CSFF-MGDH: Cross-stage Feature Fusion and Decoupled Head with Mutual Guidance for SAR Ship Detection

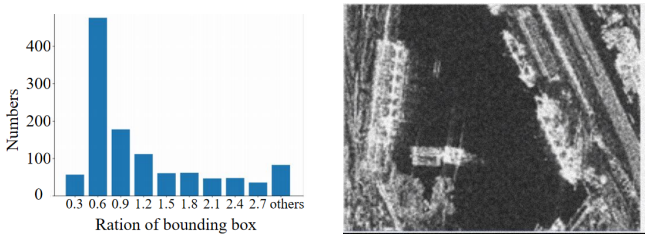
Yixin Qiao#, Xiaoxiao Yin#, Xinyuan Zhou, Shiyong Lan*, Wenwu Wang *Senior Member, IEEE*, Haohan Chen

***Abstract*—Deep learning-based synthetic aperture radar (SAR) ship detection methods have emerged as the leading techniques due to their strong feature extraction and generalization capabilities across various scenes and conditions. However, they still face challenges in distinguishing ships from complex backgrounds, especially in cases involving small or offshore vessels, dense inshore regions, or ships with textures and grayscale similar to their surroundings. To address these challenges, this paper introduces CSFF-MGDH, a novel SAR ship detector that integrates adaptive feature learning and a mutually guided decoupled head (MGDH) into the YOLOX framework. First, deformable convolution is incorporated into the backbone to overcome the limitations of standard square convolution in handling large ship deformations caused by severe noise in remote sensing images. Second, a cross-stage feature fusion module (CSFFM) is introduced to fuse features from adjacent layers, mitigating receptive field discrepancies in multi-layer feature maps caused by deformable convolution and reducing noise through local self-supervised interaction. Finally, a mutually guided decoupled head is designed to guide the regression branch using classification features, improving single-category object detection. Extensive experiments on the SSDD and HRSID datasets demonstrate that the proposed method substantially outperforms the baseline methods in detection accuracy.**

***Index Terms*—Image analysis, SAR ship detection, Deformable convolution, Adaptive feature fusion, Dutual guidance.**

I. INTRODUCTION

Recently, synthetic aperture radar (SAR) [1] has emerged as a powerful tool for maritime traffic monitoring and port management. The unique strength of SAR lies in its ability to provide detailed imagery regardless of weather conditions or lighting, which is crucial to keeping a close watch on sea traffic and overseeing port activities. However, the similar appearance between ships and background ocean clutter in SAR images makes them difficult to distinguish, and the various shapes of ships pose significant challenges to current SAR-based ship detection. The SAR ship detection dataset (SSDD) [2], as the first dataset that embodies the aforementioned characteristics (i.e., similar appearances between ships and background clutter, wide range of ship shapes), has been widely used for research on deep learning-based ship detection technology with SAR images. Fig. 1 shows the distribution statistics of different ships in the SSDD dataset, where Fig. 1(a) counts the number of ships with different aspect ratios and Fig. 1(b) shows the ships and their surrounding environment. It is clear that ships often blend in with their surroundings, making them difficult to distinguish. In addition, the size of the ships and their aspect ratios vary significantly.



(a) Aspect ratio statistic of SSDD. (b) An image from SSDD.

Fig. 1: Statistics on the distribution of ships in the SSDD. It is evident that ships closely resemble their surrounding environment (right), and they also display a diverse range of aspects (left).

SAR ship detection methods must be adaptable to ships of varying shapes and sizes. However, most existing algorithms rely on standard square convolutions with fixed sampling patterns (as shown in Fig. 3), which struggle to capture features of irregularly shaped targets. In contrast, deformable convolutions (DCNs) [3] can adaptively focus on key features by adjusting the sampling range of the convolutional kernel through learnable offsets. Furthermore, DCNv2 [4] enhances this flexibility by introducing a modulation mechanism that allows more precise control over sampling regions. Inspired by those, Cao et al. [5] proposed a multi-scale DCN to improve the detection accuracy for small targets with geometric deformation. Liu et al. [6] incorporated deformable convolutions to enhance the representation of irregularly shaped ships and employed a pyramid network of morphological and topological features to capture their intrinsic structural information.

To address the challenges posed by varying ship scales and rotation angles in SAR images, Hu et al. [7] proposed a local attention module (LAM) based on deformable convolution to enhance local feature extraction, along with a non-local attention module (NLAM) to capture global context, achieving a balance between local and global information. Bai et al. [8] designed a deformable convolution-based feature enhancement pyramid incorporating spatial enhancement and feature alignment modules to reduce scattering noise and misalignment in up-sampled deep features, thereby improving feature representation. More recently, Bao et al. [9] developed YOLO-LDFI, a lightweight SAR ship detection model that integrates linear deformable convolution (LDConv), a deformable cross-attention mechanism (DCAM), a feature-adaptive head (FA-

* Corresponding author: lanshiyong@scu.edu.cn. # Equal contribution.
Y. Qiao, X Yin, X. Zhou, S. Lan, and G. Deng are with College of Computer Science, Sichuan University, Chengdu, 610064, China. W. Wang is with the Center for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, U.K. The codes are available at https://github.com/SYLan2019/CSFF-MGDH.

Head), and an inner-EIoU loss into YOLOv11n, significantly boosting detection accuracy and efficiency in complex environments. Despite these advancements, deformable convolution-based approaches still risk expanding receptive fields beyond target regions, introducing background noise—a limitation not fully mitigated by conventional concatenation, summation, or attention mechanisms.

To address the impact of background noise on SAR images, Zhang et al. [10] proposed a YOLOX-based plug-and-play spatial patch detector module to predict the position of a ship and filter out the ocean background at the image level. Lu et al. [11] proposed an information enhancement module and a feature refinement module to enrich spatial contextual information. Tang et al. [12] used a pyramidal pooling attention network to mitigate the effect of background noise on ship detection by taking into account the saliency of ships in SAR images. Recently, Li et al. [13] proposed a method combining local context awareness and sea–land segmentation to precisely locate SAR targets and reduce scene interference, using multi-granularity knowledge distillation with target masking and spatial-channel attention for stable incremental detection. Another method, SDDFD [14], introduced smooth distribution distillation to adaptively weight spatial responses in complex backgrounds, and designed a depth-focused distillation with consideration of interlayer relationships to jointly optimize localization and classification for SAR target detection. However, these convolution-based methods struggle to handle unpredictable changes in ship shapes in the presence of natural interference found in SAR images.

The design of the detection head is crucial in object detection. For example, YOLOv3 [15] employs a traditional coupled head, where a single convolutional layer performs both classification and regression, offering simplicity but suffering from task interference. To address this, YOLOX [16] introduced a decoupled head that separates the two tasks, alleviating conflicts. However, SAR images are characterized by low resolution, high noise, and varying target orientations, making the regression task, i.e. precise localization, more challenging than the classification task, i.e. making binary decision on whether a target is a ship. A fully decoupled design neglects the complementary relationship between these tasks, thereby limiting localization accuracy.

To address these issues, we integrate DCNs into the YOLOX CSPDarkNet backbone [16], forming an adaptive feature extraction network termed DCN-CSPDarkNet. When extracting multilevel feature maps, deformable convolutions may extend receptive fields beyond target regions, introducing background noise. To alleviate this, we propose a cross-stage feature fusion (CSFF) module that employs deep features to generate dynamic masks, adaptively constraining shallow features. This design retains the flexibility of DCNs while suppressing noise through feature consistency optimization, enhancing robustness and fusion accuracy. Moreover, in SAR ship detection, factors such as low resolution, noise, and rotational variations lead to unstable local geometry, making regression weaker than classification. Fully decoupled detection heads overlook the complementarity between the two tasks. To address this, we design a mutually guided decoupled head (MGDH), where classification features generate attention maps to guide regression, and regression gradients guide classification toward spatially reliable regions. This mutual guidance reinforces key feature representation, reduces interference, and improves overall detection accuracy.

Our novel contributions can be summarized as follows.

1) We integrate the deformable convolution DCNs into the backbone network CSPDarkNet to construct a new backbone DCN-CSPDarkNet, which allows the detection of ships in different shapes and mitigates the impact of environmental noise.
2) We introduce a new CSFFM that exploits deep features to generate dynamic masks to adaptively constrain shallow features, preserving the flexibility of deformable convolutions while suppressing noise through feature consistency optimization. This helps reduce background noise caused by deformable convolutions when their receptive fields extend beyond target regions.
3) We develop a decoupled yet mutually guided head, where the classification head supervises the position regression task. This mutual guidance not only addresses the absence of task interaction in fully decoupled designs but also improves feature representation.

This paper is a significant extension of the conference version [17], with the following improvements over [17]. First, we conduct an in-depth analysis of the key challenges in the existing SAR ship detection methods, as presented in Section I and Section II. Second, we provide a detailed explanation of the implementation process of the proposed method, as presented in Section III. Third, by adding more comparative experiments, we thoroughly analyze how each module of our proposed method plays a role in the detection of ships with SAR images, and visualize the detection results in various complex environmental scenarios, as presented in Section IV. These extensions further demonstrate the robustness of our proposed method in handling ships of different shapes and complex environmental backgrounds.

II. RELATED WORKS

A. Conventional SAR Ship Detection Methods

Traditional SAR ship detection methods mainly include pixel-level and feature-based methods. Pixel-level methods [18] directly process the SAR image and detect the target by analyzing the brightness, phase, and other information of the pixels. These methods are simple and easy to implement, but they are sensitive to noise and prone to false detection in the complex background. Feature-based methods [19] perform detection using features from SAR images, such as edge, shape, and texture, but they often need to hand-craft features and are more sensitive to target changes.

The constant false alarm rate (CFAR) [18] is a traditional method commonly used for SAR image-based ship detection, which estimates the statistical properties of the background ocean clutter by statistically analyzing the neighborhood around each pixel in the SAR images. Ship targets in SAR images often induce significant changes in background characteristics, such as brightness and phase. The CFAR utilizes these

statistical properties to distinguish the target from the background and detects them by setting an appropriate threshold. In [19], the authors proposed the multi-scale rotation-invariant Haar-like CNN (MSRIHL-CNN), which combines low-level texture–edge features with high-level deep features through a multi-layer fusion strategy to achieve a more comprehensive representation of ship targets.

In [20], a moving window is utilized to extract two images from multi-view SAR images, and the cross-correlation coefficient between them is computed to generate a coherence image. Compared to the CFAR method, this approach [20] can effectively detect "invisible" ships embedded in sea surface speckle images. The work in [21] presents a technique based on the generalized likelihood ratio test for ship detection in SAR images, which takes into account the electromagnetic model applicable to sea surface clutter and scattering from ship echo signals. OR-CFAR [22], a robust constant false alarm rate detector for outliers under Gaussian clutter, proposed an adaptive threshold method for truncation of the sea clutter to eliminate high-intensity outliers in the local reference window and accurately model the probability density function of the sea clutter. Despite their simplicity of use, traditional SAR ship detection methods have limitations and challenges in dealing with complex background, diverse targets, and interfering factors.

B. Deep Learning Based SAR Ship Detector

Deep learning based object detection methods continue to emerge, offering a wide range of options and applications, which can be roughly categorized into two-stage detectors [23]–[25] and one-stage detectors [26], [27].

Two-stage detector. A two-stage detector is an object detection framework that locates and classifies objects in an image through a two-step process [23], [24]. In the first stage, the model generates a set of region proposals, i.e. candidate areas in the image that are likely to contain objects. In the second stage, these proposals are passed to a refinement network, which performs more precise classification of the object category and regression of the bounding box to fine-tune the object's location and size. This two-step process prioritizes accuracy and localization precision over computational speed, and is known for its highly precise results, especially for complex scenes.

Well-known classical two-stage detectors are the R-CNN series [24] [25] [28] [29]. Liao et al. [30] introduced a semi-supervised SAR ship detection framework based on Faster R-CNN [31], incorporating a decoder to reconstruct both labeled and unlabeled samples. By leveraging a large volume of unlabeled SAR images, the model effectively builds a robust latent space and learns more representative features of SAR targets. Wei et al. [32] proposed a multi-pooling channel attention (MPCA) mechanism that assigns adaptive weights to each channel to suppress redundant information and reduce false alarms. They further developed a top-level feature enhancement module to recover semantic information lost in high-level feature maps by integrating semantic cues from multiple feature spaces. Zhang et al. [33] presented an oriented SAR ship detection network utilizing soft thresholding and contextual information. A soft threshold quantization module was introduced to mitigate background noise interference, while additional constraints on the central coordinates and shapes of oriented bounding boxes enhance robustness to target shape variations in the ground truth.

Single-stage detector. A single-stage detector [26] performs object localization and classification in a single pass through a deep learning model. Unlike the two-stage detectors, the objective is to directly predict the bounding boxes and class labels from the input image by omitting the explicit region proposal step, thus enabling significantly faster object detection. This makes it highly suitable for SAR ship detection applications.

In [26], a single-stage detector is proposed for SAR ship images with complex noise backgrounds, by incorporating modules for anchor point refinement, target detection, and interconnecting blocks, effectively leveraging the advantages of the feature pyramid structure. In [27], the Transformer-based KSPFAM detector introduces a key scattering point aggregation module for improved localization and a context refinement module to reduce false alarms. In [34], a lightweight YOLOX variant with a MobileNetV3 backbone enhances SAR target fusion via cross-channel connections, multi-scale detection with dilated convolutions, and a lightweight attention mechanism in the Neck, further optimized by the Alpha-AIoU loss. In [19], MSRIHL-CNN employs a multi-scale, rotation-invariant Haar-like CNN that hierarchically fuses low-level texture-edge and high-level deep features for richer ship representation. AIS-FCANet [35] incorporates a frequency-spatial contextual module (FSCM) combining frequency and spatial attention to enhance spectral features and contextual understanding in cluttered scenes.

Yang et al. [36] employed a RetinaNet-based method to align backbone feature scales and introduced an adaptive IoU threshold to balance positive samples. In [37], a DCW-ELAN module within the YOLOv7 framework [38] combined coordinate attention and dilated convolutions for efficient multi-scale feature aggregation and small-target sensitivity. Zhao et al. [39] proposed a YOLOX-tiny-based denoising model integrating spatial–frequency domain information and a convolutional block attention module (CBAM) [40] for adaptive feature weighting. Shan et al. [41] used a deep dense network with attention to refine ship feature extraction, while [42] introduced YOLO-Lite, embedding channel and position attention modules to enhance localization accuracy.

Despite these advances, existing fusion strategies often neglect the inconsistency of features across different layers, and fully decoupled detection heads still struggle with regression due to unstable geometric characteristics and noise in SAR imagery. To address these issues, we propose a dynamic mask supervision mechanism within the cross-stage feature fusion module to adaptively constrain features, suppress noise interference, and ensure feature consistency. Furthermore, we design a mutually guided decoupled head that establishes collaborative optimization between classification and regression tasks, improving feature representations in critical regions.

III. METHODOLOGY

A. Method Overview

The architecture of the proposed CSFF-MGDH is shown in Fig. 2. The original image is first processed by the Focus module [16], which generates a feature map that is half the size of the original image but possesses four times the number of channels. This feature map is then fed into the improved backbone (DCN-CSPDarkNet) to derive the three-layer feature maps. These three-layer feature maps first undergo the proposed cross-stage feature fusion module (CSFFM) to facilitate local self-supervised interaction, enhancing the feature representation from lower to higher layers. Subsequently, a bottom-up fusion process is conducted to facilitate bidirectional information exchange between these three-layer feature maps. Finally, within the detection head module, the mutually guided decoupled head (MGDH) is introduced to establish a mutual guidance relationship between the classification and regression branches, thereby refining the detection accuracy. The implementation of each component is elaborated in the subsequent subsections, providing details of the network's functionality and operation.

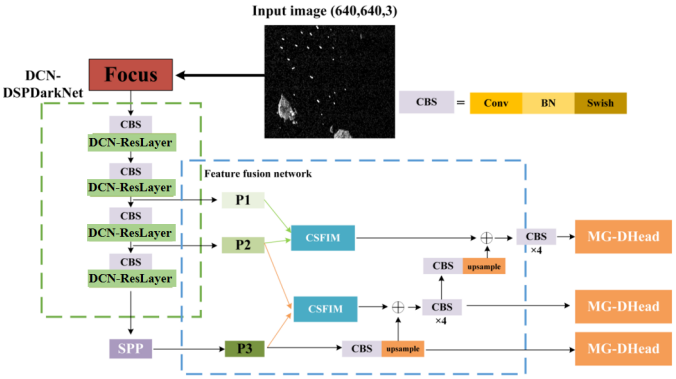


Fig. 2: Schematic diagram of our proposed method.

B. DCN-CSPDarkNet

The backbone network, CSPDarkNet [16], mainly employs 3×3 standard square convolutions. Although standard square convolutions work well for targets with regular shapes and moderate aspect ratios, they struggle with ships exhibiting extreme aspect ratios or heavy background interference, often resulting in noisy feature extraction and reduced detection accuracy. Illustrations can be found in the visualization comparison in Section V. Fig. 3 illustrates the sampling process of a ship using standard 3×3 square convolution and deformable convolution. The rectangular shape of the ship does not align well with the standard convolution kernel, resulting in incomplete coverage of the ship target and the inclusion of additional background noise, which can corrupt the ship's feature information and degrade the performance in regression of the ship's position. In contrast, deformable convolution can automatically adjust the sampling positions of the convolution kernel based on the ship's features, thereby mitigating the influence of environmental noise.

Specifically, the improved backbone DCN-CSPDarkNet is constructed based on the CSPDarkNet of YOLOX [16], in which we incorporate deformable convolution (DCN) to replace the traditional square convolution in the original ResBlock, thus forming the DCN-ResBlock. As illustrated in Fig. 2, the DCN-CSPDarkNet comprises multiple DCN-ResLayers, with each layer containing k ResBlocks ($k = 3, 9, 9, 3$). The DCN-CSPDarkNet outputs three layers of feature maps, which include the feature map P_1 from the second DCN-ResLayer, the feature map P_2 from the third DCN-ResLayer, and the feature map P_3 after the spatial pyramid pooling layer.

(a) Using a 3×3 standard convolution kernel to sample the SAR ship features.

(b) Using a 3×3 deformable convolution kernel to adaptively sample the SAR ship features.

Fig. 3: Different sampling methods used by standard convolution and deformable convolution. The standard square convolution samples at the fix positions, which may extract complex background noise when dealing with extreme aspect ratios. In contrast, the deformable convolution can adaptively adjust the positions of sampling points, thereby reducing negative impact of noise on the features.

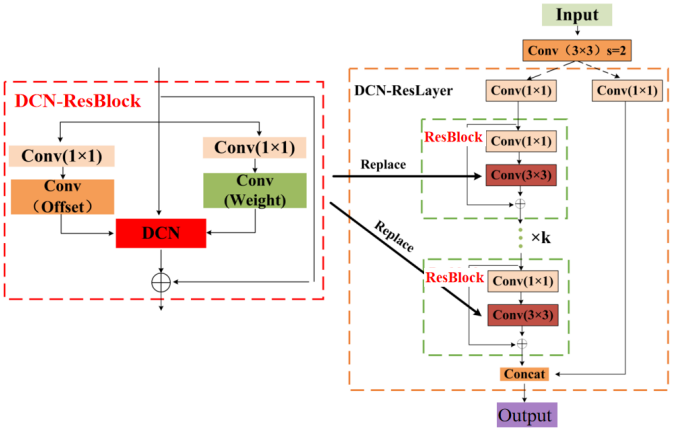


Fig. 4: Schematic diagram of DCN-ResLayer.

As illustrated in Fig. 4, before entering each DCN-ResLayer, we first apply a CBS block (i.e., 3×3 Convolution, Batch Normalization and Swish activation function) to downsample the input feature map (with H rows and W columns), reducing its size to half. Subsequently, the feature map is processed by a series of DCN-ResBlocks. Within each DCN-ResBlock, the left branch computes the sampling offsets required for DCN through a 1×1 convolution layer, while the right branch generates a feature map with dimension of $H \times W \times 9$, representing the weight information for each sampling point, through another 1×1 convolution layer. The DCN utilizes these computed offsets and weight information to adaptively sample the features of the ships. At the end

of each DCN-ResBlock, we employ a residual connection to concatenate the processed feature map with the original input feature map. After going through multiple DCN-ResBlocks consecutively, the feature map is further optimized to obtain improved ship features.

C. Cross-stage Feature Fusion Module

The cross-stage feature fusion module (CSFFM) is proposed to reduce additional noise introduced by fusing multi-layer features with different receptive fields.

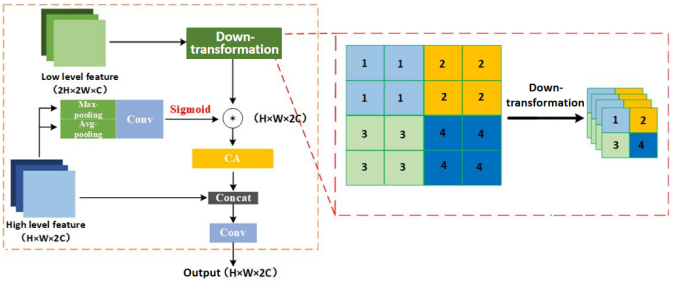


Fig. 5: Cross-stage feature fusion module.

As shown in Fig. 5, this module first performs a down-sampling operation on the low-level feature map $\boldsymbol{F}_l$, as follows:

$$\boldsymbol{F}_l' = down\_sampling\left(\boldsymbol{F}_l\right) \tag{1}$$

where $down\_sampling$ denotes the sampling operation, and $\boldsymbol{F}_l'$ denotes the feature map after downsampling, with the size of its rows and columns being half that of $\boldsymbol{F}_l$ and the number of channels being four times that of $\boldsymbol{F}_l$. As shown in the right part of Fig. 5, we transform the four local pixels of $\boldsymbol{F}_l$ into four adjacent channels. Note that we use the subscript l to represent the *low-level* feature. Likewise, we use the subscripts h, m, and w later to denote the *high-level* feature, *attention mask*, and *attention weight* map, respectively.

For the high-level feature map $\boldsymbol{F}_h$, the module employs 3×3 max pooling and 3×3 average pooling to obtain the attention mask $\boldsymbol{F}_m$, as follows:

$$\boldsymbol{F}_m = \sigma\left(Conv_3\left(\left[MaxPool_3\left(\boldsymbol{F}_h\right), AvgPool_3\left(\boldsymbol{F}_h\right)\right]\right)\right) \tag{2}$$

where σ denotes the Sigmoid function, $Conv_3$ denotes 3×3 convolution, $MaxPool_3$ and $AvgPool_3$ denote the 3×3 max pooling and 3×3 average pooling operation, respectively, and $[\cdot,\cdot]$ denotes the concatenation operation.

Subsequently, $\boldsymbol{F}_m$ is used to weight $\boldsymbol{F}_l$, and the resulting weighted feature map is fused with $\boldsymbol{F}_h$ to produce the fusion output $\boldsymbol{F}_{out}$. The above can be formulated as follows:

$$\boldsymbol{F}_{out} = Conv_{3\times3}\left(\left[\boldsymbol{F}_h, CA\left(\boldsymbol{F}_m \odot \boldsymbol{F}_l\right)\right]\right) \tag{3}$$

where CA denotes the channel attention module [40], $Conv_{3\times3}$ denotes the 3×3 convolution layer, and $\odot$ denotes the element-wise multiplication.

As depicted in Fig. 2, the proposed CSFFM leverages both channel and spatial information to implement local self-supervision on the three-tier feature maps generated by the DCN-CSPDarkNet, spanning from the lower-level to the higher-level layers. This approach can mitigate differences in receptive fields between contiguous feature maps, thereby enhancing the salient features of ships while reducing the impact of complex background noise.

D. Mutually Guided Decoupled Head

YOLOX introduced the concept of using a decoupled head in place of the traditional coupled head, implying that classification and regression tasks utilize two independent features for their respective predictions. Song et al. [12] highlighted that classification and regression tasks emphasize different areas of interest. Specifically, classification task determines the category of the target by assessing the similarity between the extracted feature and predefined categories, whereas the regression task concentrates on the discrepancies between the computed features and the ground-truth bounding boxes, aiming to refine the bounding box parameters.

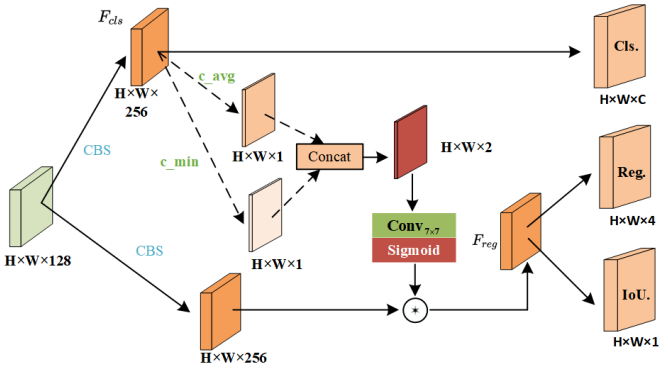


Fig. 6: Mutually guided decoupled head.

With regard to the above, the proposed MGDH is designed based on the original decoupled head as shown in Fig. 6. To achieve a lightweight model, only one CBS block is used in the input feature map to obtain the classification feature map $\boldsymbol{F}_{cls}$ and the regression feature map $\boldsymbol{F}_{reg}$. We initially obtain the low-frequency channel information and the average channel information of the classification feature map $\boldsymbol{F}_{cls}$,

$$\boldsymbol{F}_{cls} = CBS\left(\boldsymbol{F}\right), \boldsymbol{F}_{reg} = CBS\left(\boldsymbol{F}\right) \tag{4}$$

Subsequently, we concatenate these two types of information along the channel dimension to form a feature map $\boldsymbol{F}_{ma} \in \mathbf{R}^{H\times W\times2}$, as follows:

$$\boldsymbol{F}_{ma} = \left[MinPool_{channel}\left(\boldsymbol{F}_{cls}\right), AvgPool_{channel}\left(\boldsymbol{F}_{cls}\right)\right] \tag{5}$$

where $MinPool_{channel}$ and $AvgPool_{channel}$ denote the minimum and average pooling along the channel, respectively.

Next, a 7×7 standard large kernel convolution followed by a Sigmoid activation function is used to obtain an attention weight map $\boldsymbol{F}_w \in \mathbf{R}^{H\times W\times1}$. The above can be formulated as follows:

$$\boldsymbol{F}_w = \sigma\left(Conv_{7\times7}\left(\boldsymbol{F}_{ma}\right)\right) \tag{6}$$

where $Conv_{7\times7}$ denotes the 7×7 standard large kernel convolution. Finally, $\boldsymbol{F}_w$ is used to weight $\boldsymbol{F}_{reg}$ to complete the supervision of the classification branch on the regression branch.

It should be noted that during the forward pass of the network, the low-frequency channel information and the average channel information of the classification feature map are used to assign weights to the regression feature map, in order to supervise the prediction of the regression. Due to the back-propagation mechanism of the entire network, the results of the regression will also be propagated back to the classification branch after loss calculation, thereby achieving the goal of mutual guidance between classification and regression.

IV. EXPERIMENTS

A. Datasets

Our experiments utilize the SSDD dataset [2] and HRSID dataset [43]. SSDD is the first open data widely used in the study of deep learning-based ship detection techniques for SAR images. The specific details of the SSDD dataset are shown in Table I. Fig. 7 illustrates some images from the SSDD dataset. This dataset includes typical difficult-to-detect samples that require special attention in the practical application of SAR ship detection. These samples consist of small ships with indistinct features, densely parallel berthed ships within ports, large ships, ships in severe speckle noise, and ships with complex backgrounds. The SAR image samples in the SSDD dataset cover various resolutions, sensors, polarization modes, sea conditions, ship scenarios, and ship sizes. This diversity of data serves as the foundation for the construction of reliable detection models. HRSID is a dataset specifically designed for ship detection, semantic segmentation, and instance segmentation tasks in high-resolution SAR images. It comprises 5,604 images and 16,951 ship instances. The dataset encompasses diverse resolutions, polarization modes, sea conditions, and maritime scenarios, ensuring both diversity and practicality.

TABLE I: Details of the SSDD dataset

Transducers	RadarSat-2, TerraSAR-X, Sentinel-1
Polarization mode	HH, VV, VH, HV
Sampling location	Visakhapatnam and Qingdao
Scenario	inshore and offshore
Resolution	1m–15m
Number of images	1160
Number of ships	2456

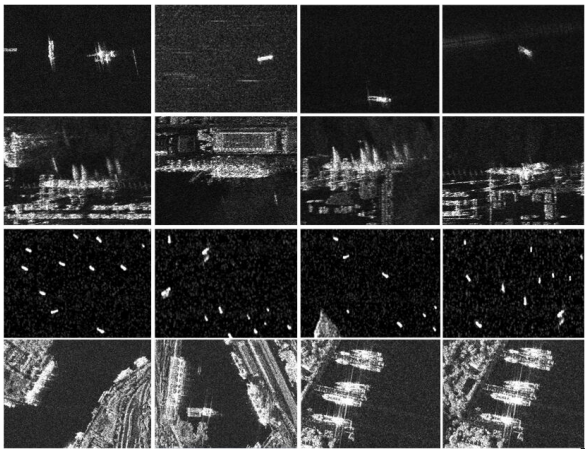


Fig. 7: Images from the SSDD dataset.

B. Evaluation Metrics

We use COCO detection metrics [44] to evaluate performance, which include AP, AP_{50}, AP_{75}, the number of model parameters ($Params$) and the inference speed in frame per second (FPS). Among them, AP_{50} represents the average precision when the intersection over union (IoU) threshold is 0.5, AP_{75} represents the average precision when the IoU threshold is 0.75 and AP is the average precision in the IoU threshold range of [0.5,0.95] and the step size of 0.05. AP_S, AP_M, and AP_L represent the detection precision for small, medium, and large objects, respectively, as defined in the COCO dataset [44]. IoU is the overlap ratio of the predicted bounding box to the ground truth bounding box, and its calculation process is as follows:

$$IOU = \frac{Area(B_{gt} \cap B_{pred})}{Area(B_{gt} \cup B_{pred})} \tag{7}$$

where B_{gt} is the ground truth bounding box and B_{pred} is the predicted bounding box. P stands for precision, R stands for recall, and the area under the PR curve is the value of the AP. The symbols $\cap$ and $\cup$ represent the intersection and union operations between two sets, respectively. The specific calculation process is as follows:

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

$$AP = \int_0^1 P(R)\, dR \tag{9}$$

where TP, FP, and FN are the number of true positive (TP), false positive (FP), and false negative (FN) samples, respectively, and $P(R)$ is the PR curve function. TP, FP and FN vary with the confidence threshold, and only when the confidence of the category is higher than the confidence threshold, the box will be calculated as a predicted box.

C. Implementation Details

The model proposed in this paper is based on the YOLOX [16] architecture and improved by mosaic to augment the input image data before training, which creates a new training sample by stitching four different images into one larger image. The images then undergo a series of preprocessing operations such as random flipping, filling, and finally scaling to the size of (640, 640, 3) for SSDD and (800, 800, 3) for HRSID dataset. The experiments are performed on 3060Ti GPU for training and evaluation with stochastic gradient descent (SGD) optimizer with a total of 300 epochs, learning rate set to 0.01, weight decay coefficient set to 0.0005, momentum to 0.9 and batch size to 4.

D. Comparison Analysis

To benchmark the proposed CSFF-MGDH, we use a variety of baseline algorithms such as SSD [45], Faster R-CNN [31], Mask R-CNN [29], YOLOX [16], FCOS [46], ATSS [47] and RetinaNet [48], which are all based on the MMdetection [49] platform. These baseline models are trained with default parameters. These object detection algorithms are compared

in terms of accuracy, inference speed (FPS), and number of parameters (Params). In addition, we compare it with the SAR ship detection algorithms YOLO-Lite [42], FEPS-Net [8], KSPFAM [27], YOLO-LDFI [9], and L-YOLOX [34].

TABLE II: Comparison of experimental results of different methods on SSDD dataset.

Method	Backbone	AP	AP_{50}	AP_{75}	FPS	Params (M)
SSD512 [45]	VGG16	64.3	94.3	75.4	43.8	24.39
Faster-RCNN [31]	ResNet50	58.8	93.9	66.6	41.7	41.12
Mask-RCNN [29]	DarkNet-53	61.6	93.9	71.4	40.5	43.07
FCOS [46]	ResNet50	52.2	89.6	56.2	78.1	31.84
ATSS [47]	ResNet50	60.2	92.1	69.0	44.4	31.89
RetinaNet [48]	ResNet50	50.5	84.9	54.9	43.7	41.7
YOLOX [16]	CSPDarknet	64.5	94.3	77.9	71.4	8.94
YOLO-Lite [42]	-	-	94.3	-	-	7.64
YOLOv7-tiny [38]	CSPDarkNet	-	96.5	-	-	13.0
FEPS-Net [8]	-	59.9	96.0	67.5	-	37.31
KSPFAM [27]	-	-	95.8	78.2	-	40.7
YOLO-LDFI [9]	-	66.4	**96.9**	-	-	-
L-YOLOX [34]	CSP-MobileNetV3_UP	65.6	94.9	78.7	48	-
CSFF-MGDH (Ours)	DCN-CSPDarkNet	**68.0**	96.7	**83.2**	53.1	22.52

- denotes that no result released. Bold/underlined text denotes the best/second-best results.

As shown in Table II, the proposed method offers an AP at 68. 0%, AP_{50} at 96. 7% and AP_{75} at 83.2%. The proposed method outperforms the baselines under all threshold conditions. For example, it improves the baseline YOLOX (by 3.5% in AP, by 2.4% in AP_{50} and by 5.3% in AP_{75}. Although our model underperforms slightly as compared with the latest YOLO-LDFI in AP_{50}, it exhibits a more stable performance in AP and AP_{75}. In addition, our method outperforms recent SAR ship detection methods (including FEPS-Net, KSPFAM, and L-YOLOX) across the AP, AP_{50}, and AP_{75} metrics. This indicates that our method can localize target regions more accurately in complex scenarios, while maintaining robust overall detection performance.

TABLE III: Comparison of experimental results of different methods on the HRSID dataset.

Method	AP	AP_{50}	AP_{75}	FPS	Params (M)
SSD512 [45]	63.6	87.9	72.8	20.5	24.39
Faster-RCNN [31]	65.0	88.5	73.8	19.2	41.12
Mask-RCNN [29]	62.7	86.8	71.1	40.1	43.07
FCOS [46]	64.7	89.0	73.6	39.5	31.84
ATSS [47]	63.7	87.3	72.1	20.9	31.89
RetinaNet [48]	64.9	88.4	73.4	-	18.9
YOLOX [16]	63.8	88.3	73.3	38.2	8.94
YOLOv7-tiny [38]	64.0	88.4	73.6	-	13.0
FEPS-Net [8]	65.2	90.7	74.3	-	37.31
KSPFAM [27]	-	87.2	70.3	-	40.7
YOLO-LDFI [9]	64.4	90.7	-	-	-
L-YOLOX [34]	65.6	89.1	75.1	-	-
CSFF-MGDH (Ours)	**66.2**	**90.8**	**76.4**	28.5	22.52

- denotes that no result released. Some baselines included in Table II are not shown here because they have no results reported on the HRSID dataset.

As shown in Table III, it can be observed that our method demonstrates outstanding performance on the HRSID dataset, significantly surpassing existing mainstream object detection algorithms. Specifically, our method achieves an average precision (AP) of 66.2, outperforming Faster R-CNN (65.9), RetinaNet (64.9), FCOS (64.7), YOLOX (63.8), YOLOv7-tiny (64.0), FEPS-Net (65.2), YOLO-LDFI (64.4), and L-YOLOX (65.6), highlighting its superior accuracy in complex scenarios. Furthermore, on the AP_{50} metric, our method achieves an impressive score of 90.8, further validating its robustness in detecting highly overlapping objects and significantly exceeding other comparative methods. On the more challenging AP_{75} metric, our method also excels with a score of 76.4, demonstrating clear superiority over other YOLO-series detectors.

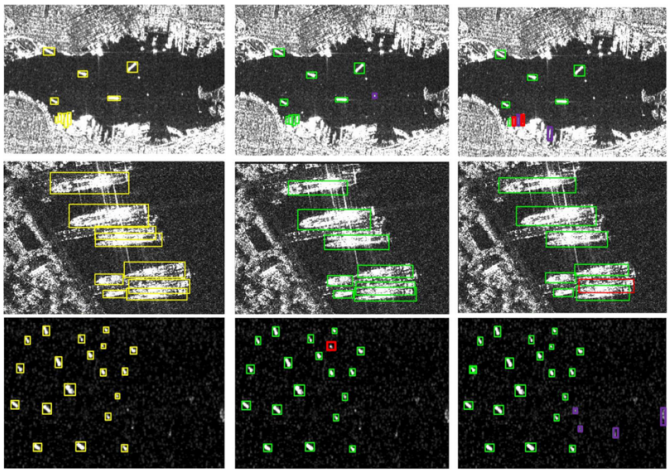


Fig. 8: Comparison of detection effect. The first column consists of real annotated images. The second column illustrates the detection effects of the proposed CSFF-MGDH, while the third column shows the detection effects of the baseline YOLOX. The yellow boxes represent the ground truth boxes, the green boxes indicate the true positive, the red boxes denote missed ship detections (false negative), and the purple boxes represent false positive.

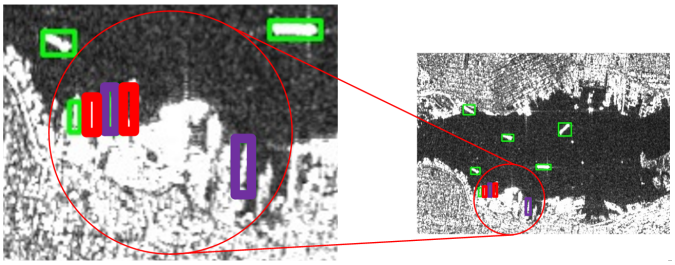


Fig. 9: A zoomed-in view of a difficult-to-detect region of a challenging inshore ship image, where the inshore ships are densely arranged with fewer and overlapping pixels. The baseline is not able to recognize the inshore ships with mixed boundaries and there are two missed detections. In addition, the inshore land background and an unrelated land background are also mis-detected as a ship.

Fig. 8 illustrates the detection effect of the different methods on three challenging SAR images in a complex noise background. The detection challenges in the first row of images lie in the dense emissions of inshore ships with a large aspect ratio, blurry boundaries between ships that are difficult to distinguish, as well as interference from numerous reefs, islands, and side lobes. The baseline method failed to detect some ships while incorrectly identifying land background as ships (the zoomed-in view is shown in Fig. 9). However, the proposed method successfully detected the challenging inshore ship samples, although there were individual cases where reefs were misidentified as ships. In the second row of images, numerous ships of similar size are densely arranged, making them difficult to distinguish. The baseline model failed to fully detect these ships, whereas our proposed method successfully identified all positive sample ships. In the third image, small ships are scattered among the sea surface clutter noise. The

proposed method missed one ship, while the baseline method misdetected many sea surface clutter noise elements in the bottom right corner as ships, resulting in numerous false positives.

E. Ablation Study

To fully validate the effectiveness of the proposed method, we performed ablation experiments on the SSDD dataset.

(1) Effectiveness of DCN-CSPDarkNet: To verify the effectiveness of adaptive feature extraction network DCN-CSPDarkNet, we compare the results using different bacebone network CSPDarkNet and DCN-CSPDarkNet, respectively. As shown in Table IV, after the replacement of CSPDarkNet with DCN-CSPDarkNet, the AP improves by 0.9%, AP_{50} improves by 0.4%, AP_{75} improves by 1.4%, the AP_S decreases by 0.7%, AP_M improves by 7.2%, AP_L improves by 10.1% and FPS decreases by 10.6%. In DCN-CSPDarkNet, the deformable convolution is used to replace the 3×3 standard square convolution and the results show that the detection accuracy improves significantly, especially for AP_M with the inference speed decreasing slightly. However, AP_S tend to decrease, which may be due to the fact that small ships are too small in area of pixels to adapt to the random receptive field of deformable convolution. Overall, the 3×3 deformable convolution with added sampling point weight information demonstrates superior adaptability in extracting SAR ship features and reducing complex background noise compared to the standard 3×3 square convolution, resulting in a significant improvement in model's detection performance.

TABLE IV: Comparison between CSPDarkNet and DCN-CSPDarkNet

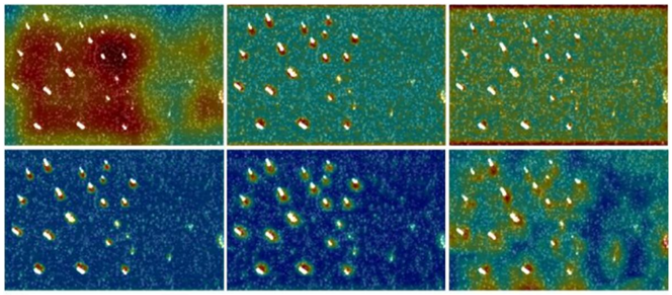
Backbone	**AP**	$AP50$	$AP75$	AP_S	AP_M	AP_L	**FPS**
CSPDarkNet	64.5	94.3	77.9	67.0	57.9	27.4	71.4
DCN-CSPDarkNet	65.4	94.7	79.3	66.3	65.1	37.5	60.8

(2) Effectiveness of CSFFM and MGDH: To show the effectiveness of the proposed CSFFM and MGDH, we add them to the backbone network DCN-CSPDarkNet, respectively. The results are shown in Table V.

TABLE V: The result of ablation experiments on the SSDD dataset.

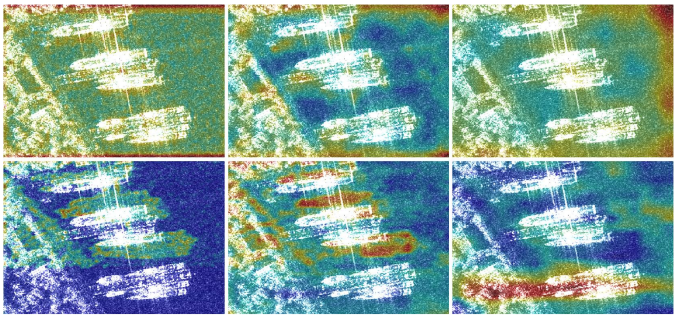
No.	**CSFFM**	**MGDH**	**AP**	$AP50$	$AP75$	**FPS**
1	×	×	65.4	94.7	79.3	60.8
2	✓	×	67.6	97.1	83.2	53.1
3	×	✓	67.0	96.7	82.5	61.1
4	✓	✓	68.0	96.7	84.2	52.2

The experiments and observations can be summarized as follows. **(1)** Experiment 1 presents the results of the adaptive feature extraction network DCN-CSPDarkNet. In Experiment 2, cross-stage fusion modules are sequentially added from top to bottom before its feature fusion network. After addition, AP increases by 2.2%, AP_{50} increases by 2.4%, and AP_{75} increases by 3.9%. However, this addition induces an extra computational cost, leading to a 7.7% decrease in the inference speed. This indicates that the CSFFM is capable of adapting to the receptive field of different multi-scale feature maps, which in turn can further adapt to the ship features. **(2)** In Experiment 3, the original decoupled detection head from Experiment 1 is replaced with the proposed MGDH, resulting in an increase of 1.6% in AP, 2.0% in AP_{50}, and 3.2% in AP_{75}, while the FPS is improved by 0.3. This suggests that the addition of this module enhances mutual supervision between the dual-branch feature maps during the backpropagation process, which somewhat strengthens the key regions of the feature map information while suppressing irrelevant noise areas, leading to more precise prediction results. Due to the lightweight design of MGDH, which eliminates one CBS convolution block compared to the original decoupled detection head, the inference speed is slightly improved. **(3)** Adding both the CSFFM and the MGDH based on Experiment 1, the accuracy is further improved, achieving an AP of 68.0%, AP_{50} of 96.7%, AP_{75} of 84.2% and a great inference speed of 52.2 FPS.



(a) First layer (b) Second layer (c) Third layer

Fig. 10: Visualization comparison of hierarchical feature maps for offshore SAR images. The first row presents the results from CSPDarkNet, while the second row shows the results from DCN-CSPDarkNet. The darker the color, the more the feature map focuses on that area.



(a) First layer (b) Second layer (c) Third layer

Fig. 11: Visualization comparison of hierarchical feature maps for inshore SAR images. The first row presents the visualization results of the feature maps from the CSPDarkNet, while the second row shows the visualization results from the adaptive feature extraction network DCN-CSPDarkNet. The darker the color, the more the feature map focuses on that area.

V. Visualization

To show the advantages of DCN-CSPDarkNet over CSPDarkNet, we visualized the last three layers of feature maps output by both DCN-CSPDarkNet and CSPDarkNet to intuitively observe the areas of interest in the images for the

two different feature extraction networks. Fig. 10 and Fig. 11 visualize the feature maps for offshore and inshore SAR images. It is not difficult to notice that the semantic information becomes increasingly rich and abstract from the first to the third layer of feature maps. In the offshore SAR images, it can be observed that the focus of CSPDarkNet is more scattered, whereas the improved backbone DCN-CSPDarkNet is more focused on the ship targets with less attention to background noise. In the inshore SAR images, there is little change in the focus on ships across different levels of feature maps from CSPDarkNet, whereas DCN-CSPDarkNet focuses on areas that match the shape of the ships. In addition, the deep feature map pays more attention to difficult-to-detect ship regions, such as confusing boundaries and ship-land adhered blocks. This shows that, DCN-CSPDarkNet is capable of adaptively adjusting the sampling range to enhance the key ship features to a certain extent and effectively mitigates the impact of the complex noise environment surrounding the ship.

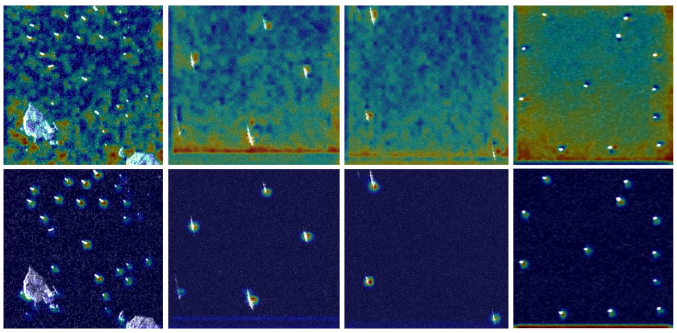
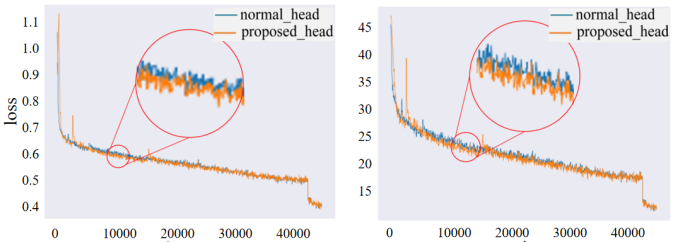


Fig. 12: Visualization comparison of SAR ship feature maps before and after CSFFM processing. The first row displays results before CSFFM processing, while the second row shows those after CSFFM processing.

We further compared the feature maps P_2 before and after CSFFM processing, as illustrated in Fig. 12. Prior to CSFFM, the feature maps contain rich details but are cluttered with disordered bright spots and noise, a side effect of the random receptive fields introduced by deformable convolutions. After CSFFM processing, however, noise is effectively suppressed, and ship features appear more concentrated and distinct. This enhancement stems from the adaptive multi-stage feature alignment mechanism of CSFFM, which retains the flexibility of deformable convolutions while reducing interference from random offsets.

Finally, we conducted an analysis of the regression loss and classification loss during the training process for both the proposed MGDH and the original decoupled detection head, in order to intuitively demonstrate their impact on the overall network. Fig. 13 shows the training curves for the classification loss and regression loss with different detection heads. It indicates that when using the proposed MGDH, the fluctuation of the classification loss is similar to that of the regression loss, suggesting a stronger correlation between the two branches. As the model converges, the classification loss for both the original decoupled detection head and the proposed MGDH is somewhat similar. However, the regression loss of the proposed MGDH is lower than that of the original decoupled



(a) classification loss (b) regression loss

Fig. 13: Comparison of training losses for different detection heads. The similarity between the classification loss and regression loss in the MGDH suggests a significant correlation between the two branches. Furthermore, the reduced regression loss highlights that the MGDH has strong ability to resist background noise.

detection head, and the smaller amplitude of loss fluctuation demonstrates that the proposed MGDH has stronger anti-interference ability and robustness. This indicates that in the MGDH, the classification branch effectively supervises the regression branch, enabling the detection head with enhanced feature decoding and prediction capabilities.

VI. Conclusion

We have presented a novel YOLOX-based SAR ship detection network, named CSFF-MGDH. Firstly, we incorporated deformable convolution into the CSPDarkNet backbone to address the limitations of conventional convolutional neural networks, which struggle to adapt to ships of varying shapes and sizes and often confuse complex background noise with ship features. This enables the backbone network to adaptively capture features of ships with varying shapes while reducing attention to ocean background noise. Secondly, we proposed a cross-stage feature fusion module to reduce the additional noise introduced during the fusion of multi-layer features with different receptive fields. This module achieves local self-supervision of feature maps between adjacent layers, effectively alleviating the differences in receptive field across multi-scale feature maps. Finally, in the decoupled detection head, we incorporated mutual guidance. The mutually guided decoupled head enables classification and regression tasks to supervise and guide each other, strengthening key features in their respective regions while further reducing the influence of irrelevant background noise. Experimental results demonstrated the superior performance of the proposed method as compared with state of the art methods. In the future, we will further study feature modeling of visual objects in strong noise and weak signal environments to improve the robustness of detection for small objects in SAR images.

References

[1] K. Tomiyasu, "Tutorial review of synthetic-aperture radar (SAR) with applications to imaging of the ocean surface," *Proceedings of the IEEE*, vol. 66, no. 5, pp. 563–583, 1978.

[2] T. Zhang, X. Zhang *et al.*, "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sensing*, vol. 13, no. 18, p. 3690, 2021.

[3] J. Dai, H. Qi *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[4] X. Zhu, H. Hu *et al.*, "Deformable ConvNets V2: More deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9300–9308.

[5] D. Cao, Z. Chen, and L. Gao, "An improved object detection algorithm based on multi-scaled and deformable convolutional neural networks," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, p. 14, 2020.

[6] Q. Liu, D. Li *et al.*, "MT-FANet: A morphology and topology-based feature alignment network for SAR ship rotation detection," *Remote Sensing*, vol. 15, no. 12, p. 3001, 2023.

[7] Q. Hu, S. Hu, and S. Liu, "BANet: A balance attention network for anchor-free ship detection in SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[8] L. Bai, C. Yao *et al.*, "Feature enhancement pyramid and shallow feature reconstruction network for SAR ship detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 1042–1056, 2023.

[9] W. Bao, S. Chen, J. Zhao, and X. Lin, "YOLO-LDFI: A lightweight deformable feature-integrated detector for SAR ship detection," *Journal of Marine Science and Engineering*, vol. 13, no. 9, p. 1724, 2025.

[10] J. Zhang, W. Sheng *et al.*, "MLBR-YOLOX: An efficient SAR ship detection network with multilevel background removing modules," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 5331–5343, 2023.

[11] Z. Lu, P. Wang, Y. Li, and B. Ding, "A new deep neural network based on SwinT-FRM-ShipNet for SAR ship detection in complex near-shore and offshore environments," *Remote Sensing*, vol. 15, no. 24, p. 5780, 2023.

[12] G. Tang, H. Zhao *et al.*, "PPA-Net: pyramid pooling attention network for multi-scale ship detection in SAR images," *Remote Sensing*, vol. 15, no. 11, p. 2855, 2023.

[13] Y. Li, L. Du, H. Liu, and Y. Guo, "Class-incremental sar ship detection and classification via context-robust exemplar replay and multigranularity knowledge distillation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 61, no. 4, pp. 9276–9289, 2025.

[14] J. Li, Z. Cui, Y. Tian, Z. Zhou, Q. Gan, and Z. Cao, "Smooth distribution and depth-focused distillation-based class-incremental learning for sar target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–16, 2025.

[15] A. Farhadi and J. Redmon, "YOLOv3: An incremental improvement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 1804, 2018, pp. 1–6.

[16] Z. Ge, S. Liu *et al.*, "YOLOX: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[17] Y. Qiao, X. Yin *et al.*, "SAR ship detector using cross-stage feature fusion and decoupled head with mutual guidance," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2025, pp. 1–5.

[18] F. C. Robey, D. R. Fuhrmann *et al.*, "A CFAR adaptive matched filter detector," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 28, no. 1, pp. 208–216, 1992.

[19] J. Ai, R. Tian *et al.*, "Multi-scale rotation-invariant haar-like feature integrated CNN-based ship detection algorithm of multiple-target environment in SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 10 070–10 087, 2019.

[20] K. Ouchi, S. Tamaki, H. Yaguchi, and M. Iehara, "Ship detection based on coherence images derived from cross correlation of multilook SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 3, pp. 184–187, 2004.

[21] P. Iervolino and R. Guida, "A novel ship detector based on the generalized-likelihood ratio test for SAR imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3616–3630, 2017.

[22] J. Ai, Q. Luo *et al.*, "Outliers-robust CFAR detector of gaussian clutter based on the truncated-maximum-likelihood estimator in SAR imagery," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 2039–2049, 2020.

[23] Z. Li, C. Peng *et al.*, "Light-head R-CNN: In defense of two-stage object detector," *arXiv preprint arXiv:1711.07264*, 2017.

[24] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, June 2018, pp. 7132–7141.

[26] S.-Q. Chen, R.-H. Zhan, and J. Zhang, "Robust single stage detector based on two-stage regression for SAR ship detection," in *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence*, 2018, pp. 169–174.

[27] Y. Yin, Z. Yang *et al.*, "Ship detection transformer in sar images based on key scattering points feature aggregation and context feature refinement," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 17 820–17 836, 2025.

[28] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[29] K. He, G. Gkioxari *et al.*, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[30] L. Liao, L. Du, and Y. Guo, "Semi-supervised SAR target detection based on an improved faster R-CNN," *Remote Sensing*, vol. 14, no. 1, p. 143, 2021.

[31] S. Ren, K. He *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[32] F. Wei and X. Wang, "SAR ship detection based on convnext with multi-pooling channel attention and feature intensification pyramid network," *Sensors*, vol. 23, no. 17, p. 7641, 2023.

[33] C. Zhang, G. Gao *et al.*, "Oriented ship detection based on soft thresholding and context information in SAR images of complex scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2023.

[34] Y. Hao, J. Wu *et al.*, "A robust anchor-free detection method for SAR ship targets with lightweight CNN," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–19, 2025.

[35] W. Xue, J. Ai *et al.*, "AIS-FCANet: Long-term AIS data assisted frequency-spatial contextual awareness network for salient ship detection in SAR imagery," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–6, 2025.

[36] R. Yang, Z. Pan *et al.*, "A novel CNN-based detector for ship detection based on rotatable bounding box in SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1938–1958, 2021.

[37] S. Zhu and M. Miao, "Lightweight high-precision SAR ship detection method based on YOLOv7-LDS," *Plos One*, vol. 19, no. 2, p. e0296992, 2024.

[38] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2023, pp. 7464–7475.

[39] C. Zhao, X. Fu *et al.*, "LPDNet: A lightweight network for SAR ship detection based on multi-level laplacian denoising," *Sensors*, vol. 23, no. 13, p. 6084, 2023.

[40] S. Woo, J. Park *et al.*, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–19.

[41] H. Shan, X. Fu *et al.*, "SAR ship detection algorithm based on deep dense sim attention mechanism network," *IEEE Sensors Journal*, vol. 23, no. 14, pp. 16 032–16 041, 2023.

[42] X. Ren, Y. Bai, G. Liu, and P. Zhang, "YOLO-Lite: An efficient lightweight network for SAR ship detection," *Remote Sensing*, vol. 15, no. 15, p. 3771, 2023.

[43] S. Wei, X. Zeng *et al.*, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120 234–120 254, 2020.

[44] T.-Y. Lin, M. Maire *et al.*, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[45] W. Liu, D. Anguelov *et al.*, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[46] Z. Tian, C. Shen *et al.*, "FCOS: Fully convolutional one-stage object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9626–9635.

[47] S. Zhang, C. Chi *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, pp. 9759–9768.

[48] T.-Y. Lin, P. Goyal *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[49] K. Chen, J. Wang *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.