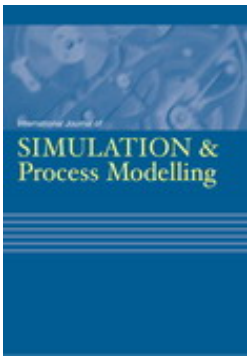


International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123
https://www.inderscience.com/ijspm

Improving maritime distress target detection through modelling and simulation with YOLOv5s and Next-ViT

Xinbo Chang, Kun Liu, Zhen Liu

DOI: 10.1504/IJSPM.2025.10070958

Article History:

Received:	22 November 2024
Last revised:	28 February 2025
Accepted:	31 March 2025
Published online:	01 September 2025

Improving maritime distress target detection through modelling and simulation with YOLOv5s and Next-ViT

Xinbo Chang

School of Automation,
Qingdao University,
Qingdao, Shandong, China
Email: changxinbo@qdu.edu.cn
Email: xbchang1211@hotmail.com

Kun Liu*

Qingdao Campus,
PLA Naval Aviation University,
Qingdao, Shandong, China
Email: liukun6606@stu.ouc.edu.cn
*Corresponding authors

Zhen Liu

School of Automation,
Qingdao University,
Qingdao, Shandong, China
and
Shandong Key Laboratory of Industrial Control Technology,
Qingdao University,
Qingdao, Shandong, China
Email: zhenliuzz@hotmail.com

Abstract: To additionally raise the accuracy of maritime distress target detection, an improved YOLOv5s model with Next-ViT is proposed through modelling and simulation. In this model, Next-ViT is applied to extract representations, followed by adopting a neck network with spatial context pyramid and Focal-GIoU loss to identify the targets. To validate its effectiveness, extensive experiments are conducted on the sub-dataset of the SeaDronesSee dataset. Contrasted to the primary YOLOv5s model, the proposed model has promoted $recall$, $mAP_{0.5}$ and $mAP_{0.5-0.95}$ by 9.2, 6.3 and 3.3 percentage points, respectively, demonstrating superior performance over existing models.

Keywords: YOLOv5s model; target detection; Next-ViT; spatial context pyramid.

Reference to this paper should be made as follows: Chang, X., Liu, K. and Liu, Z. (2025) 'Improving maritime distress target detection through modelling and simulation with YOLOv5s and Next-ViT', *Int. J. Simulation and Process Modelling*, Vol. 22, Nos. 1/2, pp.18–28.

Biographical notes: Xinbo Chang received his BE in Measurement and Control Technology and Instrumentation from Qilu University of Technology in 2023. Since 2023, he is pursuing his Master's in Control Engineering at the School of Automation, Qingdao University. His research interests include artificial intelligence and Intelligent control technology.

Kun Liu received his MS in Computer Technology from Ocean University of China in 2013. From 2013 to 2021, he worked as a Senior Computer Software Engineer at Shandong Airlines. Since 2021, he worked as an Engineer at Naval Aviation University. His research interests include computer control and artificial intelligence.

Zhen Liu received his PhD in Control Theory and Applications from Ocean University of China, Qingdao, China in 2017. He was a Joint PhD candidate at the School of Engineering, University of the West of England, UK and a Visiting Scholar at the College of Engineering, University of Kentucky, USA. He is currently a Distinguished Professor at the School of Automation, Qingdao University, China. His current research interests include intelligent control and robot, UAV control and machine vision, and cyber-physical systems. He was the recipient of the Shandong Province Taishan Scholar Special Project Fund.

1 Introduction

With the advancement in depth of ocean programs, offshore human activities are becoming frequent, and a growing number of people are enthusiastic about recreational activities such as swimming, surfing and motorboats at sea. To ensure the safety of human life and property in marine activities, search and rescue tasks for maritime distress targets are becoming more important. However, conventional oceanic search and rescue methods (Geng et al., 2025; Li et al., 2024) suffer from inefficiency and high risks, which prevent them from detecting maritime hazards in a timely manner. As artificial intelligence technologies advanced, and various unmanned intelligent devices were widely used in the search and rescue of maritime distress targets to solve the aforementioned troubles. The auxiliary function of unmanned aerial vehicles (UAV) in maritime search and rescue is studied, where an algorithm for determining target positions is proposed by Ghazali et al. (2021). Combining UAVs, aerial protection base stations, and control terminals, a marine urgent assistance system that utilised UAVs and landing surfaces is suggested by Chen et al. (2021). In recent years, target detection for UAVs have become a major direction within the domain of computer vision. A fully convolutional network was applied to aerial image detection by Han et al. (2019), and a modified version of the U-net and Google Inception v4 net detection algorithms was proposed. With the introduction of two-stage object detection algorithms based on region-based convolutional neural networks (R-CNN) (Girshick et al., 2014), Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2017), the area of object detection started to link with deep learning. Simultaneously, one-stage algorithms based on you only look once (YOLO) (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020; Khalfaoui et al., 2022) and single shot multibox detector (SSD) (Liu et al., 2016) had emerged with the demands of the times, which effectively compensated the problem of the long-time consumption of two-stage algorithms. The widely-used YOLO models also face certain challenges. The YOLOv5 series algorithm has been extensively used in target detection after multiple improvements and updates (Wang et al., 2024; Song et al., 2024; Wang et al., 2025). It is favoured for its convenience, accuracy and speed. However, when it comes to detecting sea targets from the UAV's perspective, the inconsistent size and angle of these targets lead to less-than-ideal detection results. In order to solve the aforementioned problems, an improved YOLOv5 model with next generation vision transformer (Next-ViT) is proposed in this paper through modelling and simulation.

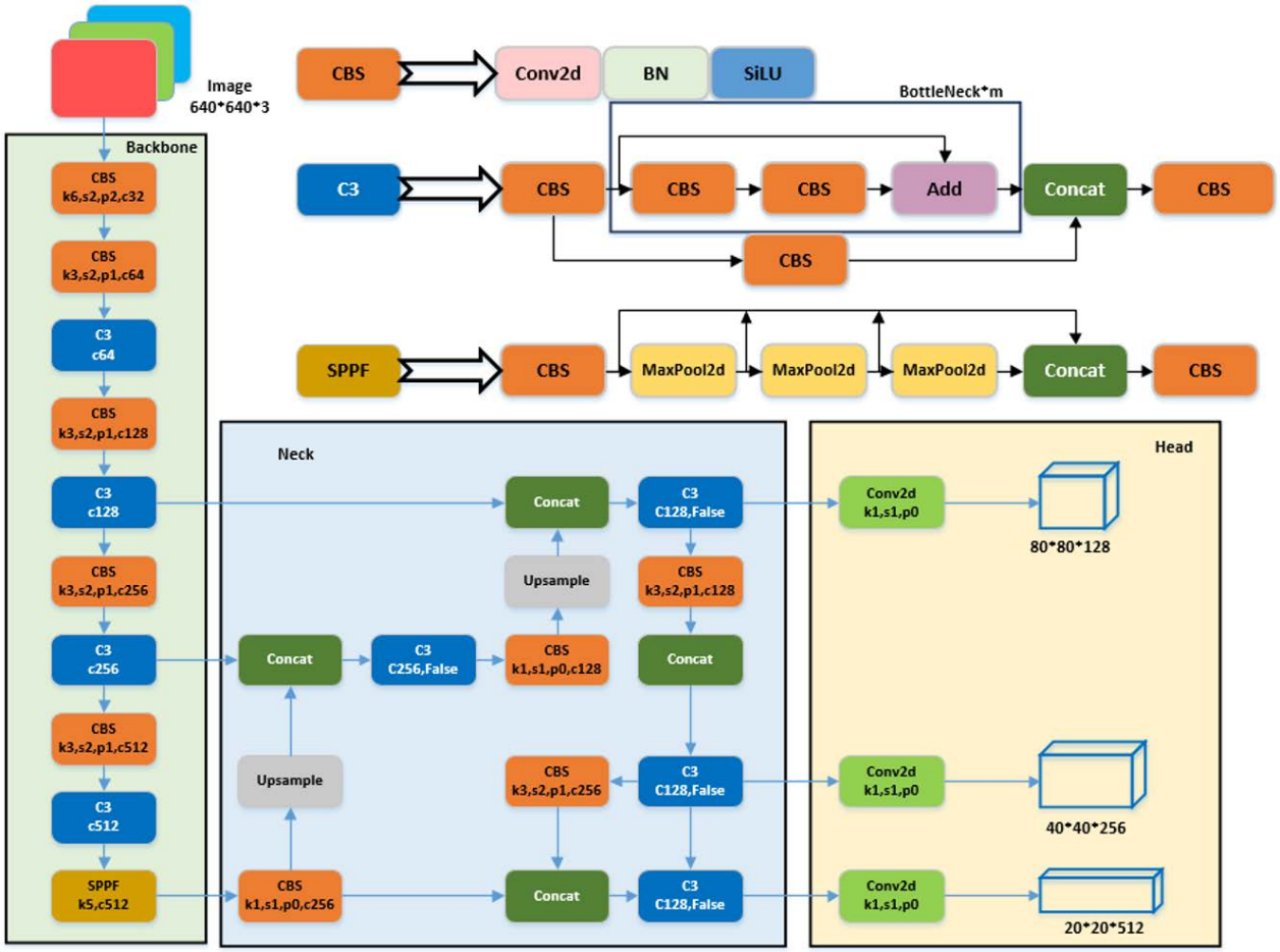
The original model of YOLOv5 is divided into different sizes according to the difference in depth and width of the network, and these sizes are delimited into YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x based on the rise in model volume. As the size of the model magnifies, the detection accuracy becomes more precise, and the network gets more flexible. The YOLOv5s network model which is relatively small and requires low hardware equipment is selected as the baseline for modelling and simulation. It enables fast inspection speeds while maintaining high inspection accuracy. Next-ViT is combined with spatial context pyramid (SCP) (Liu et al., 2024) and injected into the proposed innovative YOLOv5s model through repeatedly simulation experiments. What is more, focal generalised intersection over union (Focal-GIoU) is proposed to satisfy the ameliorated model and tasks for maritime distress targets.

The main innovations include:

- Unlike Zhang et al. (2024) and Bai et al. (2022), Next-ViT has been integrated into YOLOv5s to improve feature extraction. The conventional backbone of the YOLOv5s based on convolutional neural network (CNN) (Girshick et al., 2014) has been replaced with the transformer-based Next-ViT model. Compared to traditional CNN networks, Next-ViT utilises a self-attention mechanism to capture global contextual information, enhancing the model's perceptual ability. Additionally, a hierarchical structure and multi-scale feature fusion have been adopted to process features more effectively at different scales, improving detection accuracy for maritime distress objects.
- Compared to Liu et al. (2024) and Li et al. (2022), the SCP has been introduced into the neck network to aggregate global features. Since images exhibit long-range correlations that provide complementary information for detecting ambiguous objects, the incorporation of SCP allows global features to be effectively aggregated and integrated into each pixel. This enhancement strengthens the ability of the model to recognise objects in complex maritime environments.
- Focal-GIoU loss has been proposed to improve bounding box regression. The original complete intersection over union (CIoU) loss (Zheng et al., 2022) used in YOLOv5s has been replaced with Focal-GIoU loss, which combines FocalL1 loss (Zhang et al., 2022) and GIoU loss (Rezatoffghi et al., 2024). This novel loss function balances the optimisation contributions of high and low quality samples, leading to more robust and accurate object localisation.

The structure of the paper is as follows. In Section 2, the YOLOv5s network architecture is deeply explored, with its components, working principles, and significance in object detection being explained in detail. In Section 3, the improved YOLOv5s-NextSF model, which combines YOLOv5s-NextViT with SCP and Focal-GIoU, is introduced. In Section 4, the simulation experiments and result analysis are presented. The experimental setup, including dataset selection, evaluation metrics and parameter details, is provided, and valuable reference for subsequent research is offered. The conclusions are presented in Section 5.

Figure 1 Overall network structure block diagram of YOLOv5s (see online version for colours)



2 YOLOv5s network architecture

Mosaic data augmentation and adaptive anchor box calculation were used to innovate the focus structure. And two kinds of cross stage partial (CSP) (Wang et al., 2020) structures for backbone and neck networks were designed to enhance feature fusion capabilities. The feature pyramid networks (FPN) (Lin et al., 2017) and path aggregation network (PAN) (Liu et al., 2018) pyramid structure was adopted to handle targets of different scales better. In addition, the CIoU loss function was used by YOLOv5 to raise network detection accuracy and convergence stability.

The YOLOv5s network can be segmented into the following sections: input, backbone, neck and head layers. The overall pattern is shown in Figure 1.

2.1 Input

The input end is applied to process the input images, mainly comprising mosaic data augmentation, adaptive anchor box calculation, and adaptive image scaling. The effect of enriching the dataset is achieved by mosaic data augmentation by concatenating four random pictures through operations such as flipping, scaling, and number field changes. Adaptive anchor box calculation is foremost to preset a border, and during training, the optimal anchor box value is constructed based on the offset of the actual border position relative to the preset border. Letterbox adaptive image scaling technology is used by adaptive image scaling to prevent information loss during zooming, and images of various sizes are zoomed to identical measurements before inputting them into the network for better detection results.

2.2 Backbone

The function of the backbone network is to extract image properties and continuously shrink the feature map. The main modules include CSP bottleneck with 3 convolutions (C3) and spatial pyramid pooling fast (SPPF). The model's expressive power is improved by the C3 module, named after three CBS modules within it, through increasing the depth of convolutional layers. The CBS module can reduce the size of the feature map to one half, enabling downsampling of the feature map thereby extracting target features. The SPPF block is used to achieve the extraction of various features. Compared to the original spatial

pyramid pooling (SPP) module, the computation is diluted and the feature fusion speed is accelerated.

2.3 Neck

A feature pyramid FPN + PAN construction is adopted by the neck network, which combines shallow properties obtained from the backbone and transmits them to the head layer for detection. Wherein, upsampling is performed by the FPN layer to capture obvious semanteme characteristics, while downsampling is executed by the PAN layer from bottom to top to convey strong localisation characteristics.

2.4 Head

As the output end, the head network is responsible for predicting object detection based on the extracted properties. It includes bounding box loss function and non-maximum suppression (NMS). The problem of bbox non-coincidence is commendably solved by the operation of using CIoU loss, which extremely assists in improving the performance of prediction box regression. Meanwhile, weighted NMS is applied to boost the recognition capability of multi-objective and occluded purposes to identify the optimal object detection position (Liu et al., 2024).

3 Ameliorated YOLOv5s-NextSF network architecture

3.1 The replacement of the backbone network with Next-ViT

In the task of maritime distress target detection, the backbone network plays a crucial role in feature extraction, and its performance directly affects the detection effect of the entire model. The traditional backbone network of YOLOv5s on CNNs has certain limitations when dealing with complex and changeable maritime scenarios. To overcome these limitations and enhance the model's ability to capture the features of maritime targets, especially for targets with different scales and angles and those in complex backgrounds, the backbone network of the original model is replaced with Next-ViT, which contains three major components: next convolution block (NCB), next transformer block (NTB) and next hybrid strategy (NHS). A novel multi-head paradigm convolutional attention is used firstly by the Next-ViT model in NCB, which combines multi-head cross attention (MHCA) and multilayer perceptron (MLP) to form NCB through the metaformer architecture to obtain short-term dependency information. Secondly, after capturing local information occupying NCB, NTB is designed to learn global information. MHCA and efficient multi-head self-attention (E-MHSA) are used simultaneously in NTB to process multi-frequency information and transmit them to the MLP layer to extract more basic and obvious features, thereby achieving the effect of improving model performance. The normalisation of NCB and NTB servicing BatchNorm layer and ReLU activation function. Finally, a new hybrid strategy, the NHS, is used to heap up NCB and NTB operating the same $(n + 1) * \mathrm{L}$ hybrid paradigm, thus enhancing model capability in backward position tasks. A patch embedding layer is added to the input part of each stage for segmentation encoding. The overall pattern diagram of Next-ViT is shown in Figure 2.

Figure 2 The overall pattern diagram of Next-ViT (see online version for colours)

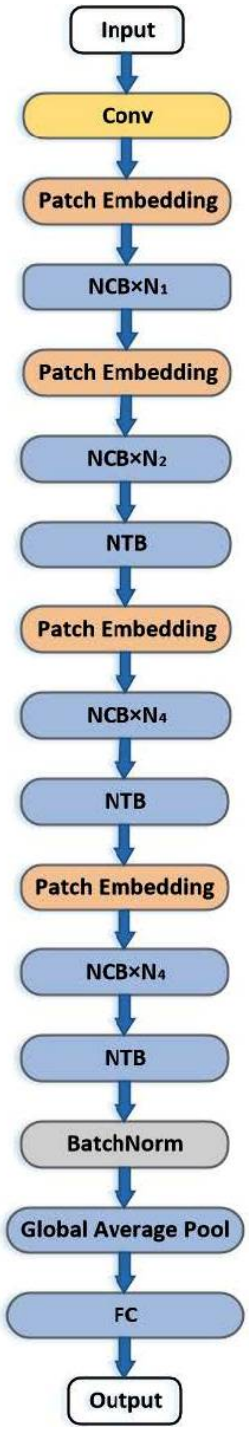
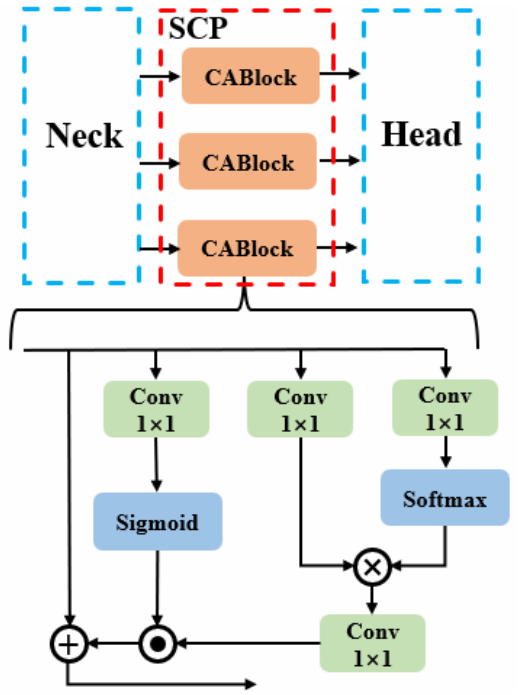


Figure 3 The introduced SCP structure and its CABlock module (see online version for colours)



3.2 Spatial context pyramid introduction in neck network

After the backbone network is replaced with Next-ViT, to further optimise the model's ability to recognise targets in complex maritime environments, the improvement of the neck network is also of great significance. The detection of maritime distress targets faces many challenges. For example, situations such as blurred targets and targets being similar to the background occur frequently. The long-range correlations in images contain rich complementary information, which can help solve these problems. In order to obtain the complementary information provided by the long-range correlation of images for fuzzy targets, SCP is led into the neck network to aggregate global features and combine them into each pixel. The core idea of SCP is that when the feature information of each pixel is large enough, there is no need to aggregate features in other spatial positions. The design ensures the difference between each pixel to enormously avoid information confusion. This module is placed after the neck network to connect the neck and head networks, and each layer consists of a context aggregation block (CABlock) and other joining parts. The context aggregation method for each pixel in CABlock is shown in equation (1):

$$O_i^j = P_i^j + a_i^j \cdot \sum_{j=1}^{X_i} \left[\frac{\exp(w_k P_i^j)}{\sum_{n=1}^{X_i} \exp(w_k P_i^n)} \cdot w_v P_i^j \right], \quad (1)$$

where the j^{th} and n^{th} pixel of the input feature maps of the i^{th} layer are indicated with P_i^j and P_i^n respectively. The j^{th} pixel of the output feature maps of the i^{th} layer are stood for O_i^j, where $j, n \in X_i$. The linear projection matrix is represented by w_k and w_v. Structurally analogous to P_i^j and O_i^j, a_i^j regulates context aggregation balance. The projection operator, generated by applying sigmoid activation to P_i^j, is mathematically formulated in equation (2):

$$a_i^j = \frac{1}{1 + \exp(-w_a P_i^j)}, \quad (2)$$

where the projection is implemented through w_a (structurally analogous to w_k and w_v).

The specific structure after bringing SCP into the neck network is shown in Figure 3.

3.3 Amelioration of loss function

The improvements to the backbone network and the neck network lay the foundation for enhancing the model's performance. However, the choice of loss function also has a crucial impact on the training effect and detection accuracy of the model. In the task of maritime distress target detection, the diversity and complexity of samples make it difficult to meet the requirements for traditional loss functions. The original CIoU loss function used in YOLOv5s has certain deficiencies in handling some scenarios and cannot fully balance the contributions of different-quality samples to model training. To optimise the training process of the model and improve the accuracy of target localisation, the loss function is ameliorated, and the Focal-GIoU loss function which is integrated with FocalL1 loss and GIoU loss is adopted to replace the original CIoU loss function.

3.3.1 GIoU loss

Based on IoU, injects the minimum rectangle that surrounds the predicted bounding box and the authentic bounding box by GIoU. The non-overlapping regions of the two bounding boxes are added to the loss function algorithm to solve the problem of non-allopatric objectives that cannot be optimised. The equation is as follows in equation (3):

$$GIoU = \frac{A \cap B}{A \cup B} - \frac{C - (A \cup B)}{C}, \quad (3)$$

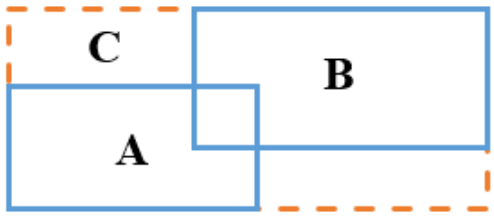
where A and B express the predicted box and the genuine box respectively, and C is the minimum rectangle that can surround the two bounding boxes. The schematic diagram of the relationship between A, B and C is shown in Figure 4.

The range of values for GIoU is $-1 \leq GIoU \leq 1$. When the predicted box and the real box completely overlap, obtain GIoU = 1, while the two bounding boxes do not intersect, come by IoU = 0. At this point, the farther the distance between the predicted box and the real box, the closer the GIoU tends to be. When GIoU is used as the loss function, equation (4) is satisfied as follows:

$$L_{GIoU} = 1 - GIoU, \quad (4)$$

and the closer the two bounding boxes are to the overlapping loss function, the more approached the value is to 0, which accords with the basic requirements of the loss function and can achieve optimisation of the model.

Figure 4 Schematic diagram of the relationship between predicted box A, true box B, and C (see online version for colours)



3.3.2 *FocalL1 loss*

The BBR loss function has the problem of imbalanced training samples, where high-quality samples with small regression errors account for a small proportion of the total sample. Based on this issue, FocalL1 loss is proposed by Zhang et al. (2022) to make high-quality samples occupy a more significant position. It meets the following conditions:

1 When the regression error approaches 0, the gradient amplitude also inclines to 0.

2 The gradient amplitude rapidly expands with the increase of regression error. When the regression error increases to a certain extent, the gradient amplitude gradually decreases.

3 A hyperparameter is existed to counterbalance the optimisation contribution of high and low quality samples to the loss function.

4 The gradient amplitude is normalised to a certain range to further balance high-quality and low-quality samples.

Based on the above conditions, β is introduced to suppress low-quality samples, but high-quality samples will also be inhibited. By introducing $\alpha = e\beta$ to normalise the gradient amplitude to [0, 1], high-quality samples can be promoted while low-quality samples are constricted. The designed family of gradient amplitude functions can be represented by equation (5):

$$g(x) = \frac{\partial L_f}{\partial x} = \begin{cases} -\alpha x \ln(\beta x), & 0 < x \leq 1; \\ -\alpha \ln(\beta), & x > 1, \end{cases} \tag{5}$$

where $\frac{1}{e} \leq \beta \leq 1$.

By integrating the above equation, FocalL1 loss that satisfies all four conditions simultaneously is obtained as shown in equation (6):

$$L_f(x) = \begin{cases} -\frac{\alpha x^2 (2\ln(\beta x) - 1)}{4}, 0 < x \leq 1; \\ -\alpha \ln(\beta) x + C, \ x > 1, \end{cases} \tag{6}$$

where $\frac{1}{e} \leq \beta \leq 1$.

3.3.3 *Focal-GIoU*

In order to focus the GIoU loss on high-quality samples, FocalL1 loss was combined with GIoU loss to form Focal-GIoU loss as shown in equation (7):

$$L_{Focal-GIoU} = IoU^{\gamma} L_{GIoU}, \tag{7}$$

in which the parameter that inhibits low-quality samples is expressed as γ, and the contact method has followed the approach of Focal-EIoU (Zhang et al., 2022).

4 Experiment and result analysis

4.1 *Experimental datasets*

The SeaDronesSee dataset is a large database published by Varga et al. from the University of Tuebingen in Germany for the analysis of ocean drone images, aimed at bridging the gap between land-based and sea-based visual systems. The dataset was published at the WACV conference in 2022 and mainly applies to object detection and tracking in computer vision algorithms. Over 54,000 frames of images are gathered and annotated, and 400,000 instances of drones from 5–260 metres and 0–90 degrees are obtained, providing numerous metadata information including height and perspective are provided by the dataset. The detection targets are divided into five categories: swimmer, boat, jet ski, life-saving-appliances, and buoy. Due to the high difficulty in collecting actual maritime distress rescue targets and the limited amount of data collected, while the fact that the categories in the SeaDronesSee dataset can effectively simulate the state of maritime distress targets, the selection of the SeaDronesSee dataset as the experimental datasets in this experiment is more in line with the facts. Due to the large size of the datasets, to accelerate the experimental progress and apply it to actual maritime rescue operations in a short period, 893 training set images and 155 validation set images were selected as sub-dataset of the SeaDronesSee dataset through stratified sampling. The validation of algorithm effectiveness is accelerated by using sub-dataset regarded as experimental datasets. The statistics of the sub-dataset are shown in Table 1.

Table 1 Dataset statistics

	Train	*Valid*
Number of images	893	155
Number of each class	swimmer: 3,666 boat: 1,308 jet ski: 228 life_saving_ appliances: 88 buoy: 434	swimmer: 624 boat: 234 jet ski: 34 life_saving_ appliances: 36 buoy: 59

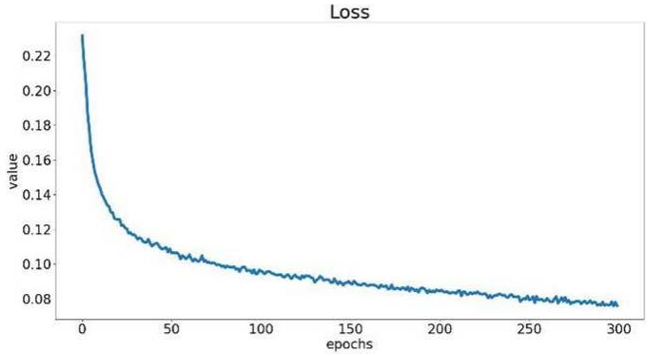
4.2 *Experimental environment and parameter settings*

The hardware environment for this tentative is CPU: 13th gen Intel(R) Core(TM) i7-13700KF (24 CPUs)

3.4 GHz, RAM: 64 GB. It is equipped with an NVIDIA GeForce RTX 4070 Ti GPU. The running environment is Python 3.9.18 under the Windows 11 operating system, and on this basis, a 2.1.2 version of the Python deep learning framework was built, which was expedited through CUDA11.8 and cuDNN8700.

The input image size is set to 640 × 640; the initial learning rate is 0.01; the learning rate momentum is 0.937; the weight attenuation is 0.0005; the optimiser is SGD; the batch size and workers are both 8. The early stop mechanism is adopted. In the experiment, the same dataset, environment, and parameters are employed to draw the loss function curve of the validation set, as shown in Figure 5. It is effortless to perceive that the model quickly fits in the first 100 epochs, and after 200 epochs, the loss value gradually converges. Therefore, training iterations are set to 300.

Figure 5 Validation set loss function curve (see online version for colours)



4.3 Model performance evaluation indicators

Precision (*P*), recall (*R*), and mean average precision (*mAP*) are selected as evaluation indicators. The precision is the proportion of correctly predicted samples that are predicted to be positive. The equation of precision is as follows in equation (8):

$$P = \frac{TP}{TP + FP}, \tag{8}$$

where *TP* stands for the number of specimens with positive predicted values and correct predictions, and *FP* shows the amount of samples with negative predicted values and correct predictions.

The ratio of correctly predicted positive samples to the total actual positive samples is called the recall rate. The expression of recall is as follows in equation (9):

$$R = \frac{TP}{TP + FN}, \tag{9}$$

in which *FN* expresses the number of samples with negative predicted values but incorrect predictions.

As is known, *mAP* is the sum of the average precision (*AP*) of a single category divided by the overall number of categories, which is a significant indicator for evaluating the detection effect. The equations of AP and mAP are as follows in equations (10) and (11):

$$AP = \int_0^1 P(R)\, dR, \tag{10}$$

$$mAP = \frac{\sum AP}{NC} \times 100\%, \tag{11}$$

where *NC* signals the total number of categories.

Moreover, $mAP_{0.5}$ refers to the average accuracy mean at an IoU threshold of 0.5, which is selected in this experiment to evaluate the detection accuracy of the model.

4.4 Results and analysis of ablation experiments

In order to verify the effectiveness of each improvement module in this experiment, seveb ablation experiments were conducted on a subset of the selected SeaDronesSee dataset. The experiment is conducted based on the YOLOv5s model. Firstly, the backbone network is replaced with Next-ViT, and then the spatial context pyramid (SCP) is added to the primitive feature pyramid (FPN + PAN) of the neck network. Finally, the initial CIoU loss function is replaced with the Focal-GIoU loss function. The adoption of this improvement method on the model is indicated by '+', and the experimental results are shown in Table 2.

After using Next-ViT as the backbone network of YOLOv5s, *recall*, $mAP_{0.5}$ and $mAP_{0.5-0.95}$ increased by 5.3, 5.2 and 2.4 percentage points separately. Based on YOLOv5s, the neck network incorporates a spatial context pyramid (SCP) composed of context aggregation to fuse global features. Although the value of $mAP_{0.5-0.95}$ slightly decreases, the values of *recall*, and $mAP_{0.5}$ increase by 1.6 and 0.2 percentage points, respectively.

Figure 6 Comparison of mAP0.5 between the YOLOv5s NextSF model and other versions of the YOLO series models (see online version for colours)

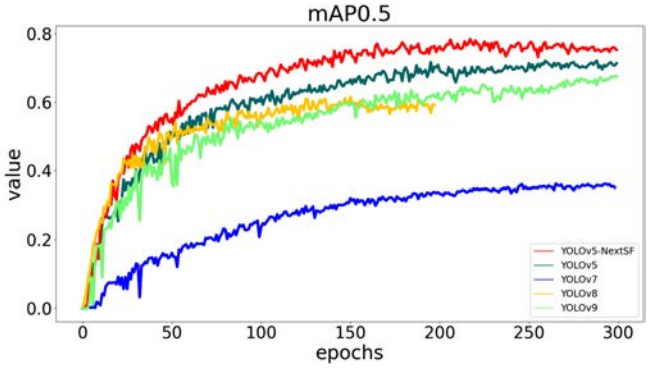
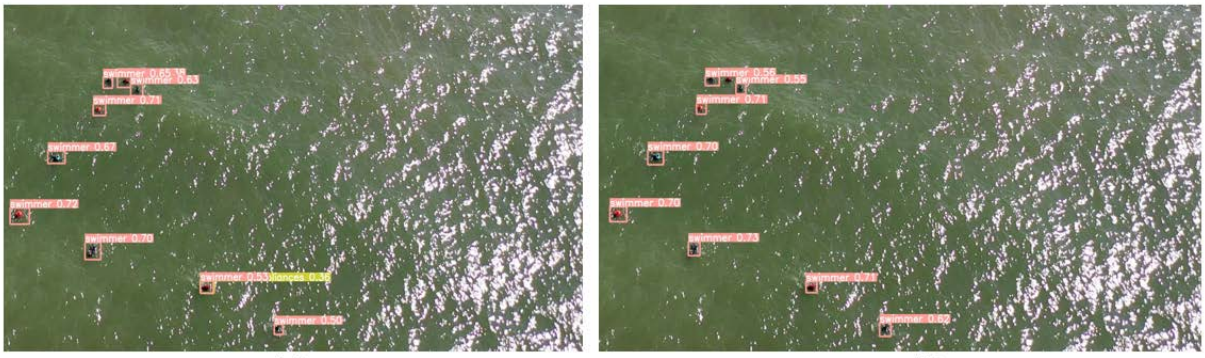


Table 2 Comparison results of ablation experiments

Serial number	*Model*	*Recall (%)*	*mAP@0.5 (%)*	*mAP@0.5:0.95 (%)*
1	YOLOv5s	66.236	72.091	37.142
2	YOLOv5s + SCP	67.823	72.306	37.067
3	YOLOv5s + NextViT	71.578	77.289	39.579
4	YOLOv5s + Focal-GIoU	67.621	72.425	37.844
5	YOLOv5s + NextViT + SCP	74.225	78.186	40.393
6	YOLOv5s + NextViT + Focal-GIoU	72.382	77.865	40.286
7	YOLOv5s + NextViT + SCP + Focal-GIoU (ours)	75.462	**78.408**	40.411

Figure 7 The first comparison of the model effects on maritime target detection before and after improvement, (a) YOLOv5s (b) YOLOv5s-NextSF (see online version for colours)

Figure 8 The second comparison of the model effects on maritime target detection before and after improvement, (a) YOLOv5s (b) YOLOv5s-NextSF (see online version for colours)



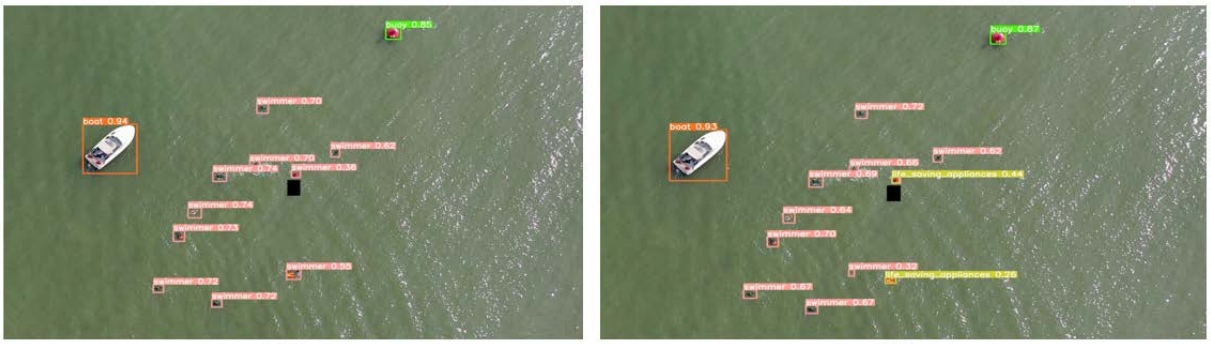
After replacing CIoU of YOLOv5s with GIoU, add FocalL1 loss and combine it with GIoU to form Focal-GIoU. Using the YOLOv5s + Focal-GIoU model, there was a slight enhancement in *recall*, $mAP_{0.5}$ and $mAP_{0.5-0.95}$ compared to the original model, with upgradations of 1.4, 0.3 and 0.7 percentage points, separately. By integrating three modules into YOLOv5s, a new model YOLOv5s + Next-ViT + SCP + Focal-GIoU is generated. Compared with the original YOLOv5s model, *recall*, $mAP_{0.5}$ and $mAP_{0.5-0.95}$ have significantly promoted by 9.2, 6.3 and 3.3 percentage points, respectively.

In order to further demonstrate the superior performance of the improved model in maritime distress target detection, objective indicators were compared between the YOLOv5s-NextSF model and other versions of the YOLO series models. Comparative experiments were conducted under the same sample and training environment. Visualise and compare the $mAP_{0.5}$ of the training results of each model as shown in Figure 6. The training performance of the enhanced YOLOv5s-NextSF model is significantly better than other versions of the YOLO series.

Figure 9 The third comparison of the model effects on maritime target detection before and after improvement, (a) YOLOv5s (b) YOLOv5s-NextSF (see online version for colours)

(a) (b)

Figure 10 The fourth comparison of the model effects on maritime target detection before and after improvement, (a) YOLOv5s (b) YOLOv5s-NextSF (see online version for colours)



(a) (b)

Figure 11 The fifth comparison of the model effects on maritime target detection before and after improvement, (a) YOLOv5s (b) YOLOv5s-NextSF (see online version for colours)

(a) (b)

4.5 Comparison of validation results

In this section, five sets of images were selected for testing on the original YOLOv5s and the improved YOLOv5s-NextSF, and the comparison of the model effects before and after improvement was visualised. The comparison of the effects of the selected five sets of images is shown in Figures 7–11.

As shown in Figure 7, due to the detection target being located at the edge of the image, the YOLOv5s network was unable to detect the boat.

As shown in Figure 8, the brightness of the image increases due to the reflection of sunlight. The YOLOv5s network mistakenly detects the sparkling sea surface as life_saving_appliance.

As shown in Figure 9, since the dense detection targets, the YOLOv5s network missed and incorrectly detected some swimmers.

As shown in Figure 10, because of the dense detection of targets, the YOLOv5s network mistakenly detects some life_saving_appliances as swimmers.

As shown in Figure 11, the background of the image becomes more complex due to the splashing waves on the

water surface after the motorboat travels rapidly at sea, which leads to the YOLOv5s network detecting one jet ski target as two.

After comparing the visualisation effects of the improvement of the YOLOv5s model, it is evident that the improved YOLOv5s model can more effectively identify detection targets. For example, in Figure 11(a), the rapid movement of the motorboat generates the formation of waves, the background near the target becomes complex and the detection object becomes blurred, resulting in incorrect detection of the number of targets in the original model. However, the modified model with CSP can obtain more complementary information for fuzzy targets, further promoting the detection ability for fuzzy targets. The revised YOLOv5s-NextSF model can effectively detect fuzzy targets [Figure 11(b)].

In contrast to the pristine YOLOv5s model, although the modified YOLOv5s-NextSF has increased *recall*, $mAP_{0.5}$, and $mAP_{0.5-0.95}$ by 9.226%, 6.317% and 2.567%, there are also some limitations. The weight size trained by the YOLOv5s-NextSF model is 69.4 MB, far exceeding the size of the original model of 13.7 MB. The improved model has a floating-point operation count of 103.49 GFLOPs, which is significantly higher than the original model's. The model designed relies on drones for operation, and the computing power that UAVs can carry is limited. Therefore, the focus of the next work is on how to lightweight the model so that the improved network model can be loaded onto onboard computers.

5 Conclusions

With the development of UAV technology in recent years, using UAV aerial images for maritime distress target detection has gradually become one of the core technologies of UAV applications. However, because of the influence of complex environments such as ocean winds and waves, traditional target recognition methods have lower recognition accuracy. In order to enhance the efficiency of maritime target recognition and extend the applicability of deep learning algorithms to the domain of maritime distress target detection, the YOLOv5s maritime distress target detection algorithm based on Next-ViT is presented in this paper. Foremost, the backbone network of YOLOv5s is replaced by Next-ViT to improve the balance between latency and accuracy. Then, SCP is introduced into the neck network to obtain complementary information on fuzzy targets by utilising the long-range correlation of images. Eventually, the CIoU loss function utilised by YOLOv5s is modified by Focal-GIoU to ensure that the contribution of diverse quality samples is equitably considered during the optimisation of the loss function. A subset of the SeaDronesSee dataset is employed in the conduct of experiments, and the result shows that the performance of maritime distress target detection can be effectively progressed by the ameliorated YOLOv5s-NextSF model. The model can be widely applied in the field of maritime distress target detection. Although the improved model has made some progress in maritime distress target detection, there still exist some problems such as high computational complexity and insufficient lightweight of the model. Thus, further research may focus on lightening the model and improving its practicability, under the limited computing resources of UAV.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under grant 61803217, the Natural Science Foundation of Shandong Province under grant ZR2023MF029, the Team Plan for Youth Innovation of Universities in Shandong Province under grant 2022KJ142, and the Taishan Scholar Special Project Fund under grant TSQN202408163.

The authors have no competing interests to declare that are relevant to the content of this article.

References

Bai, J., Dai, J., Wang, Z. and Yang, S. (2022) 'A detection method of the rescue targets in the marine casualty based on improved YOLOv5s', *Frontiers in Neurorobotics*, Vol. 16, pp.1053124–1053124, DOI: 10.3389/FNBOT.2022.1053124.

Bochkovskiy, A., Wang, C. and Liao, H.M. (2020) *YOLOv4: Optimal Speed and Accuracy of Object Detection*, ArXiv, abs/2004.10934, DOI: arXiv:2004.10934.

Chen, M., Zeng, F., Xiong, X., Zhang, X. and Chen, Z. (2021) 'A maritime emergency search and rescue system based on unmanned aerial vehicle and its landing platform', *2021 IEEE International Conference on Electrical Engineering and Mechatronics Technology (ICEEMT)*, pp.758–761, DOI: 10.1109/ICEEMT52412.2021.9602734.

Geng, W., Yi, J. and Cheng, L. (2024) 'An efficient detector for maritime search and rescue object based on unmanned aerial vehicle images', *Displays*, Vol. 87, pp.102994–102994, DOI: 10.1016/J.DISPLA.2025.102994.

Ghazali, S.N.A.M., Anuar, H.A., Zakaria, S.N.A.S. and Yusoff, Z. (2016) 'Determining position of target subjects in maritime search and rescue (MSAR) operations using rotary wing unmanned aerial vehicles (UAVs)', *2016 International Conference on Information and Communication Technology (ICICTM)*, pp.1–4, DOI: 10.1109/ICICTM.2016.7890765.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) 'Rich feature hierarchies for accurate object detection and semantic segmentation', *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp.580–587, DOI: 10.1109/CVPR.2014.81.

Girshick, R. (2015) 'Fast R-CNN', *2015 IEEE International Conference on Computer Vision (ICCV)*, pp.1440–1448, DOI: 10.1109/ICCV.2015.169.

Han, L., Tao, P. and Martin, R.R. (2019) 'Livestock detection in aerial images using a fully convolutional network', *Computational Visual Media*, Vol. 5, No. 2, pp.221–228, DOI: 10.1007/s41095-019-0132-5.

Khalfaoui, A., Badri, A. and Mourabit, I.E. (2022) 'Comparative study of YOLOv3 and YOLOv5's performances for real-time person detection', *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pp.1–5, DOI: 10.1109/IRASET52964.2022.9737924.

Liu, K., Qi, Y., Xu, G. and Li, J. (2024) 'YOLOv5s maritime distress target detection method based on Swin transformer', *IET Image Processing*, Vol. 18, No. 5, pp.1258–1267, DOI: 10.1049/ipr2.13024.

Liu, S., Qi, L., Qin, H., Shi, J. and Jia, J. (2018) 'Path aggregation network for instance segmentation', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8759–8768, DOI: 10.1109/CVPR.2018.00913.

Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) 'Feature pyramid networks for object detection', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.936–944, DOI: 10.1109/CVPR.2017.106.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C. (2016) 'SSD: single shot multibox detector', *Lecture Notes in Computer Science*, pp.21–37, DOI: arXiv:1512.02325.

Liu, Y., Li, H., Hu, C., Luo, S., Luo, Y. and Chen, C.W. (2024) 'Learning to aggregate multi-scale context for instance segmentation in remote sensing images', *IEEE Transactions on Neural Networks and Learning Systems*, pp.1–15, DOI: 10.1109/TNNLS.2023.3336563.

Li, Y., Yuan, H., Wang, Y. and Xiao, C. (2024) 'GGT-YOLO: a novel object detection algorithm for drone-based maritime cruising', *Drones*, Vol. 6, No. 11, pp.335–335, DOI: 10.3390/DRONES6110335.

Li, S., Lin, Z., Wang, H., Yang, W. and Liu, H. (2024) 'Text-guided multi-class multi-object tracking for fine-grained maritime rescue', *Remote Sensing*, Vol. 16, No. 19, pp.3684–3684, DOI: 10.3390/RS16193684.

Redmon, J. and Farhadi, A. (2017) 'YOLO9000: better, faster, stronger', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.6517–6525, DOI: 10.1109/CVPR.2017.690.

Redmon, J. and Farhadi, A. (2018) *YOLOv3: An Incremental Improvement*, ArXiv, abs/1804.02767, DOI: arXiv:1804.02767.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) 'You only look once: unified, real-time object detection', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.779–788, DOI: 10.1109/CVPR.2016.91.

Ren, S., He, K., Girshick, R. and Sun, J. (2017) '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, DOI: 10.1109/TPAMI.2016.2577031.

Rezatoffghi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. and Savarese, S. (2019) 'Generalized intersection over union: a metric and a loss for bounding box regression', *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.658–666, DOI: 10.1109/CVPR.2019.00075.

Song, X. and Tang, H. (2024) 'Blood cell target detection based on improved YOLOv5 algorithm', *Electronics*, Vol. 13, No. 24, pp.4992–4992, DOI: 10.3390/electronics13244992.

Wang, C.Y., Mark, L.H.Y., Wu, Y.H., Chen, P.Y., Hsieh, J.W. and Yeh, I.H. (2010) 'CSPNet: a new backbone that can enhance learning capability of CNN', *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp.1571–1580, DOI: 10.1109/CVPRW50498.2020.00203.

Wang, Y., Wang, B. and Fan, Y. (2025) 'PPGS-YOLO: a lightweight algorithms for offshore dense obstruction infrared ship detection', *Infrared Physics and Technology*, Vol. 145, pp.105736–105736, DOI: 10.1016/J.INFRARED.2025.105736.

Wang, R., Liu, B. and Liao, T. (2024) 'A multi-scale attention-based pedestrian detection method for roadways using the YOLOv5 framework', *Journal of Electronic Research and Application*, Vol. 9, No.1, pp.224–232, DOI: 10.26689/JERA.V9I1.9457.

Zhang, P., Zhu, P., Sun, Z., Ding, J., Zhang, J., Dong, J. and Guo, W. (2024) 'Research on improved lightweight YOLOv5s for multi-scale ship target detection', *Applied Sciences*, Vol. 14, No. 14, pp.6075–6075, DOI: 10.3390/APP14146075.

Zhang, Y.F., Ren, W., Zhang, Z., Jia, Z., Wang, L. and Tan, T. (2022) 'Focal and efficient IoU loss for accurate bounding box regression', *Neurocomputing*, Vol. 506, pp.146–157, DOI: 10.1016/j.neucom.2022.07.042.

Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q. and Zuo, W. (2018) 'Enhancing geometric factors in model learning and inference for object detection and instance segmentation', *IEEE Transactions on Cybernetics*, Vol. 52, No. 8, pp.8574–8586, DOI: 10.1109/TCYB.2021.3095305.