2024 IEEE 17th International Conference on Signal Processing (ICSP) | 979-8-3503-8738-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSP62129.2024.10846727

An enhanced YOLOv8 and its application for risk prevention in underground excavation operations

Tianzheng Liu[1,5], Kefei Li[2], Yuyuan Pu[3,4,5], Houxi Zhu,[1,5], Jianfeng Yi[3,4,5]

[1]*Beijing MTR Construction Administration Company Limited, Beijing 100068, China*

[2]*Beijing Infrastructure Investment Company Limited, Beijing 100101, China*

[3]*Beijing Metro Consultancy Corporation Limited, Beijing 100068, China*

[4]*National Engineering Laboratory for Urban Rail Transit System Safety Assurance Technology, Beijing 100068, China*

[5]*Beijing Key Laboratory for Urban Rail Transit Automated Operation System and Safety Monitoring, Beijing 100068, China*

13910107300@163.com, likefei@bii.com.cn, 18513521398@wo.cn, 46804691@qq.com, 22360513@qq.com

***Abstract*—Small target detection is used in underground underground excavation operations. At present, there are still challenges and problems in small target detection. Among them, small targets have low resolution in target detection, have few features that can be extracted, and are susceptible to environmental noise interference. In addition, it is difficult to accurately extract features for targets that span multiple local areas and have odd shapes. In response to these problems, this paper proposes an enhanced YOLOv8 object detection algorithm for safety precautions in underground excavation operations. Among them, a new multi-scale feature fusion module is proposed, and the original P2 layer of the benchmark model is added to the feature pyramid of the neck network for feature fusion to improve the detection accuracy of small targets. The Global Attention Mechanism is added to the backbone network of the benchmark model so that the model can focus on more important key features. In addition, it is proposed to introduce the deformable convolution DCNv3 into the benchmark model structure to replace the original fixed convolution kernel, making the convolution operation more flexible and adaptable, and increasing the receptive field required for the detection task. The improved algorithm improves the detection effect of small targets on the COCO dataset, and the algorithm is applied to the dark mining scenario, and it is also improved on the dark mining dataset. The experimental results prove the effectiveness of this method.**

***Index Terms*—small object detection, underground excavation operations, attention mechanism, deformable convolution**

I. INTRODUCTION

In computer vision, object detection is a classic and most difficult problem to obtain accurate results for object detection. With the significant progress of deep learning technology in the past few decades, most researchers have focused on enhancing object detection, segmentation, and classification. Object detection performance is measured by detection accuracy and inference time.

Nowadays, object detection has been widely used in the process of underground excavation operations [16], object detection technology can be used to detect whether the steel grille assumption is timely, and to determine whether the existence time of the detection hole exceeds a certain threshold. If the threshold is exceeded, an alarm will be issued to prevent the risk of dark excavation operations.

Although object detection has made great progress [19], there are still great problems and challenges. Among them, small targets contain limited information and are easily interfered by noise. Traditional feature extraction methods are mainly aimed at medium and large objects, which are not friendly to small targets. In addition, most current target detection models use traditional convolutional neural networks to extract local features. However, traditional CNN [1] may cause information loss or increase in errors when processing irregularly shaped targets, and cannot perform fine analysis of local details of the target. And it is difficult to accurately extract the features of some irregular or deformed targets. These problems will lead to a decrease in detection accuracy.

The innovations of this paper mainly include adding the P2 feature layer to the feature pyramid for feature fusion based on the Yolov8 algorithm model, adding an attention mechanism to increase the focus on small targets, and replacing the fixed convolution kernel in the traditional CNN with a deformable convolution, so that this model has adaptive spatial aggregation regulated by input and task information.

II. RELATED WORK

Single-stage detectors are now popular in real-time applications due to their excellent speed and accuracy trade-off. The most prominent architecture among single-stage detectors is the YOLO [5] series.

Since YOLOv1 [11], the YOLO series of object detectors has undergone tremendous changes in network structure, label assignment, etc. YOLOv2 [11] uses a new backbone network Darknet-19. It uses anchor boxes and is able to handle a wider range of object sizes and aspect ratios. YOLOv3 [11] uses the backbone network of Darknet-53 and improves the anchor boxes to have different scales and aspect ratios, making them more accurate and stable than previous versions of YOLO. The improvement of YOLOv4 [11] is mainly to use a new architecture of CSPNet, namely the CSPDarknet53 [3] backbone network. YOLOv5 [11] uses the improved CSPDarknet53 and introduces data enhancement technology and the concept of spatial pyramid pooling to further improve detection accuracy.

YOLOv6 [11] uses a variant of the EfficientNet [6] architecture and introduces a new anchor box, namely the dense anchor box. YOLOv7 [11] uses a new loss function called ”focal loss”, which is better for detecting small objects than the cross entropy loss function of the earlier version of YOLO. As the representative algorithm of the YOLO series, YOLOv8 [11] [4] has a backbone network that is basically the same as YOLOv5, except that the C3 module is replaced by the C2f module. It also introduces advanced anchor-free methods and is equipped with dynamic label allocation to improve the performance of the detector, which is significantly better than YOLOv5 in terms of accuracy. At present, compared with YOLOv5 with 46.5M parameters to achieve mAP 49.0, YOLOv8 achieves the best balance between model complexity and accuracy at a speed of 50.2 mAP with 25.9M parameters.

The advantage of the YOLO algorithm is that it has a fast response speed and is suitable for real-time applications. Therefore, the YOLO algorithm is used in the underground excavation process to realize the recognition of small targets, including steel grilles and detection holes.

III. METHODOLOGY

The YOLOv8 framework structure is mainly divided into the input end, backbone network, neck network and head network. The improvement of the algorithm in this paper is based on the Yolov8 framework, in which a new multi-scale feature fusion module, attention module and deformable convolution are introduced. The improved model structure is shown in the Fig. 1. The multi-scale feature fusion module introduces the P2 feature layer at the neck end of the benchmark model for feature fusion, and adds a set of upsampling connection operations and a set of convolution connection operations. The attention mechanism module adds the Global Attention Mechanism(GAM) layer after the SPPF layer of the backbone network to weight the extracted features, which is more conducive to feature fusion. The variable convolution is added to the C2f layer to replace the traditional fixed convolution to optimize the feature extraction process.

A. Multi-scale feature fusion module

The current mainstream object detectors [19] generally consist of two parts: the backbone network and the detection head. The detection head makes decisions based on the information output by the backbone network. At present, general feature extractors usually use subsampling operations to eliminate noise and reduce the resolution of feature maps to extract deeper semantic information, which will inevitably lead to the loss of some target information. For large and medium-sized objects, the impact of this information loss is relatively small. However, it almost eliminates the signals of small objects.

In the yolov8 model structure, the backbone extracts a total of five layers of features from P1 to P5. As the number of convolutions increases, the receptive field gradually increases, and more complex semantic features can be extracted. The model sends the P3, P4, and P5 layers extracted by the backbone to the neck end, and enlarges the original image through upsampling so that it can be displayed on a higher resolution display device. Then the image is reduced by downsampling to fit the size of the display area. Finally, the image is fused through the PANet feature pyramid [10] to enhance the overall feature representation.

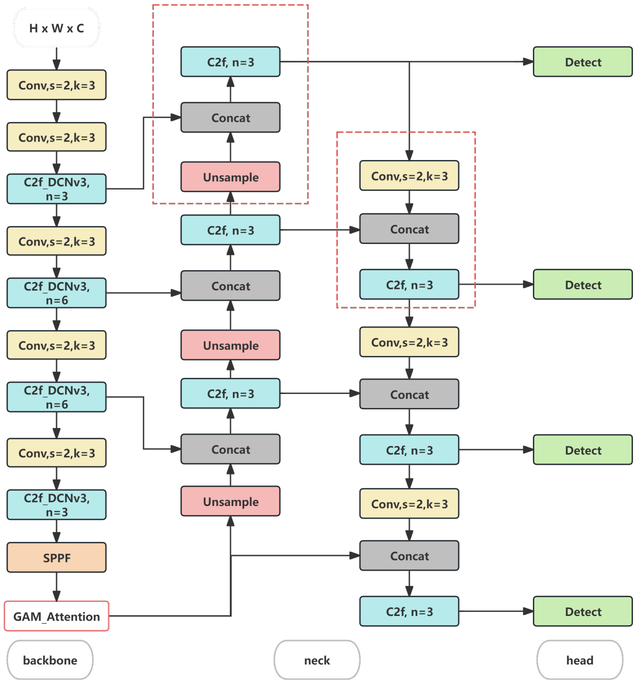


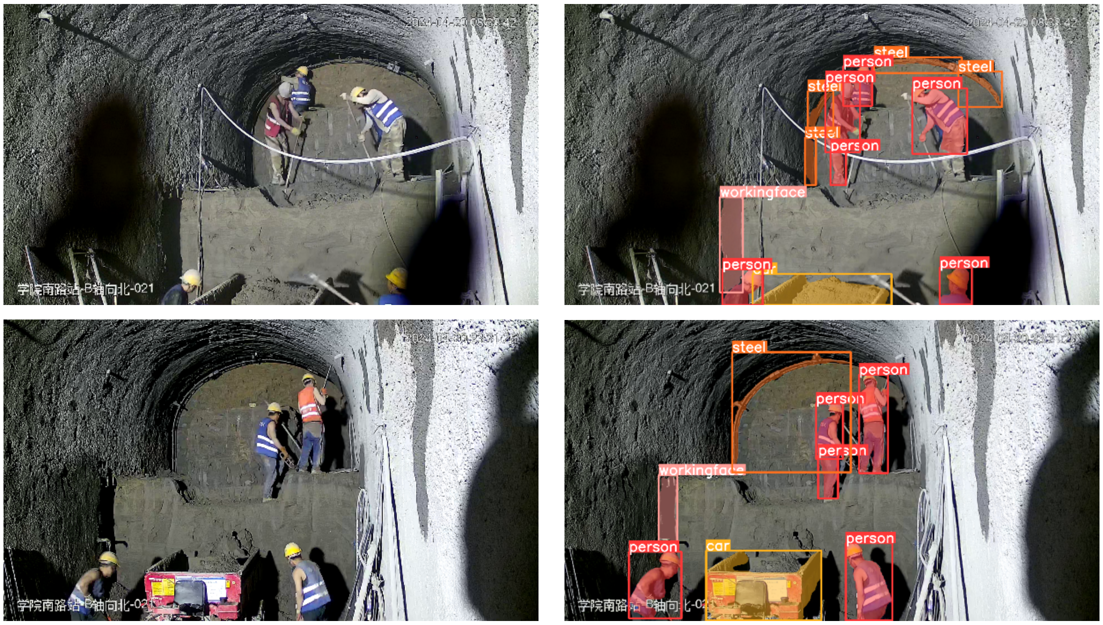
Fig. 1: Improved YOLOv8 model structure. Modifications are made on the neck and backbone ends respectively.

The P2 layer of Yolov8 has fewer convolutions, a higher resolution of the feature map, and a larger size. In order to improve the detection accuracy of small targets, the solution proposed in this paper is to add P2 to the feature pyramid at the neck end for feature fusion. That is, a set of Unsample, Concat, C2f layers and a set of Conv, Concat, C2f layers are added to the neck end to fuse the features of the P2 layer, thereby enhancing the feature representation of small targets and improving the detection accuracy of small objects. The model structure after adding is shown in Figure 1.

B. Global Attention Mechanism module

In image classification tasks, the attention mechanism can help the model focus on the most relevant areas of the image that contain the object of interest and ignore the background or other interference. Since small target detection is easily disturbed by image background and noise, which hinders subsequent detection, this problem should be reduced by adding an attention mechanism.

GAM Attention [9] as a whole consists of two modules: channel attention module and spatial attention module. The channel attention submodule uses a three-dimensional arrangement to retain information in three dimensions. Then, it uses a two-layer MLP to amplify the cross-dimensional channel-



(a) Picture before target detection

(b) Picture after target detection

Fig. 2: Small object detection effect of underground excavation pictures

space dependency. In the spatial attention submodule, in order to focus on spatial information, two convolutional layers are used for spatial information fusion, and the pooling operation is removed to further retain the feature map.

Because the backbone is in the feature extraction stage, the model should focus on more important key features. Therefore, the GAM attention layer is added after the SPPF layer of the backbone network, so that relatively important information can be filtered out from a large amount of information, the degree of attention to the underlying features and the degree of attention to small targets can be enhanced, thereby improving the detection accuracy of the model.

C. Deformable Convolution Module

Compared with traditional convolution kernels, deformable convolution [2] adds a displacement to the normal sampling coordinates. The sampling offset is flexible and can dynamically learn the appropriate receptive field. Ordinary two-dimensional convolution samples on the input feature map x and uses a regular grid R to determine the location of the sampling points, for example:

$$R=\{(-1,1),(-1,0),\ldots,(0,1),(1,1)\} \tag{1}$$

For each position on the output feature map y,

$$y\left(p_0\right)=\sum_{p_n \in R} w\left(p_n\right) \cdot x\left(p_0+p_n\right) \tag{2}$$

where p_n enumerates all positions in R, $w\left(p_n\right)$ is the weighted sum.

DCNv1 is augmented with an offset $\{\Delta p_n \mid n=1,\ldots,N\}$, where $\{N=|R|\}$, then the above equation becomes:

$$y(p_0)=\sum_{p_n \in \mathbb{R}} w(p_n) \cdot x(p_0+p_n+\Delta p_n) \tag{3}$$

The offset is obtained using a convolutional layer on the same feature map, and the sampling is performed at the offset position $p_n+\Delta p_n$.

As an extension of conventional convolution, DCNv2 modulates each sample by the learned feature amplitude. Specifically, for each position k, DCNv2 learns a modulation scalar Δm_k. This modulation scalar is used to adjust the feature amplitude at that position, thereby enhancing or weakening the influence of that position on the final output.

The latest DCNv3 [18] extends DCNv2. First, the original convolution weights are divided into depth-wise and point-wise parts, which improves the efficiency of the model. Second, a multi-grouping mechanism is introduced so that different groups on a convolutional layer can have different spatial aggregation patterns, thereby bringing stronger features to downstream tasks.

At the backbone end of yolov8, the Conv2d convolution layer in the original ConvModule module is replaced with the DCNv3 variable convolution to form the DCNv3Module, so that irregular objects and deformed targets can be more accurately extracted and detected.

IV. EXPERIMENT

First, experiments were conducted on the general dataset COCO2017 dataset and compared with other YOLO models for object detection.

The experimental results are shown in Table 1. The enhanced model improves AP50:95 by 1.6% and AP50 by 1.5% compared to the baseline model, while the number of

TABLE I: Comparison of quantitative indicators: baseline model, improved model, and other target detection models

Model	$AP_{50:95}$(%)	AP_{50}(%)	Param
YOLOv5-M r7.0	45.4	64.1	21.2
YOLOv6-S v3.0 [7][37]	44.3	61.2	18.5
YOLOv7-S af [13][38]	45.1	61.8	11.0
DAMO YOLO-S [15]	46.0	61.9	12.3
GOLD YOLO-S [12]	45.4	62.5	21.5
PPYOLOE-M [14]	43.0	60.5	7.9
RT DETR-R18 [17]	46.5	63.8	20
YOLOv8-S	44.9	61.8	11.2
Ours	46.5	63.3	10.9

TABLE II: Comparison of mAP: Area=small, medium, large (IoU=0.50:0.95, maxDets=100)

Model	Area	$AP_{50:95}$(%)
YOLOv8	small	25.9
	medium	49.8
	large	61.0
Ours	small	28.8
	medium	50.7
	large	61.6

parameters is reduced by 0.3M. Compared with other YOLO models, rt-detr-r18 has the same AP value, but the number of parameters is reduced by nearly half, achieving better detection results with fewer parameters.

The experiment also compared the average precision of different annotated areas, still on the COCO dataset [8]. Table 2 shows that the enhanced model has the greatest improvement in the effect of small target detection. This proves the enhanced model is more robust than the baseline model in small target detection.

The improved model in this paper is applied to the target detection model of underground excavation operations and verified using the underground excavation dataset. The pictures before and after the detection are shown in figure 2. It is found that the model can detect small objects such as steel grilles and detection holes well. When performing risk prevention, an alarm is triggered when the existence time of the detection hole exceeds a certain threshold.

V. CONCLUSION

Aiming at the difficulty of detecting small targets in risk prevention of dark mining operations, this paper proposes an improved YOLOv8 target detection model. To solve the problem that small target detection is easily disturbed by image background and environmental noise, an improved multi-scale feature fusion strategy is adopted. To solve the problem that small targets have low resolution and few extractable features, a strategy of adding a GAM attention mechanism to the backbone network is adopted. To solve the problem that traditional CNN may cause information loss or increase in errors when dealing with irregularly shaped or deformed targets, a strategy of replacing traditional fixed convolution layers with DCNv3 variable convolution is adopted. The accuracy of small target detection is improved on the COCO dataset, and the feasibility of the model method is verified on the underground excavation dataset.

REFERENCES

[1] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[2] Feng Chen, Fei Wu, Jing Xu, Guangwei Gao, Qi Ge, and Xiao-Yuan Jing. Adaptive deformable convolutional network. *Neurocomputing*, 453:853–864, 2021.

[3] Chaima Gouider and Hassene Seddik. Yolov4 enhancement with efficient channel recalibration approach in cspdarknet53. In *2022 IEEE Information Technologies & Smart Industrial Systems (ITSIS)*, pages 1–6. IEEE, 2022.

[4] Muhammad Hussain. Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. *Machines*, 11(7):677, 2023.

[5] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A review of yolo algorithm developments. *Procedia computer science*, 199:1066–1073, 2022.

[6] Brett Koonce and Brett Koonce. Efficientnet. *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pages 109–123, 2021.

[7] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, and Xiangxiang Chu. Yolov6 v3. 0: A full-scale reloading. *arXiv preprint arXiv:2301.05586*, 2023.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[9] Yichao Liu, Zongru Shao, and Nico Hoffmann. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv preprint arXiv:2112.05561*, 2021.

[10] Yiqun Mei, Yuchen Fan, Yulun Zhang, Jiahui Yu, Yuqian Zhou, Ding Liu, Yun Fu, Thomas S Huang, and Humphrey Shi. Pyramid attention network for image restoration. *International Journal of Computer Vision*, 131(12):3207–3225, 2023.

[11] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, 2023.

[12] Chengcheng Wang, Wei He, Ying Nie, Jianyuan Guo, Chuanjian Liu, Yunhe Wang, and Kai Han. Gold-yolo: Efficient object detector via gather-and-distribute mechanism. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] 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.

[14] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, et al. Pp-yoloe: An evolved version of yolo. *arXiv preprint arXiv:2203.16250*, 2022.

[15] Xianzhe Xu, Yiqi Jiang, Weihua Chen, Yilun Huang, Yuan Zhang, and Xiuyu Sun. Damo-yolo: A report on real-time object detection design. *arXiv preprint arXiv:2211.15444*, 2022.

[16] Zhien Zhang, Mingli Huang, and Baohua Wu. Risk analysis and control factors based on excavation of a large underground subway station under construction. *Symmetry*, 12(10):1629, 2020.

[17] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[18] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023.

[19] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.