International Journal of Computational Intelligence Systems (2023) 16:58
https://doi.org/10.1007/s44196-023-00235-4

RESEARCH ARTICLE

YOLF-ShipPnet: Improved RetinaNet with Pyramid Vision Transformer

Zhiruo Qiu[1] · Shiyang Rong[2] · Likun Ye[3]

Received: 11 January 2023 / Accepted: 1 April 2023

Abstract

In the field of ship detection, the intricate nature of ship images arises from a multitude of factors, including variations in ship orientation, color contrasts, and diverse shapes. These factors collectively contribute to the challenge of achieving high detection precision. Thus, it is necessary to investigate the application of advanced networks for ship image detection. In this paper, we have put forward an improved network called YOLF-ShipPnet, which utilizes a popular pyramid vision transformer with increased depth as the backbone for the RetinaNet network. To increase the model's generalization ability, You Only Look Once eXtreme's (YOLOX's) hue, saturation, and value (HSV) random augmentation technique is employed to simulate light and color effects on ship images during the construction of the network. Ablation experiments were conducted on the model with two popular datasets: High-Resolution Ship Collections 2016 (HRSC2016) and SAR Ship Detection Dataset (SSDD). The YOLF-ShipPnet network has been verified to improve detection precision and generalization ability in ship detection by 5.22% and 5.46%, respectively, compared to RetinaNet baseline, exhibiting strong robustness and high effectiveness. The proposed network is applicable to the field of fine-grained ship detection and achieves an accuracy improvement of 10.03% compared to the baseline network.

Keywords YOLF-ShipPnet · Ship detection · RetinaNet · PVTv2 · HSV · Depth effectiveness · Fine-grained classification

Abbreviation

YOLOX	You Only Look Once eXtreme
HSV	Hue, saturation, and value
HRSC2016	High-Resolution Ship Collections 2016
SSDD	SAR Ship Detection Dataset
YOLO	You Only Look Once
SSD	Single-stage detector
Faster RCNN	Fast region-based convolutional neural network
Mask RCNN	Mask region-based convolutional neural network
FPN	Feature pyramid network
R-CNN	Region-based convolutional neural network
NMS	Non-maximum suppression processing
Skew-NMS	Skew non-maximum suppression
PVT	Pyramid vision transformer
PVTv1	Pyramid Vision Transformer Version 1
PVTv2	Pyramid Vision Transformer Version 2
ResNet	Residual network
CNN	Convolutional neural network
RNN	Recurrent neural network
ViT	Vision transformer
TNT	Transformer iN transformer
RR-CNN	Rotated region-based convolutional neural network
RRoI	Rotated region of interest
FCN	Fully convolutional network
RPN	Region proposal network
R^2PN	Rotated region proposal network
MSCAF-Net	Multi-scale convolutional attention fusion network

Likun Ye, Zhiruo Qiu and Shiyang Rong have contributed equally to this work and should be considered co-first authors.

✉ Likun Ye
202030020427@mail.scut.edu.cn

[1] School of Management Engineering, Capital University of Economics and Business, Beijing 100070, People's Republic of China

[2] College of Information Science and Engineering, Northeastern University, Nanhu Campus, Shenyang 110006, People's Republic of China

[3] Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou International Campus, Guangzhou 511442, People's Republic of China

Published online: 19 April 2023

GHFormer-Net	Gradient harmonized transformer network
GHM-C	Gradient harmonized single-stage detector with context aggregation
GHM-R	Gradient harmonized regression
UP-ViTs	UP-vision transformer
IR-Net	Improved RetinaNet
SRA	Spatial reduction attention
FCN	Fully convolutional networks
SAR	Synthetic aperture radar
BIGSARDATA	SAR in big data era
AP	Average precision
mAP	Mean average precision

1 Introduction

Ship detection exhibits excellent application prospects in maritime trade, ship traffic control, port transportation, and national defense security. The conducted research on advanced networks is of much importance [1]. Although researchers have paid much effort into investigating ship detection, it remains challenging due to the various orientations of ships, color contrast within specific ship types, and the higher resolution requirements of the ship images. Many researchers have proposed networks for ship detection; however, they often do not incorporate a fine-grained study of specific ships. This is due to the fact that changes in external light intensity or the use of different colors on similar ships can greatly affect detection accuracy.

With the advancement in object detection algorithms for deep convolutional neural networks, two detection modes with different stage numbers have emerged and are distinguished by whether the region proposal is utilized [2]. One-stage algorithm for detection is characterized by direct access to positional coordinates and corresponding regression for target categories, which helps to reduce time complexity. Typical one-stage detectors are You Only Look Once (YOLO) [3], single-stage detector (SSD) [4], and RetinaNet [5]. Two-stage detection introduces the application of region proposal as front refinement, and regions of interest are classified and located in the latter stage [6]. Typical two-stage detectors are faster region-based convolutional neural network (faster RCNN) [7], mask region-based convolutional neural network (mask RCNN) [8], and feature pyramid network (FPN) [9]. In this paper, the base classifier is a one-stage RetinaNet, which operates through the use of a robust focal loss method. As a result, it is able to combine the benefits of a one-stage detector, such as fast speed, with those of a two-stage detector, such as high detection precision. With respect to computer vision, the horizontal frame target detection algorithm based on a region-based convolutional neural network (R-CNN) shows rich application scenarios in the classification and detection of remote sensing images. However, it may generate background noise when detecting targets with large aspect ratios. The omission of detection targets is likely to occur when performing non-maximum suppression (NMS) processing. Therefore, in recent years, a significant number of scholars have devoted themselves to the study of rotating region proposal algorithms designed by introducing anchor frame rotation angle parameters, which effectively retain the target orientation feature information as opposed to background noise [10] and improve tilt target detection accuracy with the help of skew non-maximum suppression (skew-NMS).

In this paper, we introduce YOLF-ShipPnet, a novel network architecture that utilizes the state-of-the-art Pyramid Vision Transformer Version 2 (PVTv2) as the backbone network for the RetinaNet base classifier. The rotating frame is used for global and fine-grained ship image detection. Distinguished from foregoing works, the depth of the network is increased, and we perform random data augmentation using YOLOX's HSV to improve its fine-grained classification capability for the ship dataset.

Our main contributions are:

(1) We propose the YOLF-ShipPnet network, which will be used in ship detection for commercial and military purposes. In the YOLF-ShipPnet network, we introduce the application of popular PVTv2 architecture to the construction of the backbone of the RetinaNet base classifier, fully exploring its depth effectiveness. Also, YOLOX's HSV is led into the field to manipulate random data augmentation on the ship datasets.

(2) We demonstrate that the detection precision of YOLF-ShipPnet outperforms the conventional scheme and, as the depth of the network's depth increases, it gradually shows better performance. After random data augmentation, it is shown that the model can have better generalization abilities and is effective in enhancing the designed network. With the deepened network and effective data augmentation, the proposed YOLF-ShipPnet network is applicable for fine-grained ship detection and transcends the baseline by a large margin.

The remaining parts of this paper are developed as follows. In Sect. 2, we summarize the related work concerning the development of transformer architecture and the evolution of rotated frame detection. Also, existing methods with different functionalities for ship detection are briefly summarized. In Sect. 3, we demonstrate the detailed construction of the PVTv2 backbone and the mechanism of YOLOX's HSV for data augmentation. In Sect. 4, we provide the results of the validation and ablation experiments on two datasets to prove the depth effectiveness of the PVTv2 network and the validity of the YOLOX's HSV data augmentation strategy.

It is proved that our proposed network can be applied to fine-grained ship detection. We also compare our network with other advanced networks to demonstrate its superiority. The conclusion of our research is given in Sect. 5. Reflection on our current work and future research prospect are demonstrated in Sect. 6.

2 Related Work

Transformer architecture is used in place of residual network (ResNet) [11] to form the backbone of the RetinaNet [5] network. In 2017, the Google team first proposed the transformer model, which abandoned the traditional convolutional neural network (CNN) and recurrent neural network (RNN) architecture, making the entire network structure composed completely of the attention mechanism. Transformer is the pioneer in transduction mode design with the functionality of calculating primary substitution of input and output based on the principle of self-attention [12], which is widely used in the computer vision field to manipulate image detection tasks.

The development of transformer structure can be divided into three stages, with its enhancement in function and boost in efficiency. In the first stage, the emergence of the attention mechanism enhanced the traditional CNN network with an optimized fusion of functionality. Bello et al. (2019) introduced a self-attention transformer model with two-dimensional architecture that combines convolutional feature maps and feature mapping generated by self-attention[13]. By leveraging a global perspective to analyze the entire image, the model outperforms a CNN that is limited to processing only local information, resulting in a significant improvement in accuracy for image detection tasks. Later, the transformer architecture reached the level of complete replacement of CNN with its excellent testing performance, due to the attention mechanism being used in image detection. For example, Dosovitskiy et al. (2020) introduced a vision transformer (ViT), which is directly applied to a series of image patches without any reliance on CNNs, demanding less computational power and achieving better detection performances as compared to first-class prototypes of convolutional neural networks [14]. Since then, based on ViT, a series of methods have emerged to improve and optimize the transformer structure for enhanced efficiency and effectiveness. Han et al. (2021) issued a brand-new vision architecture of a transformer called Transformer iN Transformer (TNT), which divides the local patch into sub-patch. It integrates both information and generates representation at patch granularity with the help of the outer transformer [15]. Wang et al. (2022) proposed the pyramid vision transformer (PVT), which can obtain higher output resolution when trained on denser regions of an image and reduce the cost of large feature maps' computations by employing a progressively contracting pyramid [16]. To conclude, the transformer architecture integrates the attention mechanism into the construction of the forward feedback network, which has better parallelism and global optimization capabilities. It significantly improves the execution of dense image detection in terms of efficiency and accuracy, exhibiting broad application prospects in multimodality and object identification.

Rotated frame detection is widely used when conducting ship detection. For example, Liu et al. (2016) proposed a novel ship rotation bounding box that accurately captures the true shape of ships embedded in complex backgrounds. The method involves generating representative candidate regions using a closed-form region approach, which outperforms traditional horizontal frame target detection schemes [17]. Hu et al. (2017) introduced the rotated region-based convolutional neural network (RR-CNN), which integrates a rotated region of interest (RRoI) pooling layer and a regression model equipped with a rotating bounding box to accomplish ship detection. It excels in the extraction of key features within rotated regions and thus can capture the inclined detection targets more precisely [18]. Liao et al. (2022) proposed a novel rotated region proposal network (R^2PN) to form multi-directional proposals featured by the angle information of the orientation of ships, which adopts a pooling layer activated by rotated region of interest to manipulate key feature extraction and uses bounding boxes regression to increase the accuracy of the inclined ship region proposals. The proposed network model achieves superior performance in ship detection, particularly for ships with multiple orientations [10].

Existing methods for ship detection hardly investigate into the depth effectiveness of the feature extraction networks and have limited generalization or fine-grained detection ability. Liu et al. (2023) proposed multi-scale convolutional attention fusion network (MSCAF-Net), a framework with PVTv2-B2 backbone for detecting camouflaged objects that focuses on learning features that are sensitive to context at different scales. While the efficacy of the network is evident for the reference datasets, its potential for profound exploration is constrained due to the utilization of a solitary layer of the PVTv2 network [19]. Sun et al. (2022) proposed gradient harmonized transformer network (GHFormer-Net), which utilizes PVTv2-B1 as the backbone network and incorporates gradient harmonized single-stage detector with context aggregation (GHM-C) and gradient harmonized regression (GHM-R) loss functions to improve fruit detection in low-light conditions. The experimental results demonstrate the effectiveness of the model, but the study only investigates the first layer of the network and does not explore the potential benefits of using deeper layers of PVTv2 [20]. Hao et al. (2021) proposed a unified network called UP-Vision Transformer (UP-ViTs) for systematic pruning of vision transformers and their extensions. However, their study revealed that when using UP-ViTs to prune PVTv2-B2 into UP-PVTv2-B1 on ImageNet–1 k validation, it increased the accuracy of PVTv2-B1, but was less effective than the deepened PVTv2-B2. This suggests that the lack of depth effectiveness in the network's design may have contributed to the suboptimal results [21]. Liu et al. (2017) proposed RR-CNN, which features an intensive task approach for non-maximum suppression among different classes, overcoming challenges in detecting strip-like rotated assembled objects. The network outperforms baseline models by a significant margin, but its compatibility with other rotation-based frameworks is limited [18]. Yan et al. (2019) proposed an innovative data augmentation method that utilizes simulated remote sensing ship images to augment positive training samples, thereby improving the quality of the training set. Experimental results on the ship detection dataset using Faster R-CNN demonstrate the effectiveness of the approach. However, the method is only applicable to a limited number of ship models and does not possess the ability of fine-grained classification [22]. Zhao et al. explores low-resolution fine-grained object classification and proposes a new model, which combines feature equilibrium principle and progressive interaction theory. It improves the accuracy of network when applied to low-resolution image detection, but when it comes to fine-grained classification, it only increases the baseline model by 3.4%, which is not satisfactory enough [23].

3 Model and Network

We propose a brand-new network called YOLF-ShipPnet, which incorporates deepened PVTv2 into the construction of the backbone of the RetinaNet network. The network structure of YOLOF-ShipPnet is demonstrated in Fig. 1. The RetinaNet network is a comprehensive baseline network consisting of a backbone, a neck and a head consisting of two subnets. The backbone network, namely, PVTv2, implements the convolutional feature mapping over the target image and is treated as a non-self-convolutional network. The neck part of the network is the feature pyramid network (FPN), which is utilized for multi-scale feature integration. The output of the FPN is then fed into the head part of the network, which comprises two subnets, namely the class subnet and the box subnet. Two subnets are distinguished by branch functions, one for classification and the other for regression. Specifically, the first subnet carries out object classification with a convolutional technique targeted at the backbone output, and the second subnet implements convolutional regression with the help of a bounding boxes. In YOLF-ShipPnet, considering the need for higher precision accompanied by the deepened network, we choose to use PVTv2 for its deepened network architecture.

3.1 Backbone Network: PVTv2 with Transformer Architecture

Since the introduction of ViT, there has been a large number of researches on vision transformers, roughly along two main directions: one is to improve the effectiveness of ViT in image classification; the other is to apply ViT to other image tasks, such as image segmentation and target detection. The PVT [24] introduced in this paper belongs to the latter. PVT is a simple, non-convolutional backbone that can be applied for many prediction tasks containing dense images. Unlike ViT, which employs a pure transformer architecture, PVTv2 incorporates a hybrid architecture that combines both transformer and convolutional neural network (CNN) structure. PVT overcomes the difficulty of applying a transformer to various task-oriented predictions with complex partitions, exhibiting better feature extraction performance.

PVT was originally proposed by Wang Wenhai and Xie Enze at Nanjing University and has undergone two generations of evolution, Pyramid Vision Transformer Version 1 (PVTv1) and PVTv2 [25]. Generally, PVTv1 has three main limitations. Firstly, PVTv1 treats the images as a series of non-overlapping facets, which somewhat loses the

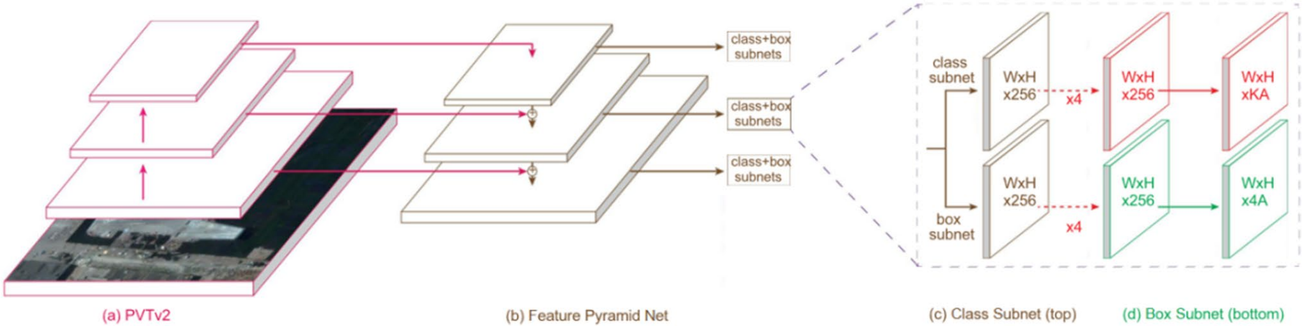


Fig. 1 The architecture of YOLF-ShipPnet network diagram of the model

characteristic of partial continuity of the images, limiting its application for fined-grained ship classification. Secondly, the size of the encoding of position in PVTv1 architecture is pre-determined and invariant for processing images of discretional size. However, the most significant drawback of PVTv1 is that its network architecture has limited depth, which harms the precision of image classification. Taking into account the fact that the detection precision of our baseline network maintains at a low level, we choose to use deepened PVTv2 with depth-wise convolution as the backbone to trade off a lightweight network for higher precision, as shown in Fig. 2. It can detect dense ship images and perform feature extraction of local features more smoothly for fine-grained classification, which is satisfactory for application on ship image detection (Table 1).

Different layers of the PVTv2 network (B0–B5) are constructed by changing the following hyperparameters:

S_i : The stride in stage i for overlapping patch embedding.

C_i : The number of channels in the output of the ith stage.

L_i : The number of encoded overlapping in the ith stage.

R_i Deceleration ratio of the ith stage Spatial Reduction Attention (SRA).

P_i : Mean pool size of linear SRA in the ith stage.

N_i : Head number of the self-attention network in the ith stage.

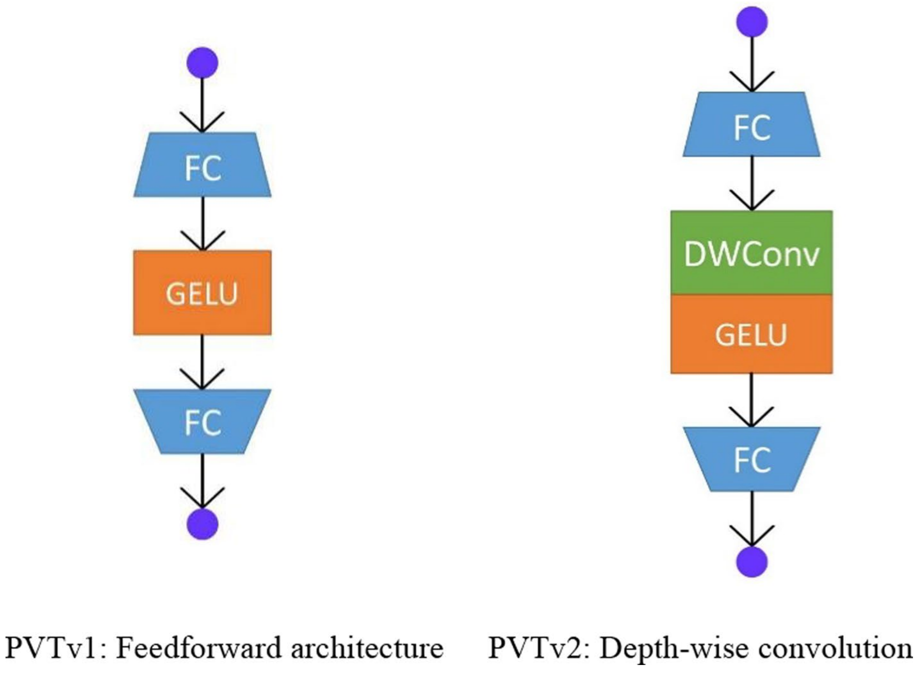
E_i : The expansion ratio of the ith stage feedforward layer.

The design of the PVTv2 network adheres to the principle that is used to construct ResNet, where the number of channel dimensions increases as the layers deepen, leading to a theoretical improvement in the detection precision of PVTv2 with increased depth. So, we suppose that the effect

Table 1 Overall network architecture of PVTv2

	Output size	Layer label	Pyramid Vision Transformer v2		
			B0	B3	B5
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Overlapping patch embedding	$S_1 = 4$ $C_1 = 32$	 $C_1 = 64$	
		Transformer encoder	$R_1 = 8$ $N_1 = 1$ $E_1 = 8$ $L_1 = 2$	$R_1 = 8$ $N_1 = 1$ $E_1 = 8$ $L_1 = 3$	$R_1 = 8$ $N_1 = 1$ $E_1 = 4$ $L_1 = 3$
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Overlapping patch embedding	$S_1 = 2$ $C_2 = 64$	 $C_2 = 128$	
		Transformer encoder	$R_2 = 4$ $N_2 = 2$ $E_2 = 8$ $L_2 = 2$	$R_2 = 4$ $N_2 = 2$ $E_2 = 8$ $L_2 = 3$	$R_2 = 4$ $N_2 = 2$ $E_2 = 4$ $L_2 = 6$
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Overlapping patch embedding	$S_1 = 2$ $C_3 = 160$	 $C_3 = 320$	
		Transformer encoder	$R_3 = 2$ $N_3 = 5$ $E_3 = 4$ $L_3 = 2$	$R_3 = 2$ $N_3 = 5$ $E_3 = 4$ $L_3 = 18$	$R_3 = 2$ $N_3 = 5$ $E_3 = 4$ $L_3 = 40$
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Overlapping patch embedding	$S_1 = 2$ $C_4 = 256$	 $C_4 = 512$	
		Transformer encoder	$R_4 = 1$ $N_4 = 8$ $E_4 = 4$ $L_4 = 2$	$R_4 = 1$ $N_4 = 8$ $E_4 = 4$ $L_4 = 3$	$R_4 = 1$ $N_4 = 8$ $E_4 = 4$ $L_4 = 3$

Fig. 2 Comparison of the depth of two versions of PVT


FC
GELU
FC
FC
DWConv
GELU
FC
PVTv1: Feedforward architecture
PVTv2: Depth-wise convolution

of depth-wise convolution of PVTv2 is still manifested in the HRSC2016 dataset. Taking into account the matching of the dataset, network complexity, and computational cost, we choose to deepen the layer of our network from B1 to B5.

3.2 Neck: Feature Pyramid Net

We apply FPN to the neck part of the network. The origin of the idea of FPN is the image pyramid in traditional image processing [26]. It aims to enhance the robustness of the model when the input images are of different sizes or various objects exist in the scenarios of target detection. FPN adopts the multi-scale feature fusion method, which considers global and local features during target detection. FPN enhances the conventional convolutional network with novel transverse connections and top-down pathways, thereby constructing a comprehensive, multi-dimensional feature pyramid from singular input images. Each layer of the pyramid can be used to detect objects with various dimensions. FPN is a powerful technique for improving multi-dimensional predictions from fully convolutional networks (FCN). It has been used to generate a range of subsequent networks such as region proposal network (RPN), deep mask object proposal, and two-stage detectors like faster R-CNN and mask R-CNN.

3.3 Head: Classification and Regression of Rotating Frame Networks

The objective of focal loss [27] is to address the issues of imbalanced class distribution and the resulting challenges in classification, particularly when the dataset contains a large number of easy background samples and a few foreground samples that are challenging to classify. Focal loss mitigates these problems and enhances the accuracy of detection by modifying the cross-entropy function, increasing the category weights α and the sample difficulty weight modulating factor $(1 - p_t)$. The focal loss function takes the following form:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma}\log(p_t). \tag{1}$$

In formula 1, $-log(p_t)$ stands for the initial cross-entropy loss function, α is the weight parameter between categories, $(1 - pt)^{\gamma}$ is the modulating factor between simple and complex samples, and γ is the focusing parameter.

One common loss function used for bounding box regression in the head part of object detection models is the L1 loss. In ship detection, L1 loss is particularly useful for accurately predicting the coordinates of the bounding box around a ship. By minimizing the mean absolute difference between the predicted and actual bounding box coordinates, L1 loss helps to improve the accuracy of the ship detection model. The formula for L1 loss is as follows:

$$L1 = \sum_{i=1}^{n} \left| y_i - f(x_i) \right|, \tag{2}$$

where y_i denotes the true label and $f(x_i)$ indicates the predicted label.

3.4 Data Augmentation Strategy: HSV [28]

HSV is a color space put forward by a.r. Smith, in 1978 inspired by the intuitive properties of the color [29], also known as the hexagonal model. In the field of data augmentation in ship detection, it is used to enhance the color contrast of the image by adjusting the intensity ratio of hue, saturation, and value channel [30]. It extracts and manifests the feature color space of the ship image corresponding to the change of light state and external color of ships. The color space of the HSV model can be visualized using a cone, as shown in Fig. 3, accompanied by target images with contrast brightness and colors. H (hue) in the cone represents the phase angle of the color, with a range of 0° to 360°. S (saturation) stands for a ratio value, which is correlated with the purity of a specific color. Following the direction of the S arrow, the purity of color witnessed a significant increase. V (value) represents the brightness of the color, ranging from 0 to 1. V value of the cone ranges from 0 at the black bottom

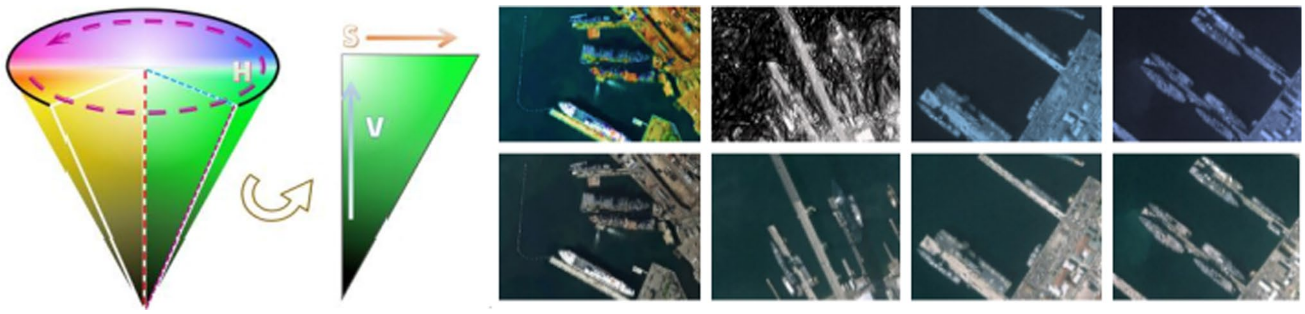


Fig. 3 HSV augmentation for color-oriented data augmentation

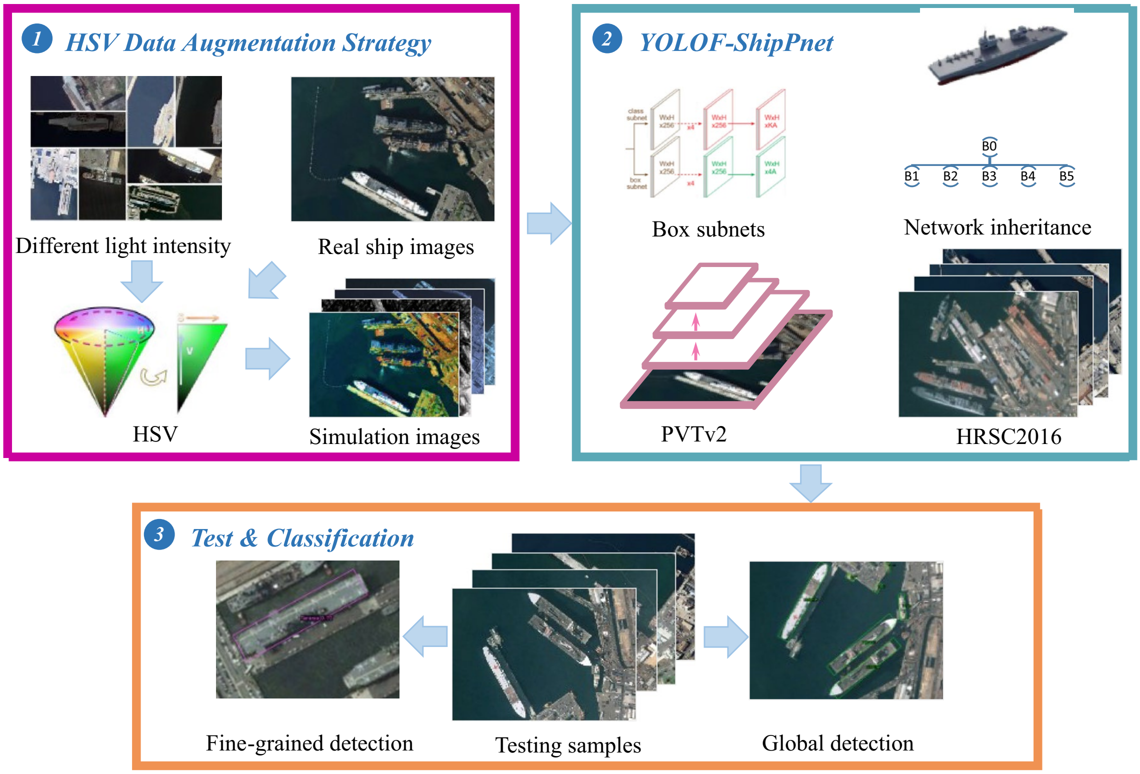


Fig. 4 Design flowchart of YOLOF-ShipPnet

point to 1 at the top white point, with higher values indicating greater brightness.

In the YOLF-ShipPnet network, HSV is used to manipulate data augmentation on the ship dataset by adequately adjusting the three-channel values of the color space, aiming to simulate the background state of ship images under various lighting conditions and also to adjust the brightness, colors, and other factors of the image to reduce the sensitivity of our proposed model to ship colors. The data augmentation strategy eliminates disturbance factors such as potential changes in light intensity and color differences of a specific ship, which significantly improves the local feature extraction ability and the robustness of the network. The efficiency of training and the performance of our network is also further enhanced with the help of the YOLOX's HSV random data augmentation technique.

3.5 YOLOF-ShipPnet

The model of YOLOF-ShipPnet is shown in Fig. 4. HSV color space is employed as a data augmentation technique to augment the light effects on ships and the external colors of certain ships. This approach produces a set of synthesized images by leveraging the HRSC2016 dataset, which helps in improving the model's training. We select PVTv2 as the backbone network, which is an enhanced transformer network with depth inheritance. After testing and refinement, our proposed network is expected to carry out global ship image and fine-grained classification.

4 Experiment and Analysis

4.1 Dataset

Ablation experiments are performed on the famous remote sensing dataset HRSC2016 [31] and the synthetic aperture radar (SAR) dataset SSDD [32] to validate the effectiveness of our proposed YOLF-ShipPnet network.

Northwestern Polytechnical University published the HRSC2016 [31] dataset in 2016. The set issued by Google Earth contains 1,061 images with 4 classes and 19 subclasses, covering 2976 instances of ships. The training, validation, and test sets incorporate 436, 181, and 444 images, respectively. The image sizes of HRSC2016 range from 300×300 to1500×900, with the majority of the images having sizes greater than 1000×600. The dataset covers 27 types of remote sensing ground objects. For a fair comparison with other networks, only ship objects are selected for our experiments.

The SSDD dataset [32] was first unveiled at the SAR In Big Data Era (BIGSARDATA) conference in Beijing in

2017. The set contains 1160 images and 2456 ships, with an average number of ships of 2.12 per image. The image sizes are around500×500. The set is partitioned into the training, validation, and test sets, with a random ratio of $7:1:2$. This dataset contains SAR images specially used for ship detection with a single ship type.

Our ablation experiments use average precision (AP) and mean average precision (mAP) for evaluation of the performance of YOLF-ShipPnet. In MMROTATE, the general definition of AP is the gross area below the precision–recall curve. *Precision* measures the accuracy of prediction, while *recall* reflects the proportion of positive samples that are successfully retrieved. So to calculate them, the quantities that shall be known in advance aretp, the number of correctly determined positive samples, and fn andfp, which are incorrectly determined negative and positive samples. Formulas 3 and 4 illustrate the calculation processes of *precision* and *recall*:

$$Precision = \frac{tp}{tp + fp}. \tag{3}$$

$$Recall = \frac{tp}{tp + fn}. \tag{4}$$

After plotting corresponding data points of *precision* and *recall* into a curve, the value of AP can be calculated by integrating the area beneath the curve. Then, mAP is derived by averaging over the $AP$of each epoch.

4.2 Configuration of Ablation Experiment and Model Training

All the experiments are conducted on a deep-learning server. The detailed configuration is shown in Table 2.

Table 2 Configuration of parameters

Parameter	Configuration
Central processing unit (CPU)	12 core Intel(R) Xeon (R) Platinum 8255C
Graphic processing unit (GPU)	RTX 2080 Ti
Operating system	Ubuntu 18.04
Programming language	Python 3.8
GPU accelerator	CUDA 10.2

Our experiments are trained on the HRSC2016 dataset. The optimizer of YOLF-ShipPnet is AdamW. The momentum coefficient is 0.9 and the weight decay coefficient is equal to 0.05. The original learning rate of the model is 0.0001. The significance of weight decay is that the learning rate gradually reduces during training and converges quickly. Also, a threshold of 72 epochs is set to ensure the convergence of the network.

4.3 Ablation Experiments

The YOLF-ShipPnet we propose employs PVTv2 as the backbone, and YOLOX's HSV is used for random data augmentation. To analyze the extent to which the proposed network elevates the performance of the model, we design a set of ablation experiments.

RetinaNet serves as the baseline object detection framework in our experimental setup, acting as a standard of comparison. It is composed of a backbone network and an FPN. The backbone network extracts image features, while the FPN produces feature maps of varying resolutions for further regression and classification. To demonstrate the effectiveness of our backbone network, we compare the performance of PVTv2 with the baseline. For depth effectiveness experiments, we explore the efficacy of PVTv2 layer by layer, comparing their *mAPs* and investigating the general trend of *mAPs* with increased depth. Random data augmentation based on HSV is also performed and compared with the baseline and the PVTv2 layer with the best performance. Additionally, we assess the fine-grained classification capability of our network and use the baseline for comparison. Finally, the generalization ability of YOLF-ShipPnet over different datasets is evaluated by replacing the original dataset with SSDD.

4.3.1 Effectiveness of PVTv2

In this section, we use only PVTv2 as the backbone for the baseline, referred to as PVTv2-B0, to evaluate the effectiveness of PVTv2. Table 3 below presents the results of the experiment after 72 epochs.

According to the results from Table 3, it can be seen that the feature extraction accuracy reaches52.50%, indicating that our baseline is reliable. Compared with the values of AP under different categories, the PVTv2 group generally has higher AP values than the RetinaNet group, and the mAP is finally improved by 0.41%. Therefore, it can be concluded that PVTv2

Table 3 Ablation experiments of PVTv2 on HRSC2016

Method	*mAP*	*AP*50	*AP*60	*AP*70	*AP*80	*AP*85	*AP*90	*AP*95
RetinaNet	0.5250	0.8380	0.8050	0.7080	0.3990	0.2140	0.1080	0.0200
PVTv2-B0	0.5291	0.8500	0.8350	0.7240	0.3860	0.2010	0.1030	0.0110

effectively enhances the ability of the feature extraction of the YOLF-ShipPnet.

4.3.2 Depth Effectiveness of PVTv2

In this section, PVTv2-B0 is used as the control group. We inherit PVTv2-B0 and modify the weights to obtain the B-series networks based on PVTv2 to prove the depth effectiveness of PVTv2. To observe the changes in indicators, Table 4 lists the effect of B0, B3, and B5: the control group, the group with moderate effect, and the group with the best effect.

According to Table 4, it can be seen that there is an upward trend in the *mAP* from B0 to B5. The overall detection accuracy of the model was improved by 2.27% and 2.78% in each step, with a total increase of 5.05%. Comparing the average accuracy of each method in Table 4, the mean precision level shows an overall upward trend from PVTv2-B0 to PVTv2-B5. PVTv2-B5 has the highest mean average precision, which is expected and demonstrates the depth effectiveness of PVTv2.

4.3.3 Effectiveness of HSV Data Augmentation on Ship Dataset

In this section, we aim to verify the contribution of data augmentation to the detection performance of our model. Based on the network involved in the above experiments, we only add *YOLOXHSVRandom* to randomly adjust the hue, saturation, and value of ship images.

Considering the inheritance relationship among the networks, our experiment adds augmentation to the baseline only to verify that it can improve the model detection accuracy without adding the PVTv2. Then, we add HSV strategy to the PVTv2 B5 to verify that the combination of deepened PVTv2 and data augmentation jointly contribute to the model performance.

Table 5 presents the results of two groups of experiments based on the baseline, with or without data augmentation. The *mAP* value increases from 52.50 to 53.84% before and after the augmentation, showing a leap of 1.34% in average precision. It indicates that adding data augmentation alone can improve the model effect.

Table 6 shows the results of the three experimental groups: PVTv2-B0, PVTv2-B5, and PVTv2-B5_Aug. By comparing the results of PVTv2-B5 with and without data augmentation, the *mAP* increases by 0.17%. We can also find that *mAPs* of PVTv2-B5 and PVTv2-B5_Aug are 5.05% and 5.22% higher than PVTv2-B0. The results demonstrate our model's robustness and show that the augmentation can further improve the detection performance of the model on the PVTv2, verifying the effectiveness of the data augmentation strategy.

4.3.4 Effectiveness of Fine-Grained Classification Experiment for Ship Dataset

The above ablation experiments validate the effectiveness of PVTv2-B5_Aug, which is 5.63% more accurate than the baseline.

In this section, the baseline and PVTv2-B5_Aug are used to detect 31 subclasses of ships to examine the ability of the YOLOF-ShipPnet network for fine-grained ship detection. Table 7 shows the results of the fine-grained experiments of baseline and PVTv2-B5_Aug on HRSC2016.

Table 4 Ablation experiments of the depth effect of PVTv2 on HRSC2016

Method	*mAP*	*AP*50	*AP*60	*AP*70	*AP*80	*AP*85	*AP*90	*AP*95
PVTv2-B0	0.5291	0.8500	0.8350	0.7240	0.3860	0.2010	0.1030	0.0110
PVTv2-B3	0.5518	0.8600	0.8440	0.7270	0.4940	0.2880	0.0780	0.0100
PVTv2-B5	0.5796	0.8570	0.8400	0.7370	0.4970	0.2910	0.1120	0.0910

Table 5 Ablation experiments of HSV data augmentation on HRSC2016

Method	Data Aug	*mAP*	*AP*50	*AP*60	*AP*70	*AP*80	*AP*85	*AP*90
RetinaNet		0.5250	0.8380	0.8050	0.7080	0.3990	0.2140	0.1080
RetinaNet	√	0.5384	0.8440	0.8210	0.7120	0.4400	0.2120	0.1070

Table 6 Ablation experiments of HSV data augmentation with increased depth on HRSC2016

Method	Data Aug	*mAP*	*AP*50	*AP*60	*AP*70	*AP*80	*AP*85	*AP*90
PVTv2-B0		0.5291	0.8500	0.8350	0.7240	0.3860	0.2010	0.1030
PVTv2-B5		0.5796	0.8570	0.8400	0.7370	0.4970	0.2910	0.1120
PVTv2-B5	√	0.5813	0.8590	0.8470	0.7450	0.5100	0.2890	0.1370

Table 7 Fine-grained classification experiments on HRSC2016

Method	Data Aug	*mAP*	*AP*50	*AP*60	*AP*70	*AP*80	*AP*85	*AP*90
RetinaNet		0.1958	0.3140	0.2920	0.2540	0.1690	0.0960	0.0230
PVTv2-B5	√	0.2961	0.4330	0.4130	0.3800	0.2800	0.1950	0.0880

Table 8 Ablation experiments on the SSDD dataset

Method	Data Aug	*mAP*
RetinaNet		0.7151
PVTv2-B5	√	0.7697

From Table 7, it can be known that both the baseline and PVTv2-B5_Aug can be used for fine-grained detection. PVTv2-B5_Aug performs better on fine-grained detection, showing an enormous leap of 10.03%.

4.3.5 Performance of YOLF-ShipPnet on the SSDD Dataset

In this section, the dataset is replaced with SSDD to verify the generalization ability of PVTv2-B5_Aug(YOLF-ShipPnet). Table 8 shows the *mAP*of baseline and PVTv2-B5_Aug on the SSDD dataset.

In comparison to the detection accuracy between the two groups, PVTv2-B5_Aug showed an improvement in performance of 5.46%. This reflects the strong generalization ability of our proposed network and indicates its potential application in other datasets.

4.3.6 Loss Curve for Training

The following plots are the training loss of the above ablation experiments. In these plots, the networks reach convergence after 72 epochs (Figs. 5, 6, 7).

4.4 Visualization of the Result

We visualize the results of baseline and PVTv2-B5_Aug on HRSC2016 to intuitively compare the detection effect before and after the model improvement.

As shown in Fig. 8, part (i) shows the visualization results of the baseline and part (ii) demonstrates the results of PVTv2-B5_Aug.

In Fig. 8, some ships that are not identified with the baseline detector are identified by PVTv2-B5_Aug, indicating that PVTv2-B5_Aug shows a better detection performance than baseline.

In Fig. 9, the detection precision of PVTv2-B5 is higher than the baseline for the same ship, indicating that PVTv2-B5_Aug can identify ships more accurately. From Fig. 10, we discover that the baseline and PVTv2-B5_Aug can detect multiple classes of ships, and PVTv2-B5 performs better in terms of identifiability and accuracy.

At the same time, PVTv2-B5_Aug on the SSDD dataset also achieves a better detection effect, which verifies the generalization ability of the model. Figure 11 demonstrates the visualization results, which show the effectiveness of PVTv2-B5_Aug on the SSDD dataset.

4.5 Comparisons Among the Advanced Networks

Table 9 shows the performance of YOLF-ShipPnet and other networks on the HRSC2016 dataset. It is observed that the *mAP* of our proposed network shows a significant leap compared to other ship detection models, further verifying the depth effectiveness of the PVTv2 backbone and the excellent performance of the YOLOX's HSV random data augmentation strategy. Then, it is observed that the networks listed can only perform global ship detection. However, our network extends the function of fine-grained classification for more specific purpose ship classification.

5 Conclusion

This paper proposes a rotation ship detection network YOLF-ShipPnet based on RetinaNet, which innovatively introduces the application of deepened PVTv2 network and HSV strategy for data augmentation. Generally, the backbone network utilizes the popular transformer structure along with the deepened PVTv2 network, which focuses on exploring the depth effectiveness in the context of ship image detection. The neck part employs the FPN model for multi-scale fusion of features. The head part takes in the combined characteristics and performs classification and regression of the rotating frame. To further improve the generalization and fine-grained classification abilities of our proposed network, we applied the random data augmentation strategy HSV on the ship datasets to complement the PVTv2 network and achieve a more cohesive and effective performance. Through a series of validation and ablation experiments, it has been confirmed that the YOLF-ShipPnet exhibits promising depth effectiveness for the detection of ships. Furthermore, the efficacy of the HSV data augmentation strategy has been demonstrated, resulting in significantly improved accuracy compared to the baseline model. The use of this strategy also makes the models less sensitive to such external factors as color or light changes. In addition, the YOLF-ShipPnet has demonstrated exceptional generalization abilities,

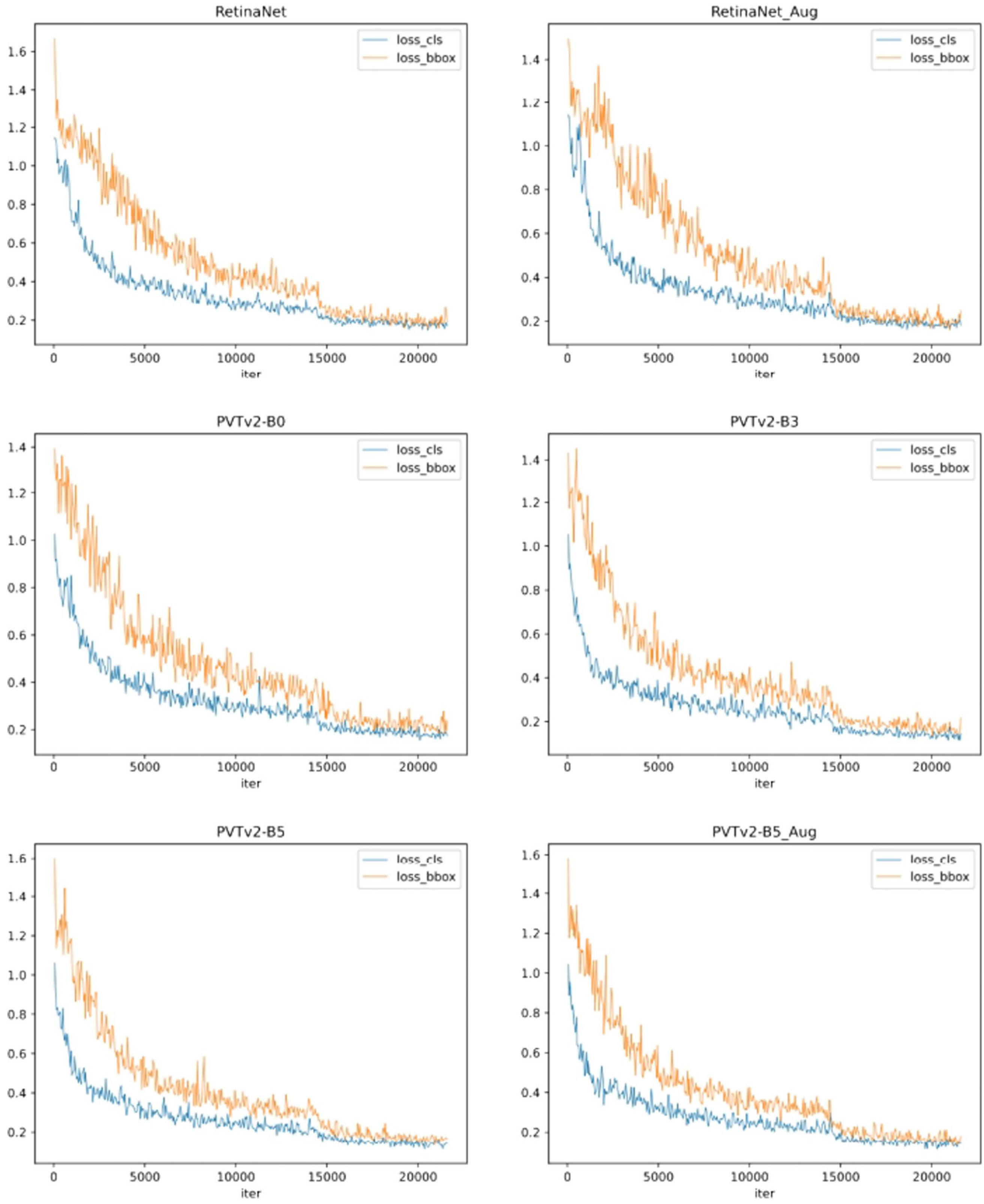


Fig. 5 Loss curve for ablation experiments on HRSC2016

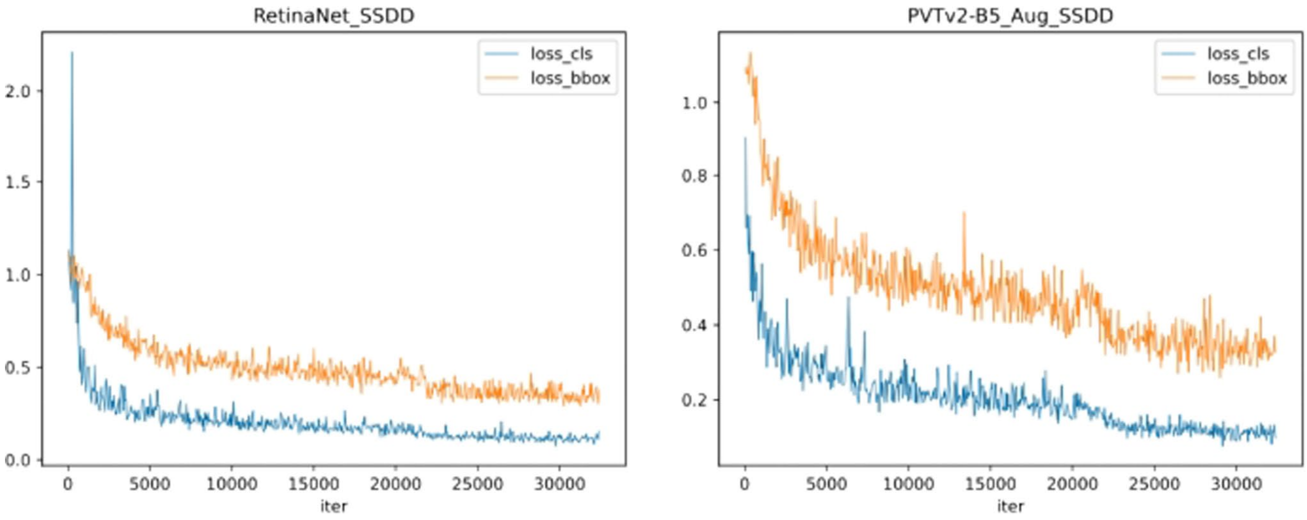


Fig. 6 Loss curve for ablation experiments on SSDD

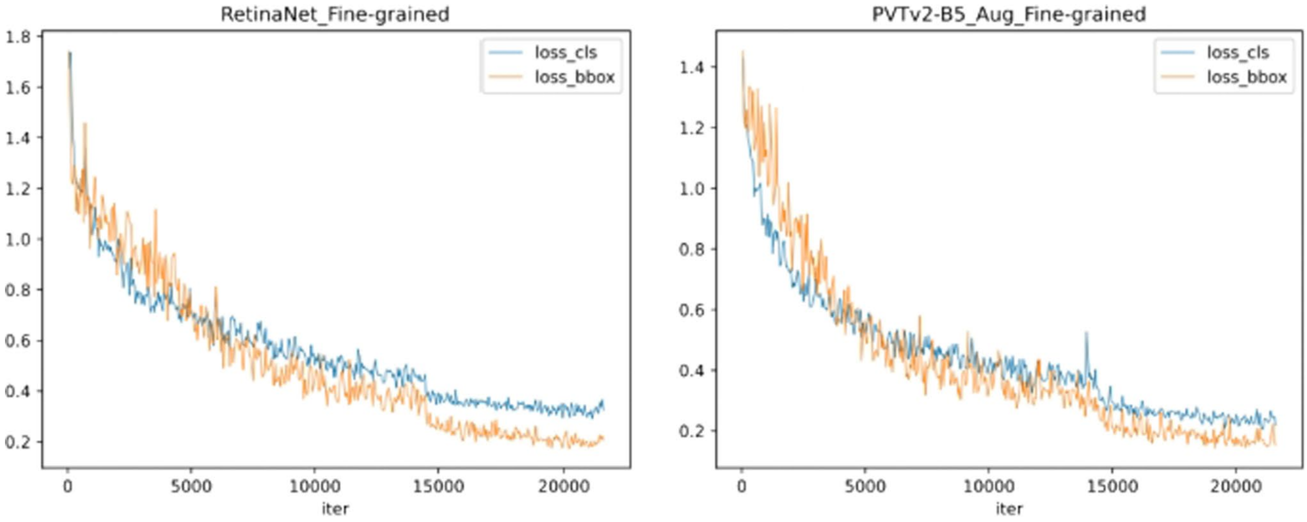
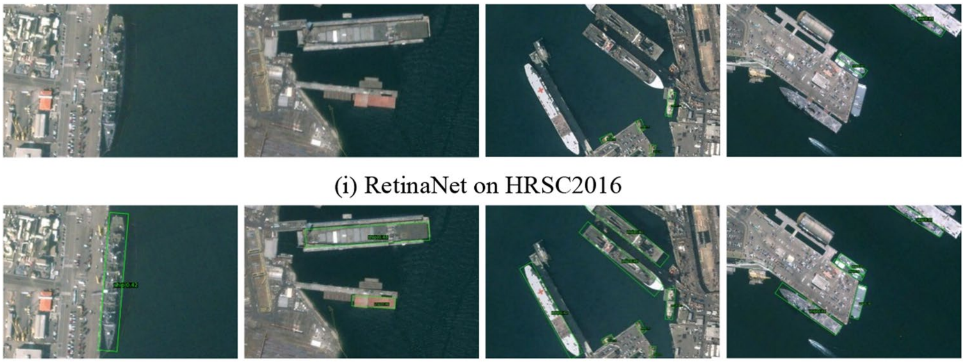


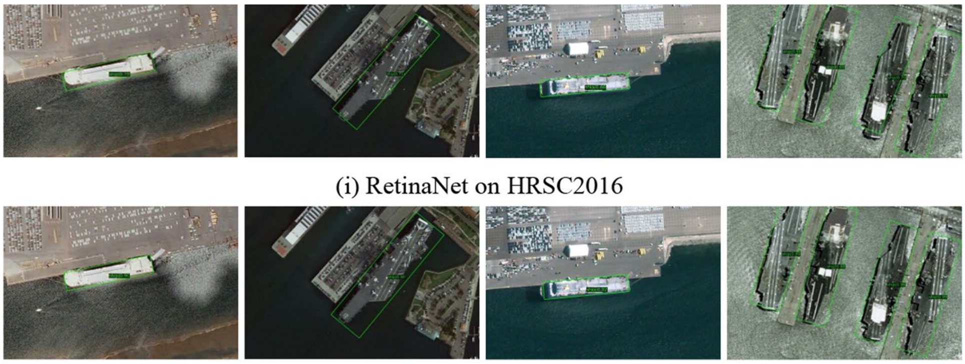
Fig. 7 Loss curve for fine-grained classification experiments on HRSC2016


(i) RetinaNet on HRSC2016

(ii) PVTv2-B5_Aug on HRSC2016

Fig. 8 Effectiveness of PVTv2-B5_Aug on HRSC2016

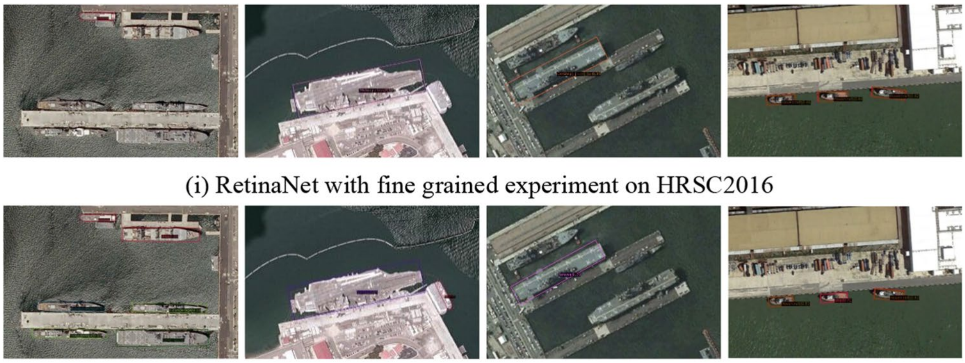
Fig. 9 Higher detection accuracy of PVTv2-B5_Aug on HRSC2016



(i) RetinaNet on HRSC2016

(ii) PVTv2-B5_Aug on HRSC2016

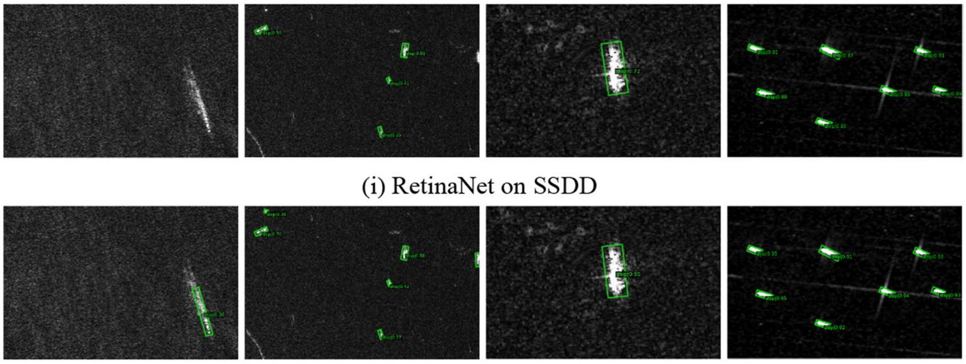
Fig. 10 Effectiveness of PVTv2-B5_Aug with fine-grained experiment on HRSC2016



(i) RetinaNet with fine grained experiment on HRSC2016

(ii) PVTv2-B5_Aug with fine grained experiment on HRSC2016

Fig. 11 Effectiveness of PVTv2-B5_Aug on SSDD



(i) RetinaNet on SSDD

(ii) PVTv2-B5_Aug on SSDD

Table 9 Performance of YOLF-ShipPnet and other networks on HRSC2016

Model	Backbone	Input_size	mAP
YOLF-ShipPnet	PVTv2	800 × 800	0.5813
IR-Net [19]	ResNet	800 × 800	0.5580
CP [20]	Fast R-CNN	800 × 800	0.5570
Kld [33]	ResNet50	800 × 512	0.5415
FR-O [34]	VGG-16	1024 × 1024	0.5413
RetinaNet [35]	ResNet50	800 × 512	0.5206

particularly for fine-grained classification, as verified using the HRSC2016 and SSDD datasets. These results suggest that the proposed network has great potential for applications in industrial ship management. Overall, our work highlights the significant strengths of the PVTv2 network for enhancement of accuracy in the depth dimension and the importance of the HSV data augmentation strategy for improving the generalization capability. The use of this network in real-world scenarios may lead to significant improvements in the efficiency and effectiveness of ship management systems.

6 Reflection and Future Work

The YOLOF-ShipPnet network has shown promising results in terms of its depth effectiveness. However, there is still room for improvement in the model's performance by further tuning the parameters associated with the number of layers, which will be investigated in future studies. While FPN has been used as the neck part of the network, other networks such as faster R-CNN may offer promising performance in ship detection tasks due to their robustness and flexibility. Therefore, it may be worthwhile to retrain the network using faster R-CNN and compare the results with the previous ones. Currently, the HSV technique is utilized as a means of random data augmentation to enhance the network's ability to generalize when presented with ship images that vary in color or lighting conditions. However, this method is limited in some circumstances, and other data augmentation strategies should be explored in the future to accommodate different application scenarios.

Acknowledgements All authors acknowledge Northwestern Polytechnical University and Naval Aeronautical and Astronautical University, Department of Electronic and Information Engineering for providing the experimental data used in this study.

Author Contributions This study was done by the authors' own efforts, from the formulation of the research idea, to the design of the research method, the design and execution of the program, the management of the data, the statistical analysis, and the final interpretation. LY conceived the innovative points of the research and was responsible for the design of ablation and fine-grained classification experiments. ZQ carried out the statistical analysis and played a key role in implementing the experiments. SR performed the modeling and construction of the network. All authors participated in drafting the original manuscript. LY and ZQ checked and revised the manuscript. These authors contributed equally to this work and should be considered co-first authors. All authors read and approved the final manuscript.

Funding Fully self-funded. The author(s) received no financial support for the research, authorship, and/or publication of this article.

Availability of Data and Materials Data and material used in the study were obtained from the public database High-Resolution Ship Collections 2016 (HRSC2016): (https://www.kaggle.com/datasets/guofeng/hrsc2016) and SAR Ship Detection Dataset (SSDD) (https://github.com/TianwenZhang0825/Official-SSDD/blob/main/README.md).

Declarations

Conflict of Interest It is considered that there is no conflict of interest or competing interests and therefore not applicable.

Ethics Approval and Consent to Participate These data have eliminated personal privacy. Therefore, this study is not relevant to personal privacy approval.

Consent for Publication All authors have read and approved the final manuscript (Once the review is completed by both parties, I hereby consent for International Journal of Computational Intelligence Systems to public this research).

References

1. Chen, W., Yao, B., Li, Y., Liu, L., Liang, J.: A real-time ship detection system for large-scale optical remote sensing image on micro-nano satellite. In: 2022 IEEE International Conference on Real-time Computing and Robotics (RCAR), pp. 450–455 (2022)
2. Zhang, A., Liao, Y., Liu, S., et al.: Mining the benefits of two-stage and one-stage hoi detection. Adv. Neural. Inf. Process. Syst. **34**, 17209–17220 (2021)
3. Diwan, T., Anirudh, G., Tembhurne, J.V.: Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimedia Tools Appl. 1–33 (2022)
4. Cheng, L., Ji, Y., Li, C., et al.: Improved SSD network for fast concealed object detection and recognition in passive terahertz security images. Sci. Rep. **12**(1), 1–16 (2022)
5. Wang, Y., Wang, C., Zhang, H., et al.: Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery. Remote Sens. **11**(5), 531 (2019)
6. Du, L., Zhang, R., Wang, X.: Overview of two-stage object detection algorithms. J. Phys: Conf. Ser. **1544**(1), 12033–12039 (2020)
7. Li, Z., Li, Y., Yang, Y., et al.: A high-precision detection method of hydroponic lettuce seedlings status based on improved Faster RCNN. Comput. Electron. Agric. **182**, 106054 (2021)
8. Xu, Y., Li, D., Xie, Q., et al.: Automatic defect detection and segmentation of tunnel surface using modified Mask R-CNN. Measurement **178**, 109316 (2021)
9. Gong, Y., Yu, X., Ding, Y., et al.: Effective fusion factor in FPN for tiny object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1160–1168 (2021)
10. Yurong, L., Haining, W., Cunbao, L., et al.: Research progress of optical remote sensing image target detection based on deep learning. J. Commun. **43**(5), 190–203 (2022)
11. Zhang, C., Benz, P., Argaw, D.M., et al.: Resnet or densenet? Introducing dense shortcuts to resnet. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3550–3559 (2021)
12. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. 30 (2017)
13. Bello, I., Zoph, B., Vaswani, A., et al.: Attention augmented convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3286–3295 (2019)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Preprint at arXiv:2010.11929 (2020)
15. Han, K., Xiao, A., Wu, E., et al.: Transformer in transformer. Adv. Neural. Inf. Process. Syst. **34**, 15908–15919 (2021)
16. Wang, W., Xie, E., Li, X., et al.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.

In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)

17. Liu, Z., Wang, H., Weng, L., et al.: Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geosci. Remote Sens. Lett. **13**(8), 1074–1078 (2016)
18. Liu, Z., Hu, J., Weng, L., et al.: Rotated region based CNN for ship detection. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp. 900–904 (2017)
19. Liu, Y., Li, H., Cheng, J., Chen, X.: MSCAF-Net: a general framework for camouflaged object detection via learning multi-scale context-aware features. In: IEEE Transactions on Circuits and Systems for Video Technology (2023)
20. Sun, M., Xu, L., Luo, R., et al.: GHFormer-Net: Towards more accurate small green apple/begonia fruit detection in the nighttime. J. King Saud Univ. Comput. Inf. Sci. **34**(7), 4421–4432 (2022)
21. Yu, H., Wu, J.: A Unified Pruning Framework for Vision Transformers. Preprint at arXiv:2111.15127 (2021)
22. Yan, Y., Tan, Z., Su, N.: A data augmentation strategy based on simulated samples for ship detection in RGB remote sensing images. ISPRS Int. J. Geo Inf. **8**(6), 276 (2019)
23. Zhao, W., et al.: Feature balance for fine-grained object classification in aerial images. IEEE Trans. Geosci. Remote Sens. **60**, 1–13 (2022)
24. Menon, G.S., Murali, S., Elias, J., Aniesrani Delfiya, D.S., et al.: Experimental investigations on unglazed photovoltaic-thermal (PVT) system using water and nanofluid cooling medium. Renew. Energy **188**, 986–996 (2022)
25. Ge, Z., Liu, S., Wang, F., et al.: Yolox: Exceeding yolo series in 2021. Preprint at arXiv:2107.08430 (2021)
26. Zhou, H., Li, Y., Chen, P., Shen, Y., Zhu, Y.: Improved FPN-based ship target detection for SAR images in complex scenes. J. Dalian Maritime Univ. 1–8 (2022)
27. Wang, Z., Xie, X., Yang, J., et al.: Soft focal loss: Evaluating sample quality for dense object detection. Neurocomputing **480**, 271–280 (2022)
28. Li, Y., Zhou, S., Chen, H.: Attention-based fusion factor in FPN for object detection. Appl. Intell. (2022). https://doi.org/10.1007/s10489-022-03220-0
29. Ge, Y., Jialong, Z., Ying, W.: Human body detection and tracking algorithm based on HSV and RGB color space. Autom. Technol. Appl. **41**(9), 17–2028 (2022). https://doi.org/10.20033/j.1003-7241.(2022)09-0017-05
30. Tellez, D., Litjens, G., Bándi, P., et al.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. **58**, 101544–101553 (2019)
31. Liu, Z., Yuan, L., Weng, L., Yang, Y.: A high resolution optical satellite image dataset for ship recognition and some new baselines. In: Proceedings of the International Conference on Pattern Recognition Applications and Methods, vol. 2, pp. 324–331 (2017)
32. Zhang, T., et al.: Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. Remote Sens. **13**(18), 3690 (2021)
33. Yang, X., et al.: Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. Adv. Neural. Inf. Process. Syst. **34**, 18381–18394 (2021)
34. Xia, G.-S., et al.: DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
35. Zhang, S., et al.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.