Wuhan University
Journal of Natural Sciences

2024, Vol.29 No.4, 338-348

Article ID 1007-1202(2024)04-0338-11 DOI https://doi.org/10.1051/wujns/2024294338

Cite this article: PENG Cheng, HE Bing, XI Wenqiang, *et al*. Improved YOLOv7 Algorithm for Floating Waste Detection Based on GFPN and Long-Range Attention Mechanism[J]. *Wuhan Univ J of Nat Sci*, 2024, 29 (4): 338-348.

Improved YOLOv7 Algorithm for Floating Waste Detection Based on GFPN and Long-Range Attention Mechanism

□ **PENG Cheng[1,2†], HE Bing[1,2], XI Wenqiang[2], LIN Guancheng[2]**

1. School of Physics and Electrical Engineering, Weinan Normal University, Weinan 714099, Shaanxi, China;

2. Engineering Research Center for X-ray Imaging and Detection of Shaanxi Provincal Universities, Weinan 714099, Shaanxi, China

Abstract: Floating wastes in rivers have specific characteristics such as small scale, low pixel density and complex backgrounds. These characteristics make it prone to false and missed detection during image analysis, thus resulting in a degradation of detection performance. In order to tackle these challenges, a floating waste detection algorithm based on YOLOv7 is proposed, which combines the improved GFPN (Generalized Feature Pyramid Network) and a long-range attention mechanism. Firstly, we import the improved GFPN to replace the Neck of YOLOv7, thus providing more effective information transmission that can scale into deeper networks. Secondly, the convolution-based and hardware-friendly long-range attention mechanism is introduced, allowing the algorithm to rapidly generate an attention map with a global receptive field. Finally, the algorithm adopts the WiseIoU optimization loss function to achieve adaptive gradient gain allocation and alleviate the negative impact of low-quality samples on the gradient. The simulation results reveal that the proposed algorithm has achieved a favorable average accuracy of 86.3% in real-time scene detection tasks. This marks a significant enhancement of approximately 6.3% compared with the baseline, indicating the algorithm's good performance in floating waste detection.

Key words: floating waste detection; YOLOv7; GFPN (Generalized Feature Pyramid Network); long-range attention

CLC number: TP399

0 Introduction

The urban expansion and population growth have led to a corresponding rise in waste production. The ocean, as the planet's largest aqueous expanse, inevitably bears the brunt of this problem. Investigations indicate that inland rivers are the primary source of marine wastes, highlighting the crucial role of addressing floating wastes in rivers to alleviate marine pollution[1]. Traditionally, the detection of floating wastes in rivers primarily relies on manual inspections. However, this method has various limitations, such as low efficiency and high costs. The deep learning-based millimeter-wave radar target detection technology exhibits greater robustness to varying illumination conditions and offers the potential for long-distance detection[2,3]. However, there are

Received date: 2024-03-02

Foundation item: Supported by the Science Foundation of the Shaanxi Provincial Department of Science and Technology, General Program-Youth Program (2022JQ-695), the Scientific Research Program Funded by Education Department of Shaanxi Provincial Government (22JK0378), the Talent Program of Weinan Normal University (2021RC20), and the Educational Reform Research Project (JG202342)

Biography: PENG Cheng, male, Ph.D., research direction: image processing and computer vision. E-mail: pengcheng@wnu.edu.cn

† Corresponding author. E-mail: pengcheng@wnu.edu.cn

still challenges including weak echoes from non-metallic wastes, susceptibility to water wave interference, and limited semantic information. In recent years, the rapid development of computer vision and machine learning technologies has spurred significant interest in image-based waste detection methods. Traditional machine learning algorithms primarily use sliding windows to extract certain features from images[4-7]. These methods overly relies on manually designed features, thereby limiting their performance and accuracy to some extent. The deep learning-based target detection algorithm has become the mainstream in target detection research due to its simple structure, and absence of the need for manually designed features. These characteristics contribute to its superior generalization and adaptability. The deep learning-based target detection models have undergone continuous evolution and improvement. Initially, R-CNN (Region-based Convolutional Neural Network) [8] and Faster R-CNN[9] adopted a two-stage detection process, which entailed higher computational costs but yielded greater accuracy. Subsequently, the emergence of single-stage detection models such as SSD (Single Shot MultiBox Detector) [10], RetinaNet[11], and YOLO[12] significantly accelerated the detection speed, allowing real-time target detection. However, floating wastes are commonly classified as small-scale targets, indicating their typically smaller dimensions in images, which makes accurate feature extraction challenging. The floating waste detection is further influenced by the complex environment, including object reflections along the riverbanks and intense light reflections off the river. Therefore, the existing deep learning algorithms may encounter issues of information loss or image blurring when detecting floating wastes, resulting in sub-optimal results.

Currently, YOLO has emerged as a classic single-stage detection algorithm, renowned for its fast processing speed and efficient memory utilization. YOLOv7[13] is a further improvement upon the original YOLO, achieving heightened detection accuracy and faster inference speeds through the incorporation of advanced structures and convolutions. These attributes are particularly beneficial for the task of small target detection. In terms of parameter reduction and enhanced computational efficiency, YOLOv7 assimilates the latest performance improvements in CNNs (Convolutional Neural Networks), specifically leveraging advancements in edge device computation, such as the RepConv (Reparameterized Convolution)[14] and the ELAN (Efficient Layer Aggregation Networks) multi-branching architecture[15]. These innovations collectively enhance the model's capacity for feature extraction. It is within this context that an improved YOLOv7 is proposed in this paper, which mainly includes the adoption of GFPN (Generalized Feature Pyramid Network) and a long-range attention mechanism to effectively address the challenges associated with floating waste detection in rivers. The key contributions of this work can be outlined as follows:

- Aiming to address challenges related to low target resolution and insufficient information, we optimize the Neck of YOLOv7 with GFPN. It effectively facilitates the transmission of feature information across different layers, resulting in the generation of richer semantic information. Additionally, the GFPN-RepC2f module is further proposed to replace the Queen-Fusion of GFPN. It not only reduces the complexity of GFPN but also leverages the rich gradient flow information of C2f and RepConv to enhance the quality of feature aggregation.
- The incorporation of the long-range attention mechanism enables the model to focus more on the target region during feature extraction, thus reducing the complex interference of the scene around rivers and the adverse effects of image noise. It also establishes the association and contextual information between targets to further improve the accuracy of the detector.
- The WiseIoU is employed to assign varying weights to the targets, ensuring a fair evaluation of the model when confronting targets of different qualities. This approach facilitates the acquisition of adaptive gradient gains, thereby reducing the adverse effects of low-quality small target samples.

The rest of this paper is organized as follows: Section 1 reviews previous related work. Section 2 presents the improved YOLOv7-based algorithm for floating waste detection in rivers. Section 3 first conducts ablation experiments, and comparison experiments, followed by visualizing the detection performance of the proposed algorithm. Section 4 summarizes the results of the work and looks forward to future research directions.

1 Related Work

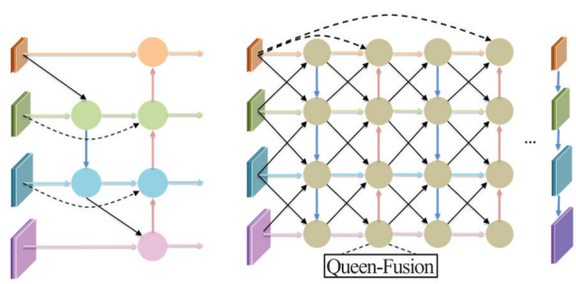
Despite the excellent performance of deep learning-based target detection algorithms on large targets, they still encounter numerous challenges in detecting small targets. Small targets often present issues such as low resolution, blurriness, and lighting variations, making them susceptible to being overlooked or misdetected.

Zand *et al*[16] addressed the detection of freely rotating objects of any size by employing CNNs and rotation-invariant feature maps, thereby enhancing the capability of small object detection. However, their focus was on remote sensing images. Zhang *et al*[17] investigated super-resolution techniques based on GAN (Generative Adversarial Networks). By utilizing a generator capable of restoring blurry small images to clear high-resolution images, they achieved more accurate detection. Nevertheless, GAN-based methods are more challenging to train, and the limited reward for generating samples during the training process may somewhat impede further improvements in detection performance. Gao *et al*[18] proposed a small target detection method named IENet that leverages both appearance and contextual information to enhance detection performance. However, in the context of detecting floating wastes in rivers, the surrounding scene is intricate, and not all contextual information proves beneficial for detection. Leng *et al*[19] explored methods of data augmentation and sample generation to enhance the diversity and richness of small target samples. It is noteworthy that this approach may potentially introduce noise or generate unrealistic samples, thereby impacting the performance of the detector. In recent years, with the development of single-stage detectors, the YOLO series has garnered significant attention due to its high accuracy and speed. YOLO has undergone continuous improvements, leading to subsequent versions with steadily advancing detection performance. Zhu *et al*[20] proposed an improved model of TPH-YOLOv5, which employed an additional prediction head and utilized the self-attention mechanism to enhance the detection performance. Another work[21] introduced a series of YOLO-Z models, varying in scale and built upon YOLOv5, which replaced the original Neck with BiFPN (Bidirectional Feature Pyramid Network) to enhance small target detection performance. However, it is not suitable for scenarios with significant changes in target scales. YOLOv7, a newer improvement in the YOLO series, introduced a more efficient ELAN module on the foundation of YOLOv5, along with a method for auxiliary head training. These enhancements endow YOLOv7 with superior computational efficiency and predictive capability. It outperforms YOLOv5 with higher accuracy and 120% faster speed under the same size and is 180% faster than YOLOX. Therefore, in comparison to other methods, YOLOv7 is chosen as the baseline in this study.

2 Improved YOLOv7-Based Algorithm for Floating Waste Detection in Rivers

2.1 Improvement of the Neck

YOLOv7 is an advanced real-time object detection algorithm based on the YOLO series of algorithms with improvements and optimizations. It has been designed to provide better support for edge devices. The architecture of YOLOv7 is primarily composed of three components: Backbone, Neck, and Head. One can find more in Refs. [22-24]. The Neck in the YOLO series integrates feature maps from various levels, achieving a multi-scale feature representation and acquiring contextual information. Its role is pivotal in bridging the Backbone and Head, significantly influencing network performance. Hence, this work emphasizes the analysis of the Neck structure. The Neck of YOLOv7 also plays a crucial role in integrating the position and detailed information from shallow features with the semantic information derived from deep features. This multi-scale detection strategy has demonstrated its effectiveness in enhancing algorithm performance. YOLOv7's Neck closely adheres to the design principles of YOLOv5, incorporating a blend of the FPN (Feature Pyramid Network) and PAN (Path Aggregation Network) structures. However, it is essential to note that despite its effectiveness, the simplicity of this method can lead to weight imbalances. BiFPN[25], as illustrated in Fig. 1(a), eliminates redundant nodes to reduce computational overhead. It also incorporates skip connections and assigns learnable weights to features, allowing the system to distinguish the importance of different features.



(a) The structure of BiFPN (b) The structure of GFPN

Fig. 1 Comparison of two feature fusion networks

While BiFPN enhances feature fusion by introducing more connections and skipping connections across different layers, research indicates that each BiFPN layer is only connected to adjacent layers, limiting the range of information transmission and fusion. This re-

sults in insufficient feature fusion for distant feature layers. To address these issues, GFPN[26] introduces the dense-link, as depicted in Fig. 1(b). In this structure, the shortest gradient distance extends from 1 layer to $1+\log_2^n$ layers, which implies more efficient information transmission for features, allowing the network to scale deeper. Another key improvement in GFPN is the Queen-Fusion module that can increase adaptability to multiscale variations. The Queen-Fusion incorporates a 3×3 convolution for cross-scale feature fusion. It not only receives input from the preceding node but also gathers input features simultaneously from nodes diagonally above and below, minimizing the risk of information loss during feature fusion. Although the GFPN-based Neck facilitates the comprehensive exchange of high-level semantic information and low-level spatial details, a satisfactory balance between accuracy and time cost has not been achieved. Numerous studies propose that multi-branch architectures generally outperform their single-branch counterparts, even though the latter are more deployment-friendly. The RepConv deals with this well, reducing the computational and parametric load while improving speed. The C2f module[27] represents an enhancement of the C3 module of YOLOv5 and borrows insights from the ELAN structure in YOLOv7. While maintaining a lightweight profile, C2f can capture more extensive gradient flow information. Building upon RepConv and C2f, we introduce the RepC2f module, replacing the original Queen-Fusion in GFPN, as illustrated in Fig. 2. The figure reveals that the RepC2f is obtained by replacing the initial ordinary convolution in the Bottleneck of C2f with RepConv, leading to improved generalization performance while reducing parameters and computational costs. The improved Neck of YOLOv7 is referred to as GFPN-RepC2f. In addition, it can be observed that the GFPN-RepC2f eliminates unnecessary upsampling operations while retaining the benefits of more efficient downsampling operations, resulting in a significant reduction in inference latency with minimal loss of accuracy.

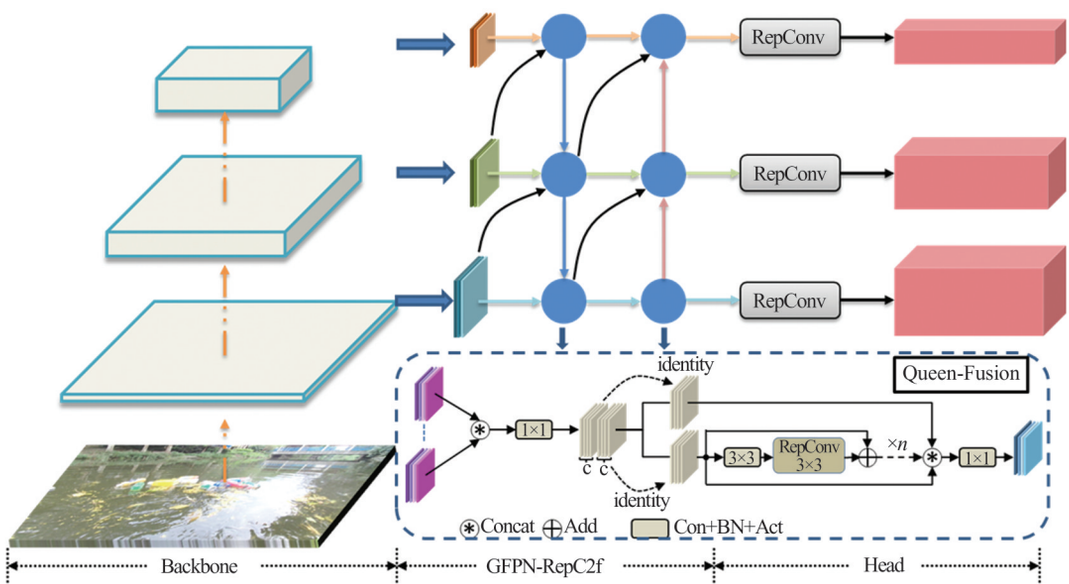


Fig. 2 YOLOv7 network with improved Neck using GFPN-RepC2f module

2.2 Neck Structure Incorporating Long-Range Attention Mechanisms

The detection of floating wastes in rivers requires the precise extraction of detailed features. However, rivers frequently contain intricate background interferences. By reasonably applying attention mechanisms, the feature extraction network can be directed specifically to focus on target areas, thus enhancing the model's detection capabilities and accuracy. Given the task of this study, which centers on detailed feature extraction and addressing complex background interferences, the application of attention mechanisms is deemed crucial. An ideal attention mechanism should exhibit the following characteristics. Firstly, it ought to efficiently capture long-range spatial information, aligning with the original self-attention concept to enhance feature representation. Secondly, the attention module should be deployed efficiently, ensuring minimal compromise to the overall inference speed of the model. Finally, the attention module should be kept concise, avoiding excessively intricate designs to maintain the model's versatility across different tasks. The superior performance of the Transformer

stems from its ability to effectively extract global features. However, this comes at the expense of increased computational cost. A FC (Fully Connected) Layer with a straightforward structure and fixed weights is effective in capturing long-range correlations, thereby generating attention maps with a global receptive field. Nevertheless, the computational complexity exhibits a quadratic relationship with the input feature maps. Fortunately, in typical CNNs, the feature maps have a low rank. As a result, dense FC for input feature maps becomes unnecessary, and leveraging this characteristic can alleviate the computational burden of FC. Drawing from this understanding, a long-range and CNN-based attention mechanism, DFC (Decoupled Fully Connected), is introduced. DFC decouples the large convolution kernels into horizontal and vertical convolutions, striking a balance between computational efficiency and global feature extraction capability[28], as depicted in Fig. 3.

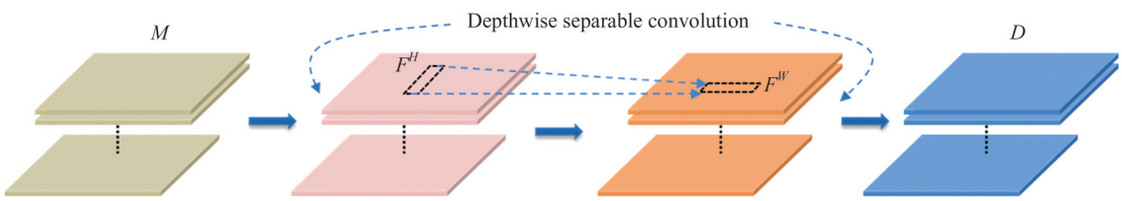


Fig. 3 The DFC attention mechanism for the decomposition of raw dense FC into horizontal and vertical directions

It can be observed from Fig. 3 that the fundamental concept of this attention mechanism involves decomposing the original dense FC into horizontal and vertical directions. This enables the acquisition of long-range features in the respective directions. Concurrently, the time-effective depthwise separable convolutions are adopted to mitigate computational demands. The computational process is delineated as follows:

$$\delta'_{hw}=\sum_{h'=1}^{H}F^{H}_{h,h'w}\odot M_{h'w},\quad h=1,2,\cdots,H;\ \ w=1,2,\cdots,W \tag{1}$$

$$\delta_{hw}=\sum_{w'=1}^{W}F^{W}_{w,hw'}\odot \delta'_{hw'},\quad h=1,2,\cdots,H;\ \ w=1,2,\cdots,W \tag{2}$$

where $M\in R^{C,H,W}$ symbolizes the input feature map, and F denotes the depthwise separable convolution kernel, which is partitioned into vertical $F^H(1\times K_H)$ and horizontal $F^W(K_W\times 1)$ directions. $D=\{\delta_{11},\ \delta_{12},\ \cdots,\ \delta_{HW}\}$ represents the resulting attention map. In the framework of DFC, the attention calculation at a specific point involves direct contributions from all feature points within its corresponding row and column. Consequently, every feature point within that region indirectly contributes to the attention computation for that specific point. By employing weight sharing, the DFC process, as expressed by equation (1), can be simplified through a straightforward convolution, thereby eliminating the need for tensor transpositions and reshaping operations that could otherwise impact the actual inference speed. Additionally, to further reduce computational costs, convolution operations are replaced with time-effective depthwise separable convolutions. When these operations are sequentially applied to the input feature map M, the computational complexity decreases from $O(H^2W+HW^2)$ in the non-decoupled scenario to $O(K_HHW+K_WHW)$, providing a beneficial solution for improving model performance and application in resource-constrained scenarios.

In GFPN-RepC2f, the Queen-Fusion module enables the interaction and fusion of feature maps across different scales, thereby enhancing the representational capability of the feature maps. This contributes to a better understanding of information at various scales within the image, playing a crucial role in the performance of the network's Neck. Examining the Queen-Fusion schematic in Fig. 2, it becomes apparent that, for the input feature map, a pointwise-convolution is initially employed to achieve the desired number of feature map channels. Subsequently, the resulting feature map is evenly partitioned into two parts, each following its respective path, and finally undergoes another pointwise-convolution to generate the output result. This methodology significantly diminishes both parametric and computational costs. Nevertheless, the pointwise-convolutions sacrifice the spatial interaction between pixels, impacting the representation of spatial information. The DFC can offer a partial restoration of this interaction. However, when DFC is processed in parallel with the Queen-Fusion, there is still an associated computational cost. To mitigate this, downsampling is applied to features in both the horizontal and vertical directions, aiming to enhance the execution speed of the DFC attention mechanism. The structure of the Queen-Fusion, when incorporated with the long-range attention mechanism, is visually depicted in Fig. 4.

2.3 Improvement of the Loss Function

In the context of object detection tasks, accuracy and stability are of paramount importance. The assessment metrics for evaluating the quality of the detection

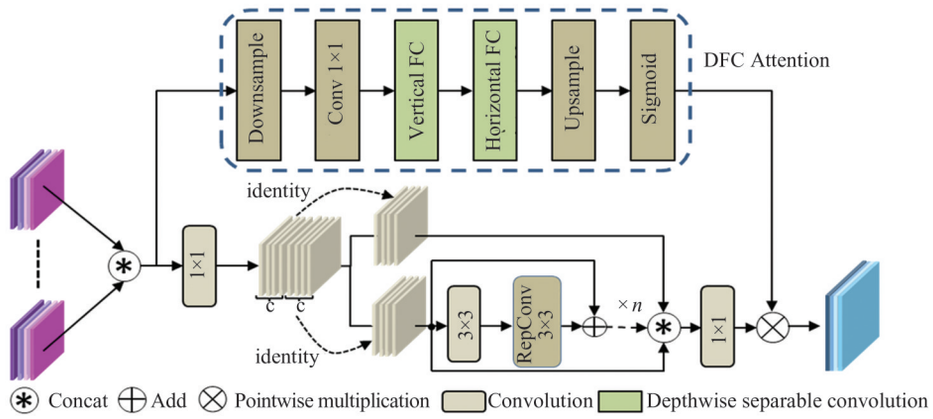


Fig. 4 Queen-Fusion structure incorporating DFC attention mechanism

algorithms typically rely on Intersection over Union (IoU) matching[29], employed to quantify the overlap between predicted bounding boxes and actual bounding boxes. YOLOv7 utilizes the CIoU[30] function for computing bounding box regression loss. The formula for this function is expressed as follows:

$$l_{\mathrm{CIoU}}=\underbrace{1-\frac{H_iW_i}{S_u}}_{\text{IoU loss}}+\underbrace{\frac{(x-x_{gt})^2+(y-y_{gt})^2}{H_g^2+W_g^2}}_{\text{Distance loss}}+\underbrace{\alpha v}_{\text{aspect ratio loss}} \tag{3}$$

where

$$v=\frac{4}{\pi^2}\left(\tan^{-1}\frac{w_{gt}}{h_{gt}}-\tan^{-1}\frac{w}{h}\right),\quad \alpha=\frac{v}{1-\mathrm{IoU}+v} \tag{4}$$

The CIoU loss function consists of three components: the conventional IoU loss, distance loss and aspect ratio loss. In equation (3), H_i and W_i represent the height and width of the intersection between the predicted and target bounding box, while $S_u=wh+w_{gt}h_{gt}-W_iH_i$ denotes the union of the two boxes, as depicted in Fig. 5. The distance loss involves the normalized distance between the predicted bounding box's center coordinates (x,y) and the target bounding box's center coordinates (x_{gt},y_{gt}). H_g and W_g are the height and width of the minimum enclosing box of the two boxes, $\rho=(x-x_{gt})^2+(y-y_{gt})^2$ is the square of the distance between the centers of the two boxes, and $c=H_g^2+W_g^2$ is the square of the diagonal length of the minimum enclosing box. Additionally, α serves as a balancing parameter, and v is utilized to assess whether the aspect ratios of the two boxes are consistent.

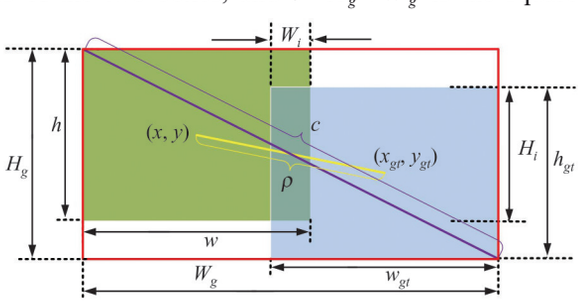


Fig. 5 The geometric relationships in CIoU losses adopted in YOLOv7

Although the CIoU loss function has shown effectiveness in improving network performance, it fails to account for the impact of low-quality samples on loss and the challenge of balancing easy and difficult samples. Addressing this concern, we adopt WiseIoU[31] as the bounding box regression loss function. This loss function incorporates a DNFM (Dynamic Non-monotonic Focus Mechanism), evaluating box quality through the use of an "outlier degree" to prevent excessive penalization by geometric factors, such as distance. The expression for this loss function is as follows:

$$l_{\mathrm{WiseIoU}}=\underbrace{1-\frac{H_iW_i}{S_u}}_{l_{\mathrm{IoU}}}+\underbrace{\exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(H_g^2+W_g^2)^*}\right)}_{\text{Distance attention}}\underbrace{\gamma}_{\text{DNFM}} \tag{5}$$

where $\gamma=\beta/\delta\alpha^{\beta-\delta}$, $\beta=l_{\mathrm{IoU}}^*/\overline{l}_{\mathrm{IoU}}\in[0,\ \infty)$. WiseIoU is comprised of three components. The initial part corresponds to the conventional IoU loss, while the second part involves distance attention and the superscript indicates that it is detached from the computational graph. The utilization of CIoU supports the notion that incorporating geometric factors, such as distance, is advantageous for the model's convergence. Moreover, when there is a substantial overlap between the predicted and target bounding boxes, the distance attention can mitigate the impact of gradient updates, guiding the network's focus toward predictions of average quality.

The last part is the DNFM, which supplements the first two parts by introducing a focus mechanism through the construction of the gradient gain (focusing coefficient), and is the core component of WiseIoU. In equation (5), $\overline{l}_{\mathrm{IoU}}$ signifies the exponentially weighted moving average with a momentum of m, and δ, α are hy-

perparameters. β serves as an "outlier degree" metric that describes the quality of the prediction boxes, where a smaller outlier degree suggests a higher-quality prediction box to which a small gradient gain is assigned by the γ function in order to focus the regression on the normal-quality boxes. Conversely, a larger outlier degree indicates lower-quality boxes and a relatively smaller gradient gain is still assigned, effectively preventing harmful gradients from low-quality samples. This process is illustrated in Fig. 6 and it can be observed that the γ function attains its maximum value within the middle range of the outlier degree. A small outlier degree (high-quality samples) or a large outlier degree (low-quality samples) results in a decrease in the γ function. This implies that samples of ordinary quality will experience a more significant increase in gradient gain, leading to an enhancement in the network's generalization performance. Additionally, the dynamic nature of $\overline{L}_{\mathrm{IoU}}$ signifies that the criteria for categorizing anchor box quality are also dynamic. It allows WiseIoU to dynamically allocate gradient gain, adapting to the current conditions for each batch of data, thereby providing greater flexibility.

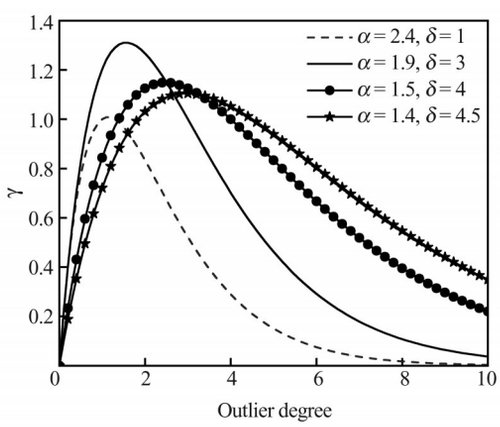


Fig. 6 The curves of the dynamic non-monotonic focusing mechanism function under various parameter conditions

3 Results

The experimental running environment is the Ubuntu20.04 system, with a computer equipped with an Intel i5-6400 CPU and RTX 3090 graphics card, and the popular pytorch framework is adopted for the deep learning model. Throughout the experiments, the SGD (Stochastic Gradient Descent) optimizer is utilized to iteratively adjust the network weights. The initial learning rate for network training is established at 0.001, accompanied by a learning momentum of 0.9 and a weight decay rate of 0.000 5. The experimental dataset is composed of two components. One part is the dataset for floating wastes in rivers captured from the perspective of an unmanned boat, as released by Orcauboat[2]. This dataset covers reflections from varying light intensity, waves and objects along the riverbank as well as observations of the targets in multiple directions and viewpoints, contributing to a wealth of samples and rigorous experimental designs, shown in Fig. 7. Another part comprises images obtained from public network downloads and on-site mobile phone photography, annotated using LabelImg. The dataset has a total of 2 500 images, systematically divided into a training set and a validation set to fulfill the experimental requirements.

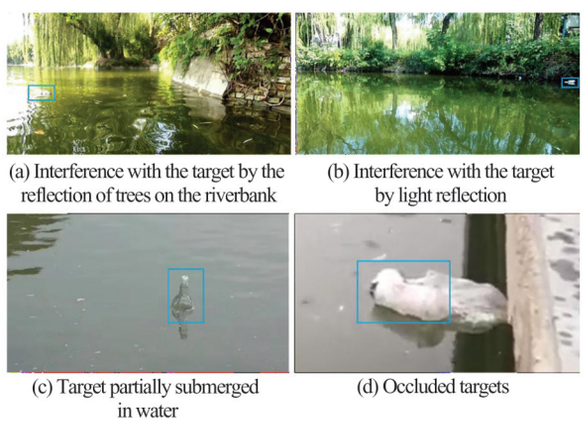


Fig. 7 Examples of different forms of interference

3.1 Ablation Experiment

In this paper, YOLOv7 is employed as the baseline and specific improvements are made to overcome its limitations. These improvements center around the incorporation of a new Neck structure, specifically the GFPN, utilizing multi-layer connections to expand the range of feature fusion and enhance feature representation. In the Queen-Fusion of GFPN, the RepC2f is introduced. This adjustment enables the module to capture more intricate gradient flow information in a lightweight design. Within the Queen-Fusion, the DFC attention mechanism is incorporated to capture long-distance dependency relationships, thereby better focusing on target features. The application of the WiseIoU loss function in bounding box regression is explored to enhance detection accuracy. To better understand the contributions of different components in the improved YOLOv7, a series of ablation experiments are conducted. Precision, Recall, mAP (mean Average Precision), and Elapsed time serve as evaluation metrics. For ease of illustration, YOLOv7, WiseIoU, GFPN, RepC2f, and DFC are denoted as A, B, C, D, and E in turn. The outcomes of the ablation experi-

ments, including the P-R (Precision-Recall) curve presented in Fig. 8, provide valuable insights into the performance changes across the evaluation metrics for the improved algorithm.

Figure 8 reveals that the incorporation of WiseIoU, GFPN, RepC2f, and the long-range attention DFC structures into the model has yielded notable improvements in several aspects. These include a heightened focus on small target features, the suppression of information loss in long-range feature transmission, and the consideration of the influence of anchors with different qualities. As a result, there has been a substantial enhancement in feature engineering. This is further reflected in the metrics presented in Table 1, where, in most cases, the inclusion of each structure has resulted in a noticeable enhancement in the model's performance. In comparison to the baseline, the mean average precision of the algorithm proposed in this paper has increased by approximately 6.3%.

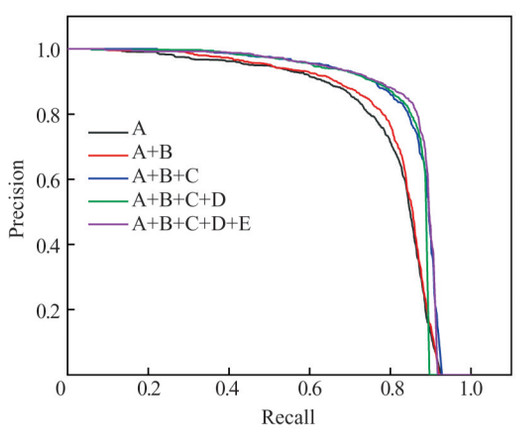


Fig. 8 Influence of different modules on P-R curves in ablation experiments

Table 1 Ablation experiment results

Algorithm	Precision /%	Recall /%	mAP@ 0.5/%	mAP@ 0.5:0.95/%	Elapsed time/ms
A	84.30	72.15	80.03	37.15	5.60
A+B	83.80	75.50	81.30	37.20	5.61
A+B+C	85.91	80.63	85.85	40.95	8.20
A+B+C+D	84.20	83.11	85.20	40.80	6.96
A+B+C+D+E	85.92	83.43	86.30	41.90	7.31

3.2 Comparison Experiment

To validate the superiority of the proposed algorithm, we conducted comparative experiments involving five detection algorithms: YOLOX[32] from Megvii, TPH-YOLOv5[20] with an additional fourth detection layer based on YOLOv5, YOLOv7, YOLOv7-BiFPN[33] with Neck replaced by BiFPN and the proposed algorithm, denoted as F, G, H, I, and J, respectively. The results are presented in Table 2, revealing that the mean average precision of these five algorithms generally exceeds 80% on the dataset. The proposed algorithm attains 86.3%, demonstrating the highest detection accuracy among the five, thereby validating the effectiveness of our algorithm. YOLOX exhibits the poorest overall performance, displaying the lowest precision metrics. Additionally, it incurs the longest processing time, primarily due to the lack of corresponding improvements in feature fusion for small targets and the decoupled head that increases complexity, impeding algorithm speed. TPH-YOLOv5, lacking enhancements in feature fusion, exhibits inadequate capabilities in multi-scale target detection, resulting in sub-optimal performance. YOLOv7 demonstrates detection performance similar to TPH-YOLOv5 but with the fastest execution speed, aligning with the network's characteristic of being faster and more efficient. This is attributed to the choice of YOLOv7 as the baseline for improvement in this study. YOLOv7-BiFPN, replacing YOLOv7's Neck with BiFPN, effectively fuses features from different levels, providing rich multi-scale feature representation and resulting in a respectable detection performance of 84.41%. However, the lack of efficient and gradient-rich C2f and reparameterization operations makes the network inferior to the performance and speed of the proposed algorithm. The proposed algorithm demonstrates superior performance, although with a lower execution speed than YOLOv7. This indicates that, despite considerations for efficient deployment and hardware-friendly convolution operations, such as the C2f, decoupled convolutions in DFC, and depthwise separable convolutions, the unavoidable increase in network layers during the improvement process leads to a reduction in execution speed. Additionally, the detection performance of YOLOv7, YOLOv5, SSD, and Fast RCNN networks was compared in Ref. [22], with detailed results beyond the scope of this paper. In conclusion, the proposed algorithm strikes a favorable balance between detection accuracy and speed for the de-

Table 2 Comparison of detection performance

Algorithm	Precision /%	Recall /%	mAP@ 0.5/%	mAP@ 0.5:0.95/%	Elapsed time/ms
F	81.01	74.93	79.50	33.81	9.56
G	82.11	76.35	80.91	36.42	6.25
H	84.30	72.15	80.03	37.15	5.60
I	82.15	81.91	84.41	38.52	8.22
J	85.92	83.43	86.30	41.90	7.31

tection of floating wastes in rivers, achieving significant improvements in both aspects.

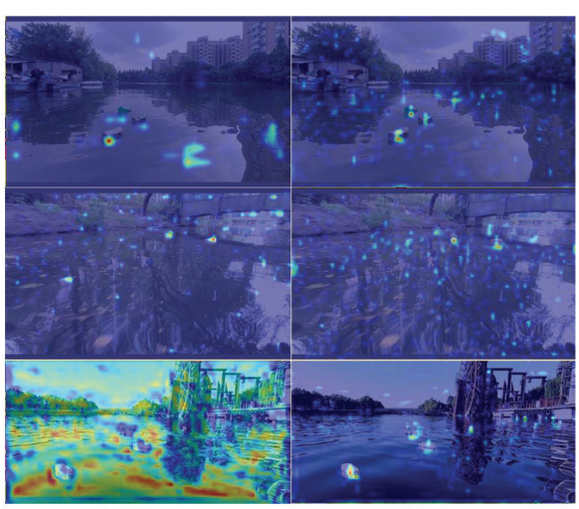
3.3 Visualisation of Detection Results

To provide a more intuitive illustration of the proposed algorithm's detection performance, three images under different interference scenarios are selected for evaluation. These scenarios include intense light reflection, interference from riverbank objects, and interference caused by tree reflections along the riverbank. Concurrently, a comparative analysis is conducted by contrasting the results of the proposed algorithm with those of the baseline, illustrated in Fig. 9. The figure depicts three typical scenarios of detecting floating wastes under different interference conditions, presented sequentially from top to bottom. A careful examination of the images reveals the enhanced accuracy of the proposed algorithm, particularly in identifying floating wastes under challenging scenarios. The consistently high detection confidence levels, surpassing 80%, underscore the reliability of the proposed algorithm. In contrast, the baseline algorithm demonstrates sub-optimal detection accuracy, marked by instances of both false positives and false negatives. Specifically, under intense light interference, the confidence level drops to 78%, false positives emerge in the presence of riverbank object interference, and false negatives manifest under interference from tree reflections along the riverbank. These findings robustly affirm the superior detection performance of the proposed algorithm.

(a) Ground-truth of three detected images (b) Detection results of YOLOv7 (baseline) (c) Detection results of the proposed algorithm

Fig. 9 Detection results in complex scenarios

To assess the effectiveness of the enhanced modules, specifically the DFC featuring a long-range attention mechanism, we employed the Grad-CAM visualization technique[34] to compare the performance differences in target recognition between the improved algorithm with DFC and the native YOLOv7. Grad-CAM technology visually represents the network's focus on the input image through a heatmap, aiding in the interpretation of key areas the network prioritizes during target identification, as depicted in Fig. 10. The results indicate that the improved algorithm produces higher heatmaps for target locations and lower heatmaps for irrelevant environmental details in non-target regions. This suggests that the attention module can effectively capture information from long-range spatial contexts, generating attention maps with a global receptive field. Consequently, the model concentrates more on the target area, enhancing the accu-



(a) YOLOv7 (b) Ours

Fig. 10 Comparison of the visual heat maps for the detection of floating wastes in rivers

racy and reliability of the detection task.

4 Conclusion

This paper provides a comprehensive examination of the attributes associated with floating wastes in rivers, encompassing challenges such as diminutive scale, low pixel count, limited information, and intricate backgrounds. These challenges frequently contribute to the inadequate performance of conventional target detection algorithms, giving rise to instances of both false positives and negatives. To tackle these challenges, we propose a detection algorithm founded on YOLOv7, incorporating the GFPN-RepC2f and a long-range attention mechanism. A sequence of experiments is executed, culminating in some findings: (1) The GFPN-RepC2f enhances information propagation capabilities, extending to deeper network layers for improved feature capture. (2) The DFC mechanism ensures fast execution on standard hardware, allowing the network to focus more on feature details and, consequently, further enhancing detection accuracy. Nevertheless, the incorporation of the attention mechanism unavoidably augments network layers and inference latency. This trade-off, prioritizing recognition accuracy over speed, is deemed advantageous for efficient waste cleanup in rivers. (3) This study provides novel insights and approaches to address the issues of waste detection, making significant contributions to environmental conservation and ecological development. Further research and optimization will focus on expanding the dataset by increasing sample quantity and incorporating more waste categories, promoting the model ' s effective deployment on waste cleanup machinery.

References

[1] Tong Y Q, Liu J F, Liu S Z. China is implementing "Garbage Classification" action[J]. *Environmental Pollution*, 2020, **259**: 113707.

[2] Cheng Y W, Zhu J N, Jiang M X, *et al*. FloW: A dataset and benchmark for floating waste detection in inland waters[C]// 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). New York: IEEE, 2021: 10933-10942.

[3] Zhou T H, Yang M M, Jiang K, *et al*. MMW radar-based technologies in autonomous driving: A review[J]. *Sensors*, 2020, **20**(24): 7283.

[4] Bansal M, Kumar M, Kumar M. 2D object recognition: A comparative analysis of SIFT, SURF and ORB feature descriptors[J]. *Multimedia Tools and Applications*, 2021, **80** (12): 18839-18857.

[5] Wei Y, Tian Q, Guo J, *et al*. Multi-vehicle detection algorithm through combining Harr and HOG features[J]. *Math Comput Simul*, 2018, **155**: 130-145.

[6] Campbell C, Ying Y M. *Learning with Support Vector Machines*[M]. Cham: Springer International Publishing, 2011.

[7] Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning[J]. *Journal of Applied Science and Technology Trends*, 2021, **2**(1): 20-28.

[8] Bharati P, Pramanik A. Deep learning techniques—R-CNN to mask R-CNN: A survey[C]//*Computational Intelligence in Pattern Recognition*. Singapore: Springer-Verlag, 2020: 657-668.

[9] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137-1149.

[10] Liu Q P, Bi J J, Zhang J W, *et al*. B-FPN SSD: An SSD algorithm based on a bidirectional feature fusion pyramid[J]. *The Visual Computer*, 2023, **39**(12): 6265-6277.

[11] Huang L C, Wang Z W, Fu X B. Pedestrian detection using RetinaNet with multi-branch structure and double pooling attention mechanism[J]. *Multimedia Tools and Applications*, 2024, **83**(2): 6051-6075.

[12] Diwan T, Anirudh G, Tembhurne J V. Object detection using YOLO: Challenges, architectural successors, datasets and applications[J]. *Multimedia Tools and Applications*, 2023, **82** (6): 9243-9275.

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

[14] Ding X H, Zhang X Y, Ma N N, *et al*. RepVGG: Making VGG-style ConvNets great again[C]//2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2021: 13728-13737.

[15] Lee Y, Hwang J W, Lee S, *et al*. An energy and GPU-computation efficient backbone network for real-time object detection[C]//2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*). New York: IEEE, 2019: 752-760.

[16] Zand M, Etemad A, Greenspan M. Oriented bounding boxes for small and freely rotated objects[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, **60**: 4701715.

[17] Zhang Y Q, Bai Y C, Ding M L, *et al*. Multi-task generative adversarial network for detecting small objects in the wild [J]. *International Journal of Computer Vision*, 2020, **128**(6): 1810-1828.

[18] Gao C, Tang W, Jin L Z, *et al*. Exploring effective methods to improve the performance of tiny object detection[C]//*Eu-*

ropean Conference on Computer Vision. Cham: Springer-Verlag, 2020: 331-336.
[19] Leng J X, Ren Y H, Jiang W, *et al*. Realize your surroundings: Exploiting context information for small object detection[J]. *Neurocomputing*, 2021, **433**: 287-299.
[20] Zhu X K, Lyu S C, Wang X, *et al*. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*). New York: IEEE, 2021: 2778-2788.
[21] Benjumea A, Teeti I, Cuzzolin F, *et al*. YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles [EB/OL]. [2021-10-01]. http://arxiv.org/abs/2112.11798.
[22] Qi L G, Gao J L. Small object detection based on improved yolov7[J]. *Computer Engineering*, 2023, **49**: 41-48(Ch).
[23] Wang X R, Xu Y, Zhou J P, *et al*. Safflower picking recognition in complex environments based on an improved YOLOv7[J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2023, **39**(6): 169-176.
[24] Kang J, Wang Q, Liu W., *et al*. Detection model of aerial photo insulator multi-defect by integrating cat-bifpn and attention mechanism[J]. *High Voltage Engineering*, 2023, **49**: 3361-3376(Ch).
[25] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection[C]//2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2020: 10778-10787.
[26] Jiang Y Q, Tan Z Y, Wang J Y, *et al*. GiraffeDet: A heavy-neck paradigm for object detection[EB/OL]. [2022-10-01]. http://arxiv.org/abs/2202.04256.
[27] Li Y T, Fan Q S, Huang H S, *et al*. A modified YOLOv8 detection network for UAV aerial image recognition[J]. *Drones*, 2023, **7**(5): 304.
[28] Tang Y H, Han K, Guo J Y, *et al*. GhostNetV2: Enhance cheap operation with long-range attention[J]. *Advances in Neural Information Processing Systems*, 2022, **35**: 9969-9982.
[29] Rezatofighi H, Tsoi N, Gwak J Y, *et al*. Generalized intersection over union: A metric and a loss for bounding box regression[C]//2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: IEEE, 2019: 658-666.
[30] Zheng Z H, Wang P, Liu W, *et al*. Distance-IoU loss: Faster and better learning for bounding box regression[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, **34**(7): 12993-13000.
[31] Tong Z J, Chen Y H, Xu Z W, *et al*. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism[EB/OL]. [2023-10-01]. http://arxiv.org/abs/2301.10051.
[32] Ge Z, Liu S T, Wang F, *et al*. YOLOX: Exceeding YOLO series in 2021[EB/OL]. [2021-10-01]. http://arxiv. org/abs/2107.08430.
[33] Wang Y, Wang H Y, Xin Z H. Efficient detection model of steel strip surface defects based on YOLO-V7[J]. *IEEE Access*, 2022, **10**: 133936-133944.
[34] Zhang Y Y, Hong D, McClement D, *et al*. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging[J]. *Journal of Neuroscience Methods*, 2021, **353**: 109098.

基于GFPN和长程注意力机制的改进YOLOv7河面漂浮垃圾检测算法

彭程[1,2†], 何冰[1,2], 席文强[2], 林关成[2]

1. 渭南师范学院 物理与电气工程学院, 陕西 渭南 714099
2. X射线成像与检测陕西省高校工程研究中心, 陕西 渭南 714099

摘要: 河面漂浮垃圾具有尺度小、像素少、信息量低和背景复杂的特点, 容易产生误检、漏检的问题, 从而导致检测效果不佳。针对这些问题,本文提出了一种基于YOLOv7的河面漂浮垃圾检测算法, 该算法融合了改进的广义特征金字塔网络(GFPN)和长程注意力机制。首先,将YOLOv7中的Neck替换为改进的GFPN网络,从而提供更有效的信息传输, 以方便扩展到更深的网络。其次, 引入了基于卷积且硬件友好的长程注意力机制, 使算法能够快速生成具有全局感受野的注意力图。最后, 算法采用WiseIoU优化损失函数, 实现自适应梯度增益分配, 缓解低质量样本对梯度的负面影响。仿真结果表明, 所提出的算法在实时场景检测任务中取得了86.3%的平均准确率, 这比基准提高了6.3%, 表明该算法在漂浮垃圾检测方面表现优异。

关键词: 河面漂浮垃圾检测; YOLOv7; GFPN; 长程注意力

□