Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems*, 17(12):3434–3445, 2016.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[8] Fen Fang, Liyuan Li, Ying Gu, Hongyuan Zhu, and Joo-Hwee Lim. A novel hybrid approach for crack detection. *Pattern Recognition*, 107:107474, 2020.

[9] Jingwei Liu, Xu Yang, Stephen Lau, Xin Wang, Sang Luo, Vincent Cheng-Siong Lee, and Ling Ding. Automated pavement crack detection and segmentation based on two-step convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, 35(11):1291–1305, 2020.

[10] Zhen Liu, Wenxiu Wu, Xingyu Gu, Shuwei Li, Lutai Wang, and Tianjie Zhang. Application of combining yolo models and 3d gpr images in road detection and maintenance. *Remote Sensing*, 13(6):1081, 2021.

[11] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023.

[12] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

[13] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11563–11572, 2020.

[14] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[15] Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, and Yoshihide Sekimoto. Rdd2022: A multinational image dataset for automatic road damage detection. *arXiv preprint arXiv:2209.08538*, 2022.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[17] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022.

[18] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.