YOLOv7-marine: An Improved YOLOv7 Model for Object Detection in Marine Environments

1st Puhui QU
Key Laboratory of Underwater Acoustic Communication and Marine Information Technology
Xiamen University
Xiamen, China
qupuhui@163.com

2rd Keyu CHEN*
Key Laboratory of Underwater Acoustic Communication and Marine Information Technology
Xiamen University
Xiamen, China
chenkeyu@xmu.edu.cn

3nd En CHENG
Key Laboratory of Underwater Acoustic Communication and Marine Information Technology
Xiamen University
Xiamen,China
chengen@xmu.edu.cn

***Abstract*—In this paper, we propose a novel target detection algorithm that addresses the challenge of difficult recognition and localization in sea surface general purpose target detection. The proposed algorithm is based on an improved YOLOv7, incorporating an efficient non-parametric attention mechanism module-SimAM into the original network, which reduces the model parameters and enhances the expressiveness of the network as well as the extraction ability of the model for important features. Additionally, we introduce a new module, CN-CSP, that merges the strengths of CSP and ConvNext, thereby improving the network's learning ability while reducing the computational overhead. Furthermore, the integration of the rssp module into the backbone of YOLOv7 enables the network to extract features in a more comprehensive and multi-scale manner. Experimental results on The Sea Surface Target Dataset indicate the superiority of the proposed algorithm, achieving detection accuracy of 78.3% with improvements of 3.1% compared to the original YOLOv7 model.**

***Index Terms*—Yolov7; ConvNext; marine target detection; attention mechanism;**

I. Introduction

The ocean holds a significant position in the global economic growth, and its exploration and utilization of marine resources is vital. Recently, deep learning techniques have gained considerable attention and have been increasingly used in various real-world applications, including object detection, video surveillance, autonomous driving, and face recognition. These deep learning-based target detection algorithms demonstrate better results than traditional methods and are characterized by faster detection and higher accuracy. Target detection algorithms can be classified into two categories: two-stage and one-stage target detection algorithms. Two-stage algorithms such as SPP-Net [1], Faster R-CNN [2], and R-FCN [3], first generate the region proposal and then perform detection. Conversely, one-stage algorithms such as SSD [4]and YOLO [5]–[10] series algorithms directly obtain the location and class information of the target and exhibit faster detection. The major contributions of this work can be summarized as follows:

DOI reference number: 10.18293/SEKE23-042.
This work was supported in part by the National Natural Science Foundation of China (62071402 and 62271425).

(1). A CN-CSP architecture is proposed, which is derived from the combination of ConvNext [11] and CSPnet [12]. This architecture not only enhances the learning capability of CNN, but also significantly improves the target detection performance.

(2). we have optimized the MPConv module in YOLOv7 [10] by incorporating the SimAM [13] module to create the Sim_MPConv module. The Sim_MPConv module effectively suppresses irrelevant information without adding any additional parameters, leading to improved performance without increasing computational complexity. It also enhances detection accuracy and recognition of small targets on the sea surface.

(3). A novel RSPP module is introduced, which integrates the residual and SPP [1] structures. This module effectively extracts rich feature information from the input image, enabling a deeper network, and results in improved detection accuracy.

II. Architecture

A. The Overview of the YOLOv7 algorithm

The YOLOv7 algorithm [10] optimizes the balance between detection accuracy and efficiency through innovative strategies. Specifically, it integrates the extended efficient long-range attention network (ELAN), model scaling using cascaded models [14], and convolutional reparameterization [15]. The YOLOv7 network comprises three modules: Backbone, Neck, and Prediction. The Backbone module contains ELAN and MPConv convolutional layers. The ELAN layer increases feature diversity by directing different feature groups to enhance learning ability without compromising gradient paths. The MPConv module has a convolutional layer and Maxpool layer, forming two branches. Their features are combined by Cat to improve feature extraction.The Neck module uses a Path Aggregation Feature Pyramid Network (PAFPN) [14] to fuse features from different levels through bottom-up paths, enabling smoother information transfer from lower to higher levels. The Prediction module adjusts the channel numbers for P3, P4, and P5 features from the PAFPN using RepVGG Blocks [15]. Finally, 1×1 convolution predicts confidence, category, and anchor boxes.

B. Integration of Efficient 3D Attention Module

The attention mechanism plays an important role in facilitating the effective identification of key regions in complex visual scenes. As shown in Eqs(1) to (3) and the left part of Figure 1 , SimAM [13] module evaluates each neuron in each network by defining a linear differentiability energy function, where t is the target neuron, x is the neighboring neuron, and λ is the hyperparameter. e lower energy indicates that the neuron is more differentiated from its neighbors, and the neuron is more important.The neurons are weighted according to importance by $1/E$ as shown in equation (4).

$$e_t^* = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{1}$$

$$\dot{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M} (x_i - \hat{\mu})^2 \tag{3}$$

$$\tilde{X} = \text{sigmoid}\left(\frac{1}{E}\right) \odot X \tag{4}$$

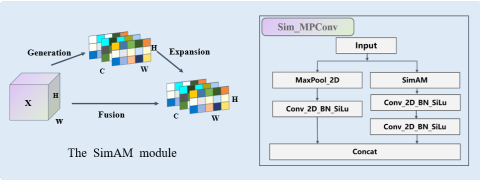

Fig. 1. The SimAM module and the Sim_MPConv structure.

In our experiments, we integrated the SimAM attention mechanism with MPConv to formulate the Sim_MPConv module, as illustrated in Figure 1 on the right. The Sim_MPConv module enhances the contribution of neurons that convey more relevant information and effectively mitigates the impact of irrelevant features, thereby strengthening the network's feature representation capability and enhancing the model's target localization accuracy while reducing the influence of background interference.

C. Efficient CN-CSP Structure

The ConvNext-Block structure originates from the ConvNext [11] architecture, and we propose a novel integration of the ConvNext-Block with the CSP structure to form the ConvNext-Block-CSP (CN-CSP), which is depicted in figure 2.

The CN-CSP structure employs the Layer Normalization (LN) layer, which stabilizes the model and reduces the oscillation of gradients during training. the structure incorporates the Gaussian Error Linear Unit (GELU) [16]function, which not only overcomes the gradient vanishing issue, but also accelerates the training speed compared to the traditional sigmoid function. And this module effectively enhances the performance of the network while also optimizing the utilization of each computing unit and reducing extraneous resource consumption.

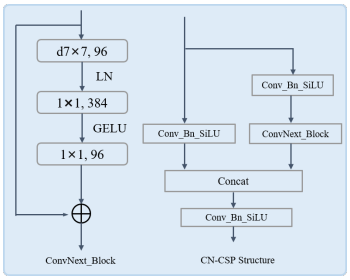

Fig. 2. The ConvNext-Block and the structure of CN-CSP.

D. The RSPP Structure

The residual edges present in the residual [17] structure,are pivotal in enabling the network to learn nonlinear representations and accelerate the training process. Moreover, the spatial pyramid pooling layer, as shown in figure 3 on the left part, effectively captures both global and spatial information of the feature map, significantly enhancing the network's generalization ability. Through the integration of these two components and additional modifications, the Residual Spatial Pyramid Pooling (RSPP) structure is formed, as depicted in figure 3 on the right. The RSPP structure enables the extraction of features more efficiently, leading to improved prediction performance without sacrificing computational efficiency.

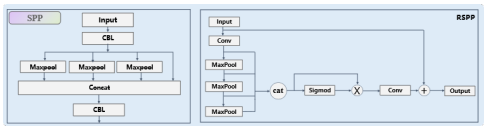

Fig. 3. The SPP Module and the RSPP structure.

E. The Yolov7-marine Architecture

As depicted in figure 4, the improved network architecture of YOLOv7 is presented. By incorporating the RSPP structure into the backbone of YOLOv7, the deep feature extraction capability of the network is improved. The MpConv structure in YOLOv7 is then advanced by adding the 3D attention mechanism, SimAM, to form Sim-MPConv, which enhances the feature extraction capability. Finally, by integrating the CN-CSP structure into the neck of YOLOv7, the learning ability of the CNN is boosted while making the network more efficient.

III. Experiments

In order to evaluate the performance of the improved YOLOv7 algorithm, we trained and evaluated the algorithm on a commonly used dataset of The Sea Surface Targets.

A. Experimental Setup

We use Common target dataset on The Sea Surface Target Dataset to conduct experiments and validate our object detection method. All our experiments did not use pre-trained models. That is, all models were trained from scratch.

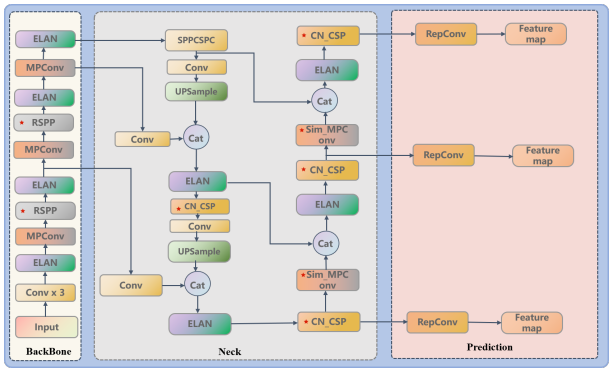

Fig. 4. Improved network model based on Yolov7 network.

The Sea Surface Target Dataset, consisting of 7150 images, encompasses ten categorical classes, specifically lighthouse, sailboat, buoy, railbar, cargoship, navalvessels, passengership, dock, submarine, and fishingboat. The dataset is partitioned into training set, validation set, and test set, utilizing a ratio of 7:2:1, respectively.

In order to rigorously evaluate the performance of the proposed method, a series of experiments were carried out with the following parameter settings. The input images were preprocessed by resizing to a resolution of 640x640 pixels. The optimization algorithm employed in the experiments was Stochastic Gradient Descent (SGD), with a learning rate of 1e-2 was applied to the model via a weight decay of 5e-4, and the learning rate was adapted via the Cosine Annealing schedule. The batch size for training was set to 20, with a total of 300 training epochs being performed.The software environment for the experiment is: operating system Linux , Python 3.10, PyTorch 1.11.0, CUDA 11.5.2, GPU: RTX 3090.

TABLE I
ABLATION EXPERIMENTS ON THE SEA SURFACE TARGET DATASET

Method	RSPP	Sim_MPConv	CN-CSP	mAP0.5(%)	Patams(M)	FLOPS(G)
G1	×	×	×	75.20	36.90	104.70
G2	×	✓	×	76.30	36.90	104.70
G3	×	×	✓	76.60	38.72	108.10
G4	×	✓	✓	76.90	38.72	108.10
G5	✓	×	×	77.10	37.58	107.00
G6	✓	✓	✓	78.30	39.05	109.80

B. Ablation Experiments

In order to evaluate the contribution of the proposed improvements to the overall performance of the model, we conducted a set of ablation experiments on the Sea Surface Target Dataset. The results of these experiments are summarized in Table 1, which indicates the utilization of the RSPP, CN-CSP, and SimAM methods. These ablation experiments were able to evaluate each of the improvements and allowed us to assess their efficacy in improving model performance.

As illustrated in Table 1, the performance evaluation of the YOLOv7 algorithm was carried out in several experiments. The first experiment showed a detection accuracy of 75.2% with the original YOLOv7 algorithm. The second experiment aimed to enhance the detection accuracy by incorporating the SimAM module into the YOLOv7 algorithm. This integration resulted in an improvement of 1.1% in the detection accuracy, without increasing the number of model parameters.The third experiment focused on integrating the CN-CSP structure into the YOLOv7 algorithm, which resulted in an improvement of 1.4% in the detection accuracy, with a small increase in the computational cost of the model. The fourth experiment combined the structures from the second and third experiments, resulting in a further improvement of 1.7% in the detection accuracy.The fifth experiment added the RSPP structure to the YOLOv7 algorithm, which resulted in an improvement of 1.9% in the detection accuracy. The final experiment incorporated the RSPP structure into the fourth experiment,resulting in a substantial improvement of 3.1% in the detection accuracy.

In conclusion, these experiments demonstrate the effectiveness of incorporating different structures into the YOLOv7 algorithm, in improving its detection accuracy.

C. Analysis

In this study, we conduct a comprehensive evaluation of the improved YOLOv7 algorithm in comparison to the current mainstream target detection algorithms on The Sea Surface Target Dataset. The experimental results are presented in Tables 2.The results reveal the superiority of the improved YOLOv7 algorithm in terms of accuracy compared to the existing methods.

As demonstrated in Table 2, the improved YOLOv7 model achieved an average accuracy of 78.3% on The Sea Surface Target Dataset, outperforming YOLOV5s (6.0) and YOLOX-s by 7.1% and 6.3%, respectively, in terms of accuracy. In addition, the improved YOLOv7 model showed a remarkable improvement in detection accuracy with regards to map0.5 and map0.5-0.95, with an increase of 3.1% and 1.5% over the original YOLOv7, respectively.

Figure 5 illustrates a comprehensive performance evaluation of various algorithms on The Sea Surface Target Dataset. The first row displays the detection of cargo ships under foggy conditions. The results reveal that the improved YOLOv7 algorithm has an accuracy of 87%, which is 5% higher than YOLOv7. The second row showcases the success of the improved YOLOv7 algorithm in identifying small targets on the sea surface that were previously missed by YOLOv7. Moreover, the improved YOLOv7 network demonstrates greater accuracy than YOLOv7 for identifying other targets. The third row demonstrates that the overall accuracy of the improved YOLOv7 model is higher than that of the YOLOv7 network under normal conditions.

IV. CONCLUSION

In this paper, we address the challenges in surface target detection by improving the YOLOv7 network architecture. The proposed improvements aim to tackle the difficulties in detecting surface targets and coping with complex surface

TABLE II
COMPARISON OF DETECTION ACCURACY OF DIFFERENT TARGET DETECTION ALGORITHMS ON THE SEA SURFACE TARGET DATASET

Methods	lighthouse	sailboat	buoy	railbar	cargoship	navalvessels	passengership	dock	submarine	fishingbooat	mAP@0.5(%)	Map@0.5-0.95(%)
Retinanet	60.34	86.62	95.60	39.32	50.25	94.20	82.41	65.41	58.64	35.48	66.83	37.30
YOLOv5s(6.1)	68.10	80.00	92.30	63.50	42.10	91.30	89.30	74.20	73.50	38.20	71.20	39.90
YOLOX-s	69.30	79.30	92.60	63.80	45.80	91.50	89.90	75.60	73.40	38.80	72.00	40.60
YOLOv7	72.60	88.70	94.40	63.80	48.60	94.80	91.30	78.60	81.60	37.10	75.20	44.10
YOLOv7-marine	**86.80**	86.20	**95.20**	**65.20**	**54.50**	**96.40**	91.00	**82.00**	81.40	**44.30**	**78.30**	**47.80**

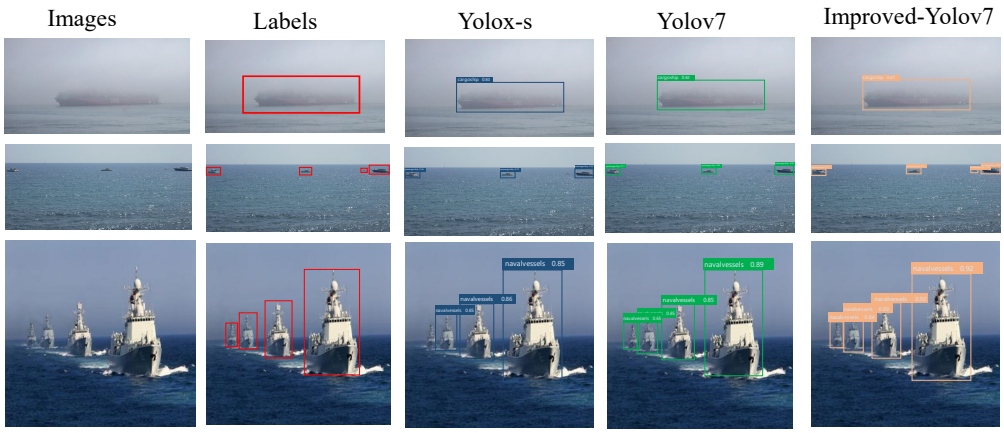

Fig. 5. Performance of different algorithms on The Sea Surface Target Dataset.

environments. The network structure is enhanced by incorporating the MPConv module from the original architecture with an efficient 3D attention mechanism to form the Sim_MPConv structure, and by adding the CN-CSP structure to the YOLOv7 neck, which enhances feature extraction capabilities. Additionally, the RSPP structure is utilized in the YOLOv7 backbone to significantly improve network prediction accuracy. Experimental results demonstrate that the proposed method effectively improves the accuracy of sea surface target detection without significantly increasing the number of model parameters or computational effort. The effectiveness of the approach has been validated on both the Sea Surface Target Dataset indicating its general applicability and potential for practical use. Further improvement in accuracy for sea surface target detection is expected in future studies.

REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

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

[3] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016.

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

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

[6] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[7] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[8] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[9] Glenn Jocher et al. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, February 2022.

[10] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[11] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[12] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020.

[13] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*, pages 11863–11874. PMLR, 2021.

[14] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 13029–13038, 2021.

[15] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021.

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.