International Journal of Innovative Computing, Information and Control
Volume **20**, Number **6**, December **2024**

ICIC International ©2024 ISSN 1349-4198
pp. **1789–1802**

RESEARCH ON SMALL TARGET DETECTION ALGORITHM COMBINING ATTENTION MECHANISM

Zitong Yan[1], Xu Li[2,*], Xinlong Wang[1] and Chunlong Yao[1]

[1]School of Information Science and Engineering
[2]Innovation and Entrepreneurship Education Center
Dalian Polytechnic University
No. 1, Qinggongyuan, Ganjingzi District, Dalian 116034, P. R. China
{220520854000600; 220520854000614}@xy.dlpu.edu.cn; yaocl@dlpu.edu.cn
*Corresponding author: lixu@dlpu.edu.cn

Received January 2024; revised May 2024

Abstract. *With the application of deep learning in target detection, the detection of regular-sized targets has matured, but the detection of small targets is still a research difficulty. Aiming at the problem that small target detection is more prone to leakage and false detection, an improved YOLOv7 network model is proposed using the VisDrone dataset as an example. The BiFormer attention mechanism is added to the neck network to help it concentrate more effectively on key areas of the image and increase sensitivity to small targets. We improve the network structure by using dynamic convolution instead of normal convolutions, so as to get more accurate features related to a particular object and improve the overall detection accuracy. Finally, a detection head is added to reduce the loss of character of small targets and increase the accuracy of detection in complex backgrounds. The findings indicate that the* $mAP_{0.5}$ *value of the improved YOLOv7 model is 54.3%, which is an improvement of 4 percentage points over YOLOv7. Ablation experiments have also demonstrated that each module brings an improvement in detection accuracy. Experiments have shown that the improved algorithm is able to detect small targets more effectively.*

Keywords: Small target detection, YOLOv7 model, Attention module, Dynamic detection heads, Dynamic convolution

1. **Introduction.** As an essential direction of image handling and computer vision, target detection is used in a broad scope of applications, and its task consists of two main parts, namely, recognizing the category of the picture and locating the position where the target is located. Small target detection, as a challenge in the area of object detection, is extensively used in vision tasks such as autonomous driving, medical field, UAV navigation, satellite positioning, and industrial inspection. In the field of target detection, early algorithms usually relied on hand-designed features, but these methods often suffered from low generality, slow processing speed and low accuracy. And the object detection algorithm on the basis of deep learning has become the leading technique in the area of target detection due to their structural simplicity and excellent detection performance, which better meets the ever-increasing target detection requirements. Currently, methods based on deep learning for the detection of targets can generally be divided into two broad classes [1]: one class consists of two-stage detection algorithms, represented by the R-CNN family, which first generate candidates and then do classification and bounding box regression on these areas; the other class is the one-stage [2] detection algorithms.

DOI: 10.24507/ijicic.20.06.1789

These are represented by the SSD and YOLO ranges. This type of algorithm predicts the object class and location directly in the image, providing faster detection.

Currently, target detection is mainly focused on natural scene images, face recognition, pedestrian detection and other related problems have been relatively mature [3]. However, due to different imaging angles and lack of effective data sets, direct application of existing algorithms to small target images captured by drones will result in mediocre results. Therefore, the study of target detection algorithms for small targets is of great significance for its application.

Compared to other target detection tasks, small target detection has a shorter history and is still underdeveloped. There are two main types of small target definition: one of them is based on the relative scale, i.e., the median of the ratio of bounding box area to image area for all instances belonging to a class falls within the range of 0.08% and 0.58% [4]; the other is a definition based on an absolute scale. For example, in 2014, in the MS COCO dataset, a standard set of data in the target detection domain, small targets are defined as targets that are less than 32×32 pixels in size [5].

Compared to large and medium-sized targets, small target objects occupy a smaller area in the image, have a lower resolution, have less available information, lack feature expression capability, and are very susceptible to background interference and noise. Therefore, small target detection has great difficulties and challenges. For the past few years, a great deal of research has been done on the detection of small targets by domestic and foreign scholars [6]. Lin et al. [7] proposed the FPN to fuse shallow and deep features through a top-down connection to detect targets of different scales. A Simplified Bidirectional Feature Pyramid Network (SBFPN) [8] was designed by Yu et al. to fuse multi-scale features more effectively. And the offset and reuse of small target information are achieved by using hopping connections in the middle layer of SBFPN. Koyun et al. [9] proposed a two-step target detection framework called "focus and detection" to improve small target detection accuracy. Aiming at the problem that image targets are unevenly distributed and small targets are easily affected by noise, Leng et al. designed a model focusing on "difficult regions": Pareto Refocus Detection (PRDet) [10], which uses a reverse attentional mechanism to distinguish congested regions from normal regions and re-detects the congested regions by detecting them based on region-specific contextual information. Aiming at the problem of limited feature information and unbalanced learning of small targets by the model, Xu et al. proposed a Dynamic Coarse-to-Fine Learning (DCFL) [11] allocator, which utilizes coarse prior matching and refined posterior constraints to dynamically allocate labels to provide suitable and extremely balanced supervision for diverse instances. In the research of single-stage detection, with its lightweight model design and fast detection capability, the YOLO series has gained wide attention in the area of target identification. Cao et al. [12] improved the YOLOv4 network by introducing the MA attention module to improve the feature extraction. In 2022, Luo et al. [13] proposed a YOLO-DRONE (YOLOD) model for UAV images, which was improved on YOLOv4 to make it more suitable for small target detection. In 2023, Han et al. [14] proposed to apply the S-ECA and AFF structures to the neck network structure of YOLOv5s to enhance the recognition of small targets. In order to facilitate UAVs to capture small targets in the scene, Liu et al. [15] introduced depth-separable convolution and multi-scale feature fusion technology into the original YOLOv5 algorithm to reduce the number of model parameters and computation, and to improve the detection accuracy of the model, so as to enhance the model's ability to detect small targets. Qi et al. [16] added the SimAM attention to YOLOv7, and at the same time reduced the pooling nuclei in the pooling layer, so as to improve the positioning precision in the dense state and solve the issue of severe target occlusion. Zhang et al. [17] used an improved Swin Transformer

(STR) module based on the YOLOv7 algorithm, which can better take advantage of the contextual information in the image, enhancing real time and robustness. In a two-stage detection study, Qu et al. [18] used expansion convolution and feature fusion together to reinforce the semantic information of deep characteristics to strengthen the detection of small targets. In 2024, Fan et al. proposed the Spiking Fusion Module (SFDM) [19], which improves the model's ability to detect targets at different scales. In addition, a simple and efficient target detector SFOD (Source-Free Object Detection) is designed by integrating this module with Spiking DenseNet and SSD detection head, which makes the source detector more adaptable to unmarked target domain data. Liu et al. proposed a new Scale and Location Sensitivity (SLS) [20] loss enabling the detector to locate the target more accurately and to improve the detection accuracy by introducing a multiscale prediction of the target by a multiscale detector head in a common U-Net.

Although previous studies have provided effective improvement solutions for enhancing recognition accuracy, small targets are numerous and present a dense distribution in images, which often leads to missed and false detection problems. In addition, these methods have limited capability in target feature extraction, which further affects the accuracy of detection. In response to the above issues, this work takes YOLOv7, an algorithm with a high detection rate in one-stage target detection, as a benchmark, and selects targets on the small target dataset VisDrone-2019 for detection, and improves the YOLOv7 model as follows:

- For the issue that small target recognition accuracy is not high, the BiFormer attention mechanism is introduced to achieve feature optimization by focusing more effectively on key regions in the image;
- The convolution is improved by introducing the ODConv (Omni-dimensional Dynamic Convolution) dynamic convolution, which is able to adjust the parameters in accordance with the specific features of the target, strengthens the adaptability of the features and enables more accurate extraction of the features related to a specific target;
- The addition of a dynamic detection head greatly enhances the model's flexibility and accuracy in dealing with complex scenes, enabling it to carry out more precise analyses of different categories of targets and backgrounds.

The first two sections of this paper present the definition of small target detection, the current state of research, and a detailed description of the models used. Section 3 focuses on the research methodology. In Section 4, we sort out the research process, the datasets used and analyze and visualize the experimental results. Our conclusions are set out in Section 5.

2. **Overview of the Model.**

In the 5-160 FPS range, the YOLOv7 has outperformed most target detectors in speed and accuracy [21]. And it introduces RepVGG (Re-param VGG) [22] model reparameterization, assisted head detection and other methods into the network architecture to help improve detection accuracy.

In this article, on the basis of the many problems existing in the identification of small targets, the YOLOv7 model has been enhanced to enhance its precision in identifying small objects. The improved model architecture diagram is illustrated in Figure 1. The main improvement operations are as follows. First, the common convolution in ELAN module is changed to dynamic convolution ODConv. In this paper, we refer to the improved module as ELAN-OD, and the structure of ELAN-OD is shown in Figure 2. This dynamic convolution is able to dynamically adjust the processing strategy according to the target features (e.g., size and shape) in the image after the image is input, so that the overall detection accuracy can be subsequently improved without increasing the amount

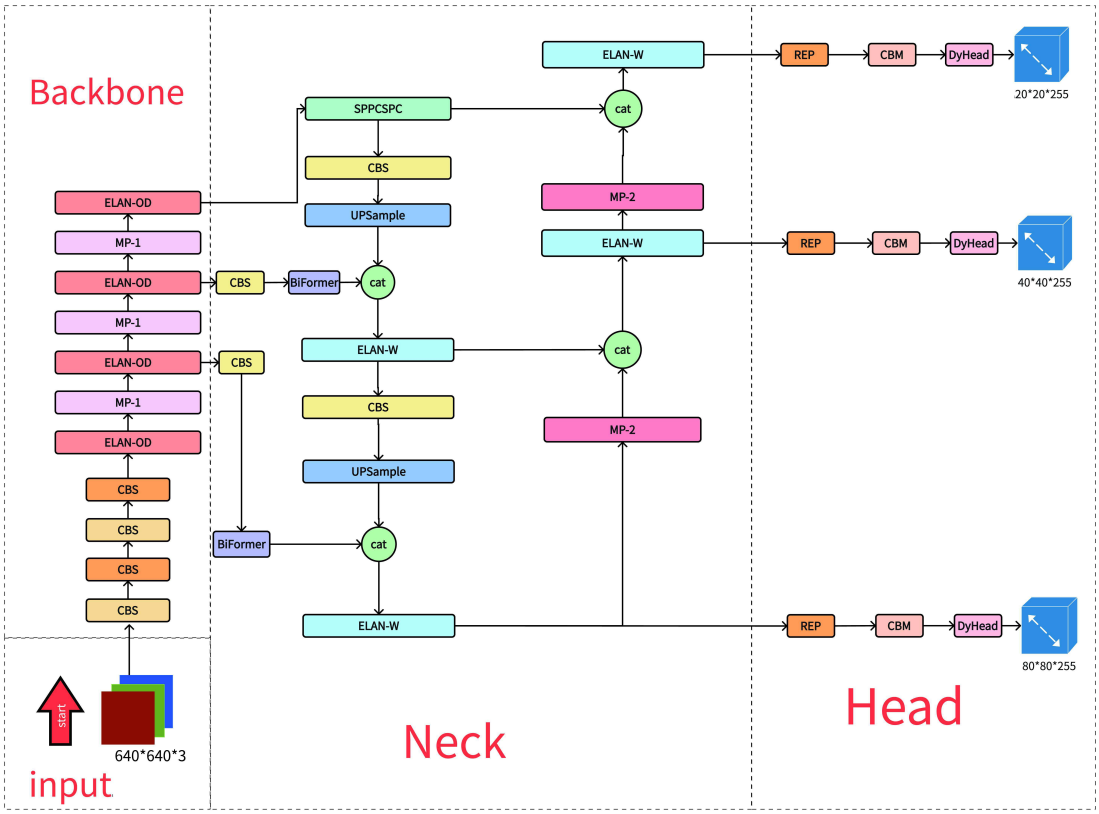

FIGURE 1. Model architecture diagram

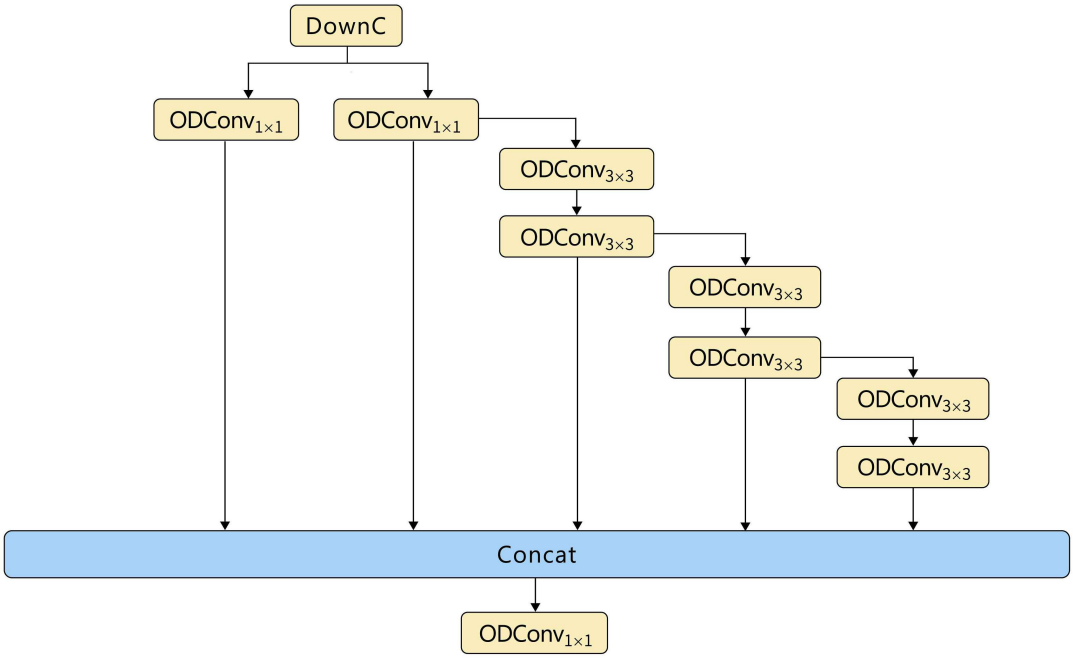

FIGURE 2. ELAN-OD module diagram

of computation. Then the BiFormer attention mechanism is added at the beginning of the Neck part of YOLOv7, through which the two-way attention mechanism can better capture the spatial relationship between different areas within the image, so that the model can deal more effectively with the links between the targets and the contextual

information between them, thereby improving the precision of the detection. Finally, add dynamic detection head to the Head section of YOLOv7. DyHead (Dynamic Head) can adapt to different input features by dynamically adjusting its convolution kernel and attention weights, which improves the model's attention to details when dealing with small targets, thus increasing recognition precision. The workflow of the improved network structure is mainly based on the following points: firstly, a 640×640 image is input, and a $160 \times 160 \times 128$ feature map is output in the backbone, after four convolutional layers (Conv+BN+SiLU), it is inputted into the improved ELAN-OD module, and the convolutional parameters are adjusted according to the specific characteristics of the target to extract the features more accurately, while keeping the input and output channels identical; Next, the network is divided into two branches for downsampling at the MP (maxpool) layer, the first branch achieves spatial downsampling by maxpool (maximum pooling) followed by convolution, and the other undergoes compression by convolution and then downsampling by convolution. The 32-fold downsampled characteristic pattern finally output by backbone is input into Neck, the number of channels is reduced by SPPCSPC, and the first two feature maps output by backbone are fed into the added BiFormer, which is then combined and then fed into the Head through the ELAN-W module, and the focus of the model on small targets is enhanced by the added dynamic detection head, and then finally three outputs of different sizes are performed to get the final result.

3. **Methodology.**

3.1. **BiFormer attention mechanism.**

Attention mechanisms are the focus of research in the field of deep learning and have a wide range of applications in artificial intelligence [23]. When observing an image, people usually focus their attention on a certain part of the image on demand, and humans learn where they should focus their attention for future images to be observed based on previously observed images. The attentional mechanism arose from this, which is a way to mimic the human visual and cognitive systems, allowing neural networks to concentrate on the relevant portions of the data as they process it, while ignoring the unimportant information. Because of the problems with small object identification, for example, small targets are easy to produce aggregation phenomenon and low resolution, and easy to be interfered by the background factors, in this paper, we add the BiFormer attention mechanism [24], which enables the model to concentrate its attention on the important locations, thereby capturing the key information of the small targets. It has good performance and high computational efficiency.

With BRA (Bi-level Routing Attention) as the fundamental building block, BiFormer uses a four-level pyramid structure to filter out unimportant key-value pairs at the coarse-grained grade, and then applies Token-to-Token attention to the rest of the candidate regions, using sparsity to reduce computation and memory. The structure of BiFormer's basic building block BRA is illustrated in Figure 3. Here: H, W and C represent the length, width and number of channels of the input characteristic map, individually, S denotes the region partition factor, and Q, K and V represent the query, key and value vectors, individually. Its main process can be summarized as follows. Input a feature map $X \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$, it is first divided into $S \times S$ areas that do not overlap, letting each region contain $\frac{HW}{S^2}$ feature vectors, by which the input feature map X is transformed into $X^r \in \mathbb{R}^{S^2\times\frac{HW}{S^2}\times C}$. The Q, K and $V \in \mathbb{R}^{S^2\times\frac{HW}{S^2}\times C}$ of the feature map are then obtained by linear mapping. Secondly, the adjacency matrix is used to build a directed graph to find the engagement relationship of different key-value pairs. Finally, token-to-token attention is used to model the local information and enhance the local information, which

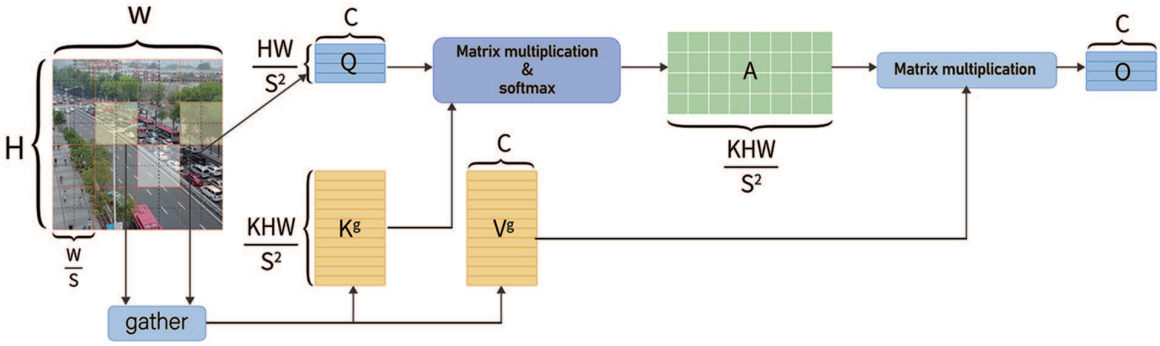

FIGURE 3. Bi-level routing attention mechanism

is calculated as the following Equation (1):

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{1}$$

where K^g and V^g are the tensor of the collected bond values, $LCE(\cdot)$ is a local context enhancement term using deep convolutional parameterization, and in the improved method in this paper, we set its parametric quantity to 5, which can save the computation effectively.

3.2. **Full-dimensional dynamic convolution ODConv.** Ordinary dynamic convolution only focuses on the number of convolutional cores, ignoring the spatial dimension of convolutional cores and other things, increased computation and time, and low detection accuracy for small targets. To address these issues, we introduce Omni-dimensional Dynamic Convolution (ODConv) [25], which reduces the extra parameters considerably and also outperforms other convolutions in terms of output features or convolutional weights. It introduces a new multidimensional attention mechanism, which learns different attention types of convolution kernel along four dimensions of nuclear space. Moreover, these types of attention are applied to the relevant convolution kernels to improve convolution feature extraction performance. The formula for ODConv is the following Equation (2):

$$y = (\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \cdots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n) * x \tag{2}$$

where $x \in R^{h \times w \times c_{in}}$, $y \in R^{h \times w \times c_{out}}$ denote the input and output characteristics, respectively; W_i indicates the ith convolution kernel $W_i^m \in R^{k \times k \times c_{in}}$, $m = 1, \ldots, c_{out}$, consisting of c_{out} filters; $\alpha_{wi} \in R$ is the attention scalar used to weight W_i; $\alpha_{si} \in R^{k \times k}$, $\alpha_{ci} \in R^{c_{in}}$ and $\alpha_{fi} \in R^{c_{out}}$ denote the three newly introduced concerns, computed along the spatial dimensions, number of input channels and number of output channels of the kernel space of the convolution kernel W_i, respectively; $\odot$ means multiplying along different dimensions of nuclear space. α_{si}, α_{ci}, α_{fi} and α_{wi} are calculated using the Multiple Attention Module $\pi_i(x)$. $*$ denotes a convolution operation.

The ODConv is implemented by employing an SE-type attention module, and then using the multi-head attention module $\pi_i(x)$ to calculate multiple types of attention. In concrete terms, the input x undergoes a channel-by-channel Global Average Pooling (GAP) operation to be compressed into a feature vector with a length of c_{in}. Then there is a Fully Connected (FC) layer and four head branches. For each of the four branches at the head, there is an FC layer with output sizes $k \times k$, $c_{in} \times 1$, $c_{out} \times 1$ and $n \times 1$ and a function, creating normalized α_{si}, α_{ci}, α_{fi} and α_{wi}, respectively. The structure of ODConv is illustrated in Figure 4.

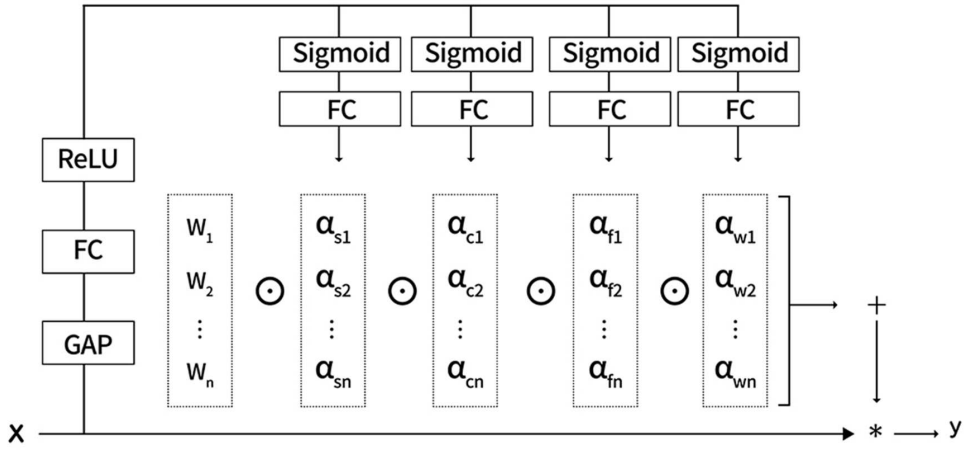

FIGURE 4. ODConv structure diagram

3.3. **Addition of detection head.** Due to the clustering of detection targets in the VisDrone-2019 dataset resulting in very serious occlusion between detection targets and the presence of some too small targets, feature information is easily lost. The original YOLOv7 detector head adopts a static feature fusion approach and selects a fixed IDetect detector head, which is not able to deal with diversified targets or scenes more effectively, especially when handling small targets.

In this paper, in order to increase the sensitivity to small targets, the original YOLOv7 is supplemented with the DyHead [26] (Dynamic Head). DyHead implements a dynamic and adaptive feature processing mechanism that greatly enhances the flexibility and accuracy of the target detection model when dealing with complex scenes. Unlike traditional static feature fusion methods, through its unique dynamic convolution and attention mechanism, DyHead is able to dynamically adjust its internal parameters based on the characteristics of the input picture. For example, it is able to effectively handle a wide range of scale variations from small, distant targets to large, close-range targets, while keeping a high degree of accuracy in complex or interference-rich backgrounds.

DyHead deploys different attention modules in each dimension of features, namely scale perception module, space perception module and task perception module, and integrates these three perceptual modules into a unified attentional mechanism. Self-attention can be expressed as Equation (3):

$$W(\mathcal{F}) = \pi(\mathcal{F}) \cdot \mathcal{F} \tag{3}$$

where $\pi(\cdot)$ is the attention function. This attention function is realized through the fully connected layer; however, this approach is computationally expensive and too costly.

Therefore, DyHead proposes to do the attention in each of the three dimensions, given the 3D feature tensor $\mathcal{F} \in R^{L \times S \times C}$ at the detection layer, and this attention function is calculated as the following Equation (4):

$$W(\mathcal{F}) = \pi_C(\pi_S(\pi_L(\mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F}) \cdot \mathcal{F} \tag{4}$$

where $\pi_L(\cdot)$, $\pi_S(\cdot)$, $\pi_C(\cdot)$ are three different attention functions acting on dimensions L, S, and C, scale, space and mission, respectively, which are sequentially applied to detect the head and can be stacked multiple times. In this experiment, four sets of $\pi_L(\cdot)$, $\pi_S(\cdot)$, $\pi_C(\cdot)$ modules are used to be stacked sequentially to give the detection head a stronger representation capability, which enhances the ability to detect small targets. The DyHead structure is shown in Figure 5.

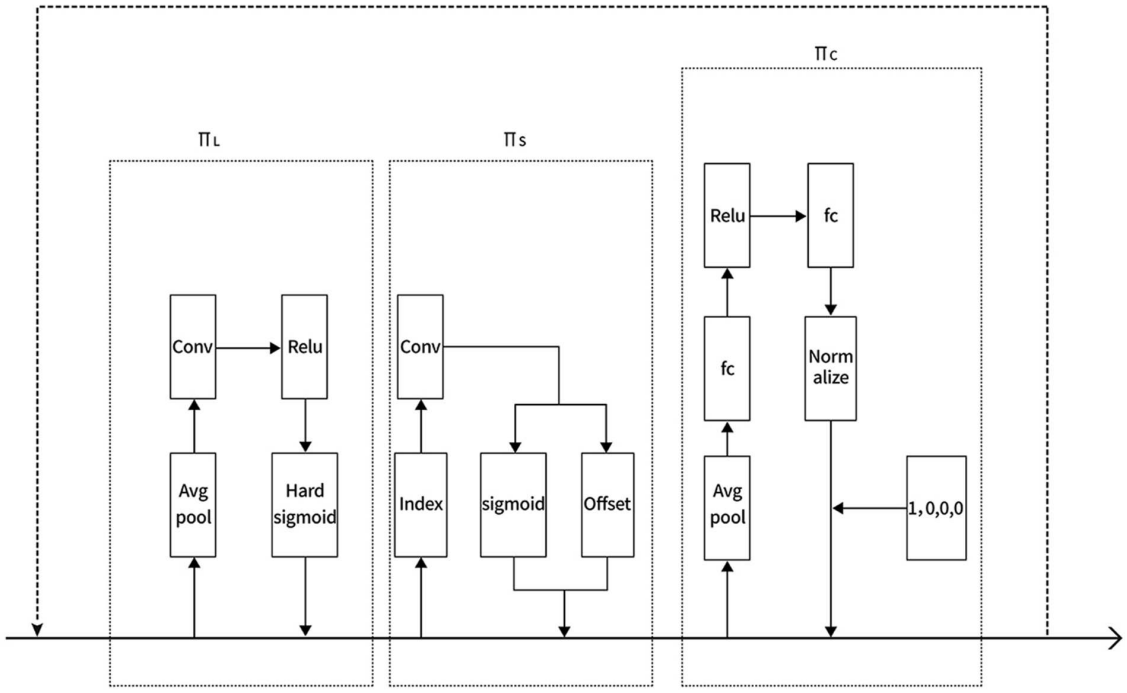

FIGURE 5. DyHead structure

4. Experiments.

4.1. **Experimental environment and parameter settings.** The system employed here is Ubuntu 22.04.1, the environment used for the experiment is python3.8, pytorch 1.13.1, Cuda11.6, the experiment uses a graphics card RTX3080Ti for the model training, the epochs of this study trained with VisDrone-2019 as the dataset are set to 300, the batchsize is 4, the initial rate of learning is set at 0.01, and the network input size is 640×640.

4.2. **Datasets.** The experimental dataset selected for this paper is the VisDrone-2019 dataset [27]. VisDrone-2019 contains 10,209 still images, including 6,471 training sets, 3,190 test sets and 548 validation sets, captured by a variety of UAV cameras in a variety of environments, weather and lighting situations. The detection targets of this dataset include 10 classes like people, cars and buses.

4.3. **Evaluation indicators.** In target detection, Precision (P), Recall (R), mean Average Precision (mAP), and Giga Floating-point Operations Per Second (GFLOPS) are usually used as performance evaluation metrics. Here P is the precision rate, R is the recall rate, mAP is the average precision rate averaged over all categories, Params is the number of parameters and GFLOPS is the number of floating point operations. $mAP_{0.5}$ represents the average detection accuracy at 0.5 for the IOU threshold for all target classes, $mAP_{0.5:0.95}$ indicates the average detection accuracy of all 10 IOU thresholds ranging from 0.5 to 0.95 with the step size of 0.05. The formula is as the following Equations (5) and (6):

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

TP (True Positive) denotes the amount of positive samples that the classification accurately predicts; FP (False Positive) denotes the amount of positive examples that the classifier failed to predict correctly; FN (False Negative) denotes the amount of negative examples falsely anticipated by the classifier.

4.4. **Ablation experiment.** In order to validate the usefulness of adding BiFormer Attention Mechanism, Dynamic Convolutional ODConv and dynamic detection head, this paper chooses YOLOv7 as the benchmark model, and evaluates the impact on the efficiency of small target detection when different methods are combined with each other through the ablation experiments under the same experimental conditions, and the ablation results are listed in Table 1.

Table 1. Ablation experiment results

	Method	$mAP_{0.5}(\%)$	$mAP_{0.5:0.95}(\%)$	P	R	Params	GFLOPS
A	YOLOv7	50.3	29.3	58.5	51.1	46.0	105.3
B	A+BiFormer	51.3	29.5	59.5	51.3	46.5	126.0
C	A+ODConv	51.4	29.7	59.3	51.7	47.3	83.5
D	A+DyHead	51.7	29.9	61.7	51.1	41.3	105.0
E	B+ODConv	52.0	30.0	60.1	52.3	47.8	106.6
F	B+DyHead	52.9	30.5	64.2	51.1	41.9	125.7
G	E+DyHead	54.3	30.9	62.6	51.7	43.2	106.3

As can be seen from Table 1, Group A is the original YOLOv7 algorithm, 50.3% and 29.3% for $mAP_{0.5}$ and $mAP_{0.5:0.95}$, respectively. Experiment B after the introduction of BiFormer Attention, $mAP_{0.5}$ and $mAP_{0.5:0.95}$ improved by 1 percentage point and 0.2 percentage points, respectively. It is shown that the BiFormer module is able to process interrelationships and contextual information between targets more efficiently, thus improving the accuracy of detection. Experiment C shows a 1.1 and 0.4 percentage point improvement in $mAP_{0.5}$ and $mAP_{0.5:0.95}$, respectively, after changing the convolution to ODConv. As shown in Experiment D, along with the decrease in the amount of parameters and the calculation, the introduction of the dynamic detection head improves the model's $mAP_{0.5}$ and $mAP_{0.5:0.95}$ by 1.4 percentage points and 0.6 percentage points, respectively. Experiment G shows that adding all three modules significantly improves the precision. Practice shows that the improved method proposed in this paper is effective for small target detection in complex scenes.

4.5. **Comparison experiment.** Under the condition of ensuring that the configuration environment is consistent with the initial training parameters, this paper compares the improved model with other models to verify the superiority of this paper's algorithm. The results are summarized in Table 2. The mainstream detection algorithms are described as follows.

- Fast R-CNN [28]: Only one feature extraction network is used to extract the feature of the image, and then the region of interest on the feature map is directly used for classification and bias regression.
- YOLOv5s [29]: YOLOv5s employs a lightweight network architecture that uses a deep learning approach to identify and locate multiple targets in an image through feature extraction and target detection.
- TPH-YOLOv5 [30]: A new detection head has been added to accurately locate targets in high-density scenes; CBAM has also been integrated into YOLOv5 to help the network find regions of interest in a wide range of images.

- YOLOXL [31]: YOLOv3-SPP is chosen as the base model, and a new label assignment method was selected in combination with some recent advanced inspection techniques.
- YOLOv8 [32]: A novel SOTA model is provided that is capable of detecting targets at multiple scales with good accuracy and recall. However, due to the deeper network structure and more parameters of YOLOv8, the training time will be longer compared to YOLOv7, and higher computational resources and training data are required to achieve better performance.

TABLE 2. Comparison experiments of different detection algorithms

Method	$mAP_{0.5}(\%)$	$mAP_{0.5:0.95}(\%)$	GFLOPS
Fast R-CNN	35.2	18.3	206.7
YOLOv5s	35.1	19.1	108.1
TPH-YOLOv5	42.9	22.6	129.8
YOLOXL	46.9	28.4	155.2
YOLOv7	50.3	29.3	105.3
YOLOv8	47.0	28.5	130.0
The algorithms in this paper	54.3	30.9	106.3

As shown in Table 2, this paper's improved algorithm has significantly higher detection accuracy than other mainstream target detection algorithms on VisDrone-2019 dataset. The $mAP_{0.5}$ and $mAP_{0.5:0.95}$ attained 54.3% and 30.9%, respectively, which is more suitable for the accuracy needs of small target detection in different scenarios.

4.6. **Visualization of test results.** Figures 6 to 9 show the recognition results of the YOLOv7 algorithm and the improved algorithm proposed in this paper in different scenarios, marked by white boxes. Overall, the experimental comparison results prove that in small target detection, the method in this paper effectively reduces the leakage and false detection rate, and greatly increases detection precision.

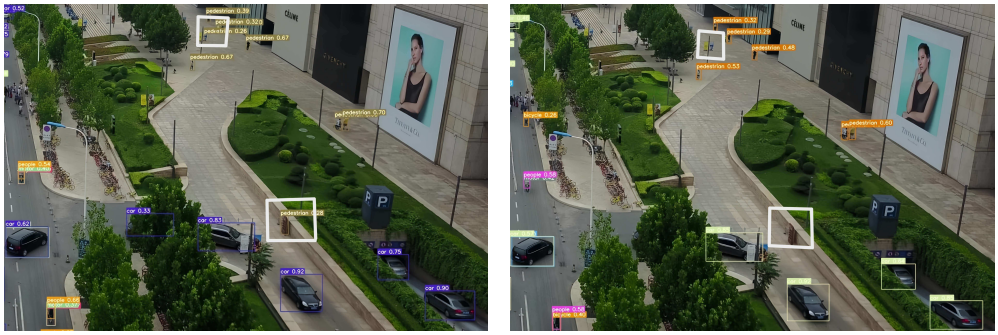

FIGURE 6. Comparison of YOLOv7 (left) and improved algorithm (right) in daytime road scenes

5. **Conclusions.** This paper investigates the identification of small targets and develops improvements to the YOLOv7 model. Add BiFormer attention mechanism to effectively extract the feature information of small targets in pictures; Secondly, the ability to dynamically tune the convolution kernel using dynamic convolution ODConv to more accurately extract features relevant to a particular target; Finally, adding a DyHead enhances the model's attention to detail, allowing it to maintain a high level of accuracy in complex

FIGURE 7. Comparison of YOLOv7 (left) and improved algorithm (right) in low light mountain scene

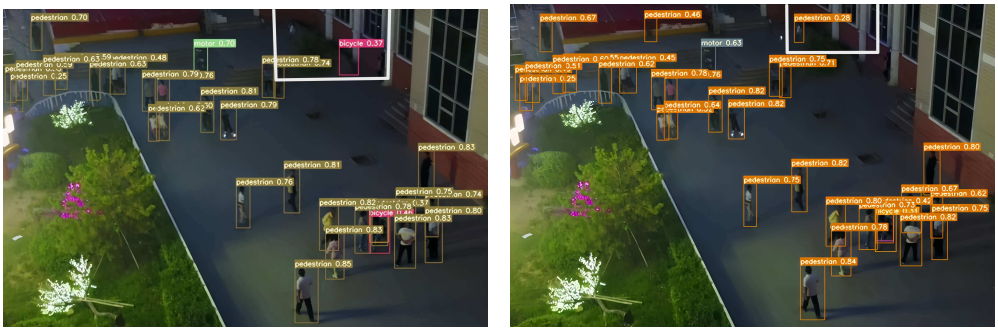

FIGURE 8. Comparison of YOLOv7 (left) and the improved algorithm (right) in dense scenes

FIGURE 9. Comparison of YOLOv7 (left) and improved algorithm (right) in nighttime scenes

or overly distracting backgrounds. Experimental results indicate that our method outperforms the YOLOv5, YOLOXL and YOLOv7 algorithms. However, the algorithm in this paper still has some shortcomings, such as improving the method to improve the network complexity, the detection effect in the dense state is general, there is still room for improvement, and need to follow up with further research and improvement.

Acknowledgment. This work was supported by the Scientific Research Fund for the Higher Education Institutions of Liaoning Province of China under Grant LJ212410152070.

REFERENCES

[1] Y. L. Gu and X. X. Zong, A review of object detection study based on deep learning, *Modern Information Technology*, vol.6, no.11, pp.76-81, 2022.

[2] X. Hou, T. Shan and J. Xue, Analysis of typical algorithms for target detection by deep learning and their current applications, *Overseas Electronic Measurement Technology*, vol.41, no.6, pp.165-174, DOI: 10.19652/j.cnki.femt.2103503, 2022.

[3] W. Hua and Q. Chen, A survey of small object detection based on deep learning in aerial images, *Overseas Electronic Measurement Technology*, 2023.

[4] C. Chen, M. Y. Liu, O. Tuzel et al., R-CNN for small object detection, *Computer Vision – ACCV 2016: The 13th Asian Conference on Computer Vision*, Taipei, Taiwan, 2016.

[5] T. Y. Lin, M. Maire, S. Belongie et al., Microsoft COCO: Common objects in context, *Computer Vision – ECCV 2014: The 13th European Conference*, Zurich, Switzerland, pp.740-755, 2014.

[6] L. Qi and J. Gao, Small object detection based on improved YOLOv7, *Computer Engineering*, vol.49, no.1, pp.41-48, 2023.

[7] T. Y. Lin, P. Dollár, R. Girshick et al., Feature pyramid networks for object detection, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2117-2125, 2017.

[8] N. Yu, H. Ren, T. Deng and X. Fan, Stepwise locating bidirectional pyramid network for object detection in remote sensing imagery, *IEEE Geoscience and Remote Sensing Letters*, vol.20, pp.1-5, 2022.

[9] O. C. Koyun, R. K. Keser, I. B. Akkaya and B. U. Töreyin, Focus-and-Detect: A small object detection framework for aerial images, *Signal Process. Image Commun.*, vol.104, 116675, 2022.

[10] J. Leng, M. Mo, Y. Zhou et al., Pareto refocusing for drone-view object detection, *IEEE Transactions on Circuits and Systems for Video Technology*, vol.33, no.3, pp.1320-1334, 2022.

[11] C. Xu, J. Ding, J. Wang et al., Dynamic coarse-to-fine learning for oriented tiny object detection, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7318-7328, 2023.

[12] Z. M. Cao, Y. Han, L. J. Kong et al., Multi-scene small object detection with modified YOLOv4, *Journal of Physics: Conference Series*, 2022.

[13] X. Luo, Y. Wu and L. Zhao, YOLOD: A target detection method for UAV aerial imagery, *Remote Sensing*, vol.14, no.14, 3240, 2022.

[14] J. Han, X. Yuan, Z. Wang and Y. Chen, UAV dense small target detection algorithm based on YOLOv5s, *Journal of Zhejiang University (Engineering Edition)*, vol.57, no.6, pp.1224-1233, 2023.

[15] Z. Liu, X. Gao, Y. Wan et al., An improved YOLOv5 method for small object detection in UAV capture scenes, *IEEE Access*, vol.11, pp.14365-14374, 2023.

[16] X. Qi, R. Chai and Y. Gao, Reconstructing SPPCSPC with optimised down-sampling for small target detection algorithm, *Computer Engineering and Applications*, pp.1-11, 2023.

[17] X. Zhang, Z. Zhu, Y. Guo et al., Multi-scale remote sensing small target detection based on cosSTR-YOLOv7, *Electro-Optics and Control*, pp.1-9, 2023.

[18] J. S. Qu, C. Su, Z. W. Zhang et al., Dilated convolution and feature fusion SSD network for small object detection in remote sensing images, *IEEE Access*, vol.8, pp.82832-82843, 2020.

[19] Y. Fan, W. Zhang, C. Liu et al., SFOD: Spiking fusion object detector, *arXiv Preprint*, arXiv: 2403.15192, 2024.

[20] Q. Liu, R. Liu, B. Zheng et al., Infrared small target detection with scale and location sensitivity, *arXiv Preprint*, arXiv: 2403.19366, 2024.

[21] C. Y. Wang, A. Bochkovskiy and H. Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7464-7475, 2023.

[22] X. Ding, X. Zhang, N. Ma et al., RepVGG: Making VGG-style ConvNets great again, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.13733-13742, 2021.

[23] J. Xu, Q. Jia, T. Qiu et al., Research and application of intelligent detection technology for bridge girder bottom appearance defects by suspended bridge inspection vehicle, *International Journal of Innovative Computing, Information and Control*, vol.20, no.1, pp.15-30, 2024.

[24] L. Zhu, X. Wang, Z. Ke et al., BiFormer: Vision transformer with bi-level routing attention, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10323-10333, 2023.

[25] C. Li, A. Zhou and A. Yao, Omni-dimensional dynamic convolution, *arXiv Preprint*, arXiv: 2209.07947, 2022.

[26] X. Dai, Y. Chen, B. Xiao et al., Dynamic head: Unifying object detection heads with attentions, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7373-7382, 2021.
[27] D. Du, P. Zhu, L. Wen et al., VisDrone-DET2019: The vision meets drone object detection in image challenge results, *2019 IEEE/CVF International Conference on Computer Vision Workshop*, pp.213-226, 2019.
[28] R. Girshick, Fast R-CNN, *Proc. of the IEEE International Conference on Computer Vision*, pp.1440-1448, 2015.
[29] D. Wang and D. He, Channel pruned YOLOv5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning, *Biosystems Engineering*, vol.210, pp.271-281, 2021.
[30] X. Zhu, S. Lyu, X. Wang et al., TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenario, *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp.2778-2788, 2021.
[31] Z. Ge, S. Liu, F. Wang et al., YOLOX: Exceeding YOLO series in 2021, *arXiv Preprint*, arXiv: 2107.08430, 2021.
[32] H. Lou, X. Duan, J. Guo et al., DC-YOLOv8: Small-size object detection algorithm based on camera sensor, *Electronics*, vol.12, no.10, 2323, 2023.

Author Biography

Zitong Yan received Bachelor of Engineering degree in Software Engineering from Anyang Normal University in 2022. She is currently pursuing a master's degree at Dalian Polytechnic University. Her main research areas include deep learning and image processing.

Xu Li received B.S. degree in Computer Science from University of Science and Technology Anshan in 2003. She received M.E. and Ph.D. degrees in Computer Application Technology from Yanshan University in 2006 and 2010, respectively. She is currently an associate professor in the Innovation and Entrepreneurship Education Center, Dalian Polytechnic University, Dalian, China. Her current research interests include natural language processing and deep learning.

Xinlong Wang received Bachelor of Engineering degree in Computer Science and Technology from Zaozhuang University in 2022. He is currently pursuing a master's degree at Dalian Polytechnic University. His main research areas include deep learning and natural language processing.

Chunlong Yao received B.S. and M.S. degrees in Computer Science from Northeast Heavy Machinery Institute in 1994 and 1997, respectively. He received Ph.D. degree in Computer Software and Theory from Harbin Institute of Technology in 2005. He is currently a professor in the School of Information Science and Engineering, Dalian Polytechnic University, Dalian, China. His current research interests include data mining and intelligent information system.