Marine Vessel Target Detection Algorithm Based On Improved YOLOv5

Chen Gao[1,2], Jiyong Xu[1,2,3,*], and Ruixia Liu[1,2]
[1] School of Mathematics and Statistics, Qilu University of Technology (Shandong Academy of Sciences)
Jinan, 250014, China
[2] Shandong Artificial Intelligence Institute
[3] Shandong Provincial Computing Center (National Supercomputing Center In Jinan)
Jinan, 250014, China
[e-mail:xujy@sdas.org]
[*]Corresponding author: Jiyong Xu

Received May 7, 2024; revised August 5, 2024; accepted September 23, 2024; published October 31, 2024

Abstract

Considering the intricate and ever-changing nature of the marine environment and the diverse range of sizes for targets involved in marine ship target recognition, which present challenges in detecting specific targets, a marine ship target detection algorithm has been developed based on an enhanced iteration of YOLOv5. Initially, the integration of dynamic snake convolution (DySnakeConv) into the feature extraction network and subsequent enhancement of the C3 module based on this integration were implemented. This integration enables dynamic adjustments based on the input image size, adaptive fusion of feature sequences, and resolution of accuracy and continuity issues during the recognition process. Subsequently, a novel hybrid encoder (FSI) was devised, utilizing target scale characteristics to enhance the extraction capability of multi-scale information, facilitating effective detection and recognition of objects within images. Finally, we selected the Shape-IOU bounding box loss function to mitigate fixed target frame issues and enhance target detection accuracy. Experimental evaluations were conducted utilizing the Infrared Maritime Ship dataset. The results demonstrated that our enhanced model achieved a prediction accuracy of 93.8% and an average precision (mAP) value of 93.89%, surpassing YOLOv8s by 1.2% and 1.8%, respectively. Moreover, there was an increase in recall rate by 2% compared to YOLOv8n while reducing parameters from 10,473,392 to 6,549,901 only. The computational load decreased by 6.3 GFLOps compared with YOLOV8n, resulting in better performance in ocean target detection and recognition.

Keywords: Dynamic snake convolution, hybrid encoder, loss function, target detection, YOLOv5.

This work is supported by the Natural Science Foundation Innovation and Development Joint Fund Project of Shandong Province under Grant NO. ZR2023LZH009 and Key R & D Project of Shandong Province under Grant NO. 2020CXGC010501.

1. Introduction

With the national emphasis on and promotion of the marine economy in the "14th Five-Year Plan" and the acceleration of efforts to build a strong marine nation, there has been a significant increase in the number and types of ships. This trend has led to heightened demands for early warning systems for vessels at sea. However, China's current technical capabilities in marine information monitoring and early warning require further enhancement, particularly in visual information acquisition by marine buoys. The limitations in this capacity hinder the diversity and accuracy of marine information monitoring. Furthermore, environmental factors such as fog, sea mist, and strong winds can substantially impact the quality and clarity of target images collected at sea, potentially resulting in information loss and misidentification of marine images. Traditional target detection algorithms, which typically rely on features and classifiers, may introduce redundancies in the detection window and are not well-suited for ship detection under these conditions. Therefore, it is imperative to explore deep learning-based models for marine target detection algorithms.

Conventional target detection methods typically involve using a sliding window traversal technique to identify potential regions, followed by manually designing a feature operator. Subsequently, the operator is used to refine the feature set, and then a classifier is developed to categorize the features and ultimately select the optimal box. In contrast, target detection methods based on deep learning leverage convolutional neural networks (CNN) to autonomously extract feature details from images. These methods then identify target information based on the extracted features, resulting in the acquisition of more comprehensive semantic features [1]. This approach enhances adaptability in detecting changes in mandates.

Deep learning-based target detection can be classified into two categories: "two-stage" [2] and "single-stage" [3,4,5,6]. These categories are based on the presence or absence of an explicit Region of Interest (RoI) extraction process. Single-stage models, pioneered by the YOLO series [7] and further developed in models like SSD [8], directly forecast the category and location of the target object through regression, albeit with slightly lower accuracy. The YOLO family of algorithms, from YOLOv2 to YOLOv10 [9-17], has enhanced the accuracy and detection speed of single-stage target detection models. On the other hand, "two-stage" target detection approaches exemplified by the R-CNN [18] series offer higher detection accuracy but at the cost of slower computation speed. Subsequent models such as SPPNet [19], Fast R-CNN [20], and Faster R-CNN [21] have been introduced to address excessive computation issues. Nevertheless, current advancements in algorithms still fall short of meeting demands for real-time detection.

The complex, diverse, and unpredictable nature of the marine environment poses significant challenges in identifying and monitoring marine objects. Target detection, a fundamental task in visual recognition, involves the identification and localization of a target within an image. Currently, target detection systems commonly rely on two standard architectures: convolutional neural network (CNN)-based and Transformer-based [22]. In the field of ship detection at sea, widely used algorithms include the R-CNN family, SSD, YOLO, RetinaNet [23], and DETR [24]. The Transformer-based object detector (DETR) [25,26,27,28,29] streamlines the target detection process and facilitates end-to-end target detection. More recently, RT-DETR [30] has been proposed for real-time target detection while DINO [31] has demonstrated substantial advancements in results. The primary contributions are outlined as follows:

• A novel hybrid encoder (FSI) has been developed to address the varying scale characteristics of the target. This approach enhances the network's capability to extract multi-scale information and effectively converts multi-scale features into a sequence of image features. Consequently, it enables accurate detection and identification of objects within images.

• Incorporating Dynamic Snake Convolution (DySnakeConv [32]) into the feature extraction network and enhancing the C3 module based on Dynamic Snake Convolution enables dynamic adjustments based on the input image size. This adaptation allows for adaptive fusion of feature sequences and helps to address precision and consistency challenges encountered during the recognition process.

• The study has calculated the optimal scaling factor for the Shape-IOU [33] loss function, tailored to the specific dataset. This calculation addresses the issue of a fixed target frame and enhances the detection of small targets.

2. Basic principles of YOLOv5

The YOLO series of algorithms is currently widely utilized for target detection and demonstrates exceptional performance in real-time maritime scenarios. YOLOv5 is a single-stage object detection algorithm based on single-stage deep learning, comprising four main components: the Input, the Backbone, the Neck, and the Predictor. This study adopts YOLOv5s as a benchmark, incorporating data augmentation, the deep convolutional network CSPDarknet53 architecture, the SPPF module, the FPN [34] structure, and the anchor frame predictor module. The architecture of Yolov5 is illustrated in **Fig. 1**.

Commonly utilized path aggregation blocks in object detection models include FPN, NASFPN [35], BiFPN [36], ASFF [37], and SFAM [38]. YOLOv5 typically employs a CSP structure and SPP layer for feature extraction, and PANet [39] for fusion across different scales. The input module resizes the input image to the required fixed size and can also perform data augmentation, adaptive anchor boxes, and multi-scale training to improve model performance. The backbone network extracts features from the input image and enhances feature representation through fusion to provide more precise information for target detection. The neck network integrates the extracted features and transfers them to the head network for prediction, generating feature maps of varying sizes. Its feature fusion capabilities contribute to enhancing the model's detection accuracy.

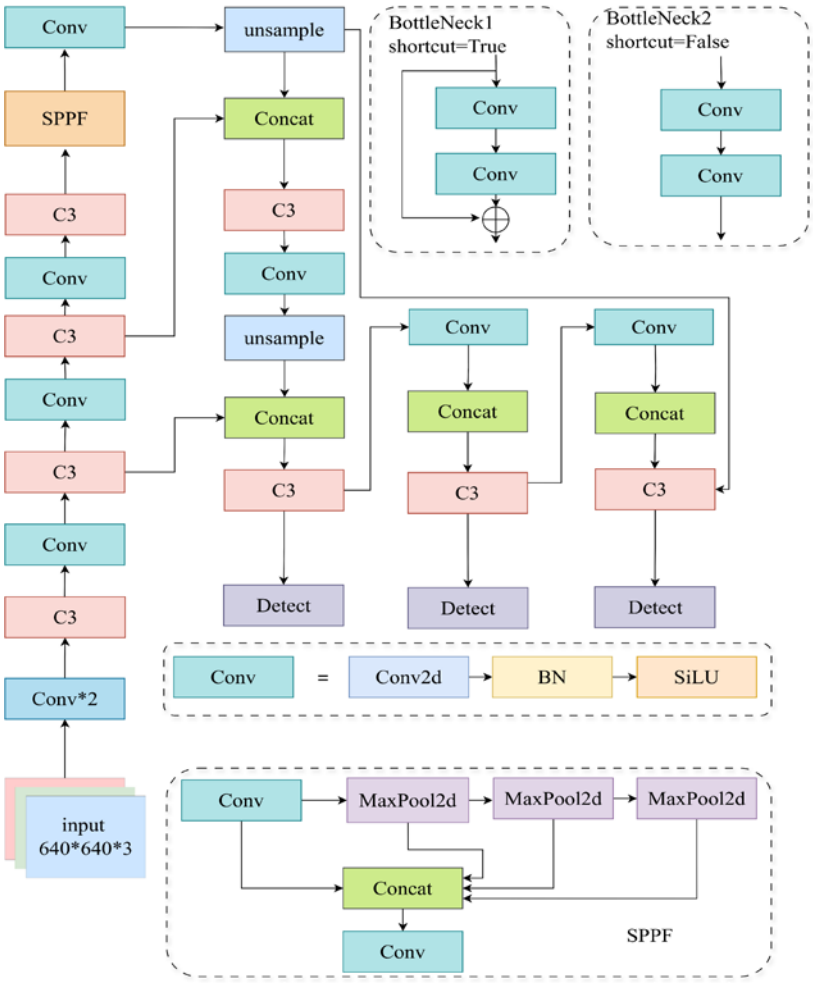


Fig. 1. YOLOv5 basic principles.

3. Improved YOLOv5 detection model

3.1. Overall network architecture

In this paper, an algorithm called Feature Sequence Interaction (FSI-YOLO) is proposed based on the enhanced YOLOv5n target detection network. The architecture of the proposed algorithm is presented in **Fig. 2**.

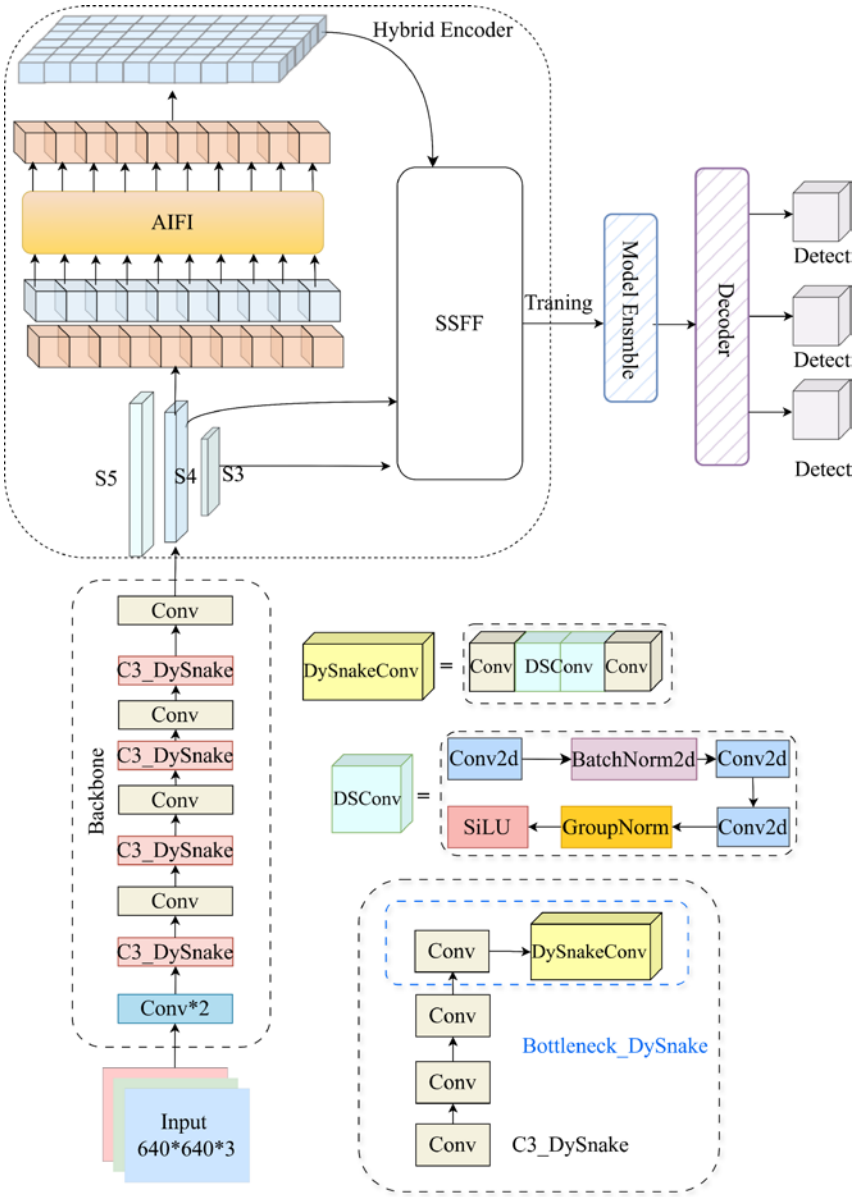


Fig. 2. FSI-YOLO structure.

The YOLOv5n model has been enhanced for improved detection of small offshore vessels, which can be challenging to accurately identify. This enhancement involved the integration of the C3 module and DySnakeConv to form the new C3_DySnake module, replacing a portion of the traditional Conv and enabling adjustment of convolutional parameters. By employing a scale transformation method to modify image size, the convolutional parameters can capture the structural characteristics of targets from different perspectives, thereby enhancing detection accuracy. Furthermore, the neck network (Neck) has been strengthened with a novel hybrid encoder (FSI), primarily consisting of the AIFI module and SSFF module to effectively convert multi-scale features into image features. The feature sequences are then fused adaptively using DySnakeConv, with the AIFI module enhancing correlation among different scales and reducing computational redundancy. Additionally, the SSFF module combines

spatial and scale features to consolidate essential information at various scales while focusing on marine target characteristics. This method improves the network's ability to extract multi-scale information and efficiently convert multi-scale features into a series of image features, resulting in accurate object detection and identification within images. Finally, the implementation of Shape-IOU bounding box loss function instead of CIOU reduces background interference by evaluating overlaps, thereby improving bounding box regression accuracy and strengthening model resilience. The research utilized the publicly available Infrared Maritime Ship dataset [40].

3.2. Hybrid Encoder

Currently, the deformable attention mechanism reduces computational costs to some extent. However, the fusion of multi-scale features leads to a notable increase in the input sequence length, imposing a substantial computational burden on the encoder. To address these challenges, this paper introduces a novel hybrid encoder FSI network architecture. The proposed model comprises two primary components: the intra-scale feature interaction (AIFI) module and the attentional scale sequence fusion (SSFF) module. These components efficiently convert multiscale features into image feature sequences to provide complementary information for marine vessel target recognition. The AIFI module enhances connectivity across different scales and reduces computational redundancy by facilitating intra-scale interaction of high-level semantic features. On the other hand, the SSFF module fuses attention scales across sequences, which aggregates crucial information at various scales by fusing spatial and scale features. Transformer, a robust tool for processing sequence data, is utilized to extract feature information and accelerate the processing of multi-scale features. The study presents various Transformer variants with diverse encoder types: (a) a traditional connection lacking a Transformer encoder, (b) an insertion of a single-scale Transformer encoder comprising a layer of Transformer blocks. Each scale's features leverage a shared encoder for intra-scale feature interactions, linking the resulting multi-scale features. Additionally, (c) cross-scale feature fusion is introduced, where cascaded multiscale features are inputted into the encoder for feature interaction. Lastly, the new hybrid encoder FSI continuously optimizes intra-scale interaction and cross-scale fusion of multi-scale features. Initially, a single-scale Transformer encoder is introduced to enable internal interaction of features within each scale for intra-scale interaction, followed by cross-scale fusion. The four encoders are depicted in **Fig. 3**, where SSE represents the single-scale encoder module and MSE represents the multi-scale encoder module.

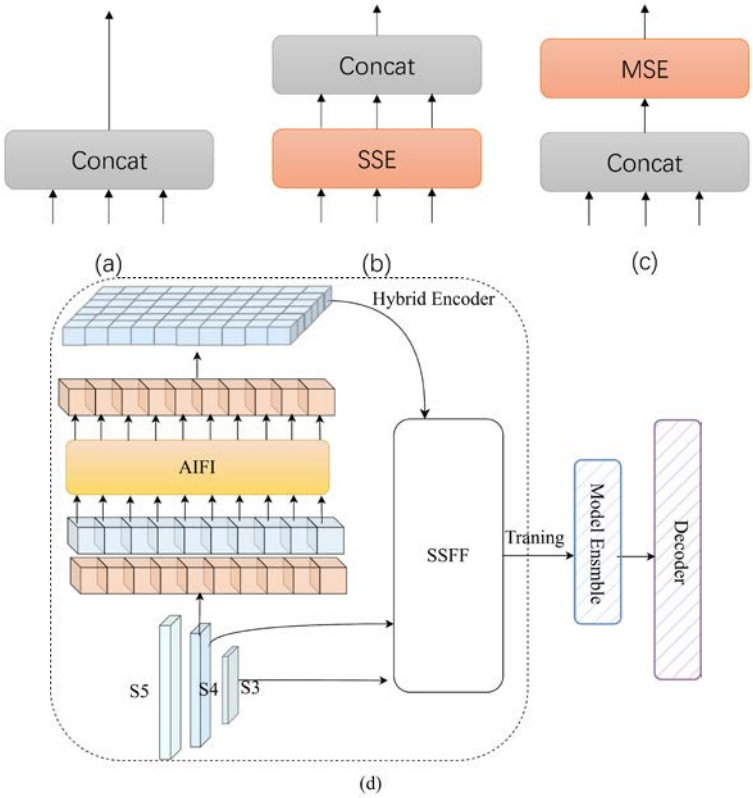


Fig. 3. Existing encoders and our encoders.

3.3. Dynamic snake convolution

Traditional convolutional feature extraction is well-known for its robustness and parameter sharing. However, it faces challenges in accurately capturing the structural characteristics of various targets. On the other hand, dynamic serpentine convolution (DSConv) is characterized by its adaptive nature, allowing it to effectively capture structural features by focusing on elongated and zigzagging local structures. This enables DSConv to better accommodate various target shapes. Additionally, DSConv incorporates offsets for deformations, which can be used to select the processing location for each target based on the input feature maps to learn deformations. Using an iterative approach, DSConv sequentially processes each target by selecting subsequent positions to observe, ensuring consistent attention without excessively expanding the sensory field due to significant deformation offsets.

In the original YOLOv5 feature fusion network, the C3 module has a fixed input size and only supports the same input resolution as the training image. It utilizes a fully connected layer for prediction, resulting in reduced processing efficiency. The newly designed C3_DySnake module addresses this limitation by being able to adapt to input images of various sizes, thereby enhancing its adaptability to inputs of different sizes and improving ship target detection performance and speed. The main improvements of this module include: 1) Replacement of the traditional convolution module with the DySnakeConv convolution layer, which can dynamically adjust the position of the convolution kernel. This comprises two traditional Conv convolutions and two DSConv modules. 2) Employing DSConv convolution to capture diverse features of the target image. Further details are illustrated in **Fig. 4**.

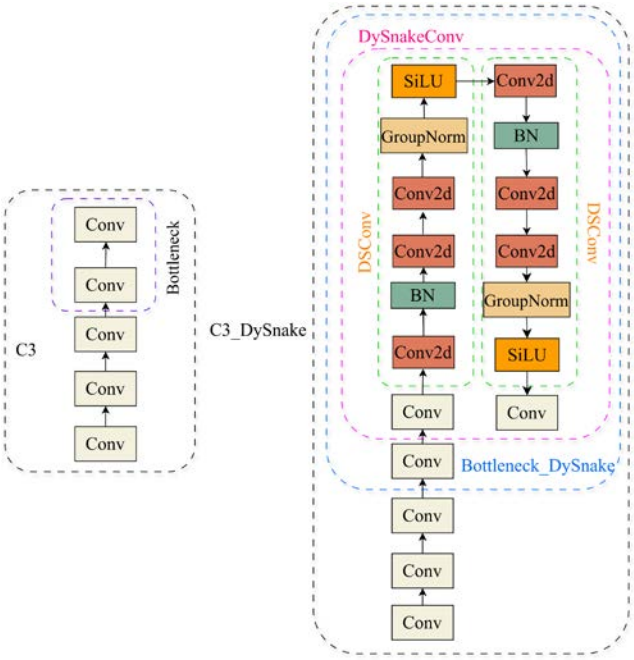


Fig. 4. Comparison between C3 and C3_DySnake.

Dynamic Snake Convolution introduces a novel approach to replace traditional convolution methods. It dynamically adjusts the shape and position of the convolution kernel based on the characteristics of the input image, allowing for adaptive modifications of the anchor frame. While DySnakeConv entails a significant computational burden, experiments indicate that leveraging the parallel processing capabilities of GPUs can accelerate the dynamic convolution process. Furthermore, conducting layered computations on distinct feature layers ensures efficient utilization of computational resources, enabling effective detection of various targets. The specific workflow involving Feature Selection and Integration (FSI) and DySnakeConv is depicted in **Fig. 5** below.

Fig. 5. Network Diagram.

3.4. Loss function

In the field of target detection, the loss function is used to measure the difference between the predicted value of the model and the ground truth. However, traditional IOU loss functions often do not consider the impact of anchor frames on regression performance. In marine environments, ship targets are often small in size, making it difficult for conventional loss functions to effectively detect them while being susceptible to background disturbances. To address these limitations, we have adopted Shape-IOU Loss [33], which better handles ship target boundaries by optimizing spatial overlap between predicted bounding boxes and ground truth annotations. This approach minimizes shape discrepancies and is particularly advantageous for objects with complex shapes or varying aspect ratios. Simultaneously, Shape-IOU Loss mitigates background interference and enhances model resilience through

overlap assessment. The performance improvement can be attributed to the Shape-IOU loss function's ability to prioritize overall overlap between predicted and ground truth bounding boxes, resulting in better alignment and reducing the likelihood of misclassifying partially occluded or irregularly shaped objects. Additionally, using Shape-IOU as our chosen loss function helps achieve a more consistent training process by focusing on the end-goal metric (IOU), which directly correlates with detection performance. To accommodate target samples of different sizes, we experiment with various Shape-IOU metrics for computing the loss function while introducing a scale factor to accurately evaluate model performance across different scales. In summary, incorporating the Shape-IOU bounding box loss function has significantly enhanced our target detection model precision. By specifically addressing the overlap between predicted and ground truth bounding boxes, this loss function effectively mitigates inherent limitations associated with conventional loss functions.

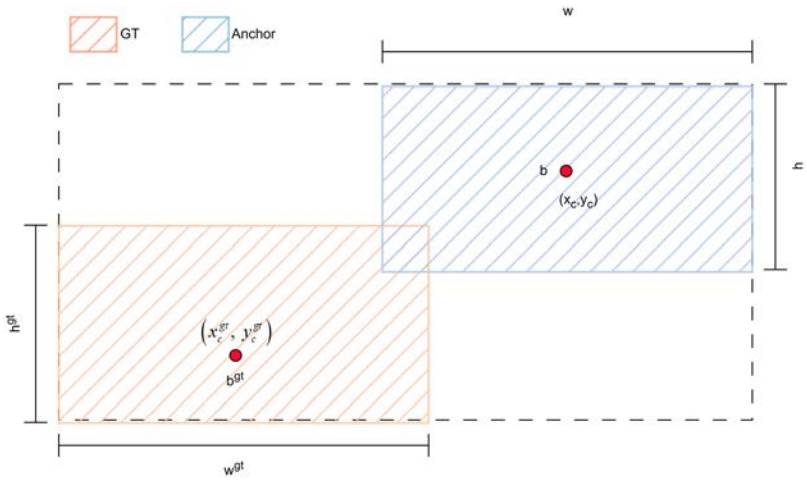


Fig. 6. Shape-IOU principle.

When considering the geometric constraints between the ground truth (GT) box and the prediction box, Shape-IOU calculates the loss by adjusting the scale of the anchor box itself. This adjustment enhances the accuracy of anchor box regression. It has been experimentally determined that Shape-IOU is linked to a scale factor, which represents the scale of the target in the dataset. In our experiment, we observed that prediction accuracy peaks when the scale factor is 0.8, but there is a slight decrease in mean average precision (mAP) compared to a scale factor of 0.0. However, as opposed to a scale factor of 0.0, prediction accuracy significantly decreases as the scale factor increases. As the scale factor progressively increases, both prediction accuracy and other related metrics decline to varying extents. The relationship can be expressed as follows:

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|}, \tag{1}$$

$$\Omega^{shape} = \sum_{t=\omega,h} \left(1 - e^{-\omega t}\right)^{\theta}, \theta = 4, \tag{2}$$

$$L_{Shape-IoU} = 1 - IoU + distance^{shape} + 0.5 \times \Omega^{shape}, \tag{3}$$

Where B and Bgt represent the prediction box and GT box respectively, where the scale factor (scale) is associated with the target's scale in the dataset. w and refer to the width and height of the anchor box. θ refer to the cost of the shape and its value is unique for each dataset, in which the parameter is set to 4. Ω^{shape} is the shape cost. $distance^{shape}$ is the distance of the shape. $L_{Shape-IoU}$ is the bounding box regression loss.

4. Experiments and analysis

4.1. Experimental settings and datasets

The hardware configuration of this experiment is NVIDIA Tesla V100 SXM2 (32GB) GPU for hardware configuration, Ubuntu 20.04 as the operating system, and Pytorch as the deep learning framework. The experiment was conducted under the environment of CUDA 11.6, conda 23.7.2, Python 3.8.12, and Pytorch 2.1.0 for GPU training. To optimize the utilization of computing resources during network training, a batch size of 16, a learning rate of 1e-4, 8 workers, and 300 epochs were set. Additionally, the Adam algorithm and SGD optimizer were employed for model optimization during training.

In this study, we conducted a comparative analysis of various datasets, focusing on their types, images, and labeling information. Following the comparison, we identified the Infrared Maritime Ship dataset as the most suitable option. Detailed comparison results are presented in **Table 1**.

Table 1. Comparison of different datasets

Dataset	Types	Images	Instances
Maritime Ship	5	4760	15265
Ship-detection	5	7690	21057
Infrared Maritime Ship	**7**	**8402**	**21638**

In this study, we utilize the Infrared Maritime Ship Target Detection Database in a real maritime defense scenario to evaluate the effectiveness of the infrared target detection algorithm under real-world conditions. The database covers a range of scenarios, time periods, and resolutions within maritime environments, including sea, harbor, and coastal settings with various types of vessels such as cruise ships, bulk carriers, warships, sailing boats, kayaks, container ships, and fishing boat targets, totaling 9400 images.

The dataset contains images of different sizes: 384*288, 640*512, and 1280*1024. Once the dataset has been selected, it undergoes preprocessing procedures aimed at enhancing training data diversity by generating additional samples through flipping and cropping techniques.

4.2. Criteria for outcome evaluation

To evaluate the practical effectiveness of the proposed model, this study utilizes metrics such as precision (P), recall (R), mean average precision (mAP) across all categories, number of parameters, and GFLOPs. These metrics are employed to assess the performance of the improved algorithmic model. Specifically, accuracy (P) and recall (R) are calculated using the

formulas (4)-(6) provided below:

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{5}$$

In Eq. (4) and Eq. (5), TP represents the correctly predicted target in the detection result, FP denotes the incorrectly predicted target, and FN refers to targets not detected by the model. The mean Average Precision (mAP) combines precision (P) and recall (R) for n categories, offering a more comprehensive evaluation of the network's performance. The average precision mAP is defined as follows:

$$mAP = \frac{\sum_{i=1}^{\mathrm{n}} \int_0^1 P(R)dR}{n} \tag{6}$$

When evaluating the performance of a model, it is important to consider both its detection effect and the size of the model. The size of the model is determined by the number of parameters, which is a key metric. The number of parameters in the model has a positive correlation with its fitting capacity; in other words, a higher number of parameters indicates better fitting ability but also implies a greater demand for computational and storage resources.

4.3. Analysis of experimental results

4.3.1. Detection algorithm comparison experiment

To demonstrate the effectiveness of the FSI-YOLO model, this study has employed a more advanced target detection model for comparison. The experimental results for target detection in the marine infrared dataset are presented in **Table 2**.

In terms of detection accuracy, each detection metric of the FSI-YOLO model outperforms the current network model, with mAP@0.5 reaching 93.89%, mAP@0.5:0.95 at 63.9%, and the R index at 90.3%. Compared to the latest detection models, FSI-YOLO exhibits superior detection accuracy and outperforms existing models.

Table 2. Comparison of different detection algorithms

Models	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	Params/M
Faster R-CNN	82.1	75.3	80.8	60.5	41.7
SSD	83.5	78.7	84.1	61.1	24.4
YOLOv3-tiny	84.6	78.6	83.6	61.3	8.69
YOLOv5s	93.1	89.9	89.2	62.5	7.03
YOLOv7-tiny	85.4	79.3	84.5	62.7	6.02
YOLOv8s	92.6	88.8	92.1	63.5	11.1
TPH-YOLOv5	93.5	90.1	91.2	62.8	36.9
YOLOv9s	**92.1**	**89.7**	**93.4**	**63.5**	**9.93**
YOLOv10s	**92.8**	**90.5**	**93.3**	**63.7**	**8.07**
FSI-YOLO(ours)	**93.8**	**90.3**	**93.9**	**63.9**	**10.4**

Table 2 illustrates that FSI-YOLO demonstrates superior performance in precision and recall compared to both YOLOv9 and YOLOv10. Additionally, the mean average precision of FSI-YOLO is 93.9%, which significantly outperforms that of YOLOv9 and YOLOv10. This indicates that FSI-YOLO exhibits enhanced detection accuracy and recall across a variety of target detection tasks. In contrast, Yolov9 shows diminished detection accuracy in complex backgrounds and with multi-scale targets, thereby positioning FSI-YOLO as an optimal solution for such scenarios. On the other hand, Yolov10 imposes excessive computational demands. To address this issue, the proposed model incorporates the FSI module, which reduces processing intensity while achieving a noteworthy improvement in computational accuracy compared to the aforementioned YOLOv9 and YOLOv10 models.

Table 3 presents the specific accuracy, recall, mAP50, and mAP50:0.95 values for each behavioral type. The highest detection rates are observed for bulk carriers, container ships, and warships across all three behavioral types. Furthermore, these first three behavioral types are more frequently labeled, exhibit higher shape similarity, and pose greater challenges in terms of detection. As a result, the accuracy for these types is slightly weaker compared to the last three behavioral types. Nevertheless, the anticipated detection effect is still achieved.

Table 3. Infrared Maritime Ship's FSI-YOLO training and testing results

class	P/%	R/%	mAP50/%	mAP0.5:0.9/%
all	93.8	90.3	93.9	63.9
liner	93.5	84.8	89.6	52.2
bulk carrier	97.3	94.4	97.2	75.8
warship	96.3	99.4	99.2	80.9
sailboat	92.9	89.1	92.1	54.4
canoe	85.5	83.3	85.3	49.2
container ship	98.3	97.9	98.4	81.5
fishing boat	92.4	82.7	88.6	47.2

In the context of the Shape-IOU loss function, the selection of the scale (or scale factor) significantly influences detection accuracy, which is dependent on the scale of the target object in the dataset. A comparative experiment was conducted to determine the most appropriate scale factor. Five groups with different scale factors were established for evaluation, as shown in **Table 4**. The experimental results indicate that when the scale is set to 0.0, the mAP@0.5 value of the improved model reaches 93.9%, exhibiting superior mean accuracy compared to other groups. It has been concluded that choosing a scale factor of 0.0 leads to significant improvements in results for this experimental dataset.

Table 4. The impact of different scales on the detection performance

scale	P /%	R/%	mAP@0.5/%	mAP@0.5:0.95/%
0.0	**93.8**	**90.3**	**93.9**	**63.5**
0.4	92.3	88.5	92.9	61.5
0.6	91.5	88.2	92.6	62.1
0.8	89.8	88.6	93.1	62.3
1.0	89.5	86.4	92.3	60.5

4.3.2. Ablation experiment

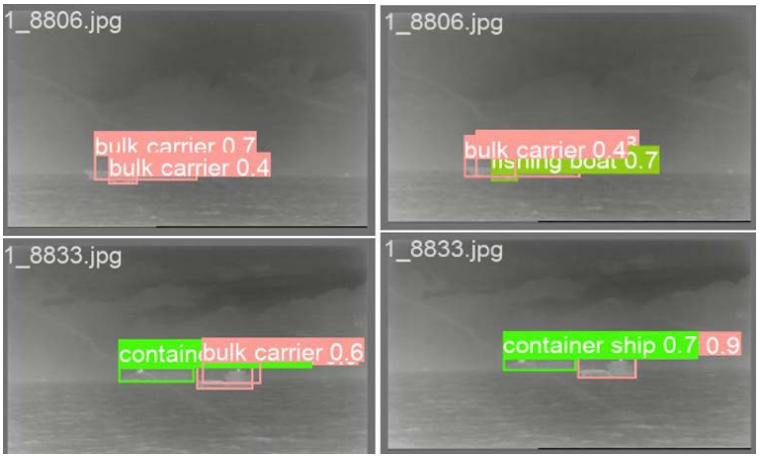
The purpose of this study is to gain a deeper understanding of the impact of improved modules and their combination with other improved modules on enhancing the performance of the original model. In this paper, we present the results of ablation experiments. **Table 5** summarizes the recognition accuracy, demonstrating that each improvement module has led to a noticeable enhancement in the model's performance, as seen in its increased accuracy (P), recall (R), and average precision (mAP). The AIFI and DySnake modules have improved model accuracy and recall rate, respectively, validating their effectiveness in detecting small vessels. The incorporation of three forms of attention within the DySnake module has resulted in a significant increase in accuracy, recall, and average precision. Although combining multiple modules (e.g., AIFI, SSFF, and DySnake) has potential to enhance model accuracy, it may also have implications for frames per second (FPS) and model size. Ultimately, integrating diverse modules allows us to utilize and enhance their individual strengths to optimize overall model performance.

Table 5. Ablation experiment

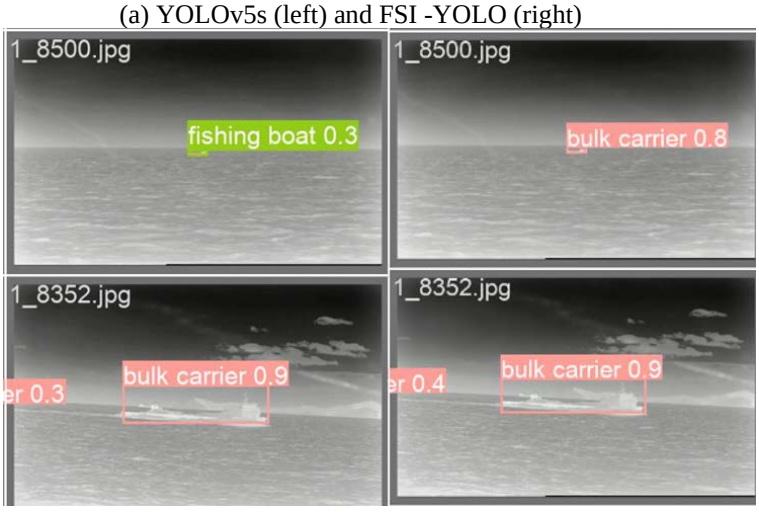
Models	P/%	R/%	mAP@0.5 /%	mAP@0. 5:0.95/%	**GFLOPs**	**Parameters**
Baseline	92.4	89.1	89.2	62.5	**15.8**	**7029004**
+SSFF	92.6	87.5	91.7	59.5	**17.9**	**7267596**
+AIFI	92.9	89.2	92.1	62.9	**15.7**	**7162636**
+DySnake	92.8	89.3	92.5	62.7	**19.5**	**9849928**
+AIFI+DySnake	93.2	89.3	92.2	61.8	**19.2**	**9204228**
+AIFI+DySnake +Shape_IOU	93.0	89.5	92.6	63.3	**19.2**	**9204228**
+ SSFF+DySnake +Shape_IOU	93.4	90.1	93.2	63.5	**19.1**	**8551924**
+AIFI+SSFF(FSI)+Shape_IOU	**92.4**	**89.3**	**92.9**	**61.9**	**18.0**	**7410604**
+AIFI+SSFF(FSI)+DySnake+Shape_IOU	**93.8**	**90.4**	**93.9**	**63.9**	**22.6**	**10473392**

It is evident from **Table 5** that the AIFI and SSFF modules have the capability to reduce the complexity of calculations and the number of parameters needed. The FSI, which encompasses these two modules, effectively reduces the model's complexity. In contrast, the network incorporating the DySnake module shows higher calculation complexity. While the calculation complexity of the network containing DySnake is relatively high, it does improve calculation accuracy and mAP accuracy, especially in the ablation experiments conducted by the SSFF module in the FSI hybrid encoder. DySnake addresses accuracy and continuity issues in the recognition process, demonstrating an improvement in mAP@0.5 accuracy by 1.8% and a dynamic adaptation to targets of varying sizes. The accuracy of FSI module decreases without DySnake, although approximately 30% fewer parameters are required. Thus, it can be inferred that employing FSI represents an effective approach for reducing model complexity. The combination of FSI and Snake in this discussed model effectively reduces complexity while increasing target detection accuracy even with an increase in model parameters within acceptable limits. Furthermore, the Shape-IOU Loss is effective for detecting ship targets

against complex backgrounds such as ocean background through measuring shape overlap between predicted frames and real frames thereby enhancing detection efficacy while mitigating background interference impact. **Fig. 7** illustrates how two different algorithms affect target recognition – (a) depicts YOLOv5s algorithm's detection result; (b) showcases FSI-YOLO algorithm's detection result.



(a) YOLOv5s (left) and FSI -YOLO (right)



(b) YOLOv5s (left) and FSI -YOLO (right)

Fig. 7. Existing encoders and our encoders.

5. Discussion

This paper presents an enhanced target detection model, FSI-YOLO, which is specifically designed to address the challenges associated with marine target detection. These challenges include environmental complexity, performance issues, and computational burden. The model incorporates a unique hybrid encoder, FSI, which integrates dynamic serpentine convolution and Shape-IOU loss techniques to significantly improve the accuracy and efficiency of target detection. Firstly, a novel hybrid encoder FSI is designed to effectively handle multiscale features, enabling the model to efficiently detect and recognize objects in images. Secondly, the dynamic snake convolution is selected as a replacement for traditional convolution, allowing adaptive adjustment of anchor frames for recognizing different targets. Additionally,

the C3 module is enhanced based on the DySnake convolution to perform adaptive feature sequence fusion. Finally, Shape-IOU Loss is adopted as the loss function for the FSI-YOLO model, which not only enhances accuracy but also improves detection efficiency. Experimental results on the Infrared Maritime Ship dataset demonstrate that the FSI-YOLO model outperforms YOLOv5 in terms of target detection with an accuracy rate of 93.8%. The FSI-YOLO model employs a hybrid encoder to manage multiscale features, thereby enhancing target detection in complex environments. Nevertheless, it is essential to enhance the diversity of training and validation data by utilizing a range of datasets to improve the model's generalization ability and achieve highly accurate detection in extreme and variable marine settings. While FSI-YOLO demonstrated satisfactory performance in experimental scenarios, its real-time operational capabilities on constrained devices at the ocean's edge may still be limited. Further research should focus on improving the computational efficiency of dynamic convolution or exploring alternative approaches with reduced computational requirements to ensure real-time performance in practical applications and facilitate broader deployment across diverse fields. It is anticipated that continuous improvement and optimization will enable FSI-YOLO to achieve greater accuracy and efficiency in target detection in complex environments, thereby providing robust support for maritime safety monitoring and other related sectors.

References

[1] Aloysius, Neena, and M. Geetha, "A review on deep convolutional neural networks," in *Proc. of 2017 international conference on communication and signal processing (ICCSP)*, pp.588-592, 2017. Article(CrossRefLink)

[2] Cai, Zhaowei, and Nuno Vasconcelos, "Cascade R-CNN: Delving Into High Quality Object Detection," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6154-6162, 2018. Article(CrossRefLink)

[3] Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal Loss for Dense Object Detection," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp.2999-3007, 2017. Article(CrossRefLink)

[4] Tian, Zhi et al., "FCOS: Fully Convolutional One-Stage Object Detection," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.9626-9635, Oct. 2019. Article(CrossRefLink)

[5] Jiao, Licheng et al., "A Survey of Deep Learning-Based Object Detection," *IEEE Access*, vol.7, pp.128837-128868, 2019. Article(CrossRefLink)

[6] Xu, Shangliang et al., "PP-YOLOE: An evolved version of YOLO," *arXiv preprint arXiv:2203.16250*, 2022. Article(CrossRefLink)

[7] Redmon, Joseph, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.779-788, 2016. Article(CrossRefLink)

[8] Liu, Wei, Dragomir Anguelov, Dumitru Erhan et al., "SSD: Single Shot MultiBox Detector," in *Proc. of Computer Vision–ECCV 2016: 14th European Conference*, Lecture Notes in Computer Science, vol.9905, Springer, pp.21-37, 2016. Article(CrossRefLink)

[9] Redmon, Joseph, and Ali Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7263-7271, 2017. Article(CrossRefLink)

[10] Redmon, Joseph, and Ali Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018. Article(CrossRefLink)

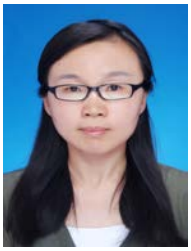
[11] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv:2004.10934*, 2020. Article(CrossRefLink)

[12] Jocher, Glenn et al., ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Zenodo, 2022. Article(CrossRefLink)
[13] Chuyi, Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie et al., "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications," *arXiv preprint arXiv:2209.02976*, 2022. Article(CrossRefLink)
[14] Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.7464-7475, 2023. Article(CrossRefLink)
[15] Jocher, G., Chaurasia, A., & Qiu, J, Ultralytics YOLO (Version 8.0.0), 2023. [Computer software] https://github.com/ultralytics/ultralytics
[16] Wang, Chien-Yao, I-Hau Yeh, Hong-Yuan Mark Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *arXiv preprint arXiv:2402.13616*, 2024. Article(CrossRefLink)
[17] Wang, Ao, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, Guiguang Ding, "YOLOv10: Real-Time End-to-End Object Detection," *arXiv preprint arXiv:2405.14458*, 2024. Article(CrossRefLink)
[18] Girshick, Ross et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proc. of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.580-587, 2014. Article(CrossRefLink)
[19] He, Kaiming et al., "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol.37, no.9, pp.1904-1916, 2015. Article(CrossRefLink)
[20] Girshick, Ross, "Fast R-CNN," in *Proc. of 2015 IEEE International Conference on Computer Vision (ICCV)*, pp.1440-1448, 2015. Article(CrossRefLink)
[21] Ren, Shaoqing, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.6, pp.1137-1149, 2017. Article(CrossRefLink)
[22] Vaswani, Ashish, Noam Shazeer, Niki Parmar et al., "Attention Is All You Need," in *Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017. Article(CrossRefLink)
[23] Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.42, no.2, pp.318-327, 2020. Article(CrossRefLink)
[24] Carion, Nicolas, Francisco Massa, Gabriel Synnaeve et al., "End-to-End Object Detection with Transformers," in *Proc. of 16th European Conference on Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, vol.12346, pp.213-229, Springer, 2020. Article(CrossRefLink)
[25] Chen, Qiang, Xiaokang Chen, Jian Wang et al., "Group DETR: Fast DETR Training with Group-Wise One-to-Many Assignment," in *Proc. of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.6610-6619, 2023. Article(CrossRefLink)
[26] Liu, Shilong, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang, "DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR," *arXiv preprint arXiv:2201.12329*, 2022. Article(CrossRefLink)
[27] Li, Feng et al., "DN-DETR: Accelerate DETR Training by Introducing Query DeNoising," in *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.13609-13617, 2022. Article(CrossRefLink)
[28] Sun, Peize et al., "Sparse R-CNN: End-to-End Object Detection with Learnable Proposals," in *Proc. of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.14449-14458, 2021. Article(CrossRefLink)
[29] Huang, Yueming, and Guowu Yuan, "AD-DETR: DETR with asymmetrical relation and decoupled attention in crowded scenes," *Mathematical Biosciences and Engineering*, vol.20, no.8, pp.14158-14179, 2023. Article(CrossRefLink)
[30] Zhao, Yian et al., "DETRs Beat YOLOs on Real-time Object Detection," *arXiv preprint arXiv:2304.08069*, 2023. Article(CrossRefLink)

[31] Zhang, Hao, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum, "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection," in *Proc. of ICLR 2023*, 2023. Article(CrossRefLink)
[32] Qi, Yaolei, Yuting He, Xiaoming Qi, Yuan Zhang, Guanyu Yang, "Dynamic Snake Convolution based on Topological Geometric Constraints for Tubular Structure Segmentation," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp.6070-6079, 2023. Article(CrossRefLink)
[33] Zhang, Hao, and Shuaijie Zhang, "Shape-IoU: More Accurate Metric considering Bounding Box Shape and Scale," *arXiv preprint arXiv:2312.17663*, 2023. Article(CrossRefLink)
[34] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick et al., "Feature Pyramid Networks for Object Detection," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.936-944, 2017. Article(CrossRefLink)
[35] Ghiasi, Golnaz, Tsung-Yi Lin, and Quoc V. Le, "NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection," in *Proc. of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7036-7045, 2019. Article(CrossRefLink)
[36] Tan, Mingxing, Ruoming Pang, and Quoc V. Le, "Efficientdet: Scalable and Efficient Object Detection," in *Proc. of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.10781-10790, 2020. Article(CrossRefLink)
[37] Liu, Songtao, Di Huang, Yunhong Wang, "Learning Spatial Fusion for Single-Shot Object Detection," *arXiv preprint arXiv:1911.09516*, 2019. Article(CrossRefLink)
[38] Zhao, Qijie, Tao Sheng, Yongtao Wang et al., "M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network," in *Proc. of AAAI Conference on Artificial Intelligence*, vol.33, no.01, pp.9259-9266, 2019. Article(CrossRefLink)
[39] Liu, Shu et al., "Path Aggregation Network for Instance Segmentation," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8759-8768, 2018. Article(CrossRefLink)
[40] Li, Chunliu,Wang Shuigen, Yantai Arrow Optoelectronic Technology Co., Ltd. in Shandong Province, China. http://openai.raytrontek.com/apply/Sea_shipping.html/

Chen Gao is pursuing his master's degree in the Department of Mathematics and Artificial Intelligence, at Qilu University of Technology (Shandong Academic of Science), Jinan, China. He received his Bachelor's degree from Linyi University, Linyi, China, in 2022. His research interests include computer vision and artificial intelligence.

Jiyong Xu is an associate researcher at Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Science). He received his master's degree from Chongqing University in 2007. His research interests include artificial intelligence, smart cities, Internet health care, and smart health care for the elderly.



Ruixia Liu (Member, IEEE) received a degree from the Shanxi University of Science and Technology, Xi'an, China, in 2004, and the Ph. degree in information science and engineering from the Shandong University of Science and Technology, Qingdao, China, in 2017. She has been an Associate researcher with the Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China. Her main research interests include medical artificial intelligence and image processing.