Expert Systems With Applications 217 (2023) 119560

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Sw-YoloX: An anchor-free detector based transformer for sea surface object detection

Jiangang Ding, Wei Li *, Lili Pei, Ming Yang, Chao Ye, Bo Yuan

School of Information Engineering, Chang'an University, Xi'an, Shaanxi 710064, China

ARTICLE INFO

Keywords:
Sea surface object detection
Sw-YoloX
Transformer
YoloX
Self-training classifier

ABSTRACT

To cope with the challenge of blurred images of sea surface objects caused by the complex and undulating sea surface environment, we propose Sw-YoloX, which can utilize the global modeling ability to encode the key semantics of sea surface objects, thereby obtaining global features that cannot be captured by CNN. Then the convolutional block attention module (CBAM) and atrous spatial pyramid pooling (ASPP) module are integrated in the neck of the detector, and the decoupled head is used as the prediction part. In addition, we also integrate multiple training strategies to effectively improve the detector performance, such as simple optimal transport assignment (SimOTA) strategy and multi-model integration. Finally, we construct the XM-10000 dataset for validation based on sea surface monitoring data in Xiamen, China. With end-to-end training, Sw-YoloX achieves higher performance than baseline and mainstream detector, with F1-Score is 78.1, mean average precision (mAP) is 54.4, and average recall (AR) is 72.0. This research, which has now been deployed in the coastal defense department in Xiamen, China, has important implications for searching for survivors and preventing smuggling.

1. Introduction

The oceans cover approximately 71 % of the Earth's total area. Efficient sensing of sea surface objects is important for navigational safety, surface rescue and defence security. The effective detection of sea surface objects such as living, floating objects and ships has always been a recognised challenge in the field of object detection. High-resolution optical technology is now used in a wide range of sensors. A sea defence security site typically sets up dozens of high resolution cameras for real-time monitoring of sea surface objects, the sheer volume of data and the amount of processing involved makes automatic interpretation of sea surface images essential. How to efficiently and quickly detect sea surface images has become a common demand in the industry at this stage.

Currently, most deep neural network detectors are designed for natural scene images. Unlike traditional detection tasks, sea surface object detection task present four main challenges, which are intuitively illustrated by some cases in Fig. 1:

1) The background of the image is complex. Depending on the weather and sea conditions, the background colour of the sea surface varies greatly, such as brown seawater on foggy days, blue seawater on sunny days and yellow seawater on cloudy days. The same object with different imaging backgrounds has different outline textures and image clarity.
2) More serious environmental disturbances. When detecting objects close to the shore, the images always contain confusing geographical elements. Objects can be obscured and disturbed by reefs, shore vegetation, buildings and signboard. This increases the complexity of the neural network in separating the sea surface objects from the background.
3) Object size uncertainty. The large span of distance between the surface object and the camera results in dramatic object scale variations. When the object is far away from the camera, such as when the distance is more than 1 km, the object will have fewer effective pixels in the image and its outline will be blurred even when the focus is adjusted.
4) Uncertainty in the attitude of the object. Due to wind and waves, the attitude of live and floating objects on the sea surface varies considerably compared to large ships. This places a high demand on the generalization of the detector.

* Corresponding author.
E-mail addresses: 2020124034@chd.edu.cn (J. Ding), grandy@chd.edu.cn (W. Li), peilili@chd.edu.cn (L. Pei), 2021024014@chd.edu.cn (M. Yang), 2021124038@chd.edu.cn (C. Ye), yuanbo_chd@chd.edu.cn (B. Yuan).

https://doi.org/10.1016/j.eswa.2023.119560
Received 26 June 2022; Received in revised form 25 August 2022; Accepted 12 January 2023
Available online 13 January 2023

To solve the above problems, we propose the Sw-YoloX detector. The overview of the detection pipeline is shown in Fig. 2. Through the comparison of multiple sets of experiments, the detector in this paper has stronger robustness and higher detection accuracy, and can better adapt to object detection tasks in complex sea surfaces. Our contributions are listed as follows:

1) As living and floating objects data under different sea states are scarce and expensive, we constructed the XM-10000 benchmark dataset for the detection task in this paper based on actual sea surface measurements from January to March 2022 in Xiamen, China.
2) To find the region of attention in high-resolution sea surface images, we use swin transformer (Liu et al., 2021a) as the backbone of Sw-YoloX. Using its global information modelling capability to solve the difficult problem of uncertainty of object attitude.
3) To address the uncertainty of object scale, we use the decoupling head and SimOTA positive sample allocation strategy of YoloX (Ge et al., 2021), eliminate the prior through the anchor-free mechanism, thereby improving the prediction potential of the transformer.
4) To increase the efficiency of object semantic information transfer in the feature fusion phase, we added ASPP (Chen et al., 2017) and CBAM Woo et al. (2018) modules to the neck of the Sw-YoloX.
5) To further improve the performance of the detector, we use some heuristic training strategies to improve the generalization ability of the detector at a relatively small cost.

2. Related work

2.1. Research on sea surface object detection based on traditional methods

Currently, the methods used for sea surface object detection fall into two main categories. The first category is traditional image processing techniques based on manually designed features. Szpak and Tapamo (2011) calibrate moving ships in the ocean based on background subtraction and real-time approximation of curve evolution of level sets, tracking moving ships in dynamic backgrounds. Guo et al. (2014) proposed an initial contour extraction method based on visually saliency prior shapes for ship recognition. Xu et al. (2018) constructed a 3D feature detector based on multiple polarization features of radar signals, which enhanced the accuracy of small object detection on the sea surface. Zhou and Jiang (2019) built a FAR controllable detector based on a decision tree, which improved the detection accuracy of weak objects on the sea surface. In the same year, Gu (2019) proposed a small floating object detection method based on multi-feature and principal component analysis, which achieved a relatively advanced detection rate. Zhang et al. (2020) proposed a polarization detector based on a complex Gaussian model, which effectively improved the detection performance of the original detector for weak objects on the sea surface. By analyzing the normalized cross-correlation values of higher-dimensional segments, Chernomorets et al. (2021) realize the task of detecting sea surface videos without prior information. Xu et al. (2021) realized small object detection on the sea surface based on radar data through feature engineering and fast convex hull algorithm. Bai et al. (2021) proposed an object detector for tiny floating objects on the sea surface based on local sparse coefficients, which effectively improved the detection performance of radar data. Zhang et al. (2021b) solved the problem of covering small objects on the sea surface by radar echoes such as ships and reefs through an oversampling high-resolution range planing algorithm. While traditional algorithms have made great progress, they require hand-designed features and do not allow for end-to-end detection. Furthermore, since the artificially designed features are not robust enough to diverse inputs, they cannot be applied to complex sea surface perception tasks.

2.2. Research on object detection based on deep neural network

In recent years, deep learning, represented by convolutional neural networks, has been widely used for machine vision pattern recognition tasks, which has opened up opportunities for automated processing of optical data on the sea surface. Several well-known benchmark datasets, such as MS COCO (Lin et al., 2014), PascalVOC (Everingham et al., 2010) and DOTA (Xia et al., 2018), have greatly contributed to the development of object detection applications. CNN-based object detectors have been widely proposed: Salman et al. (2022) developed an automatic detection and diagnosis system for prostate cancer based on the fusion of the scoring results of the Yolo algorithm and ISUP. Zhang et al. (2022) proposed an anchor-free YOLOv3 network for mammography quality detection using a combination of GIoU loss function and Focal loss. Peng et al. (2022) proposed a new global prior-guided cross-fusion module, constructed a new network structure with a global fusion strategy, and achieved state-of-the-art (SOTA) performance. Majid et al. (2022) achieves high detection accuracy and reliability based on attention mechanism and Grad-CAM method coordination and a custom deep learning framework for fire detection. In the same year, Chalavadi et al. (2022) used hierarchical dilated convolutions (mSODANet) for

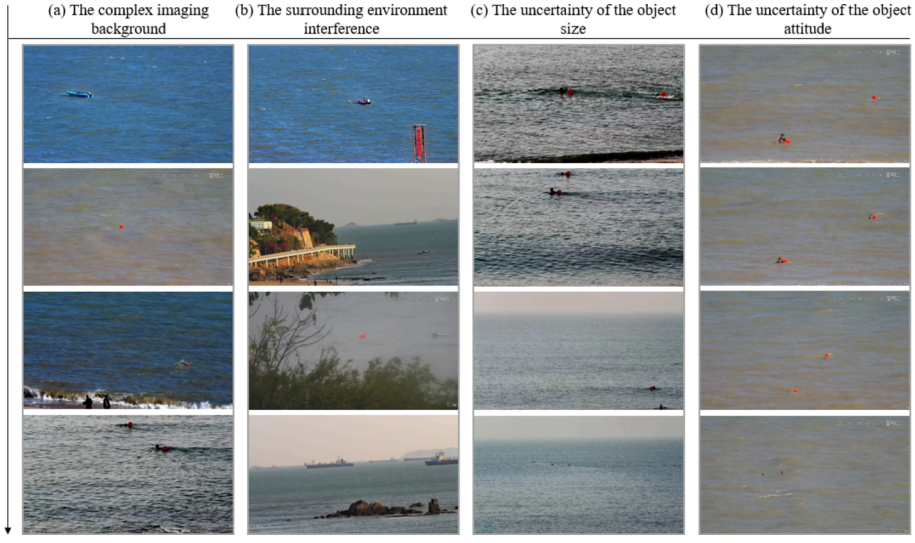

Fig. 1. Intuitive case to explain the four main problems in the task of sea surface object detection.

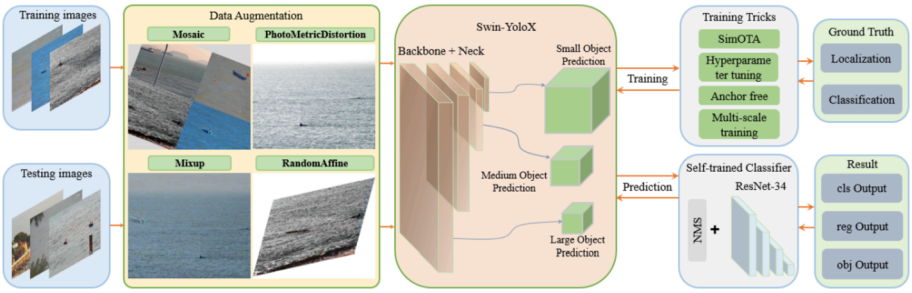

Fig. 2. The overview of working pipeline using Sw-YoloX. We apply the anchor-free mechanism to sea surface object detection, and improve the detector performance by improving the backbone and the neck. In addition, we employ other training strategies to make Sw-YoloX stronger.

multi-scale object detection in aerial images, which effectively captured contextual information in images. Each convolutional layer of the CNN-based detector is continuously integrating the previous information, so only by continuously deepening the neural network can it have a larger receptive field. To address this shortcoming, transformer-based detectors (Vaswani et al., 2017) are increasingly being proposed: Carion et al. (2020) proposed a detector named DETR based on transformer attention, which opened up a new chapter in object detection. Wang et al. (2021) proposed the SwinGD detector for automatic grape picking based on the swin transformer. Heo et al. (2022) proposed an occlusion-aware spatial attention transformer for occluded object recognition, applying the transformer to occluded object detection. By extending the local features of the CNN to the global, the design architecture of the original detector is changed and the detection accuracy is improved.

2.3. Research on sea surface object detection based on deep neural network

In order to solve the problems of traditional detection algorithms in the field of sea surface object detection, related research has gradually been carried out based on deep neural networks. Sutikno et al. (2018) used neural network for the detection of illegal fishing on the sea. Qin et al. (2021) proposed a Yolov3-based detector to improve the accuracy and real-time detection of ships. Sun et al. (2021) using Yolov4 as the baseline, proposed a sea surface object detection algorithm based on an optimized feature fusion network, which achieved certain results in the detection of long-distance and small objects under complex meteorological conditions. Guo et al. (2021) designed a feature pyramid fusion module and a head enhancement module based on the centernet detector to improve the ship detection performance in complex backgrounds. Zhang et al. (2021a) proposed a fast object detection scheme based on Yolov4 and the dark and bright channel prior (DBCP) algorithm. The detection accuracy of ships on the sea surface is improved. In the same year, Liu et al. (2021b) proposed an improved Yolov4 algorithm to improve the detection accuracy of unmanned surface vehicles. Liu et al. (2021c) proposed an improved Yolov5 algorithm to improve the detection accuracy of surface ships. The above research shows that the accuracy and speed of many detectors have reached the SOTA performance in the industry, proving that it is necessary to apply deep learning to the task of sea surface detection. However, object detection in complex and undulating sea environment is a difficult task. To break the performance bottleneck of existing detectors, we combine the transformer and anchor-free mechanism to propose Sw-YoloX, which is more suitable for the task of sea surface object detection.

3. Data and methods

3.1. Datasets and research objects

Living and floating objects data for different sea states are scarce and expensive. We constructed the XM-10000 benchmark dataset based on actual sea surface measurements from January to March 2022 in Xiamen, China, in response to the problems of poor image resolution and less valid data in the existing sea surface object dataset, which cannot meet the task of sea surface object detection in complex scenarios. The dataset consists of 12,000 images and contains three types of objects: living, floating objects and boat. It covers different weather, different backgrounds, different sea surface environments, different imaging fields of view and different object styles, which can better meet the generalization and diversity. Fig. 3 shows the normalized size distribution and the location distribution in original images of all objects in the dataset. With the help of visualization in Fig. 3, it can be seen that the position of the ground truth box is distributed randomly and uniformly, and the object scale changes drastically. There are about more than 7500 faint or tiny objects in the XM-10000 benchmark dataset, and each object occupies around 5.0×10^{-4} % of an image, which puts forward higher requirements on the performance of the detector.

Besides, to prevent the detector from overfitting during training, we perform data augmentation on the original dataset. If ordinary methods such as random flip and rotation are simply applied, the improvement of detector accuracy is limited. To solve the above problem, we combined four efficient data enhancement methods, namely mosaic, mixup, photometric distortions and random affine, as shown in Fig. 2. Mosaic is an improved version of cutmix that stitches and rotates four images, greatly enriching the background information and effectively increasing the number of effective pixels in an image. Mixup makes the detector significantly more resistant to fogging background and noise

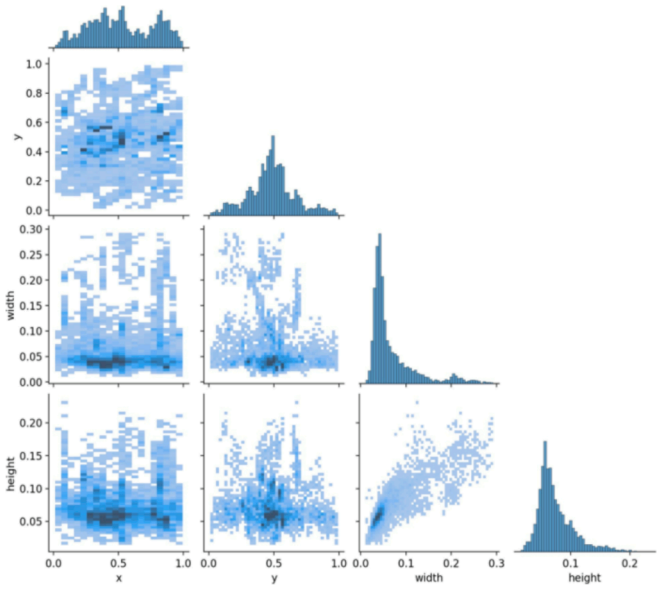

Fig. 3. Visualization of the Xm-10000 benchmark dataset. X, Y, Width, and Height represent the coordinates of the center point and the width and height of the ground truth box, respectively.

interference. Photometric distortions can be adjusted to simulate different weather conditions by adjusting hue and saturation. Random affine can expand the different attitudes of surface objects.

3.2. Use swin transformer as the backbone of Sw-YoloX

Currently, the Attention-based mechanism of the transformer achieves SOTA performance on a variety of vision tasks including image classification, object detection and semantic segmentation. The swin transformer is a hierarchical architecture for handling dense prediction problems. In general, pure transformer have a slow convergence rate and high training overhead, while CNN is not as efficient in feature extraction as transformer. To exploit the global modelling capabilities of the transformer and to avoid consuming too much memory, we used a combination of CNN and transformer, choosing the swin transformer as the backbone of the detector and using the CNN as the head of the detector. This not only reduces computational costs, but also facilitates the deployment of detectors. The forward propagation process of the backbone of Sw-YoloX is shown in Fig. 4.

Before entering the stage block, the image is downsampled by the patch partition structure into a token of 4×4 size. The patch partition is made up of slicing operation, layer normalisation and fully connected layer. Each stage contains a Patch Merging operation and several swin transformer Blocks. Each swin transformer block uses windows multi-head self-attention (W-MSA) and shifted windows multi-head self-attention (SW-MSA) to compute local self-attention and generate different proportions of tokens respectively. It consists of layer norm (LN), multi-layer perception (MLP), multi-head self-attention and residual connection. The residual structure is used to address the degeneracy of the deep transformer by allowing the network to focus only on the part of the current training that varies. The LN is used to calibrate the data distribution and facilitate training. The MLP can enhance the expressive power of the detector through non-linear mapping. The large range of information is converted throughout the detector by means of a local self-attention module, which is calculated as shown below:

$$\hat{X}^{l} = WMSA\left(LN\left(X^{l-1}\right)\right) + X^{l-1} \tag{1}$$

$$X^{l} = MLP\left(LN\left(\hat{X}^{l}\right)\right) + \hat{X}^{l} \tag{2}$$

$$\hat{X}^{l+1} = SWMSA\left(LN\left(X^{l}\right)\right) + X^{l} \tag{3}$$

$$X^{l+1} = MLP\left(LN\left(\hat{X}^{l+1}\right)\right) + \hat{X}^{l+1} \tag{4}$$

In addition, the W-MSA and SW-MSA structures are used in pairs in the swin transformer. W-MSA divides the token into a bunch of windows that do not overlap each other, and calculates attention for each window internally. In order to extract attention between neighbouring windows while remaining consistent with the computational complexity of W-MSA, SW-MSA shifts the entire window to the right and down by half of its own size. The entire token is then divided into different feature blocks as shown in Fig. 5. Finally, a Mask matrix of the same size as the original feature map is used to label the different feature regions, preventing the calculation of attention between non-adjacent feature blocks.

3.3. The neck of the Sw-YoloX

The neck of Sw-YoloX is shown in Fig. 6. The whole structure is designed to perform a deep fusion of the feature layers output by swin transformer.

In order to increase the detection accuracy of small objects and reduce the number of parameters, the two feature layers with the largest scales are fused in this paper. We then followed path aggregation network (Liu et al., 2018) of YoloX and added the CBAM lightweight module before upsampling and downsampling, which consists of two separate sub-modules: the channel attention module (CAM) and the spatial attention module (SAM). The forward calculation process for the entire network is shown in Fig. 7. Where $F \in R^{H\times W\times C}$ is the input feature, $M_C \in R^{1\times 1\times C}$ is the one-dimensional channel attention, $M_S \in R^{H\times W\times 1}$ is the two-dimensional spatial attention, F' is the channel attention feature, and F'' is the final output feature. The entire CBAM is calculated as follows.

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{5}$$

$$M_S(F) = \sigma\left(f^{7\times 7}([AvgPool(F); MaxPool(F)])\right) \tag{6}$$

$$F' = M_C(F) \times F \tag{7}$$

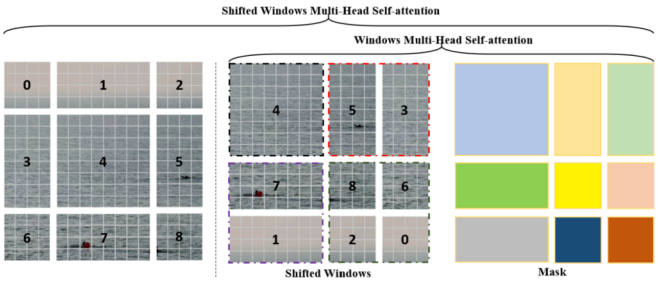

Fig. 5. Shifted Windows and its corresponding Mask matrix.

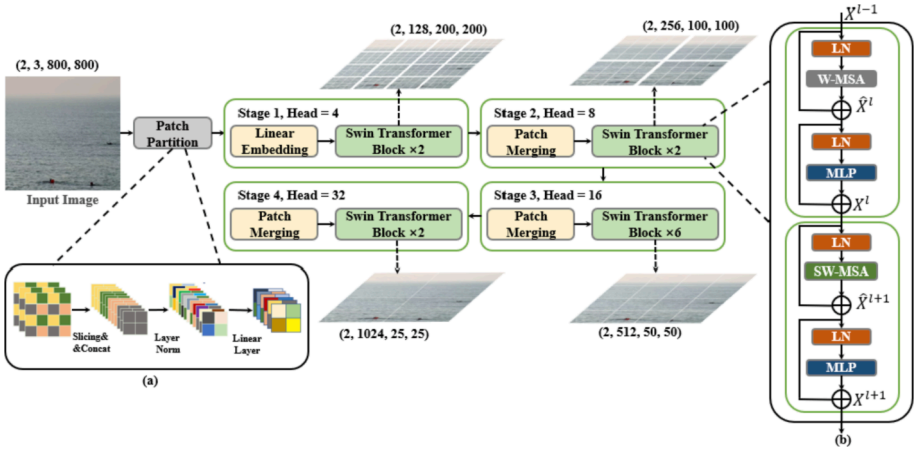

Fig. 4. The structure of the backbone of the Sw-YoloX. (**a**) The detailed structure of the Patch Partition block; (**b**) The detailed structure of the swin transformer block.

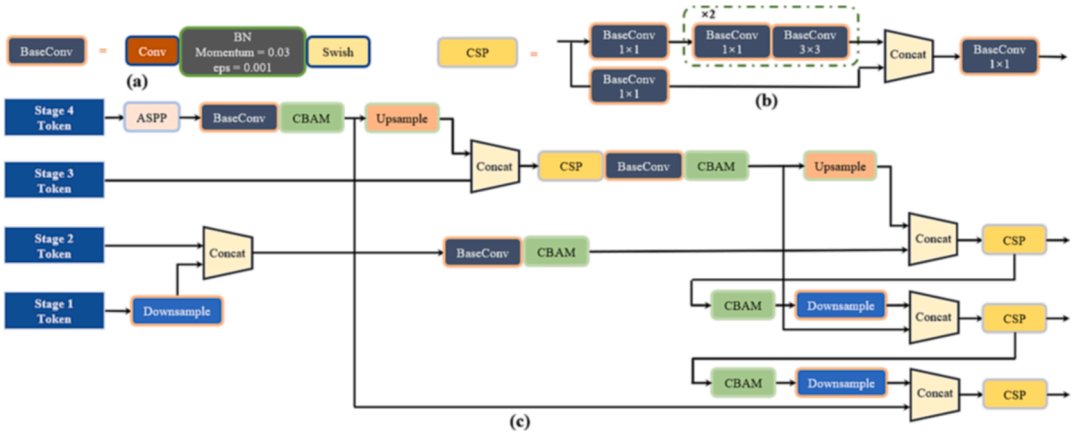

Fig. 6. The structure of the neck of the Sw-YoloX. (**a**) The detailed structure of the BaseConv; (**b**) The detailed structure of the CSP (Wang et al., 2020) block; (**c**)The detailed structure of the neck.

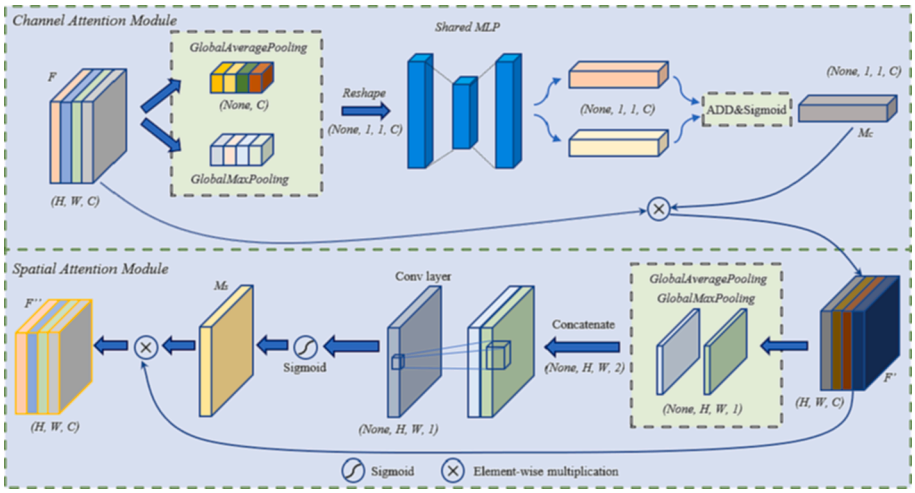

Fig. 7. The structure of the CBAM framework.

$$F^{'} = M_C(F) \times F \tag{8}$$

The seamless integration of CBAM allows the detector to adaptively adjust the fusion weights of features, highlighting effective features of sea surface objects and attenuating the effects of sensor noise and other confusing information. In addition, we added the ASPP structure after the output of the last stage and performed end-to-end training together.

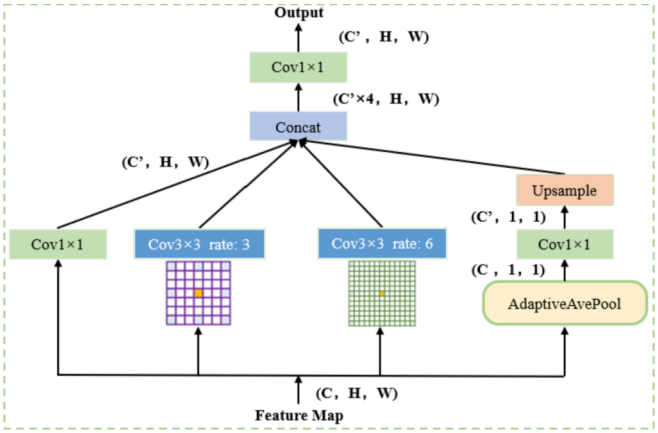

Fig. 8. The structure of the ASPP framework.

The ASPP is shown in Fig. 8.

We obtained the features of the different receptive fields through four branches in parallel. The middle two branches use atrous convolutions with expansion factors of 3 and 6, respectively. The information from the four branches is further fused by concatenate and 1×1 convolution. Each 1×1 convolution can be interpreted as a weighted feature fusion for each channel. Subsequent experiments demonstrate that ASPP is effective in enhancing the detector's ability to obtain multiscale contexts. However, too many ASPP structures may lead to an increase in model complexity and hence slower inference, so we only string ASPP structures on the output of the last stage.

3.4. Using the YoloX detector as the baseline

YoloX integrates the industry's most advanced detection technology and outperforms most detectors in terms of performance. Therefore, we have chosen YoloX's decoupled head and SimOTA strategy, combined with the swin transformer, to combine the most suitable detectors for the task of sea surface object detection. Conventional detectors output the classification task, the regression task and the confidence value together, but in reality, the three are completely different in principle. Predicting the three separately will have a significant improvement on detection accuracy. The head used in this paper is shown in Fig. 9.

We connect the output of the neck to three decoupled heads, each with two branches for the classification task and the regression task

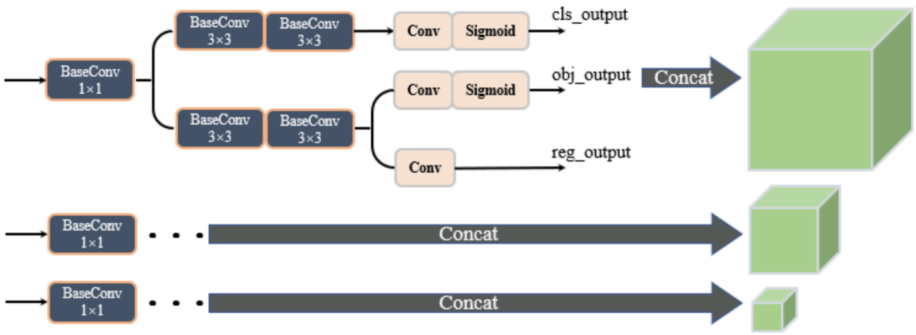

Fig. 9. The structure of the decoupled head framework.

respectively. On the branch of the regression task, the confidence output of the object is also included. The four values of the output of the regression task are the coordinates of the upper left corner and the length and width values of the ground truth box respectively. At the same time, the efficient positive sample selection method and the anchor-free based loss calculation method can be better adapted to the task of detecting sea surface objects with uncertain object size. The entire SimOTA allocation strategy is divided into 4 steps, as shown in Fig. 10:

1) Generate a positive sample candidate region. A yellow square box with 2.5 times the length of the grid cell is created with the centroid of the current GT Box as the centre. If the centre point of the grid cell is within the blue box and the yellow box, then the grid cell belongs to the positive sample candidate region, that is, all the grid cells within the red box.
2) Calculate the IoU of each prediction box in the positive sample candidate region with the current GT box and rank the obtained IoU values from smallest to largest. The top candidate_K IoU values are summed up and denoted as dynamic_k.
3) Calculate the cost value of each prediction box and the current GT box to get the cost matrix. This matrix represents the cost relationship between the current GT box and the predicted box. Through the Cost matrix, the detector can adaptively find positive samples for each GT Box. Its calculation formula is as follows:

$$c_{ij} = L_{ij}^{cls} + \lambda L_{ij}^{reg} \tag{9}$$

4) Arrange the values of the cost matrix in ascending order. The first dynamic_k prediction boxes with the smallest cost value are taken as the final positive samples of the current GT box, and the remaining prediction boxes are used as negative samples.

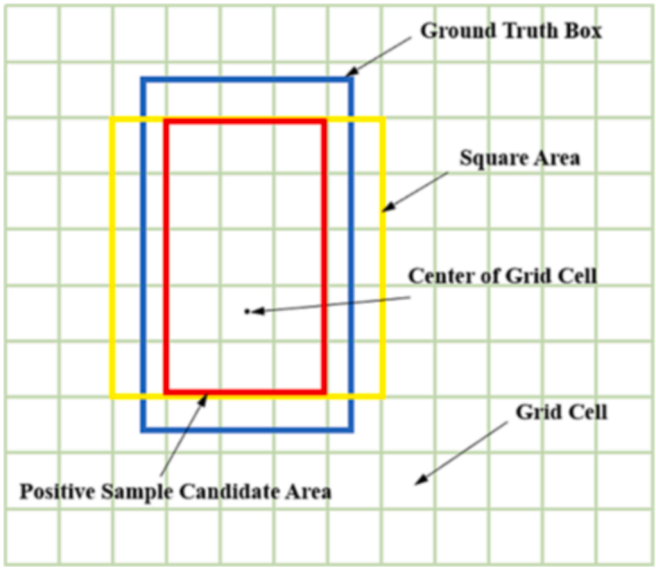

Fig. 10. Visualization of SimOTA Algorithm. The green grid represents the receptive field on the original image for each pixel on the feature map. The blue box, the yellow box and the red box represent the ground truth box, the algorithm-specified area and the candidate area, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Experiments and discussion

4.1. Training strategies

To further improve the performance of the detector, in addition to using multiple data augmentation, we also combine multi-scale training, hyperparameter tuning, and multi-model integration to improve the generalization ability of the detector at a relatively small cost. Experimental results show that these heuristic training strategies can effectively enhance the adaptability of the detector to drastic changes in the imaging background and the classification performance of difficult categories.

Self-trained classifier for multi-model integration. We deployed the Sw-YoloX trained on the XM-10000 benchmark dataset in practice and found that the detector, while having excellent positioning capability, was poor at classification. Subsequently, we visualized the normalized confusion matrix as shown in Fig. 11, and observed that the low classification accuracy for floating objects and living is the key to the overall performance of the detector. To improve the classification accuracy of difficult samples, we propose an additional self-training classifier containing background categories. We constructed the training set by cropping the real bounding box and resizing and padding each image block to 64×64, and selected ResNet34 (He et al., 2016) as the classifier. As shown in Fig. 11, with the help of the self-training classifier, our detector achieves an improvement of about 0.9 % to 1.1 % in the AP values on the validation set.

Training details. Based on the sufficient samples of the Xm-10000 benchmark dataset, this paper divides the ratio of training set, validation set and test set into 8:1:1. Since the swin transformer is difficult to train, we initialize the backbone by loading pre-trained weights, and the other convolutional layers are initialized randomly using the Kaiming distribution (He et al., 2015). In the first five epochs, the detector has zero understanding of the data distribution, which would make the model unstable or even over-fit if the preset learning rate is directly applied. Therefore, we use the warmup learning rate scheduling strategy in the first 5 epochs to prevent the gradient optimization route from being shifted. In the last five epochs, the properties of the data learned by the detector tend to stabilize, and this stability would be destroyed if a larger learning rate is also used. To further approach the optimal point of the objective function, we use a learning rate scheduling strategy based on cosine annealing in the last 5 epochs. In addition, we

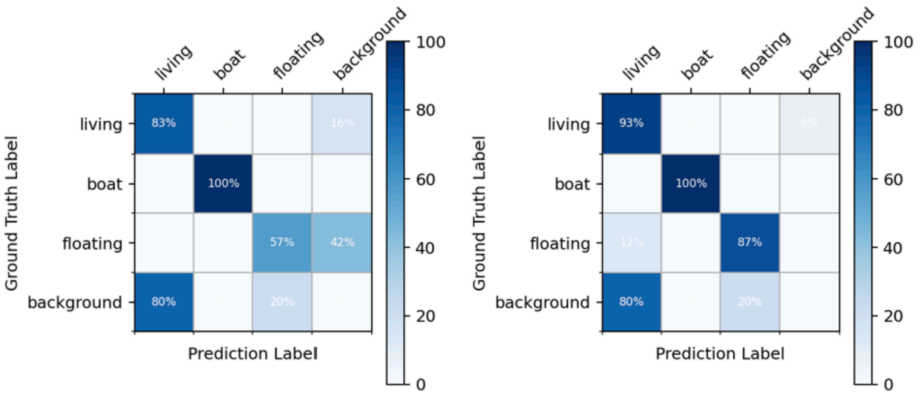

Fig. 11. The two normalized confusion matrices were made at IoU threshold of 0.65, score threshold of 0.01. They represent the accuracy of the detector on the validation set before and after adding the self-training classifier.

performed multi-scale training on Sw-YoloX, i.e. scaling each batch generated by the data loader to 1.0, 0.83 and 0.67 times its original size. For different detectors within the control group, we performed the same multi-scale testing.

Experimental environment and hyperparameters. In order to ensure the accuracy of the ablation experiments, the training and test configurations were kept consistent for all detectors in this paper. The experimental environment configuration information is as follows: the operating system is Win10, the processor is i7-11800, the GPU is Nvidia Quadro P5000, the version of Pytorch is 1.9.0, the version of CUDA is 11.1.74, the version of CuDNN is X64_8.2.1.32, the TensorRT version used in this paper for inference acceleration is 7.2.3.4. Other important dependencies include Cudatookit version 11.1.1 and Torchvision version 0.10.0. After repeated tuning, the optimal hyperparameter settings of the Sw-YoloX are as follows: The optimizer is SGD, momentum = 0.9, learning rate = 0.005, weight_decay = 0.0005, epoch = 100, batchsize = 2. The input image size is (800, 800), resize while keeping the original size, and the padding has a grey scale of 114.

Engineered deployment. In practice, we use an SDK written in C++ to implement forward inference. In order to maximize hardware

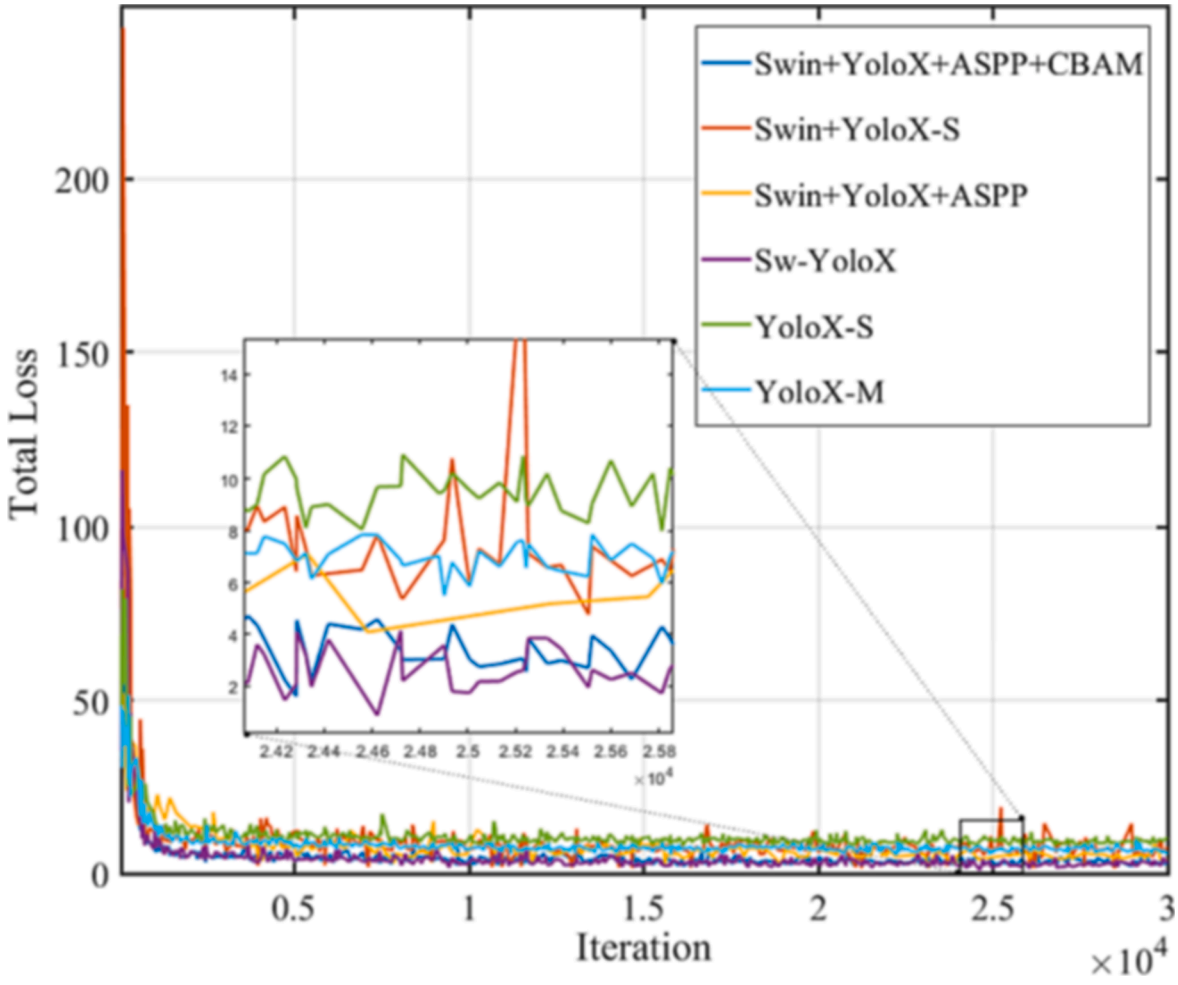

Fig. 12. Loss update curve of ablation experiments.

performance, this paper uses TensorRT to do computational graph optimization of the detector and half-precision floating-point data for forward inference. After actual testing, it is found that by calling the engine generated by TensorRT, the FPS of the original detector can be increased by about twelve times or more.

4.2. Ablation experiments

To further validate the innovation, six sets of ablation experiments were designed to evaluate the improvements to the detector performance. We visualised the loss curves for the experimental procedure, as shown in Fig. 12. Due to the sufficient amount of data, most of the detector training within the control group converged at epoch = 10, so we truncated the later loss values that were not meaningful. As can be seen from the Fig. 12, the original YoloX is not suitable for the current detection task. Under the improved strategy in this paper, the final loss is reduced to 3/7 of the initial value. Lower loss predicts higher detection accuracy. As shown in Table 1, compared with other detectors, Sw-YoloX has a great improvement in convergence speed and fitting ability. It should be noted that the total loss of Sw-YoloX here does not have a self-training classifier, because the self-training classifier is a module that needs to be trained separately.

We then quantitatively evaluated each detector. We selected a number of metrics that were important in the object detection task as criteria for the experimental evaluation. For detection accuracy, we use mAP (mean Average Precision, the IoU threshold value ranges from 0.5 to 0.95, with a stride of 0.05), AP50 (Average precision value when the IoU threshold is 0.5), AP75 (Average precision value when the IoU threshold is 0.75), AP-Area (Average precision of objects at different scales), AR (Average Recall) and F1-score as the evaluation metrics. For detection speed, we use FPS (Frames Per Second), Flops (floating-point operations per second) and Model size as evaluation metrics. The results of the ablation experiments are shown in Table 2.

Compared with the baseline, the improvement measures in this paper have a stable improvement to the performance of the detector. After replacing the backbone with swin Transformer, the performance of the detector has a major breakthrough. This is certainly an increase in computation due to the dynamic computation of self-attention, but the 4.6 % improvement in mAP and the 2.9 % improvement in AP50 make the increase in computation worthwhile. Through continuous improvement, the final mAP of the detector is 54.5, and the AP50 is 85.3. The results of ablation experiments show that our improved strategy can effectively improve the fitting accuracy of the detector, resulting in better performance. Although the GFLOPs and model size of the final detectors have increased, for the sea defense sector, multiple detectors need to be deployed via CUDA multi-threaded streams in one GPU. Therefore, higher FPS will incur additional overhead on hardware resources such as CPU and GPU. Moreover, after we accelerate through Tensor and use C++ to inference, the actual FPS can reach more than 10, which exceeds the requirement of 5 FPS. Therefore, this study has practical value.

4.3. Comparisons with the mainstream detector

In order to further verify the superiority of the Sw-YoloX detector, we conducted additional comparative experiments to compare the performance of the current mainstream high-precision detectors with Sw-YoloX. The results are shown in Table 3.

Among them, YoloX is a representative anchor-free detector and the baseline of this paper. RetinaNet (Lin et al., 2017) is a one-stage detector based on fcoal loss and has better performance on public datasets. Yolov5 is the current one-stage detector with SOTA performance. DETR is a new generation object detector based on transformer Head. Cascade RCNN (Cai and Vasconcelos, 2018) is a cascade detector that surpasses two-stage algorithms such as Faster RCNN (Ren et al., 2015) in public datasets, and it makes sense to replace its backbone with swin transformer and use it as a control group. As can be seen, the Sw-YoloX outperforms all types of detectors in the control group in all performance metrics. Although it has some FPS loss, it meets the application requirements and does not have any impact on the actual deployment. In addition, we have selected several representative detectors to visualize their attention in comparison trials, as shown in Fig. 13.

It can be seen that for sea surface images, the attention of conventional CNN detectors cannot be effectively focused. Although it is able to detect the general area of the object, it is also susceptible to noise from imaging clutter and complex environments. The detector with the anchor-priori mechanism is not well adapted to the large span of object scales and the large differences in object attitude. The anchor-free detector alone is based on weak perception of objects with an imaging distance of more than 800 m, and is not able to accurately classify living and floating objects. However, Sw-YoloX can not only better separate object and background, but also can adapt to large-span object geometric scales to achieve accurate localization of distant objects. Finally, we select some representative detection results for display, as shown in Fig. 14.

5. Conclusions

Based on the transformer and anchor-free mechanisms, combined with a number of improvement schemes and heuristic training strategies, this paper proposes a new detector named Sw-YoloX that is good at complex sea surface object detection task. From the conclusions obtained from the ablation experiments and the comparison experiments, it can be seen that Sw-YoloX has a significant improvement in accuracy compared with the original detector, and has a greater accuracy advantage than the traditional CNN + anchor base structure. In practice, our detector has been successfully deployed in Xiamen, China for the detection of drowning and stowaways on the sea surface, and the reliability and robustness of the algorithm have been demonstrated. Although Sw-YoloX has high detection accuracy, its memory overhead is higher, which makes its detection speed lower than that of lightweight networks. In the field of real-time video detection or mobile deployment, Sw-YoloX still has room for further pruning and optimization.

In future work, the convolutional layers or fully connected layers with lower weights can be further eliminated through the pruning algorithm. At the same time, a new lightweight network can also be proposed with reference to the architecture of Sw-YoloX to achieve accurate detection of objects in complex sea surfaces. We hope that the resulted detector, namely Sw-YoloX, inspires future work and serves as a baseline for sea surface object detection. In addition, the idea of

Table 1
Average loss quantification from 20,000 to 30,000 iteration in ablation experiments.

YoloX-S (Baseline)	YoloX-M	Swin-S	ASPP	CBAM	Training strategies	Convergence speed (Iteration)	Average loss (Iteration from 20,000 to 30,000)
√						1.6×10^4	9.06
	√					1.5×10^4	7.05
√		√				1.5×10^4	7.72
√		√	√			1.2×10^4	5.88
√		√	√	√		1.1×10^4	4.10
√		√	√	√	√	$\mathbf{9.5\times10^3}$	**3.86**

Table 2
Ablation experiments on Xm-10000 validation set.

YoloX-S (Baseline)	CSP-DarkNet	Swin-S	CBAM	ASPP	Self-training Classifier	Training strategies	mAP (%)	AP50 (%)	Flops (GFLOPs)	Params (M)
√	√						40.6(+0.0)	78.9(+0.0)	26.9	9.1
√		√					45.2(+4.6)	81.8(+2.9)	140.06	57.74
√		√	√				46.9(+1.7)	82.7(+0.9)	155.02	66.83
√		√	√	√			47.1(+0.2)	83.2(+0.5)	163.50	80.99
√		√	√	√	√		50.9(+3.8)	84.2(+1.0)	166.82	97.51
√		√	√	√	√	√	**54.4(+3.5)**	**85.3(+1.1)**	166.82(↓)	97.51(↓)

Table 3
Comparison with the mainstream detector.

Method	Backbone	Average Precision/IoU			Average Precision/Area			Average Recall	F1-score	FPS
		0.50:0.95	0.50	0.75	Sma.	Med.	Lar.			
YoloX-S	CSPDarkNet	40.6	78.9	35.0	38.1	35.4	52.2	53.2	63.6	39.7
YoloX-M	CSPDarkNet	41.3	80.8	36.4	38.9	36.0	53.1	55.4	65.7	23.2
RetinaNet	ResNet-50	35.3	72.8	32.4	28.8	31.7	49.6	43.9	54.8	16.0
DETR	ResNet-50	28.6	77.9	14.0	31.0	23.9	27.9	44.7	56.8	14.8
Yolov5	CSPDarkNet	39.8	78.5	34.9	36.4	35.5	52.0	51.5	62.2	27.9
Cascade RCNN	Swin-S	44.7	80.0	37.5	42.6	34.8	53.5	68.5	73.8	4.8
YoloX-S + Swin	Swin-S	45.2	81.8	37.6	41.7	36.3	53.9	68.8	74.8	15.6
Ours (YoloX -S)	Swin-S	**54.4**	**85.3**	**42.5**	**45.0**	**37.7**	**55.1**	**72.0**	**78.1**	14.5(↓)
	Swin-B	**55.0**	**85.7**	**42.7**	**45.8**	**37.9**	**55.2**	**72.9**	**78.7**	9.3(↓)

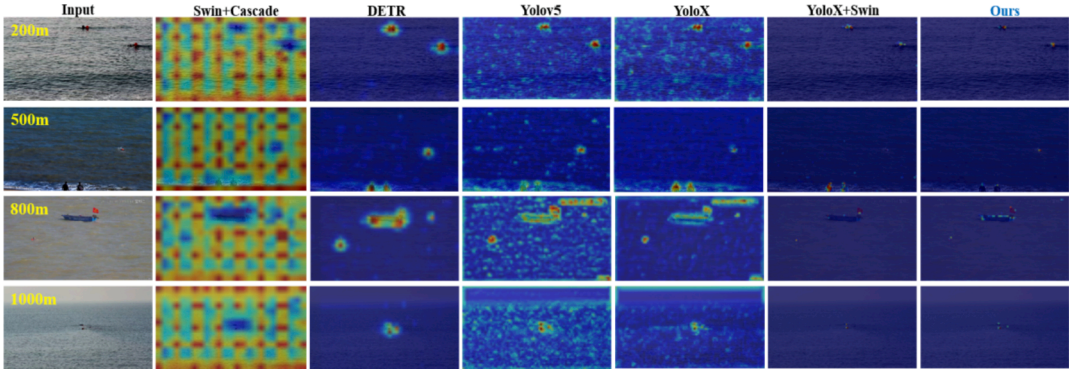

Fig. 13. Visualization of the attention performance of the detector. Detectors including Cascade RCNN, DETR, Yolov5, YoloX representing SOTA performance in their respective domains.

Fig. 14. Some visualization results from our Sw-YoloX on XM-10000, different category use bounding boxes with different color. The first row is the detection result of the CCD camera during the day, and the second row is the detection result of the infrared image at night. The Sw-YoloX detector has impressive high detection accuracy for objects blurred by motion, tiny objects, and objects with large scale spans.

combining transformer and anchor-free architecture can also be simultaneously applied to other object detection fields.

CRediT authorship contribution statement

Jiangang Ding: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing. **Wei Li:** Conceptualization, Methodology, Validation, Supervision, Project administration, Funding acquisition. **Lili Pei:** Methodology, Validation, Writing – original draft, Writing – review & editing. **Ming Yang:** Conceptualization, Software. **Chao Ye:** Methodology, Writing – original draft. **Bo Yuan:** Software, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the editor and anonymous reviewers for their constructive comments, which helped to improve the quality of this paper. This work was supported by the Fundamental Research Funds for the Central Universities, CHD [grant number: 300102249301; 300102249306] and the National Natural Science Foundation of China [grant number: 51978071].

References

Bai, X., Xu, S., Guo, Z., & Shui, P. (2021, October). *Enhanced local sparsity coefficient-based sea-surface floating target detection*. In 2021 International Conference on Control, Automation and Information Sciences (ICCAIS), Xi'an, China.

Cai, Z., & Vasconcelos, N. (2018, June). *Cascade r-cnn: Delving into high quality object detection*. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). *End-to-end object detection with transformers*. In European conference on computer vision, Springer, Cham.

Chalavadi, V., Jeripothula, P., Datla, R., & Ch, S. B. (2022). mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions. *Pattern Recognition, 126*, Article 108548. https://doi.org/10.1016/j.patcog.2022.108548

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence, 40*, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chernomorets, D. A., Golikov, V., Balabanova, T. N., Prokhorenko, E. I., Bolgova, E. V., & Chernomorets, A. A. (2021). Correlation properties of sea surface images on video stream frames. *International Journal of Nonlinear Analysis and Applications.* https://doi.org/10.22075/ijnaa.2021.25012.2883.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision, 88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. a*rXiv preprint arXiv:2107.08430.* https://doi.org/10.48550/arXiv.2107.08430.

Gu, T. (2019). Detection of small floating targets on the sea surface based on multi-features and principal component analysis. *IEEE Geoscience and Remote Sensing Letters, 17*, 809–813. https://doi.org/10.1109/LGRS.2019.2935262

Guo, W., Xia, X., & Xiaofei, W. (2014). A remote sensing ship recognition method based on dynamic probability generative model. *Expert Systems with Applications, 41*, 6446–6458. https://doi.org/10.1016/j.eswa.2014.03.033

Guo, H., Yang, X., Wang, N., & Gao, X. (2021). A CenterNet++ model for ship detection in SAR images. *Pattern Recognition, 112*, Article 107787. https://doi.org/10.1016/j.patcog.2020.107787

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December). *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile.

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition.

Heo, J., Wang, Y., & Park, J. (2022). Occlusion-Aware Spatial Attention Transformer for Occluded Object Recognition. *Pattern Recognition Letters, 159*, 70–76. https://doi.org/10.1016/j.patrec.2022.05.006

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017, October). *Focal loss for dense object detection*. In Proceedings of the IEEE international conference on computer vision.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014, September). *Microsoft coco: Common objects in context*. In European conference on computer vision, Springer, Cham.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021, October). *Swin transformer: Hierarchical vision transformer using shifted windows*. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada.

Liu, T., Pang, B., Zhang, L., Yang, W., & Sun, X. (2021b). Sea Surface Object Detection Algorithm Based on YOLO v4 Fused with Reverse Depthwise Separable Convolution (RDSC) for USV. *Journal of Marine Science and Engineering, 9*, 753. https://doi.org/10.3390/jmse9070753

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018, June). *Path aggregation network for instance segmentation*. In Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA.

Liu, T., Zhou, B., Zhao, Y., & Yan, S. (2021, July). *Ship detection algorithm based on improved YOLO V5*. In 2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE), Dalian, China.

Majid, S., Alenezi, F., Masood, S., Ahmad, M., Gündüz, E. S., & Polat, K. (2022). Attention based CNN model for fire detection and localization in real-world images. *Expert Systems with Applications, 189*, Article 116114. https://doi.org/10.1016/j.eswa.2021.116114

Peng, P., Yang, K. F., & Li, Y. J. (2022). Global-prior-guided fusion network for salient object detection. *Expert Systems with Applications, 198*, Article 116805. https://doi.org/10.1016/j.eswa.2022.116805

Qin, Z., Han, L., Shi, B., Zhang, X., & Xu, Y. (2021, April). *Improved Detection and Recognition of Sea Surface Ships Based on YOLOv3*. In The 4th International Conference on Electronics, Communications and Control Engineering, New York, NY.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems, 39*, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Salman, M. E., Çakar, G.Ç., Azimjonov, J., Kösem, M., Cedĭmoğlu,, İ., & H.. (2022). Automated prostate cancer grading and diagnosis system using deep learning-based Yolo object detection algorithm. *Expert Systems with Applications, 201*, Article 117148. https://doi.org/10.1016/j.eswa.2022.117148

Sun, X., Liu, T., Yu, X., & Pang, B. (2021). Unmanned surface vessel visual object detection under all-weather conditions with optimized feature fusion network in YOLOv4. *Journal of Intelligent & Robotic Systems, 103*, 1–16. https://doi.org/10.1007/s10846-021-01499-8

Sutikno, S., Wibawa, H. A., & Sasongko, P. S. (2018). Detection of Ship using Image Processing and Neural Network. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 16*, 259–264. https://doi.org/10.12928/telkomnika.v16i1.7357

Szpak, Z. L., & Tapamo, J. R. (2011). Maritime surveillance: Tracking ships inside a dynamic background using a fast level-set. *Expert systems with applications, 38*, 6669–6680. https://doi.org/10.1016/j.eswa.2010.11.068

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020, June). *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, Seattle, WA, USA.

Wang, J., Zhang, Z., Luo, L., Zhu, W., Chen, J., & Wang, W. (2021). SwinGD: A robust grape bunch detection model based on Swin Transformer in complex vineyard environment. *Horticulturae, 7, Article 492*. https://doi.org/10.3390/horticulturae7110492

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September). *Cbam: Convolutional block attention module*. In Proceedings of the European conference on computer vision (ECCV).

Xia, G. S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., & Zhang, L. (2018, June). *DOTA: A large-scale dataset for object detection in aerial images*. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA.

Xu, S., Ma, Y., & Bai, X. (2021, October). *Small Target Detection Method in Sea Clutter Based on Interframe Multi-feature Iteration*. In 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China.

Xu, S., Zheng, J., Pu, J., & Shui, P. (2018). Sea-surface floating small target detection based on polarization features. *IEEE Geoscience and Remote Sensing Letters, 15*, 1505–1509. https://doi.org/10.1109/LGRS.2018.2852560

Zhang, L., An, B., & Chen, Y. (2021a). Sea-surface Object Detection based on YOLO and Image Restoration. *World Scientific Research Journal, 7*, 25–33. https://doi.org/10.6911/WSRJ.202110_7(10).0004

Zhang, L., Li, Y., Chen, H., Wu, W., Chen, K., & Wang, S. (2022). Anchor-free YOLOv3 for mass detection in mammogram. *Expert systems with applications, 191*, Article 116273. https://doi.org/10.1016/j.eswa.2021.116273

Zhang, Y., Shu, Q., & Jiang, T. (2020). A GLRT-based polarimetric detector for sea-surface weak target detection. *IEEE Geoscience and Remote Sensing Letters, 19*, 1–5. https://doi.org/10.1109/LGRS.2020.3030711

Zhang, K., Shui, P. L., & Feng, Y. (2021b). Detection of Sea-Surface Small Targets Masked by Range Sidelobes of Large Objects. *IEEE Transactions on Aerospace and Electronic Systems, 58*, 1446–1461. https://doi.org/10.1109/TAES.2021.3116120

Zhou, H., & Jiang, T. (2019). Decision tree based sea-surface weak target detection with false alarm rate controllable. *IEEE Signal Processing Letters, 26*, 793–797. https://doi.org/10.1109/LSP.2019.2909584