IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 12, 2022, accepted February 22, 2022, date of publication February 24, 2022, date of current version March 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154474

Improved YOLOv4 Based on Attention Mechanism for Ship Detection in SAR Images

YUNLONG GAO[1,2], ZHIYONG WU[1], MING REN[1,2], AND CHUAN WU[1]
[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]Key Laboratory of Airborne Optical Imaging and Measurement, Chinese Academy of Sciences, Changchun 130033, China

Corresponding authors: Yunlong Gao (gaoyl15@mails.jlu.edu.cn) and Zhiyong Wu (wuzy@ciomp.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61401425.

ABSTRACT Ship detection in synthetic aperture radar (SAR) images is an important and challenging work in the field of image processing. Traditional detection algorithms usually rely on handmade features or predefined thresholds, the different performance is obtained with varying degrees of prior knowledge, and it is difficult to take advantage of big data. Recently, deep learning algorithms have found wide applications in ship detection from SAR images. However, due to the complex backgrounds and multiscale ships, it is hard for deep networks to extract representative target features, which limits the ship detection performance to a certain extent. In order to tackle the above problems, we propose an improved YOLOv4 (ImYOLOv4) based on attention mechanism. Firstly, to achieve the best trade-off between detection accuracy and speed, we adopt the off-the-shelf YOLOv4 as our basic framework because of its fast detection speed. Secondly, a thresholding attention module (TAM) is introduced to suppress the adverse effect of complex backgrounds and noises. Besides, we embed channel attention module (CAM) into improved BiFPN as the feature pyramid network (FPN) to better enhance the discrimination of the multiscale target features. Finally, the decoupled head with two parallel branches improves the performance of classification and regression. The proposed method is evaluated on public SAR dataset and the experimental results demonstrate that it has higher efficiency and feasibility than other mainstream methods, yielding the accuracy of 94.16% at intersection over union of 0.5 and 58.19% at intersection over union of 0.75.

INDEX TERMS Ship detection, SAR, attention, decouple head, YOLOv4.

I. INTRODUCTION

With the continuous improvement of space remote sensing imaging technology, high-resolution and wide-scale remote sensing images are becoming more and more enriched and facilitate a large range of applications. Remote sensing applications make remote sensing images into plug and play products, which are widely used in all aspects of social and economic life, such as traffic control [1], [2], geological and mineral exploration [3], environment monitoring [4], and urban construction [5]. As the key target of marine monitoring and wartime attack, the detection of ships has an important practical value for both civil and military fields [6]–[10]. In recent years, many researches in this field have prioritized synthetic aperture radar (SAR) images and ship detection in SAR images has become one of the most important remote sensing applications [11]–[16]. Compared with optical sensors, SAR is an active microwave remote sensing imaging sensor, which has the all-day and all-weather surveillance capabilities, making it possible to continuously monitor targets at sea [17]–[20]. Therefore, it is very important to study the ship detection in SAR images.

Many studies have been carried out about ship detection in remote sensing images in recent years [21]–[24]. Traditional feature extraction methods are usually based on handmade features such as scale-invariant feature transform (SIFT) [25], histogram of oriented gradient (HOG) [26] and local binary patter (LBP) [27], followed by shallow classification modules, e.g., support vector machine (SVM) [28], extreme learning machine (ELM) [29], and Adaboost [30]. Most of the traditional algorithms show great performance for ideal-quality images. However, they are highly dependent on manual feature extraction and availability of prior knowledge such as predefined thresholding and the distributions of sea clutters, let alone the influence of complex backgrounds and noises. As a result, their generalization

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

ability is weak, and the detection performance is far from satisfactory.

In recent years, driven by extensive remote sensing images, deep learning methods have achieved great success in object detection. State-of-the-art deep learning-based ship detection methods include one-stage and two-stage detectors. The one-stage detectors directly convert the object detection into a regression problem which is fast running. You only look once (YOLOv1) [31] as the end-to-end algorithm for object detection processes the input images only once, and this reduces the computational redundancy and improves the detection speed; Single Shot Detector (SSD) [32], RetinaNet [33], YOLOv2 [34], YOLOv3 [35], and the latest YOLOv4 [36] are the typical one-stage detection algorithms; In two-stage detectors, the first stage generates a set of candidate proposals while filtering out the majority of negative locations, the second stage classifies the proposals into background or foreground. Region CNN (R-CNN) [37] introduces deep learning methods to the field of object detection and outperforms most of the traditional detection methods; Subsequently, a series of two-stage algorithms are proposed, such as Faster R-CNN [38], Mask R-CNN [39], and Cascade R_CNN [40]. Compared with the one-stage detectors, the two-stage detectors offer high positioning accuracy with low running speed.

With the rapid development of SAR sensors, the volumes of SAR images are getting larger and the data are easier to obtain which lead to the possibility of deep learning algorithms for SAR object detection. However, some challenges still exit: 1) complex backgrounds on land and strong backscatters usually result in missing detections and false alarms, and 2) ships are often clustered and the shapes of targets in SAR images have an extreme aspect ratio. Most of all, small ship objects restrict deep networks to extract representative target features, which further limits the ship detection performance. Researchers in deep-learning community for ship detection in SAR images have made a lot of attempts to exploit CNN-based ship detection frameworks. Based on the original Faster R-CNN, researchers have made some typical improvements such as adding hard negative mining [41] and dense connection [42]. There are also some methods dedicated to building a more complex structure to improve the performance for some tough problems like dense small ships [43]. Zhao *et al.* proposes a cascade coupled convolutional network with attention mechanism to detect ships which shows a promising result for small objects [44]. A novel dense pyramid network with attention weighting is utilized and solves the problem of multiscale ship detection [45]. Besides, some training techniques such as training from scratch are also introduced in the SAR ship detection problem, and the final results outperform other pretrained ship detectors [46]. To achieve real-time ship detection in SAR images, some methods based on one-stage detectors have been gradually explored. For instance, Wang *et al.* [47] applies the end-to-end RetinaNet to SAR ship detection, and constructs a multi-resolution and complex background dataset, achieving a high detection accuracy. Du *et al.* [48] uses two identical sub-networks to extract features from the input SAR image and the corresponding saliency map at the same time, then the salient features are integrated to the deep CNN features. Zhang *et al.* [49] introduces a channel attention module and a spatial attention module in the high-speed and high-precision SAR ship detection network and obtains very excellent detection performance. As far as we know, most of the researches either focus on high-accuracy or high-speed, and only a few researches focus on both. However, both of two indicators are very import for SAR ship detection.

In this paper, we propose a novel one-stage ship detector named improved YOLOv4 (ImYOLOv4) based on attention mechanism [50] for accurate ship detection in SAR images. Firstly, to achieve the best trade-off between detection accuracy and speed, we adopt the off-the-shelf YOLOv4 as the inspiration of our basic detection framework. Secondly, we design a thresholding attention module (TAM) that is embedded in very first layer of the network to perform denoising in the image-level. The TAM block can adaptively learn a set of thresholding according to the global information of the image to suppress noises, avoiding the invalid data flow of the network. Besides, in order to improve the detection performance of multiscale ships, we obtain the optimal sizes of multiscale anchors by K-means [51] clustering according to the SAR dataset, and we improve the state-of-the-art feature pyramid network (FPN) BiFPN [52] with channel attention module (CAM) to complete the fusion operations. Finally, we use a decoupled head structure to deal with the ship classification and bounding box regression tasks separately. Based on these novel techniques above, our experiments on the public SAR Ship Detection Dataset (SSDD) [53] show that ImYOLOv4 could significantly improve the detection performance on the ship targets with multiscale sizes in front of complex backgrounds.

The main contributions of this paper are as follows:

(1) A novel one-stage ship detector named ImYOLOv4 based on attention mechanism is proposed which meets the requirement for both high-accuracy and high-speed detection.

(2) We design an embedded TAM block to perform denoising due to the considerations of complex backgrounds and strong backscatters for SAR ship detection.

(3) We integrate the CAM block with BiFPN module as the feature pyramid structure to better complete the fusion operations for the salient feature maps. The CAM block helps ImYOLOv4 pay more attention to the targets of interest, which ensures the effectiveness of detecting small ships.

(4) We replace the YOLO's head with a decoupled head to deal with the ship classification and bounding box regression tasks separately, the decouple head is validated on public SAR dataset and the comparison results confirm its improvement of detection performance.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work that are close to our method. Section 3 introduces the framework of our proposed

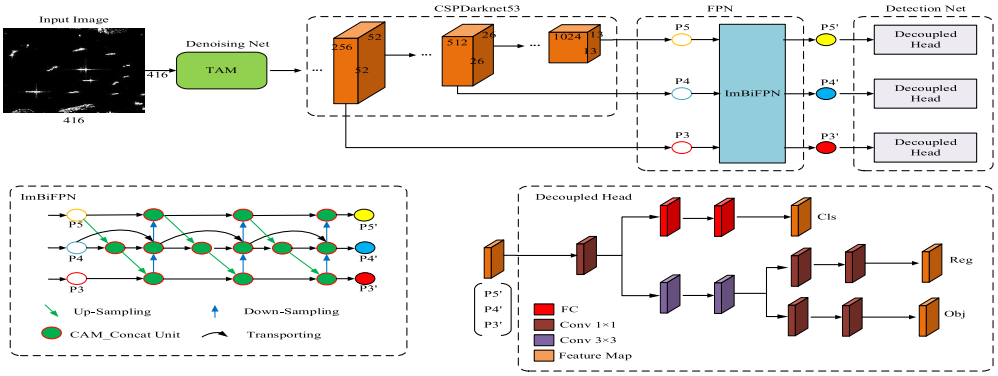


FIGURE 1. End-to-end framework of ImYOLOv4.

method in detail. Dataset and implementation settings are described in Section 4. A series of experiments and results are presented in Section 5. Finally, we summarize this paper in Section 6.

II. RELATED WORK

Deep learning-based methods have made a significant advancement in the field of SAR ship detection. Based on deep learning, researchers have introduced methods that have shown good performance in order to get better detection results. In terms of the better balance between high-accuracy and high-speed, Ma *et al.* [54] designs an Accelerated-YOLOv3 method which aims to reduce the computational time with relatively competitive detection accuracy by constructing a new architecture with less layers and channels. Chang *et al.* [55] proposes an enhanced GPU based deep learning method called YOLOv2-reduced to detect ship from SAR images, and the authors prove the method can make a big leap forward in improving the detection performance. These models with fewer number of layers sacrifice the accuracy to achieve a trade-off between detection accuracy and speed. The latest YOLOv4 has the highest accuracy in real-time target detection algorithms and offers us the use for reference, and the experiments show its best-practice for ship detection in SAR images.

In order to achieve accurate detection under poor image quality and complex backgrounds, some improvements have been proposed. Han *et al.* [56] studies how the detection performance varies from images with different complexity, backgrounds, surroundings, and quality. Fu and Wang [11] designs a fast ship detection method which consists of two cascade deep convolutional networks: scene classification network (SCN) and single shot detector (SSD), the SCN can quickly eliminate the sub-images that may not contain ships, and then the remaining sub-images are input into the SSD to implement refined ship target detection. Sun *et al.* [57] introduces a category-position module based on attention mechanism to improve the positioning performance in complex scenes by generating guidance vectors. Wang *et al.* [58] proposes a mask to guide attention maps, which performs well in the instance segmentation field. Masks are used to enhance ship position information in ship detection field and to eliminate the influence of complex backgrounds. These improvements usually bring a large amount of redundant information that greatly affect the detection efficiency. Different from the related works, we design a lightweight embedded TAM based on attention mechanism to filter the adverse effect of noises.

To ensure the ability of detecting multiscale ships, Lin *et al.* [59] proposes a new network architecture based on the Faster R-CNN by using squeeze and excitation mechanism to enhance the salient features of ship targets. Kang *et al.* [60] discloses a contextual region-based convolutional neural network with multilayer fusion, the framework fuses the deep semantic and shallow high-resolution features, improving the detection performance for small-sized ships. Sun *et al.* [61] introduces a novel bi-directional feature fusion module to the YOLO framework to efficiently aggregate multiscale features which can be helpful for detecting multiscale ships. Cui *et al.* [45] designs a feature pyramid network integrating dense attention mechanism, which made the features extracted by the network contain rich resolution and semantic information, and the proposed method proved to be suitable for multiscale ship detection. A receptive pyramid network extraction strategy and attention mechanism are also proved to be effective in the ship detection task, but the processing efficiency is low due to the complex model structure [62]. Although the CNN-based detection algorithms can automatically capture the features of ships, the detection performance of these existing methods still needs to be improved. In this paper, the proposed ImYOLOv4 integrates the CAM block with BiFPN module as the feature pyramid structure to better complete the fusion operations for multiscale ship detection, and the salient feature maps will not make the deep CNN features disappear. The details of ImYOLOv4 model are introduced in Section 3.

III. METHODOLOGY

The proposed method will be described in detail in this section. First, the overall framework of ImYOLOv4 is

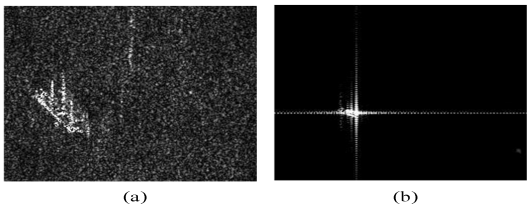


FIGURE 2. Complex backgrounds and strong backscatters disturb the detection of ships.

introduced. Afterwards, the mechanism of every key module will be explained. Other strategy validated efficient for detection such as K-means clustering for anchor box will be described at last.

A. OVERALL FRAMEWORK

The overall scheme of the proposed method and the network architecture of ImYOLOv4 are illustrated in Figure 1. Firstly, the resized input image (taking 416 as an example) is send into the TAM to perform denoising operations. Next, we adopt CSPDarknet53 [36] as the backbone to extract feature maps at three different branches. Then, the multi-scale feature maps are feed into FPN structure to obtain fused features. Specifically, the outputs (P3, P4, and P5) of CSPDarknet53 are transported to the ImBiFPN module to generate corresponding salient feature maps (P3', P4', and P5'). In ImBiFPN module, we apply up-sampling and down-sampling operations by the factor of 2 and merge the feature maps of same spatial resolution via concatenation, given to the fact that different inputs should have different weights, we design the CAM_Concat Unit by using CAM to obtain channel-wise coefficient tensor while concatenating. In the end, the decoupled head with two parallel branches is used to predict a 3D tensor detection result of bounding box, object, and classifications. The whole detection pipeline of ImYOLOv4 is in a single network, so it can be optimized end-to-end directly.

B. THRESHOLDING ATTENTION MODULE

The radar receives echo signals from ground, including ground-based clutter and detection targets because of its unique imaging technique. As a coherent imaging system, SAR inevitably generates speckle noises from the complex backgrounds, resulting in the missing detection of weak ship targets. Besides, the metal materials and the superstructure of the ships usually produce strong backscatters which will reshape the ship appearances in the SAR images and interfere with the detection process. Figure 2(a) and 2(b) show the noises mentioned above respectively.

Considering the adverse effect of these noises, we design an embedded TAM block to perform denoising in the image-level. In TAM block, we integrate the thresholding algorithm and attention mechanism to automatically learn a set of thresholding which can be used to transform the near-zero to zero for signal reconstruction. Compared with the traditional SAR feature enhancement methods, TAM does not require high expertise in signal process and its lightweight architecture has additional advantage of lower computational complexity and memory consumption.

As for a SAR image obtained by radar system, it can be decomposed as follows:

$$Y = X + N \tag{1}$$

where X is the considered scene, N is noise matrix of the same size as X which denotes the difference between the reconstructed image and real scene. Considering the sparsity of SAR image, we can recover the considered scene by dealing with the following optimization problem:

$$\hat{X} = \min_{X}\{\|Y - X\|_2^2 + \mu \|X\|_1\} \tag{2}$$

the optimization problem can be solved by iterative thresholding algorithm, however, the number of iterations has a great impact on the sparsity and precision of the considered scene. Inspired by LeakyReLu [63] activation function, we would like to optimize the function by equation (3):

$$\hat{X} = \begin{cases} Y - \mu, & Y > \mu \\ 0, & |Y| \le \mu \\ \dfrac{1}{\alpha}(Y + \mu), & Y < -\mu \end{cases} \tag{3}$$

where μ is the thresholding used to filter the noises, α gives us a non-zero gradient so that useful negative features can be well preserved.

Figure 3 illustrates the detailed architecture of TAM block which is designed upon the transformation mapping between the input $X \in \mathrm{R}^{C\times H\times W}$ and its reconstruction feature map $\hat{X} \in \mathrm{R}^{C\times H\times W}$. We adopt the channel attention module to generate a channel-wise thresholding tensor $\mu \in \mathrm{R}^{C\times 1\times 1}$. Specifically, we first squeeze the input along the spatial dimension $H \times W$ by using both average pooling and max pooling operations to obtain two channel tensors of $\mathrm{R}^{C\times 1\times 1}$, then, we merge the two tensors via element-wise summation and forward the output s to a network which consists of two fully connected (FC) layers. To reduce the complexity of TAM, the activation size of the first FC layer is set to $\mathrm{R}^{C/r\times 1\times 1}$, where r is the reduction ratio. A sigmoid function is also employed at the end of network as a simple gating mechanism to get a scaled output tensor z of (0,1). Finally, to prevent the thresholding from being neither negative nor too large, we obtain the product μ by element-wise multiplication from the scaled tensor z and the global information tensor s. Therefore, the thresholding is expressed as:

$$\mu = \frac{1}{2}(z \bullet s) \tag{4}$$

C. FEATURE PYRAMID NETWORK

For deep learning-based detection methods, FPN [64] plays an important role in solving the multiscale problems and acts as a feature extractor with the consideration of the low-level high-resolution and high-level low-resolution semantic meaning. In general, more intensive sampling can get

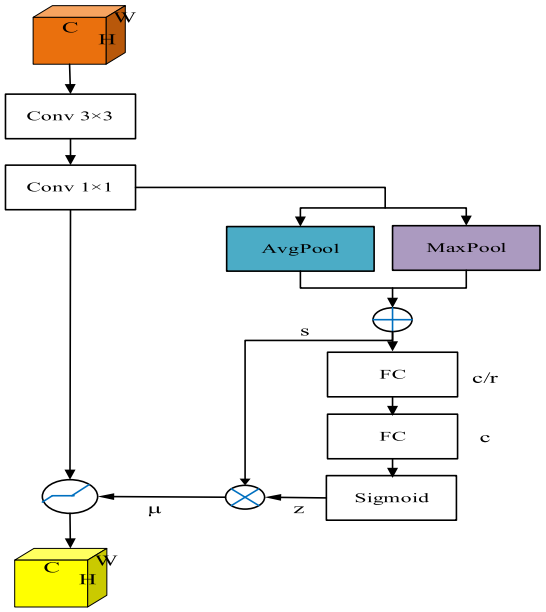


FIGURE 3. The architecture of TAM.

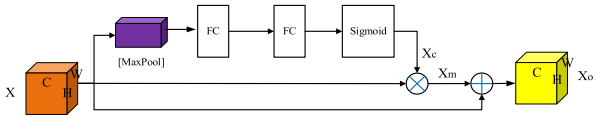


FIGURE 4. Structure of CAM block.

more detailed features, while more sparse sampling can more clearly reflect the overall trend. Fusing features of different scales can capture ample semantic information which help improve the accuracy of ship detection.

After the multiscale feature maps are extracted by CSPDarknet53 network, we forward them to the ImBiFPN structure to complete the fusion operations for salient feature maps. As depicted at the left-bottom of Figure 1, there are two main data flows in ImBiFPN, the bottom-up down-sampling and top-down up-sampling pathways. And the CAM_Concat Unit completes the feature fusion of the same spatial resolution. In the process of concatenating, we apply CAM block to automatically learn the channel-wise attention coefficients which denote the significant degree of different inputs. As shown in Figure 4, we first squeeze the concatenated feature map along the spatial dimension $H \times W$ by using max pooling operation to focus on what is important in the given input. Then, two FC layers and a simple gating mechanism via sigmoid function are employed to obtain the final channel attention map Xc. Finally, we also add a residual input for the consideration of preventing the problem of gradient-vanishing. After element-wise multiplication and summation operations, we generate the refined output Xo of CAM block:

$$X_o = X + X_m = X + X_c \bullet X \tag{5}$$

In summary, there are two differences between BiFPN and our ImBiFPN. The one is that the input of ImBiFPN is 3-level multiscale feature maps obtained by CSPDarknet53 network, while the input of BiFPN is 5-level features, the same goes for the output of both FPN structures. The second is that we design a weights generator by using CAM block to assign the different importance of inputs while concatenating. These improvements reduce the network parameters while maintain the BiFPN performance.

D. DECOUPLED HEAD

In object detection, the conflict between classification and regression tasks is a well-known problem. The two different tasks which share almost the same parameters in YOLO head could hurt the detection process. This is inspired by the nature insight that for one instance, the features in some salient area may have rich information for classification, while these around the boundary may be good at bounding box regression. Based on that case, we design a decoupled head with two branches to solve the object functions from different spatial dimensions. As depicted at the right-bottom of Figure 1, we first use a convolutional layer with kernel size 1×1 to perform the dimension reduction. Then, in the up branch, a two-layer fully connected network is employed to obtain the classification-specific output Cls. While in the down branch, two shared 3×3 convolution and two 1×1 convolution operations are used to obtain the regression-specific outputs Reg and Obj. Finally, the outputs of two branches are merged into a tensor for the task of ship prediction.

E. K-MEANS CLUSTERING

Anchor box mechanism for object detection was proposed to solve the problem of multitarget in one predicted box and has been used in many detectors. There are 9 predefined anchor boxes in our method for different scale detection. K-means clustering is adopted on the overall SSDD data to automatically find the prior boxes. Most ships in SAR images are small and weak targets, which occupy few pixels and have lower contrast. If we use the standard Euclidean distance of the conventional K-means algorithm, the bounding boxes with larger scale generate more error than the smaller scale boxes, which will lead to missed detections of small and sparse ships. What we want in the final detection are the priors that will lead to high intersection over union (IoU) scores, thus, the distance metric in this paper can be expressed as:

$$\begin{aligned} &\text{d(anchor box, cluster centroid)} \\ &\qquad = 1 - \text{CIoU(anchor box, cluster centroid)} \end{aligned} \tag{6}$$

where d(anchor box, cluster centroid) is the new distance metric that needs to be minimized, and CIoU(anchor box, cluster centroid) means the CIoU [65] values of the anchor box and different cluster centroids. The specific size of anchor boxes for three scales are shown in Table 1. The optimal cluster centroids obtained by K-means are significantly different than previous hand-picked anchor boxes and have better performance for both precision and recall on SAR ship detection.

TABLE 1. Detailed information of scaled anchor boxes.

Feature layer	Size	Anchor boxes	Number
Feature map-13	13×13	(69,34), (73,61), (89,100)	13×13×3
Feature map-26	26×26	(39,27), (42,99), (52,47)	26×26×3
Feature map-52	52×52	(12,16), (21,45), (34,55)	52×52×3

TABLE 2. Statistical distribution of the ship size.

Size	Min (Pixel)	Max (Pixel)	Number	Percentage
Small Ship	4×6	32×32	35695	59.96%
Medium Ship	32×32	96×96	23660	39.74%
Large Ship	96×96	207×109	180	0.30%

IV. DATASET AND IMPLEMENTATION SETTINGS

A. DATASET

The dataset used in this paper is a SAR dataset for ship detection published by the Digital Earth Laboratory of the Aerospace Information Research Institute, Chinese Academy of Sciences. SSDD is generated from 102 Gaofen-3 [66] images and 108 Sentinel-1 [67] images. As for Gaofen-3, the resolution of these images involves 3m, 5m, 8m and 10m with Strip-Map (UFS), Fine Strip-Map 1 (FSI), Full Polarization 1 (QPSI), Full Polarization 2 (QPSII) and Fine Strip-Map 2 (FSII) imaging mode, respectively. The Sentinel-1 imaging modes include S3 Strip-Map (SM), S6 SM and IW-mode.

The SSDD has 43819 ship chips and 59535 ship targets in total. The pixel of each image is 256 × 256. The ship targets are marked in a similar format to Pascal VOC [68]. The statistical distribution of the ship size over the SSDD is presented in Table 2, where "Size", "Min" and "Max" mean ship pixels, minimum ship size and maximum ship size, respectively. "Number" represents the total number of ships, "Percentage" denotes the percentage of the ship in whole ship targets.

From Table 2 and Figure 5, we can see that the dataset has the following characteristics. Firstly, there are multiscale SAR ships in these chips, and the size conversion range is large. Small ships and medium ships account for a large proportion of whole targets. Secondly, there are complex backgrounds in the ship chips. Some of ships are on the open sea, some in the port. All of these have brought difficulties to ship detection, and put forward higher requirements for the performance of ship detection. In the experiment, we split the training, validation and testing set randomly according to rate of 7:2:1. The training set and the validation set are used for training models and the testing set is used for testing models.

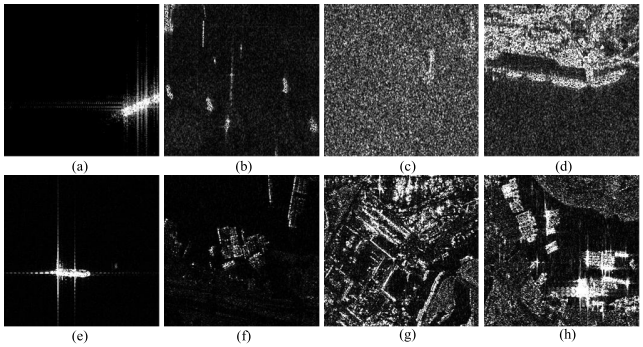


FIGURE 5. Samples of ship chips. (a), (b), (c) and (d) are from Gaofen-3 images. (e), (f), (g) and (h) are from Sentinel-1 images.

B. EVALUATION METRICS

In order to quantitatively evaluate the detection performance of ImYOLOv4, we adopt four widely used criteria, namely, precision, recall, mAP (mean Average Precision) and F1 score. The precision measures the value of detections that are true positives and the recall measures the value of positives over the number of ground truths.

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

where TP, FP and FN represent the number of true positives, false positives and false negatives.

As for detection, a higher precision and a higher recall are both expected. However, the two metrics are a pair of contradictory indicators. It means that a higher precision will result in a lower recall and a higher recall will result in a lower precision. F1 score is then used which can comprehensively combine precision and recall. A higher F1 score indicate a more ideal detection performance. F1 score is defined based on the harmonic average of precision and recall:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

Precision, recall and F1 score are all calculated based on the single point threshold. AP can solve the limitations of single point threshold and get an indicator that reflects the global performance. AP is obtained by the integral of the precision over the interval from recall=0 to recall=1, that is, the area

TABLE 3. AP50 with different parameter values of TAM.

AP50(%)	r=1	r=2	r=4	r=8	r=16	r=32
$\alpha = 0$	91.24	91.29	91.81	91.96	92.67	91.92
$\alpha = 0.05$	93.43	93.77	93.96	94.00	94.12	93.87
$\alpha = 0.1$	93.38	93.57	93.95	94.07	**94.16**	94.03
$\alpha = 0.15$	93.20	93.80	93.82	93.89	94.08	93.74
$\alpha = 0.2$	92.78	93.19	93.28	93.57	93.92	93.46
$\alpha = 0.25$	92.52	93.10	93.27	93.53	93.76	93.22

TABLE 4. AP75 with different parameter values of TAM.

AP75(%)	r=1	r=2	r=4	r=8	r=16	r=32
$\alpha = 0$	57.24	57.59	57.91	57.96	57.97	56.92
$\alpha = 0.05$	57.43	57.82	58.04	58.04	58.11	57.97
$\alpha = 0.1$	57.68	57.87	58.05	58.17	**58.19**	58.03
$\alpha = 0.15$	57.33	57.50	57.72	58.09	58.11	57.73
$\alpha = 0.2$	57.18	57.39	57.71	58.00	58.02	57.46
$\alpha = 0.25$	56.82	57.14	57.47	57.58	57.76	57.29

under the precision-recall (PR) curve.

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

C. IMPLEMENTATION SETTINGS

All experiments are implemented using the deep learning framework Pytorch and executed on a PC with TITAN XP GPU (11G memory), the PC operating system is Ubuntu 16.04. At the beginning of network training, we use the parameters pre-trained on ImageNet to initialize the network. Then, we utilize the end-to-end training strategy to train our model, in which the gradient descent algorithm is used to fine-tine the network weights. The weight decay and momentum are set to be 0.0001 and 0.9. The reduction parameter r and α used for gradient preserved in TAM block are set to 16 and 0.1 which will be explained in the following experiments. Smooth-L1 [36] Loss function is applied to calculate classification loss and a total of 2k iterations are performed for training our ImYLOLv4 model.

V. EXPERIMENTS AND RESULTS

A. PERFORMANCE OF TAM

In this section, we first examine the impact of parameters r and α and select the best combination of parameters for TAM module. The parameter r is designed to decrease the calculation complexity of the fully connected layers and α guarantees that most neurons won't be dead during the training process. We measure the AP50 (IoU=0.5) and AP75 (IoU=0.75) in the case of different parameter values and list the results in Table 3 and Table 4. As we can see from the results, adding the parameters brings the improvements in both AP50 and AP75 compared with condition when $r = 1$ and $\alpha = 0$. And we can find out that the combination of $r = 16$ and $\alpha = 0.1$ obtains the best detection precision. The reduction parameter r avoids overfitting caused by too many training parameters to a certain extent, and α expands the values of the activation function in the part of less than the thresholding -μ, which further demonstrates that avoiding neurons being dead is more important than obtaining sparsity.

To verify the effectiveness of TAM, we conduct experiments comparing the detection performance between YOLOv4, ImYOLOv4 without TAM (DeTImYOLOv4) and ImYOLOv4. For a fair comparison, we set the other hyper-parameters consistent in the experiments. And the results are displayed in Table 5. As we can see from the results, adding the TAM block brings 3.18%, 0.05, 2.29% and 8.15% increment in AP50, F1 score, precision and recall versus DeTImYOLOv4, and outperforms YOLOv4 by 0.47%, 0.01, 2.00% and 1.00% in AP50, F1 score, precision and recall, respectively. When IoU is set to 0.75, adding the TAM block brings 9.49%, 0.05, 7.60% and 3.47% increment in AP75, F1 score, precision and recall versus DeTImYOLOv4, and outperforms YOLOv4 by 7.77%, 0.03, 6.08% and 0.81% in AP75, F1 score, precision and recall, respectively. Specifically, we present some denoising results of ImYOLOv4 to further demonstrate the validity of TAM. We visualize the spatial response of the input and output feature map of TAM block by heatmap where the blue color denotes low spatial response, and the red indicates a high response. We resize the heatmaps to the same size of the SAR image and the results are shown in Figure 6. By comparing Figure 6(b), (e), and Figure 6(c), (f), we can see that the complex background triggers very low response and the irrelative information brought by background can be effectively suppressed because of TAM. While the noises are suppressed, ImYOLOv4 can focus on and extract more discriminative features of targets, which is very helpful for the ship detection.

In addition, as shown in Table 5, we also compare our TAM with some state-of-the-art attention modules, such as ECA [69], BAM [70] and CBAM [71]. We replace the TAM block with attention modules while keeping other subnets consistent to ImYOLOv4. By analyzing the results, TAM and ECA obtain better performance than the other two modules, this is mainly because that BAM and CBAM are proposed based on optical images and irrelative spatial feature would be falsely enhanced for SAR images. The TAM block can adaptively learn the channel-wise thresholding according to the global information of the image, and the experiment results demonstrate its suitability for SAR ship detection task.

B. PERFORMANCE OF FPN

We also conduct an experiment to validate the performance of FPN. FPN from YOLOv3 (YOLOv3FPN), PANet [36],

TABLE 5. Comparison results with other attention modules.

Method	IoU=0.5				IoU=0.75			
	Precision	Recall	F1	AP50	Precision	Recall	F1	AP75
YOLOv4	91.54%	89.95%	0.91	93.69%	62.56%	61.48%	0.62	50.42%
DeTImYOLOv4	91.25%	82.80%	0.87	90.98%	61.04%	58.82%	0.60	48.70%
ImYOLOv4	**93.54%**	**90.95%**	**0.92**	**94.16%**	**68.64%**	**62.29%**	**0.65**	**58.19%**
DeTImYOLOv4+ECA	92.16%	88.89%	0.90	91.06%	66.87%	62.17%	0.64	55.53%
DeTImYOLOv4+BAM	91.08%	60.60%	0.73	84.19%	56.24%	48.29%	0.52	40.33%
DeTImYOLOv4+CBAM	91.17%	66.35%	0.77	84.38%	55.38%	46.29%	0.50	38.24%

TABLE 6. Performance of different FPNs.

Method	IoU=0.5			IoU=0.75			FPS
	AP_{50}	Precision	Recall	AP_{75}	Precision	Recall	
ImYOLOv4+YOLOv3FPN	92.83%	92.42%	87.96%	51.82%	64.96%	60.16%	57
ImYOLOv4+PANet	93.76%	92.48%	88.14%	58.15%	66.92%	60.12%	48
ImYOLOv4+BiFPN	94.05%	93.55%	88.06%	58.02%	66.56%	61.37%	36
ImYOLOv4	**94.16%**	**93.54%**	**90.95%**	**58.19%**	**68.64%**	**62.29%**	**42**

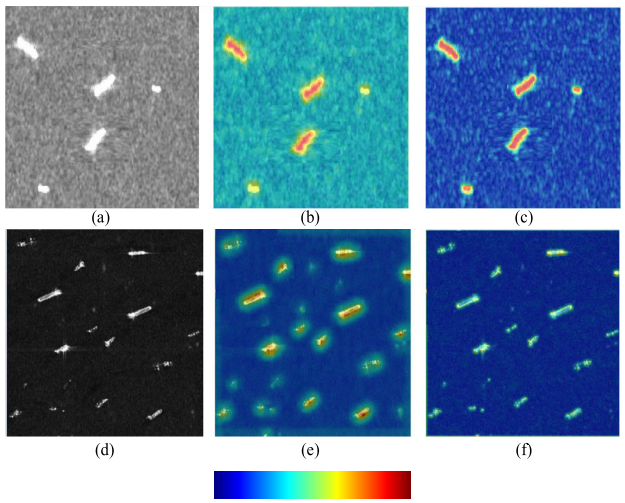


FIGURE 6. Visualization of the intermediate features. (a), (d) are the input SAR images. (b), (e) denote the heatmaps of the input images, and (c), (f) denote the corresponding heatmap outputs of TAM.

BiFPN are embedded into ImYOLOv4 as substitutions of FPN respectively. YOLOv3FPN simply contains an up-sampling pathway for fusing the features at different resolutions. PANet is originally applied in the field of image segmentation, which increases a down-sampling pathway on the basis of YOLOv3FPN. BiFPN introduces a weighted feature fusion strategy to better balance the feature information of different resolutions. The comparison results are listed in Table 6. As it is seen in Table 6, different feature fusion methods bring different detection performance. And our FPN and BiFPN achieve better performance for salient feature extraction which contributes to ship detection. Apart from the precision, we also evaluate the models by the running speed. Unlike BiFPN, our FPN uses CAM block as the weights generator, and the improvement makes our FPN achieve better accuracy and efficiency trade-offs.

C. PERFORMANCE OF DECOUPLED HEAD

In this part of experiments, we design several variants of decoupled head and make comparison to the YOLO head baseline. The variants are described as follows:

1) YOLO-Head (baseline): The coupled head is widely used in YOLO series detectors, the classification and regression tasks are solved by the single network.

2) Decoupled-Head (ours): The head splits the classification and regression on a fully connected head and a convolution head respectively.

3) Decoupled-Conv-FC-Head: The head splits the classification and regression on a convolution head and a fully connected head respectively.

4) Decoupled-FC-Head: Double fully connected heads which have the same structure as the up branch of our Decoupled-Head.

TABLE 7. Performance of decoupled head.

Method	IoU=0.5			IoU=0.75		
	AP_{50}	Precision	Recall	AP_{75}	Precision	Recall
YOLO-Head (baseline)	92.93%	91.72%	88.69%	51.52%	61.46%	45.76%
Decoupled-Head (ours)	**94.16%**	**93.54%**	**90.95%**	**58.19%**	**68.64%**	**62.29%**
Decoupled-Conv-FC-Head	91.87%	92.55%	88.46%	50.42%	62.56%	61.48%
Decoupled-FC-Head	91.67%	91.74%	87.45%	50.24%	58.93%	54.78%
Decoupled-Conv-Head	91.20%	91.85%	89.91%	52.88%	60.53%	55.76%

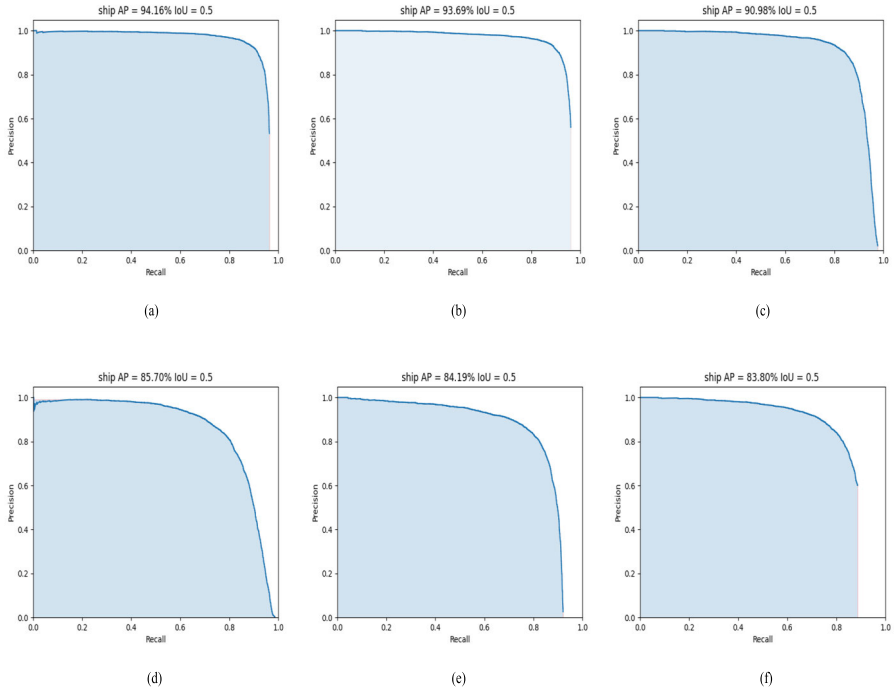


FIGURE 7. Precision-recall curves of detectors. (a)-(f) denotes ImYOLOv4, YOLOv4, YOLOv3, RetinaNet, CenterNet, and Faster-RCNN, respectively.

5) Decoupled-Conv-Head: Double convolutional heads which have the same structure as the down branch of out Decoupled-Head.

The comparison results between the variants are listed in Table 7. From the results, we can observe that decoupled head has a better performance than the single network baseline for ship detection, this is mainly because that classification and regression focus on the different problems, and different branches used for different tasks are conducive to the improvement of performance. This significant observation motivates us to rethink the architecture of the decoupled head. By comparing the variants of decoupled head, we can conclude that the fully connected head is more suitable for classification while the convolutional head has more advantage on the task of regression.

D. COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, we compare our ImYOLOv4 model with some state-of-the-art object detection models on SSDD, including RetinaNet, CenterNet [72], YOLOv3, YOLOv4,

and Faster-RCNN. The experimental results are displayed in Table 8, and Figure 7 shows the precision-recall curves of all the detectors.

As shown in Table 8, our ImYOLOv4 model outperforms one-stage detector RetinaNet by 8.46% AP50 and 16.67% AP75, YOLOv3 by 3.18% AP50 and 10.04% AP75, and YOLOv4 by 0.47% AP50 and 7.77% AP75, respectively. Compared with two-stage detector Faster-RCNN, ImYOLOv4 achieves 10.36% AP50 and 36.36% AP75 increments. Moreover, our model surpasses anchor-free detector CenterNet by 9.97% AP50 and 25.28% AP75. In addition,

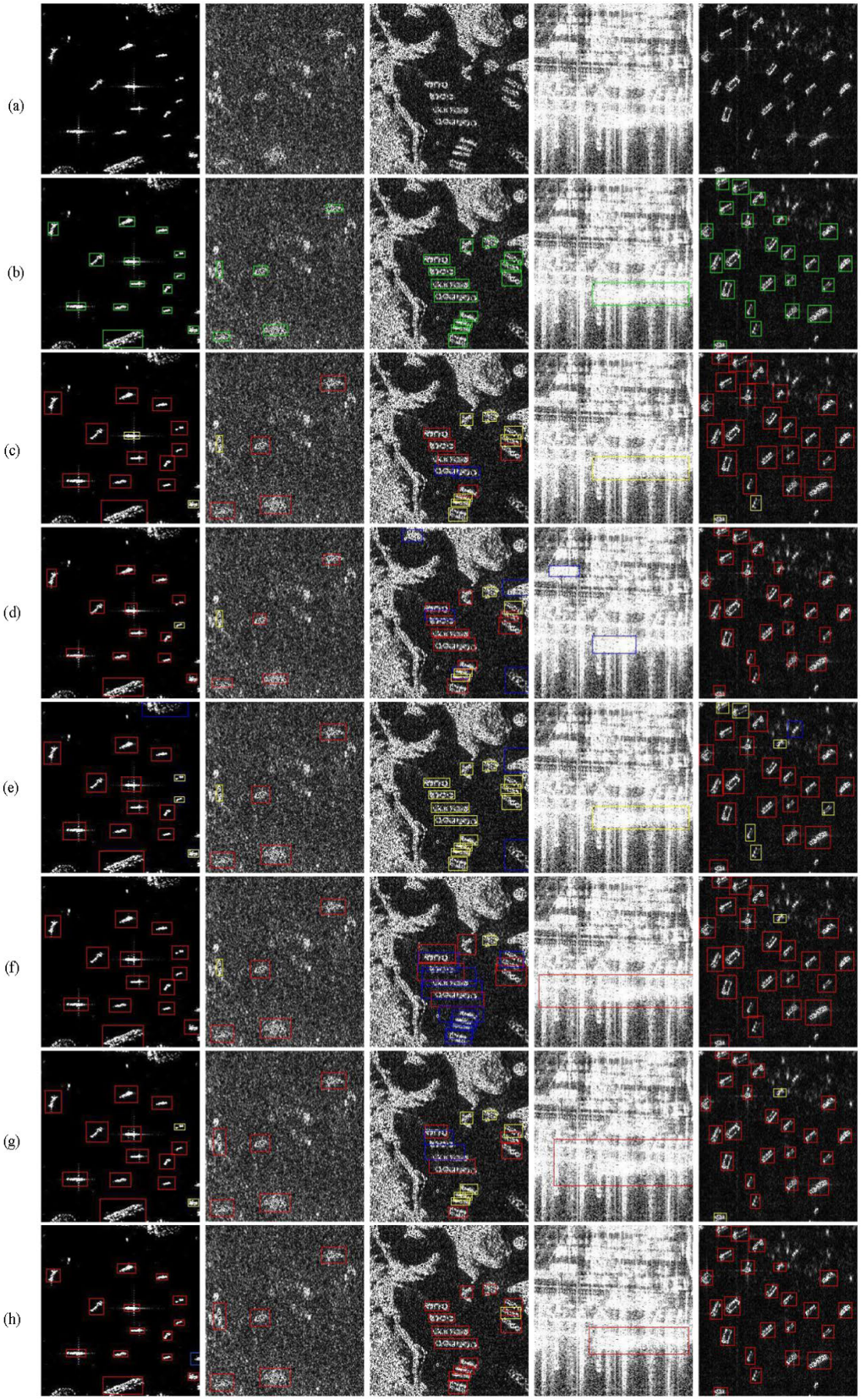


FIGURE 8. Detection results of detectors.

TABLE 8. Detection results of detectors.

Method	IoU=0.5				IoU=0.75				FPS
	AP_{50}	AP_L	AP_M	AP_S	AP_{75}	AP_L	AP_M	AP_S	
RetinaNet	85.70%	81.27%	96.20%	85.58%	41.52%	39.59%	64.18%	40.25%	39
Faster-RCNN	83.80%	63.53%	94.57%	69.23%	21.83%	40.01%	42.06%	5.59%	16
CenterNet	84.19%	15.68%	89.46%	79.74%	32.91%	4.23%	44.77%	26.14%	**78**
YOLOv3	90.98%	61.79%	95.96%	90.72%	48.15%	21.18%	62.65%	39.25%	61
YOLOv4	93.69%	74.80%	96.42%	91.28%	50.42%	25.64%	64.67%	40.00%	50
ImYOLOv4	**94.16%**	**83.64%**	**96.95%**	**93.33%**	**58.19%**	**42.49%**	**68.19%**	**50.24%**	42

as reflected by Figure 7, our method possesses a higher precision and recall curve than the state-of-the-art methods, which further shows the superiority of ImYOLOv4 over the others. When it comes to the running speed, our ImYOLOv4 is slower than CenterNet, YOLOv3, and YOLOv4 with 42 fps, but it is faster than RetinaNet and Faster-RCNN. In short, ImYOLOv4 achieves the better trade-off between detection accuracy and running speed, and we believe that the efficiency and simplicity of our method will benefit ship detection applications in the future research.

To further demonstrate the effectiveness in dealing with multiscale ship detection of ImYOLOv4, we divide the SSDD into three sub-datasets according to Table 2 and calculate evaluation metrics AP_L, AP_M. AP_S for large, medium, and small objects, respectively. From the results shown in Table 8, we can find out that the models present different detection abilities for multiscale ships. This is mainly because that the shapes of the ships in SSDD have a relatively extreme aspect ratio, and with the deepening of the network layers, the features of ships become weak, especially small-sized ships, so the detection accuracy is hard to guarantee. Moreover, to achieve a better performance, the models should take into account the effect of the complex backgrounds and noises. We embed TAM block to perform denoising operations and design the FPN structure to extract salient feature maps of small ships, which ensure the effectiveness of detecting small ships in front of complex backgrounds.

E. ANALYSIS ON MISSING SHIPS AND FALSE ALARMS

To show the detection performance of ImYOLOv4 vividly, we test it in some typical SAR images and the detection results are displayed in Figure 8. The different environment conditions include quiet sea, sea with waves, inshore land, backscatters noises and small ship cluster. And the rectangle box with different color represents different detection result, the rectangle with green, red, blue, and yellow color denotes the ground truth, detection target of detectors, false alarm and missing target, respectively. In Figure 8, (a) is the original SAR image and (b) represents the ground truth. (c)-(h) denotes the detection results of RetinaNet, Faster-RCNN, CenterNet, YOLOv3, YOLOv4, and ImYOLOv4, respectively. It is clear that our ImYOLOv4 model can distinguishes the ship targets better than the state-of-the-art methods even though the interference of complicated conditions. Although our method achieves excellent performance on SSDD, a few missing ships and false alarms still exist. As shown in the first and third column of (h) row, non-ship object is recognized as ship target due to similar features, and some ships are detected as one target because of their close distance. For missing ships, non-NMS [73] may improve the performance by adjusting the scores of other detection boxes so that close targets are not eliminated in the process. And sea-land semantic segmentation method [74] could serve as a supplement in image preprocessing which will benefit for the false alarms.

VI. CONCLUSION

In this paper, we propose a one-stage ship detector named improved YOLOv4 (ImYOLOv4) based on attention mechanism for accurate ship detection in SAR images. First, to achieve high accuracy of ship detection, we adopt YOLOv4 as the basic framework and apply CSPDarknet53 to extract multiscale feature maps. Then, the TAM module is designed based on attention mechanism to enhance the representational power of the network by dynamic feature denoising and recalibration. In addition, we construct a new FPN structure which combines the meaningful semantic information to solve with the problem of multiscale ship detection. Finally, we design a decoupled head with two branches to solve the conflict between classification and regression tasks. Extensive experimental results demonstrate that ImYOLOv4 has a promising performance on detecting ships in SAR images, while achieving a fast speed. We hope this report could help scholars get better experiences in future researches.

REFERENCES

[1] P. Gao, T. Tian, L. Li, J. Ma, and J. Tian, "DE-CycleGAN: An object enhancement network for weak vehicle detection in satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3403–3414, 2021.

[2] R. Chen, X. Li, and S. Li, "A lightweight CNN model for refining moving vehicle detection from satellite videos," *IEEE Access*, vol. 8, pp. 221897–221917, 2020.

[3] A. M. Johansson, M. M. Espeseth, C. Brekke, and B. Holt, "Can mineral oil slicks be distinguished from newly formed sea ice using synthetic aperture radar?" *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4996–5010, 2020.

[4] B. Brisco, M. Mahdianpari, and F. Mohammadimanesh, "Hybrid compact polarimetric SAR for environmental monitoring with the RADARSAT constellation mission," *Remote Sens.*, vol. 12, no. 20, p. 3283, Oct. 2020.

[5] T. Luti, P. De Fioravante, I. Marinosci, A. Strollo, N. Riitano, V. Falanga, L. Mariani, L. Congedo, and M. Munafò, "Land consumption monitoring with SAR data and multispectral indices," *Remote Sens.*, vol. 13, no. 8, p. 1586, Apr. 2021.

[6] C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain CFAR method for ship detection in HR SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 529–533, Apr. 2017.

[7] S. Wang, M. Wang, S. Yang, and L. Jiao, "New hierarchical saliency filtering for fast ship detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 351–362, Jan. 2017.

[8] X. Leng, K. Ji, X. Xing, S. Zhou, and H. Zou, "Area ratio invariant feature group for ship detection in SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2376–2388, Jul. 2018.

[9] K. Sun, Y. Liang, X. Ma, Y. Huai, and M. Xing, "DSDet: A lightweight densely connected sparsely activated detector for ship target detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 14, p. 2743, Jul. 2021.

[10] R. Yang, Z. Pan, X. Jia, L. Zhang, and Y. Deng, "A novel CNN-based detector for ship detection based on rotatable bounding box in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1938–1958, 2021.

[11] X. Fu and Z. Wang, "Fast ship detection method for SAR images in the inshore region," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3569–3572.

[12] X. Ke, X. Zhang, T. Zhang, J. Shi, and S. Wei, "SAR ship detection based on an improved faster R-CNN using deformable convolution," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3565–3568.

[13] T. Li, Z. Liu, R. Xie, and L. Ran, "An improved superpixel-level CFAR detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 184–194, Jan. 2018.

[14] W. Dai, Y. Mao, R. Yuan, Y. Liu, X. Pu, and C. Li, "A novel detector based on convolution neural networks for multiscale SAR ship detection in complex background," *Sensors*, vol. 20, no. 9, p. 2547, Apr. 2020.

[15] X. Wang and C. Chen, "Ship detection for complex background SAR images based on a multiscale variance weighted image entropy method," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 184–187, Feb. 2017.

[16] G. Yang, J. Yu, C. Xiao, and W. Sun, "Ship wake detection for SAR images with complex backgrounds based on morphological dictionary learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1896–1900.

[17] L. Liu, Y. Gao, F. Wang, and X. Liu, "Real-time optronic beamformer on receive in phased array radar," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 387–391, Mar. 2019.

[18] W.-L. Du, Y. Zhou, J. Zhao, X. Tian, Z. Yang, and F. Bian, "Exploring the potential of unsupervised image synthesis for SAR-optical image matching," *IEEE Access*, vol. 9, pp. 71022–71033, 2021.

[19] C. Mao, L. Huang, Y. Xiao, F. He, and Y. Liu, "Target recognition of SAR image based on CN-GAN and CNN in complex environment," *IEEE Access*, vol. 9, pp. 39608–39617, 2021.

[20] B. Shi, Q. Zhang, D. Wang, and Y. Li, "Synthetic aperture radar SAR image target recognition algorithm based on attention mechanism," *IEEE Access*, vol. 9, pp. 140512–140524, 2021.

[21] M. Zhu, G. Hu, S. Li, H. Zhou, S. Wang, Y. Zhang, and S. Yue, "ROS-Det: Arbitrary-oriented ship detection in high resolution optical remote sensing images via rotated one-stage detector," *IEEE Access*, vol. 9, pp. 50209–50221, 2021.

[22] M. Yang, C. Guo, H. Zhong, and H. Yin, "A curvature-based saliency method for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1590–1594, Sep. 2021.

[23] L. Chen, W. Shi, and D. Deng, "Improved YOLOv3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images," *Remote Sens.*, vol. 13, no. 4, p. 660, Feb. 2021.

[24] Y. Dong, F. Chen, S. Han, and H. Liu, "Ship object detection of remote sensing image based on visual attention," *Remote Sens.*, vol. 13, no. 16, p. 3192, Aug. 2021.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, Nov. 2004.

[26] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[27] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 32–39.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[29] J. Luo, C.-M. Vong, Z. Liu, and C. Chen, "An inverse-free and scalable sparse Bayesian extreme learning machine for classification problems," *IEEE Access*, vol. 9, pp. 87543–87551, 2021.

[30] H.-Z. Zhou and G. Yu, "Research on fast pedestrian detection algorithm based on autoencoding neural network and adaboost," *Complexity*, vol. 2021, Mar. 2021, Art. no. 5548476.

[31] J. Redmon, S. Divvala, and R. Girshick, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[32] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[33] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[34] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.

[35] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[36] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[37] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[39] Z.-W. Cai, Q.-F. Fan, R.-S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. ECCV*, vol. 4, 2016, pp. 354–370.

[40] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[41] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era: Models, Methods Appl. (BIGSARDATA)*, Nov. 2017, pp. 1–6.

[42] J. Jiao, Y. Zhang, H. Sun, X. Yang, X. Gao, W. Hong, K. Fu, and X. Sun, "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.

[43] Y. Mao, Y. Yang, Z. Ma, M. Li, H. Su, and J. Zhang, "Efficient low-cost ship detection for SAR imagery based on simplified U-Net," *IEEE Access*, vol. 8, pp. 69742–69753, 2020.

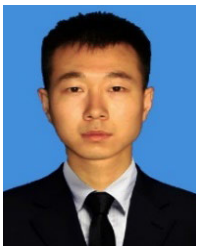
[44] J. Zhao, Z. Zhang, W. Yu, and T.-K. Truong, "A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images," *IEEE Access*, vol. 6, pp. 50693–50708, 2018.

[45] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[46] Z. Deng, H. Sun, S. Zhou, and J. Zhao, "Learning deep ship detector in SAR images from scratch," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4021–4039, Jun. 2019.

[47] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "Automatic ship detection based on retinanet using multi-resolution gaofen-3 imagery," *Remote Sens.*, vol. 11, no. 5, p. 531, 2019.

[48] L. Du, L. Li, D. Wei, and J. Mao, "Saliency-guided single shot multibox detector for target detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3366–3376, May 2020.
[49] X. Zhang, T. Zhang, and J. Shi, "High-speed and high-accurate SAR ship detection based on a depthwise separable convolution neural network," *J. Radars*, vol. 8, no. 6, pp. 841–851, 2019.
[50] J. Kong, Y. Gao, Y. Zhang, H. Lei, Y. Wang, and H. Zhang, "Improved attention mechanism and residual network for remote sensing image scene classification," *IEEE Access*, vol. 9, pp. 134800–134808, 2021.
[51] Q. Ou, L. Gao, and E. Zhu, "Multiple kernel K-means with low-rank neighborhood kernel," *IEEE Access*, vol. 9, pp. 3291–3300, 2021.
[52] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.
[53] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," *Remote Sens.*, vol. 11, no. 7, p. 765, Mar. 2019.
[54] M. Alkhaleefah, S.-C. Ma, T.-H. Tan, L. Chang, K. Wang, C.-P. Ko, C.-S. Ku, C.-A. Hsu, and Y.-L. Chang, "Accelerated-YOLOv3 for ship detection from SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2021, pp. 3030–3032.
[55] Y.-L. Chang, A. Anagaw, L. Chang, Y. Wang, C.-Y. Hsiao, and W.-H. Lee, "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, vol. 11, no. 7, p. 786, Apr. 2019.
[56] L. Han, T. Zheng, W. Ye, and D. Ran, "Analysis of detection preference to CNN based SAR ship detectors," in *Proc. Inf. Commun. Technol. Conf. (ICTC)*, May 2020, pp. 307–312.
[57] Z. Sun, M. Dai, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, 2021.
[58] J. Wang, X.-H. Yu, and Y.-S. Gao, "Mask guided attention for fine-grained patchy image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1044–1048.
[59] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019.
[60] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, p. 860, 2017.
[61] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, p. 4209, Oct. 2021.
[62] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.
[63] G. Masetti and F.-D. Giandomenico, "Analyzing forward robustness of feedforward deep neural networks with LeakyReLU activation function through symbolic propagation," in *Proc. PKDD/ECML Workshops*, 2020, pp. 460–474.
[64] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
[65] Q. Zhang, "System design and key technologies of the GF-3 satellite," *Acta Geodaetica et Cartographica Sinica*, vol. 46, no. 3, pp. 269–277, 2017.
[66] L. Huang, B. Liu, B.-Y. Li, W.-W. Guo, W.-H. Yu, Z.-H. Zhang, and W.-X. Yu, "OpenSARShip: A dataset dedicated to sentinel-1 ship interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 195–208, Jan. 2018.
[67] M. Everingham, A. Zisserman, and C.-K.-I. Williams, "The 2005 PASCAL visual object classes challenge," in *Proc. MLCW*, 2005, pp. 117–176.
[68] A.-R. Sutanto and D.-K. Kang, "A novel diminish smooth L1 loss model with generative adversarial network," in *Proc. IHCI*, vol. 1, 2020, pp. 361–368.
[69] N. Vosco, A. Shenkler, and M. Grobman, "Tiled Squeeze-and-excite: Channel attention with local spatial context," 2021, *arXiv:2107.02145*.
[70] J. Park, S. Woo, J.-Y. Lee, and I.-S. Kweon, "BAM: Bottleneck attention module," in *Proc. BMVC*, 2018, p. 147.
[71] S. Woo, J. Park, J.-Y. Lee, and I.-S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, vol. 7, 2018, pp. 3–19.
[72] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.
[73] J. Chu, Y. Zhang, S. Li, L. Leng, and J. Miao, "Syncretic-NMS: A merging non-maximum suppression algorithm for instance segmentation," *IEEE Access*, vol. 8, pp. 114705–114714, 2020.
[74] S. Wen, W. Tian, H. Zhang, S. Fan, N. Zhou, and X. Li, "Semantic segmentation using a GAN and a weakly supervised method based on deep transfer learning," *IEEE Access*, vol. 8, pp. 176480–176494, 2020.



YUNLONG GAO received the B.S. degree in computer science and technology, and the M.S. degree in computer software and theory from Jilin University, in 2015 and 2018, respectively. He is currently pursing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include object detection and image processing technology.

ZHIYONG WU received the Ph.D. degree in mechatronic engineering from the Chinese Academy of Sciences. He is currently a Research Fellow and a Doctoral Supervisor with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include machine vision technology and optical fiber communication.

MING REN received the B.S. and M.S. degrees in mechanical engineering from Harbin Engineering University, in 2017 and 2020, respectively. He currently plans to pursue the Ph.D. degree in pattern recognition. His research interests include object detection and computational imaging.

CHUAN WU received the Ph.D. degree in mechatronic engineering from the Chinese Academy of Sciences, in 2003. He is currently a Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include target tracking and image processing technology.