BBox-Free SAR Ship Instance Segmentation Method Based on Gaussian Heatmap

Fei Gao, Fengjun Zhong, Jinping Sun, *Member, IEEE,* Amir Hussain, and Huiyu Zhou

***Abstract*—Recently, deep learning methods have been widely adopted for ship detection in synthetic aperture radar (SAR) images. However, many of the existing methods miss adjacent ship instances when detecting densely arranged ship targets in inshore scenes. Besides, they suffer from the lack of precision in the instance indication information and the confusion of multiple instances by a single mask head. In this paper, we propose a novel center point prediction algorithm, which detects the center points by finding a long distance variation relationship between two points. The whole prediction process is anchor-free and does not require additional bounding box (BBox) predictions for non-maximum suppression (NMS). Therefore, our algorithm is BBox-free and NMS-free, solving the problem of low recall rate when conducting NMS for densely arranged targets. Furthermore, to tackle the deficiency of position indication information in localization tasks, we introduce a feature fusion module with feature decoupling (FD). This module uses classification branch to provide guidance information for localization branch, while suppressing the influence of the gradient flow mixing, effectively improving the algorithm's segmentation performance of ship contours. Finally, through principal component analysis (PCA) of the Gaussian distribution covariance matrix, we propose a loss function based on the distance between centroids and the difference of angle, called centroid and angle constraint (CAC). CAC guides the network in learning the criterion that a single dynamic mask head is only valid for a single instance. Experiments conducted on polygon segmentation SAR ship detection dataset (PSeg-SSDD) and high resolution SAR images dataset (HRSID) demonstrate the effectiveness and robustness of our method.**

***Index Terms*—BBox-free, feature decoupling (FD), instance segmentation, ship detection, synthetic aperture radar (SAR).**

I. Introduction

WITH the advantage of all-weather, all-day imaging, synthetic aperture radar (SAR) plays a key role in water traffic surveillance, fisheries monitoring, marine vessel management and intelligence acquisition [1]–[3]. As the fundamental application of SAR image, SAR ship target detection has received extensive attention in recent years.

Traditional SAR ship detection methods include constant false alarm rate (CFAR) [4], generalised likelihood ratio test (GLRT) [5], and visual significance [6]. Typically, these methods involve four stages: terrestrial masking, pre-processing, pre-screening and target identification [7]. Due to the complexity of the detection steps and the reliance on a priori statistical properties of the image, these methods are not robust enough when faced with changes in imaging radar or the imaging scene. In addition, the parameters of these methods need to be adjusted accordingly before the detection and their detection speed dissatisfies with demands of practical application [8].

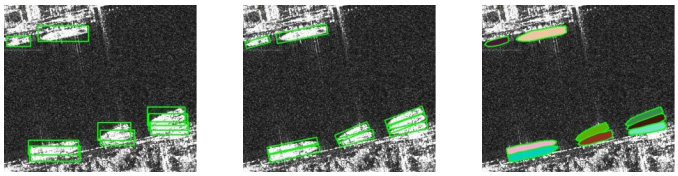

Fig. 1. Illustrative results. (a) BBox detection. (b) RBox detection. (c) Instance segmentation.

In recent years, with the development of deep learning technology, algorithms based on deep convolutional neural network (DCNN) have achieved great success in the field of optical target detection. Compared with traditional algorithms, DCNN has greatly improved in terms of detection accuracy, detection speed and robustness. This novel technology is also becoming widely adopted in the field of SAR ship target detection [9]–[16].

Ship target detection in SAR images encompasses three distinct levels of detection tasks: vertical bounding box (BBox) detection, rotated bounding box (RBox) detection, and instance segmentation. Fig. 1 visualizes the detection results of three task levels. The BBox contains massive background pixels, and adjacent targets exhibit significant overlap when densely arranged. The RBox detection predicts the angle of bounding box for refined detection. However, it still contains background pixels and faces problems of boundary discontinuity and angle periodicity [17]. Instance segmentation performs pixel-level segmentation for each target, and can completely suppress background pixels. As the demand for detection accuracy continues to rise, instance segmentation has garnered increasing attention within the field of SAR ship target detection.

The current mainstream instance segmentation algorithms can be categorized into anchor-based and anchor-free algorithms. The majority of anchor-based algorithms operate as

This work was supported in part by the National Natural Science Foundation of China under Grant 62371022.

Fei Gao, Fengjun Zhong and Jinping Sun are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: feigao2000@163.com; 18231033@buaa.edu.cn; sunjinping@buaa.edu.cn).

Amir Hussain is with the Cyber and Big Data Research Laboratory, Edinburgh Napier University, EH11 4BN Edinburgh, U.K. (e-mail: A.Hussain@napier.ac.uk).

Huiyu Zhou is with the Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K. (e-mail: hz143@leicester.ac.uk).

two-stage algorithms, involving the proposal of regions of interest (ROI) and subsequent ROI alignment [18]–[20]. In the ROI proposal phase, there is a balance maintained between positive and negative samples, leading to these algorithms achieving superior detection accuracy compared to anchor-free algorithms [21]. Anchor-free algorithms, in contrast, do not require the configuration of anchor hyperparameters, and the network structure can be entirely composed of convolution operations [22], [23]. Consequently, these algorithms offer a simpler and more flexible network structure, resulting in faster inference.

The most fundamental and important step in target detection is the distinction of multiple targets within a single image. Regardless of whether it is anchor-based or anchor-free algorithms, these algorithms use non-maximum suppression (NMS) to eliminate redundancy in the detected BBoxes, taking the center points of remaining BBoxes as the object center points [24]–[26]. During NMS, the choice of an appropriate intersection over union (IoU) threshold is of paramount importance. Given the proximity of ships in densely arranged scenarios, the IoU between the BBoxes of adjacent ships tends to be high. A low IoU threshold can lead to the identification of adjacent ships as a single target, while a high IoU threshold may result in the labeling of a single ship as multiple targets. Detecting densely arranged ships poses a significant challenge in the realm of SAR ship instance segmentation.

Apart from the challenges of densely arranged ship detection, existing instance segmentation algorithms face the problem of unable to perceive the relative position relationship between pixels and targets during mask segmentation. The instance segmentation network consists of two parallel task branches, namely classification branch and localization branch. Tian et al. [22] propose to generate a relative distance feature map from the classification branch to guide the localization branch. In the field of SAR ship instance segmentation, ship targets exhibit significant variations in scale, aspect ratio and angle [8]. Learning an absolute distance to determine whether a pixel belongs to the target is infeasible. Furthermore, the spatial attention positions of the classification and localization branches are misaligned, resulting in a shift in the location of network attention. This, in turn, complicates the exchange of information between the two branches [27], [28].

In addition to the two challenges mentioned above, Dice Loss [29], the most widely used mask segmentation loss function, cannot provide more refined guidance for network training. Most of the existing instance segmentation methods use Dice Loss as the mask segmentation loss function [18], [22], [23], [26], [30]. However, Dice Loss only focuses on the intersection ratio between the predicted mask and the ground truth. During training process, mask segmentation errors can occur at any position but have the same IoU with ground truth. The same intersection ratio means same loss value for Dice Loss, which prevents the network from learning finer mask segmentation. As a result, the network is unable to distinguish between the case where a single mask head makes an error in splitting a single instance and the case where a single mask head takes effect on multiple instances.

Gaussian heatmap contains rich semantic information and has been widely used in the field of target detection [25], [26], [31], [32]. All these methods use the local peak points of the Gaussian heatmap for center point proposal and combine it with NMS for center point de-redundancy. The branches of their networks are parallel to each other. They fail to fully utilise the trends in heat values and instance features embedded in the Gaussian heatmap. Their application of Gaussian heatmaps is limited to center point proposal, where they simply employ the Gaussian peak points as the center point proposals and subsequently combine them with BBox-based NMS for de-redundancy. In contrast, we make a comprehensive utilization of Gaussian features. During center point detection, different from BBox-based NMS, we achieve center point de-redundancy without BBox and NMS based on Gaussian distribution long-range variation rule. During mask segmentation, different from their parallel network structure of branches, we fuse localization branch features with classification branch features and introduce feature decoupling module to solve the problem of spatial location mismatch between two tasks. As a result, instance indication information with instance attributes is introduced to the localization branch. During training stage, different from commonly used Dice Loss, we introduce CAC to mask segmentation based on Gaussian distribution, guiding the network to learn the criterion that a single mask head only takes effect on a single instance. Our contributions can effectively improve the performance of instance segmentation from the perspective of center point detection, mask segmentation, and training loss function.

To address the challenge of detecting densely arranged ship targets, our method encodes each ship target as an elliptical Gaussian heatmap and regresses the heatmap by classification branch. Subsequently, by leveraging the mathematical characteristic that heat values decrease as pixels move farther from the target, we design a center point extraction and de-redundancy algorithm. The entire algorithm flow is anchor-free, BBox-free and NMS-free, improving the ability of the network to detect densely arranged targets.

To address the problem that the localization branch is unable to perceive the relative distance between the pixel and the target, we propose to fuse the feature map of classification branch into the localization branch. By feature fusion, the ship features and the relative distance information from classification branch can guide the mask segmentation process. Further, to solve the problem of spatial mismatch between two task branches, we propose a feature decoupling (FD) module to separate the gradient flow of different task branches. By applying FD to the feature fusion process, the gradient flow from the localization branch will be suppressed to prevent the spatial attention position of the network from swinging.

In order to address the problem that Dice Loss cannot perform a more refined evaluation of the predicted mask, we conduct principal component analysis (PCA) on the covariance matrix of the Gaussian distribution to obtain the orientation vector. Combining the centroid position and orientation vector, we design the centroid and angle constraint (CAC), which guides the network to learn the criterion that a single dynamic mask head is only valid for a single instance.

The main contributions of our research are summarized as

follows.

1) The center point detection algorithm based on the elliptical Gaussian distribution is proposed. The center point detection process relies entirely on the mathematical characteristics of the elliptical Gaussian distribution. The entire detection process is anchor-free, BBox-free, NMS-free, and exhibits a robust detection performance, particularly in scenes involving densely adjacent ship targets.
2) The spatial decoupled feature fusion method is proposed. We design FD module to achieve spatial decoupling between different branches with conflicting spatial locations. With FD, the classification branch is no longer influenced by the gradient flow from the localization branch, which causes the problem of task reversal. The classification branch can directly provide guidance information for the localization branch, helping it to perceive the relative position relationship between pixels and targets.
3) The CAC is designed based on the principle of PCA. This constraint addresses the limitation of the Dice Loss in assessing disparities in centroid location and orientation between masks. It serves to instruct the network to learn the criterion that a single dynamic mask head is only valid for a single instance.

The rest of this paper is divided into four parts. Section II describes the related work. Section III details the methodology of this paper. Section IV describes the experiments conducted on the polygon segmentation SAR ship detection dataset (PSeg-SSDD) [8] and high resolution SAR images dataset (HRSID) [33]. Section V concludes this paper.

II. Related Work

A. Anchor-Based Instance Segmentation Algorithms

Anchor-based instance segmentation algorithms use anchors as the target samples, requiring pre-setting hyperparameters such as the size and aspect ratio of the anchor.

The two-stage algorithm Mask R-CNN [18] is the most classic anchor-based instance segmentation algorithm. It extends the BBox detection algorithm to instance segmentation field by additionally designing a mask prediction branch based on Faster R-CNN [30]. Later researchers propose a series of R-CNN networks based on Mask R-CNN as the baseline. Cascade R-CNN [19] solves the problems of overfitting and inference-time quality mismatch between detector and test hypotheses by cascading several detection networks with different IoU thresholds. Mask Scoring R-CNN [20] designs an additional mask IoU branch to correct the deviation between mask quality and mask Score. Hybrid Task Cascade for Instance Segmentation [34] cascades multi-task information flow of different stages to improve network's performance. PointRend [35] optimizes the segmentation of object edges through point prediction. Instances as Queries [36] introduces Query into the instance segmentation field, achieving the flow of effective information through continuous cascading.

In the field of SAR ship target detection, Wu et al. [10] improve the instance segmentation accuracy through the interaction of target detection branch and instance segmentation branch. Su et al. propose HQ-ISNet [11] to improve the resolution of feature maps through a high-resolution feature pyramid network. Sun et al. [37] propose a multi-scale feature pyramid network (MS-FPN) to achieve the simultaneous detection and instance segmentation of marine ships in SAR images. Zhang et al. [38] propose a full-level context squeeze-and-excitation ROI extractor to extract feature subsets for the single level of feature. Zhang et al. [39] find existing models do not achieve mask interaction or offer limited interaction performance and propose a mask attention interaction and scale enhancement network (MAISE-Net). In addition to these, based on ROI [40]–[44], the researchers have proposed a series of improved algorithms from the perspectives of context compression, situational information interaction, etc., which contribute to the field of SAR ship instance segmentation.

Since the anchor-based algorithm can easily achieve a balance of positive and negative samples during ROI proposal process, this type of algorithm generally leads in segmentation accuracy. However, the cumbersome setting of anchor's hyperparameters and time-consuming ROI extraction make it weaker than anchor-free algorithms in generality and real-time performance.

B. Anchor-Free Instance Segmentation Algorithms

Compared with anchor-based instance segmentation algorithms, anchor-free instance segmentation algorithms do not require additional operations such as ROI extraction and can be designed as a fully convolutional one-stage algorithms with faster inference speed. Since the proposal of Focal Loss by He et al. [21], which solves the problem of the imbalance between positive and negative samples in training for anchor-free one-stage algorithms, the anchor-free one-stage algorithms start to match the detection accuracy of anchor-based two-stage algorithms.

Unlike anchor-based algorithms that use anchors for object localization, anchor-free algorithms usually use an additional network branch to predict the center points to achieve object localization. The fully convolutional FCOS [24] is the most widely used center point prediction algorithm, which encodes the center-ness of the instance target based on the distance of the pixel from the BBox border. A series of instance segmentation algorithms are proposed using FCOS as the baseline. Polar Mask proposed by Xie et al. [45] describes the distance between the sampled points on the contour and the center point in polar coordinates, generating masks through the connectivity of the sampled points. EmbedMask [46] learns semantic segmentation and pixel embedding, assigning semantic segmentation results to different instances based on the embedding distance between pixels and candidate boxes. CenterMask [47] designes a SAG-Mask branch and uses the spatial attention feature maps to predict masks on each detected BBox. CondInst [22] proposes the conditional convolution method, which encodes each instance in the parameters of the convolution head, making the mask head flexible and lightweight. Based on Yolov7, Yasir et al. [48] redesigned the structure of the one-stage fast detection network to improve high resolution SAR image segmentation one-stage detection.

In addition to using FCOS as baseline, some researchers use CenterNet [25] for center point prediction. For example, Gao et al. [26] use a standard Gaussian heatmap combined with BBox for center point prediction. In order to suppress redundant center points of the same instance, both center-ness based and Gaussian heatmap based algorithms need to perform NMS with BBox as the basic unit. The BBox-based NMS suffers from low recall rate when facing scenes with dense arrangement of instances.

Apart from center point, grid can also be used for instance localization. For example, the SOLO series [49], [50] assigns each instance to different grids, and the corresponding grid is responsible for segmenting this instance. Grid-based instance localization does not require additional BBox prediction, but the size of the grid determines the network's ability to discriminate the smallest targets. Single grid may contain multiple targets, leading to target omission.

C. Mismatch of Spatial Locations between Different Task Branches

In the field of target detection, the mismatch between the classification task and the localization task in terms of spatial location is a well known problem [27], [28], [51]. Most algorithms therefore use decoupled detection heads [18], [19], [22]. However, decoupled heads share the same feature maps, which can hinder the propagation of features and still have the problem of inconsistent targets during training [27].

Based on the attention mechanism, Gao et al. [26] and Yang et al. [52] propose feature decoupling networks to assign separate feature maps for different task branches, achieve the triage of different branch gradients. But their method can only assign features of different channels, which cannot achieve feature assignment of different spatial locations compared with the state-of-the-art attention mechanism [53]–[55].

The problem of spatial location mismatch between classification and localization tasks not only exists in the process of acquiring feature maps for different task branches, but also exists in multi-task feature fusion. Most of the existing feature fusion methods directly fuse and convolve feature maps from different task branches [56]–[58]. This way of feature fusion without any selection will cause the gradient streams of different task branches to flow into each other and suffer from the problem of inconsistent training objectives. Therefore, the spatial decoupling of the feature maps from different branches is also required in the process of feature fusion.

D. The Loss Function of Instance Segmentation

In contrast to semantic segmentation, instance segmentation requires distinguishing between different instances of the same category. During the mask segmentation of a single instance, instance segmentation, like semantic segmentation, classifies each pixel into two categories: foreground (positive samples) and background (negative samples). Thus, instance segmentation can be seen as a binary classification problem.

Binary cross-entropy (BCE) [59], [60] is the most common binary loss function that measures the information about the variability between two probability distributions. When the number of negative samples is much larger than the number of positive samples, the negative samples dominate, making the model tend to classify pixel points as negative samples [59], [61]. Therefore, BCE needs to weight the positive and negative samples to balance the difference in the number of positive and negative categories [21], [62], [63].

To deal with the problem of positive and negative sample imbalance, Milletari et al. propose the Dice Loss [29] to assess the similarity of two samples in the form of IoU. Due to the regional relevance of Dice Loss using IoU, which can well solve the problem of positive and negative sample imbalance, Dice Loss has become the most commonly used loss function for instance segmentation. To further alleviate the difference in the number of difficult and easy samples during training, Zhao et al. [64] and Prencipe et al. [65] propose weighting the difficult and easy samples. Wang et al. [66] concern that the background region also contains a large amount of information, and propose a weighted soft dice loss to mine the information in the background region. Gao et al. [26] concern that Dice Loss cannot effectively distinguish the position relationship between ships with the same degree of overlap, and propose centroid-distance-based loss to evaluate the difference of masks in the center-of-mass distance and centre region.

In the field of SAR ship detection, the directionality of the ship mask is an important ship feature. None of the above loss functions can evaluate the orientation difference of the mask, which is the focus of the proposed CAC in this paper.

III. Methodology

In this section, we first introduce the overall structure of our method. After this, three important parts of our method are detailed, namely, center point prediction method, FD and CAC. Finally, we detail the loss function of our method.

A. Network Structure

Fig. 2 illustrates the overall architecture of the proposed method in this article. The whole network consists of three parts.

1) Feature Extractor: In order to obtain feature maps of high resolution and high semantic information, we use DLA-34 [67] as feature extractor. DLA-34 extracts semantic information through the fusion of feature maps at different stages and scales. DLA network architecture includes two feature fusion methods, iterative deep aggregation (IDA) and hierarchal deep aggregation (HDA). IDA performs cross stage fusion, fusing feature maps under different stages. HDA fuses the features of multiple blocks within a stage through a tree like network structure. Through tree hierarchical fusion, DLA-34 can generate a single-scale feature map F_{out} with a downsampling rate of 4.

2) Classification Branch: The learning of Gaussian distribution is a regression problem. The more sample points the network has, the better the network learns the distribution. Therefore, the classification branch first upsamples the feature map F_{out} by a factor of 2 to obtain a feature map with a downsampling rate of 2. In order to accomplish both the

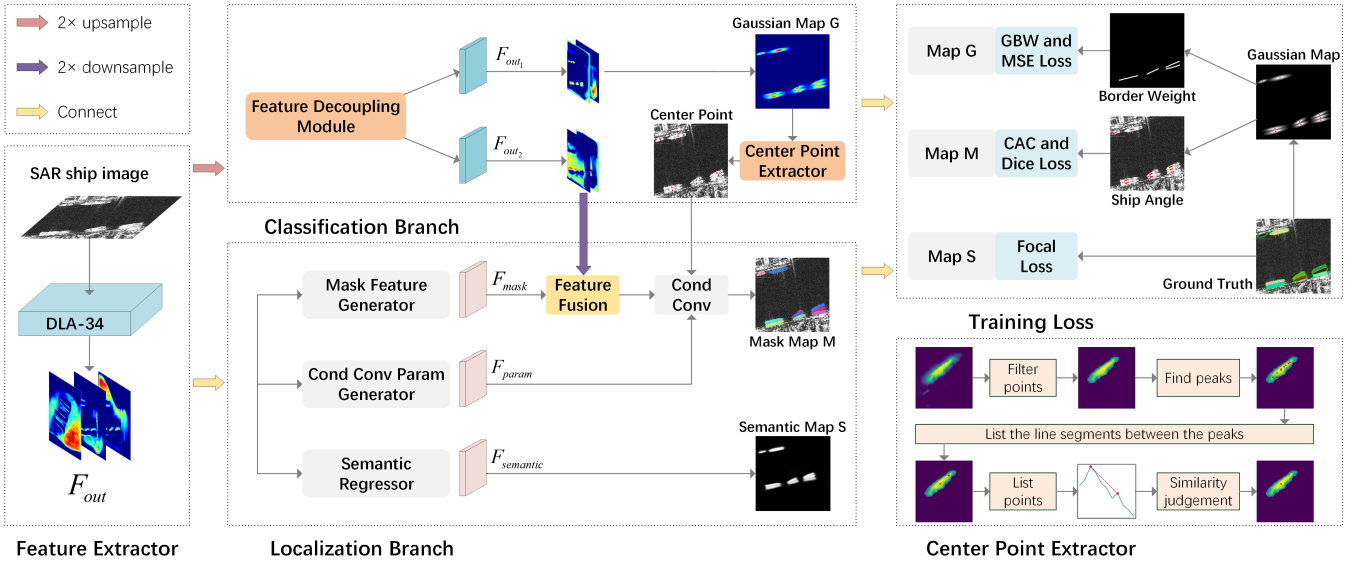

Fig. 2. Overall architecture of our method. The network structure can be divided into three parts: the feature extractor, the classification branch and the localization branch. Cond Conv means conditional convolution. Map *G*, *M*, *S* are the Gaussian map, mask map and semantic map. The training loss supervises *G*, *M*, *S* to learn the ground truth in the training process. The center point extractor combines the characteristics of Gaussian distribution to predict the center point on Map *G*.

Gaussian regression task (focusing on the central location of the instance) and the feature fusion task (focusing on the edge location of the instance) for spatial location conflicts, we use the FD module to obtain two sets of feature maps, F_{out_1} and F_{out_2}. F_{out_1} focuses on the centre of the ship and is used to generate the Gaussian heatmap G. Center point extractor performs center point prediction on map G to provide center point location for conditional convolution. F_{out_2} focuses on the edges of the ship, and provides the relative distance and edge information of the ship to the localization branch.

3) Localization Branch: The localization branch focuses on the edge position of ships and performs mask segmentation of different ship instances. The localization branch can be divided into three sub networks: mask feature generator, conditional convolution parameter generator and semantic regressor. The mask feature generator merges F_{out_2} into F_{mask} in the form of channel-wise addition. To ensure consistency of scale, F_{out_2} needs to be downsampled by a factor of 2. After feature fusion, conditional convolution, guided by center point location information, is performed to obtain instance segmentation result, namely mask map M. During conditional convolution, the conditional convolution parameter generator generates parameters for the mask head. The mask head consists of two 1x1 8-channel convolution cores and one 1x1 single-channel convolution core, so it contains a total of 169 parameters [22], and F_{param} is a 169-channel feature map. During network training, the semantic regressor serves as an auxiliary role to help localization branch to learn edge features of ships by regressing semantic map S.

B. Gaussian Heatmap Center Point Prediction Method

1) Elliptical Gaussian heatmap encode method: Ship targets can be encoded as elliptical Gaussian distributions according to the ship BBox annotations. The empirical formula of Gaussian distribution is

$$f(X) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-u)^T\Sigma^{-1}(X-u)\right) \quad (1)$$

where Σ is the two-dimensional covariance matrix, X is the coordinate vector (x, y) and u is the center point coordinate $(\bar{x}, \bar{y})$. To normalize the heat value of center point to 1, the Gaussian distribution formula is modified to

$$G(X) = \exp\left(-\frac{1}{2}(X-u)^T\Sigma^{-1}(X-u)\right) \quad (2)$$

Maximum likelihood estimation of the ship mask is performed to obtain the mean and covariance matrix of the elliptical Gaussian distribution. The specific formulas are

$$g_{x,y} = \begin{cases} 1, & (x,y) \in Mask_{ship} \\ 0, & (x,y) \notin Mask_{ship} \end{cases} \quad (3)$$

$$u = \frac{\sum_{x,y}(g_{x,y} \cdot X)}{\sum_{x,y} g_{x,y}} \quad (4)$$

$$\Sigma = \frac{\sum_{x,y} g_{x,y}^2 (X-u)(X-u)^T}{\sum_{x,y} g_{x,y}^2} \quad (5)$$

where $Mask_{ship}$ is the set of ship mask's coordinates and $g_{x,y}$ denotes the value of the image at (x, y).

2) Center point extraction method: The current mainstream one-stage target detection methods use BBox-based NMS for dense detection. The center point probability is predicted on a C channels feature map, and K points with highest probability are found as candidate center points in a top K manner. NMS is then performed based on the overlap ratio of BBox to remove redundancy. This approach links the accuracy of instance segmentation to the accuracy of BBox detection, which is the upper limit of the accuracy of instance segmentation.

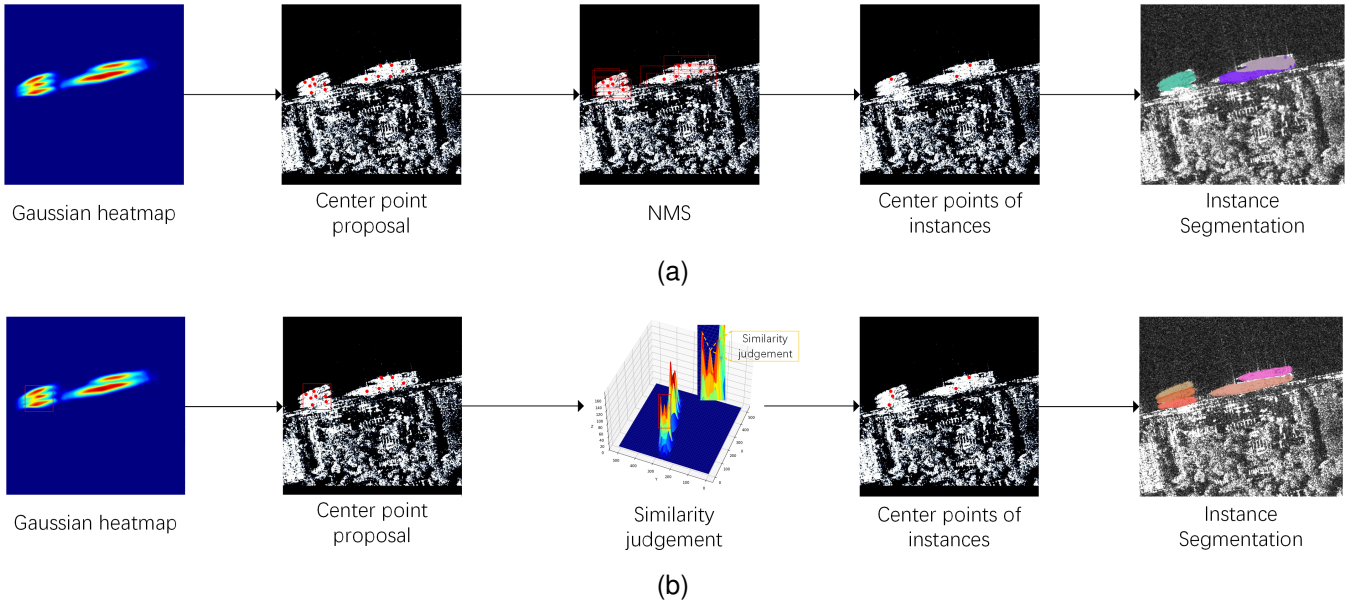

Fig. 3. Visualization of the process from center point detection to instance segmentation of different methods. (a) NMS-based method. (b) Our method. Due to the inherently high IoU of BBox from neighboring ships, it is difficult for NMS-based methods to detect neighboring ships. In contrast, our method is able to localize neighboring ships well based on the trend of Gaussian variation over long distances.

However, BBox detection is the lowest task level in the target detection domain [8], and its annotation contains massive background pixels. Therefore, the NMS process will lead to the detection of adjacent targets as a single instance due to the high degree of overlap within the BBox, as exemplified in Fig. 3(a). To solve this problem, this paper designs a Gaussian center point extraction algorithm based on the mathematical properties of Gaussian distribution. By looking for patterns of Gaussian variation over long distances between two center points, our method is able to ascertain whether these two center points originate from the same instance. As illustrated in Fig. 3(b), our method can suppress redundant center points from the same instance while preserving the center points of adjacent instances.

Center point extraction is performed on the Gaussian heatmap G of $1 \times \frac{H}{2} \times \frac{W}{2}$. The peak points are first found as candidate center points in the following steps.

1) Set a threshold T and filter out points with heat values less than T to obtain a heatmap $\tilde{G}$.
2) A MaxPool with 3×3 kernel is used to obtain the peak map G_{peak} by traversing over the Gaussian heatmap $\tilde{G}$ in stride of 1.
3) Find the coordinate point where the heatmap $\tilde{G}$ and the peak map G_{peak} have the same heat value. Use these coordinate points as candidate center points.

The above steps can be expressed by the following formula:

$$\tilde{G}(x,y) = \begin{cases} G(x,y) & , G(x,y) \geq T \\ 0 & , G(x,y) < T \end{cases} \tag{6}$$

$$G_{peak}(x,y) = \text{MaxPool}\left(\tilde{G}(x,y)\right) \tag{7}$$

$$CandPoints = \left\{(x,y) | \tilde{G}(x,y) == G_{peak}(x,y)\right\} \tag{8}$$

where $G(x,y)$ is the heat value at (x,y) and *CandPoints* denotes the set of candidate points.

In the Gaussian heatmap obtained by network regression, there may be more than one peak point on single Gaussian target. So we need to discriminate whether the candidate center points originate from the same Gaussian target and carry out de-redundancy. The specific algorithm process is described in Algorithm 1.

Algorithm 1 Candidate points de-duplication.

```
//Input candidate peak points set
input Candidate Point Set X = {x_1, x_2, ..., x_n}
    //Sorting X from largest to smallest
    X' ← Sort from largest to smallest X
    //The first in X' defaults to the ship's center point
    Y ← x'_1
    //De-duplication
    for i = 2 : n do
        if not from the same target (x'_i, Y) then
            Y ← x'_i
        end if
    end for
output Center Point Set Y
```

To discriminate whether x'_i and $\mathbf{Y}$ come from the same Gaussian target, we combine the mathematical properties of Gaussian distribution and match the similarity between the candidate center point x'_i and each center point in the set of Y according to the change pattern of heat value. The specific judgment criteria are as follows.

1) If there is a point with a heat value of 0 on the line segment between two points, then the two points belong to different targets.

2) If the difference between the heat value of candidate point and the minimum heat value on the line segment is less than M, then the two points belong to the same target.
3) If the average heat value of points on the line segment is greater than the heat value of candidate point, then the two points belong to the same target.
4) If none of the above three conditions are satisfied, then the two points belong to different targets.

Our method executes the above four criteria in sequence to discriminate whether the candidate point and the confirmed center point belong to the same Gaussian target. The first criterion states that if there is a point with a heat value of 0 between two points, then they belong to different targets. This is because any point with a heat value less than T is set to 0, so there must be a point with a heat value of 0 between two non-adjacent targets. The second criterion states that the heat value of candidate point from a different target should be at least M greater than the minimum heat value between two points. This is because a local peak is a small fluctuation that occurs during the transition from 1 to 0 in heat value, while the peak of different Gaussian targets is a long-distance transition from 1 to 0 and from 0 to 1, and its peak heat value should be significantly higher than the valley. The third criterion states that the heat value of the candidate point from a different target should be greater than the average heat value of points on the line segment. This is a supplement to the second criterion, requiring a long-distance gradient descent and ascent relationship between two points. Finally, a candidate point that dissatisfy the first three criteria is defaulted to come from a different Gaussian target.

It should be emphasized that the threshold M in the second criterion has a fundamental difference from the IoU threshold in NMS. The IoU threshold in NMS requires that the overlap ratio between two BBoxes must be less than the IoU threshold, which has an inherent limitation for detecting adjacent targets. In contrast, the threshold M in this article is based on the long-distance relationship between two different Gaussian targets. For adjacent targets, it requires that the heat value of the center point must be higher than a certain threshold above the valley, essentially requiring the network to divide two different targets with sufficient confidence, which is a requirement for the network's object detection ability.

In summary, our approach has the following advantages.

1) Our center point detection method operates independently of BBox and NMS, eliminating the necessity for designing supplementary BBox regression branches within the instance segmentation network. Therefore, the BBox annotation is not required.
2) Our method relies on the long-range variation pattern of Gaussian distribution, enabling effective discrimination between neighboring targets.
3) Our method decouples the accuracy of instance segmentation from the detection accuracy of BBox and is able to exceed the upper limit of detection accuracy of BBox.

C. Feature Decoupling Module

The instance segmentation consists of multiple sub-tasks, typically including classification and localization. Song et al. [27] find that there is a spatial mismatch between the classification and localization tasks. The former is more concerned with the center of the object and the latter is more concerned with the object edges. To demonstrate the spatial mismatch between the two tasks in SAR ship instance segmentation, we extract the classification branch feature maps (F_{out_1} in Fig. 2) and the localization branch feature maps (F_{mask} in Fig. 2). The most representative feature maps of two branches are selected by top-1 and then superimposed onto the SAR image by weighted addition. Fig. 4 visualizes the spatial positions that the classification and localization tasks focus on. Fig. 4(a) is the original image, Fig. 4(b) is the heatmap of the spatial positions that the classification task focuses on, where the red color indicates that the classification task is highly concerned with the center of the ship. Fig. 4(c) is the heatmap of the spatial positions that the localization task focuses on, where the yellow or green color indicates that the localization task is highly concerned with the edges of the ship.

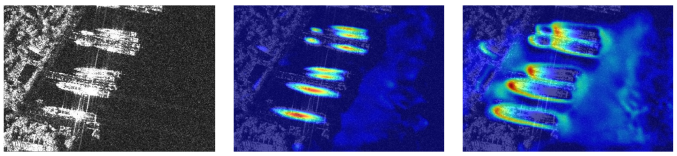

Fig. 4. Visualization of spatial attention positions for different tasks. (a) Original image. (b) Attention positions for classification task. (c) Attention positions for localization task.

Combining features from different tasks is an effective way to improve network's performance. By feature fusion, single branch can not only utilize its own features but also incorporate useful information from other task branches. In this article, we just use channel-wise addition to achieve feature fusion. And what we focus on is to resolve spatial location mismatch when fusing features from different tasks. We propose a FD module with an encoder-decoder structure, which is designed to achieve task-aware spatial decoupling. Fig. 5 illustrates the structure of the FD module.

The FD module consists of an encoder-decoder branch and a residual branch. The encoder-decoder branch is composed of an encoder and two decoders. The encoder first performs channel-wise mean and adaptive average pooling on the feature map to obtain its global information in the height, width, and channel dimensions. The specific process can be described as follows,

$$F_{H_{zip}} = \text{AdaptiveAvgPool2d}\left(\text{Mean}\left(F_{in}\right)\right) \tag{9}$$

$$F_{W_{zip}} = \text{AdaptiveAvgPool2d}\left(\text{Mean}\left(F_{in}\right)\right) \tag{10}$$

$$F_{C_{zip}} = \text{AdaptiveAvgPool2d}\left(F_{in}\right) \tag{11}$$

where AdaptiveAvgPool2d and Mean denote the adaptive average pooling and channel-wise mean, respectively. $F_{in} \in \mathbb{R}^{C\times H\times W}$ denotes the input feature. $F_{H_{zip}} \in \mathbb{R}^{1\times H\times 1}$, $F_{W_{zip}} \in \mathbb{R}^{1\times 1\times W}$ and $F_{C_{zip}} \in \mathbb{R}^{C\times 1\times 1}$ denote the feature information in the height, width and channel dimensions.

By adaptive average pooling, the encoder can obtain long-range dependencies in the three-dimensional directions. Then,

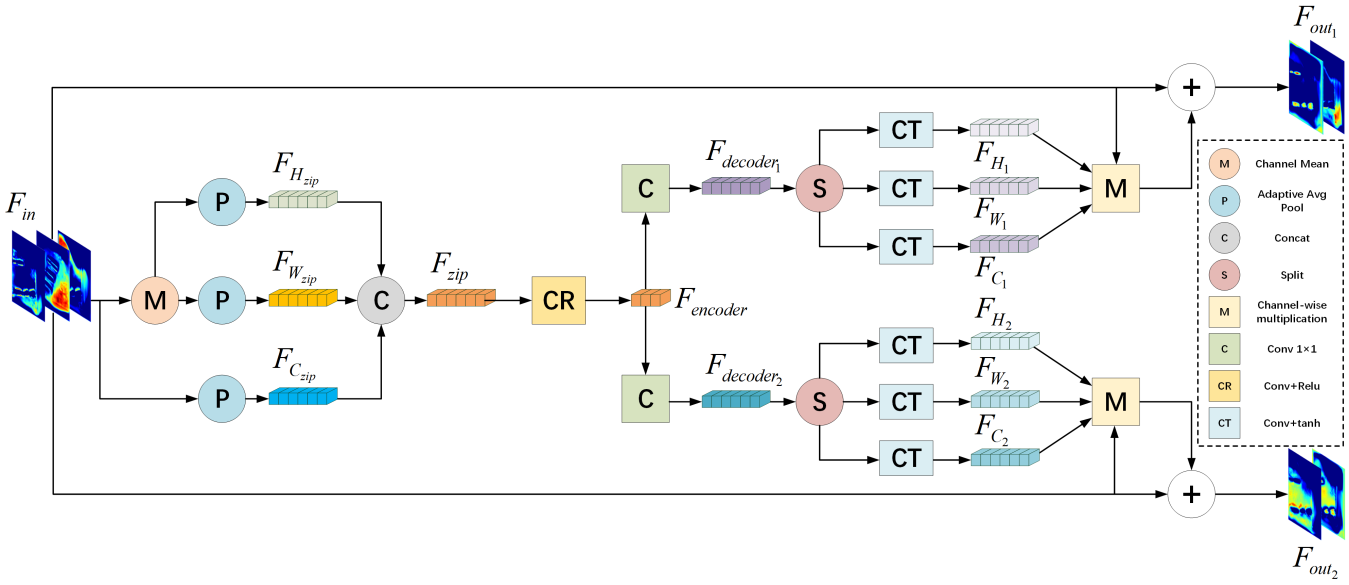

Fig. 5. Overall architecture of FD module. F_{in} is the high-resolution feature map after 2× upsampling. F_{out_1} and F_{out_2} are feature maps suitable for the classification and localization tasks, respectively, after feature decoupling.

the information of three dimensions is concatenated and interacted with each other through convolution so that the encoder can obtain global information. To suppress useless features during encoding and make full use of useful information, the global information is compressed and encoded during convolution. In this paper, we set the compression ratio r to 2 [26], [52] to prevent the loss of useful features during the compression process. This process can be illustrated as follows,

$$F_{zip} = \text{Concact}\left(F_{H_{zip}}, F_{W_{zip}}, F_{C_{zip}}\right) \tag{12}$$

$$F_{encoder} = \text{Relu}\left(\text{Conv}\left(F_{zip}\right)\right) \tag{13}$$

where Concact represents channel-wise concatenate operation. $F_{zip} \in \mathbb{R}^{(C+H+W)\times 1\times 1}$ denotes the global feature vector, and $F_{encoder} \in \mathbb{R}^{\frac{C+H+W}{r}\times 1\times 1}$ denotes the feature vector after compression. The spatial weights and channel weights of the two tasks are obtained through two decoders. The decoding process begins with a convolutional operation to recover the dimensionality of the feature map, which is then split into three-dimensional weights according to the length of C, H and W. These weights are refined through convolution and then passed through the activation function tanh to scale the value domain to $(-1, 1)$. Finally, the residual branch is introduced so that the weight value domain is transformed to $(0, 2)$. Thus, the FD is able to attenuate useless features while highlighting features of interest for different tasks. This process can be expressed as follows,

$$F_{decoder} = \text{Conv}\left(F_{encoder}\right) \tag{14}$$

$$F_H, F_W, F_C = \tanh\left(\text{Conv}\left(\text{Split}\left(F_{decoder}\right)\right)\right) \tag{15}$$

$$F_{out} = F_{in} + F_{in} \times F_H \times F_W \times F_C \tag{16}$$

D. Centroid and Angle Constraint Loss Function

In the mask segmentation training process, Dice Loss is used as the loss function, and the formula of Dice Loss is as follows:

$$DiceLoss = 1 - \frac{2\left|Mask_{gt} \cap Mask_{pred}\right|}{\left|Mask_{gt}\right| + \left|Mask_{pred}\right|} \tag{17}$$

where $Mask_{gt}$ represents the ground truth of the mask and $Mask_{pred}$ represents the predicted mask. $\cap$ denotes the intersection of two sets and $|\bullet|$ denotes the size of the set.

It can be seen that when Dice Loss is used as the loss function for mask segmentation, it only focuses on the IoU between the segmentation result and the ground truth. In cases where the IoU is the same, we would prefer the predicted mask to have the same orientation and a closer centroid distance to the ground truth. This is not reflected by the Dice Loss.

Fig. 6 illustrates two instance segmentation scenarios. The blue mask represents the ground truth, while the red mask symbolizes errors in the segmentation process. In Pred_1, the red mask is attached to the blue mask, indicating that a single mask head effectively covers a single instance, but over-segmentation occurs. In Pred_2, the red mask is distant from the blue mask, signifying that a single mask head takes effect on multiple instances. For a dynamic segmentation network, a

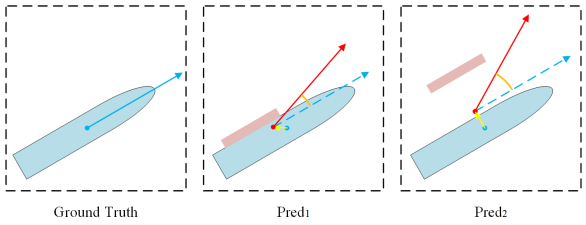

Fig. 6. The blue mask represents the ground truth of the ship instance, and the red mask represents errors made during the segmentation process. The blue dots represent the centroid of ground truth and the red dots represent centroid of predicted mask. The blue arrow represents the direction of the long axis of the ground truth, the red arrow represents the direction of the long axis of predicted mask.

key criterion is that a single mask head should exclusively take effect for the corresponding instance. Thus, the segmentation quality of Pred_1 should surpass that of Pred_2. As both cases share the same IoUs, Dice Loss is incapable of quantitatively distinguishing between them. To enable the network to discern these differences, we propose the centroid and angle constraint (CAC).

CAC consists of centroid constraint and angle constraint. Based on the elliptical Gaussian distribution encoding method, we can obtain the centroid position and covariance matrix of the mask. Assuming the predicted centroid position of the mask is (x_{pred}, y_{pred}) and the ground truth centroid position is (x_{gt}, y_{gt}), the relative distance between them is calculated as follows:

$$distance = \frac{\sqrt{\left(x_{gt} - x_{pred}\right)^2 + \left(y_{gt} - y_{pred}\right)^2}}{\max\left(H, W\right)} \tag{18}$$

where *H*, *W* represent the height and width of the feature map. The range of the calculated relative distance is $(0, \sqrt{2})$, which is used as the exponent of the natural constant *e* to obtain the centroid weight:

$$w_{centroid} = \exp\left(distance\right) \tag{19}$$

By performing principal component analysis (PCA) on the covariance matrix of the Gaussian distribution, the long axis direction of the Gaussian distribution can be obtained [68], which is basically consistent with the long axis direction of the ship. Therefore, CAC uses the long axis direction of the Gaussian distribution as the direction of the ship. We first perform singular value decomposition on the covariance matrix Σ, as shown in the following equation:

$$\Sigma = QAV^T \tag{20}$$

where Q and V represent two orthogonal matrices and A is the diagonal matrix. Based on equation (5) we decompose the second order matrix formula as follows:

$$\begin{aligned}
\Sigma &= \frac{\sum_{x,y} g_{x,y}^2 \left(X-u\right)\left(X-u\right)^T}{\sum_{x,y} g_{x,y}^2} \\
&= \frac{\sum_{x,y} g_{x,y}^2 \begin{bmatrix}\Delta x \\ \Delta y\end{bmatrix}\left[\Delta x \Delta y\right]}{\sum_{x,y} g_{x,y}^2} \\
&= \frac{\sum_{x,y} g_{x,y}^2 \begin{bmatrix}\Delta x^2 & \Delta x\Delta y \\ \Delta x\Delta y & \Delta y^2\end{bmatrix}}{\sum_{x,y} g_{x,y}^2} \\
&= \begin{bmatrix}\Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2}\end{bmatrix}
\end{aligned} \tag{21}$$

where $\Sigma_{i,j} = \sum_{x,y}\left(g_{x,y}^2 \Delta x^{4-i-j} \Delta y^{i+j-2}\right) / \sum_{x,y} g_{x,y}^2$, $\{i,j\} = \{1,2\}$ and $X - u = (\Delta x, \Delta y)$. The covariance matrix Σ is a real matrix as the variables in the above formula are all real. According to the above equation, $\Sigma_{1,2}$ and $\Sigma_{2,1}$ are equal, so Σ is a real symmetric matrix and can be diagonalised orthogonally:

$$\Sigma = QAQ^{-1} \tag{22}$$

The diagonal elements of the diagonal matrix A are the eigenvalues of the covariance matrix Σ, and each row of the orthogonal matrix Q corresponds to the corresponding eigenvector. The eigenvector corresponding to the minimum eigenvalue is the direction of the major axis of the Gaussian distribution.

Through the above steps, we can obtain the unit directional vectors v_{pred} and v_{gt} for the predicted mask and the ground truth mask, respectively. And we can obtain the cosine value of the angle between the two vectors by performing vector multiplication. Using this cosine value as the exponent of the natural constant e to obtain the weight of the angle constraint:

$$w_{angle} = \exp\left(1 - v_{pred} \cdot v_{gt}\right) \tag{23}$$

Combining the centroid distance weight with the angle weight, we obtain the CAC weight. And the final mask prediction loss is obtained by weighting the Dice Loss with the CAC:

$$w_{CAC} = w_{centroid} \times w_{angle} \tag{24}$$

$$Loss_{mask} = w_{CAC} \times DiceLoss \tag{25}$$

E. Loss Function

Our method consists of three sub-tasks, namely Gaussian heatmap regression task (classification task), mask segmentation task (localization task) and auxiliary semantic segmentation task. The loss function is designed for each of these three tasks in this paper.

We use mean square error (MSE) as the loss function for the Gaussian heatmap regression task, and the MSE is weighted by the Gaussian and boarder weight (GBW). The Gaussian weight is used to balance the positive and negative sample, and the boundary weight is used to enhance the supervision of the boundary regression for dense neighbouring ship targets. GBW can be described as follows:

$$w_{border}(x,y) = \exp\left(-\frac{\left(d_1(x,y) + d_2(x,y)\right)^2}{2}\right) \tag{26}$$

$$GBW_{x,y} = 1 + \beta_1 w_{gauss}(x,y) + \beta_2 w_{border}(x,y) \tag{27}$$

where $d_1(x,y)$ and $d_2(x,y)$ represent the distance to the two ships closest to the coordinate (x,y). $w_{gauss}(x,y)$ is the Gaussian heat value at (x,y) obtained from equation (2). The formula adds 1 to the weight is to ensure that the weight of the other region is not zero. β_1 and β_2 are the weights of w_{gauss} and w_{border}, which are set to 10 and 1 in this paper. Weighting MSE with GBW, $Loss_{gauss}$ can be described as follows:

$$Loss_{gauss} = \frac{1}{HW}\sum\nolimits_{x,y} GBW_{x,y}\left(G_{x,y} - \bar{G}_{x,y}\right)^2 \tag{28}$$

where H and W are the height and width of the feature map, $G_{x,y}$ and $\bar{G}_{x,y}$ are the predicted and ground truth values of the Gaussian distribution at the coordinate (x, y).

The mask segmentation uses the CAC designed in this paper, and the semantic segmentation uses Focal Loss as the loss function. The loss functions of the three tasks are multiplied with the corresponding weights and summed to obtain the overall loss function:

$$Loss = \alpha_1 Loss_{gauss} + \alpha_2 Loss_{mask} + \alpha_3 Loss_{semantic} \quad (29)$$

where α_1, α_2 and α_3 are the weights of each of the three tasks. In this paper, we set these three weights to 1000, 1 and 1 respectively.

IV. Experiments

This section presents the detailed results of experiments conducted on PSeg-SSDD and HRSID. First, the datasets and evaluation metrics used in the experiments are introduced. Then, We compare our method with other mainstream segmentation methods to verify the effectiveness and robustness of the proposed method. Subsequently, ablation experiments are conducted to verify the effectiveness of each module. Finally, we conduct the visualization experiments of center point prediction, FD and CAC.

A. Dataset Description and Experimental Settings

In this experiment, the performance and robustness of our method are evaluated on two representative SAR ship datasets, PSeg-SSDD [8] and HRSID [33]. PSeg-SSDD contains 1160 SAR ship images with multiple resolutions, polarizations, and sea conditions, including 2456 ship targets. The image resolution ranges from 1m to 15m, ensuring the robustness of the network to different scales of ships. According to the official partition method, 928 images are used as the training samples and 232 images as the testing samples. Among the testing set, 186 images belong to the offshore scene and 46 images belong to the inshore scene, with an offshore-to-inshore image ratio of 8:2. The distribution of the training set is also similar, and this imbalance in the number of offshore and inshore images further increases the difficulty of the network in learning inshore scene. HRSID contains 5604 SAR ship images and 16951 ship targets. The images in HRSID are cropped out into 800 $\times$ 800 pixels, and these cropped images are divided into the training set and testing set with the proportion of 13:7. Among the testing set, 1593 images belong to the offshore scene and 369 images belong to the inshore scene. There is still an imbalance between the number of images of inshore and offshore scenes.

During training, we used parameters pre-trained on ImageNet to initialise DLA-34 and used random initialisation for each subsequent branch of the network. Adam was used as the training optimization algorithm, with the initial learning rate set to 0.0001 and the learning rate reduced by a factor of 10 at epoch 80 and 120, for a total of 140 training epochs. The experimental part of the code in this paper is implemented using the Pytorch framework [69], and the MMdetection [70] framework and the AdelaiDet framework [71] are used for comparison experiments. The experimental platform is configured with Ubuntu 20.04 system, 32G RAM, and 3080ti is used for training and inference.

TABLE I
MS COCO Metrics

Metrics	Meaning
AP	IoU=0.50:0.05:0.95
AP_{50}	IoU=0.50
AP_{75}	IoU=0.75
AP_S	AP of small objects: area $< 32^2$
AP_M	AP of medium objects: $32^2 <$ area $< 64^2$
AP_L	AP of large objects: area $> 64^2$

B. Evaluation Metrics

Microsoft Common Objects in Context (MS COCO) evaluation metric [72] is the most common performance evaluation metric in the field of instance segmentation. The core of the MS COCO is the intersection ratio of the mask prediction to the mask ground truth. Based on the IoU threshold, the precision and recall of the target can be calculated:

$$precision = \frac{TP}{TP + FP} \quad (30)$$

$$recall = \frac{TP}{TP + FN} \quad (31)$$

where *TP* is the number of true positives, *FP* is the number of false positives and *FN* is the number of false negatives. According to precision and recall, average precision (AP) can be calculated as follows:

$$AP = \int_0^1 p(r)dr \quad (32)$$

where *r* represents the recall rate and $p(r)$ represents the precision when recall rate is r. Mean average precision (mAP) represents the average precision of multiple categories. In the field of SAR ship detection, only existing one category, which is ship, so AP and mAP are equal.

According to different IoU thresholds and object pixel areas, MS COCO has various evaluation metrics to comprehensively evaluate the accuracy of object detection. Table I lists these evaluation metrics.

Floating-point operations (Flops) and parameters (Params) are the prevailing evaluation metrics for assessing the computational and parametric aspects of a model. Flops represent the count of floating-point computations, with each multiplication or addition counted as one Flop. This metric allows us to gauge the computational time complexity of a network model. Params, on the other hand, represent the overall number of trainable parameters in the model. It serves as a measure of the model's size.

C. Comparison With Other Methods

To verify the effectiveness and robustness of our method, we implement experiments on two datasets of PSeg-SSDD and HRSID. AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L are used to evaluate the performance of different methods, i.e., Mask R-CNN, Cascade Mask R-CNN, Mask Scoring R-CNN, Yolact [73], CondInst, SparseInst [74], HQ-ISNet, Anchor-Free SAR Ship Instance Segmentation With Centroid-Distance Based Loss (AFSS-Inst) and RTMDet. Flops and Params are used to evaluate the computational and parametric quantities of the model. We conduct our experiments according to the division of inshore and offshore scenes. Table II demonstrates the quantitative performance of different methods. An analysis of the experimental results is provided below.

1) Comparison of Flops and Params: The computational consumption of our method is only second to Yolact and AFSS-Inst and the parameters of our method is second only to AFSS-Inst, which are benefited from the BBox-free and single-scale architecture. Our method does not have branches for BBox regression and all inference is done at a single scale. In contrast, AFSS-Inst uses the lightweight GhostNet [75] as the feature extraction network, which greatly reduces the consumption of network computation and the number of parameters.
2) Comparison of inshore scene: As evidenced by our experimental results on PSeg-SSDD, our method has a competitive performance. As shown in Table II, our method achieves improvements of 7.9%, 4.7%, 7.3%, 14.5%, 7.7%, 18.5%, 8.3%, 4.3%, 4.3% and 5.6% over Mask R-CNN, Cascade Mask R-CNN, Mask Scoring R-CNN, Yolact, CondInst, SparseInst, RTMDet, HQ-ISNet, AFSS-Inst and SRNet on AP, respectively. HRSID has a more complex background clutter and a large number of small ships, making it more challenging. Similarly, our method outperforms other compared methods of 8.5%, 6.9%, 15.3%, 10.5%, 2.1%, 2.9%, 14.3% and 3.0% over Mask R-CNN, Cascade Mask R-CNN, Yolact, CondInst, RTMDet, HQ-ISNet, AFSS-Inst and SRNet on AP, respectively. This is mainly because our proposed Gaussian center point localization method can suppress scattering noise and locate neighboring ship instances from complex scenes based on Gaussian long-range trends. At the same time, feature fusion with FD enables the network to extract sharper ship edge features from decoupled features and CAC guides the mask head to focus only on the corresponding ship instance. Our method dramatically improves the segmentation accuracy of inshore ships and has the highest AP_{50} on both datasets. Only the AP_{75} on HRSID is second to HQ-ISNet and SRNet, which use the complex network structure and has the highest computational consumption.
3) Comparison of offshore scene: Compared to inshore scene, offshore scene is easier to segment ships due to the lack of interference from inshore harbours and islands. It can be observed from Table II that our method achieve increment of AP by ranging from 0.5% to 15.3% on PSeg-SSDD and ranging from 0.1% to 9.4% on HRSID. Besides, our method has the highest AP_{50} and the AP_{75} of our method is only second to HQ-ISNet and SRNet, which indicates the superiority of our method. Based on long-range Gaussian trends, our Gaussian center point localization method is able to accurately localize ships from scattering noise from ships and the sea surface. As a result of decoupling features by FD, the classification feature maps provide more fine-grained ship positioning features for the localization feature maps. Furthermore, CAC guides the mask head to eliminate interference from other instances. As a result, our method is able to segment smoother masks while achieving more accurate ship localization.

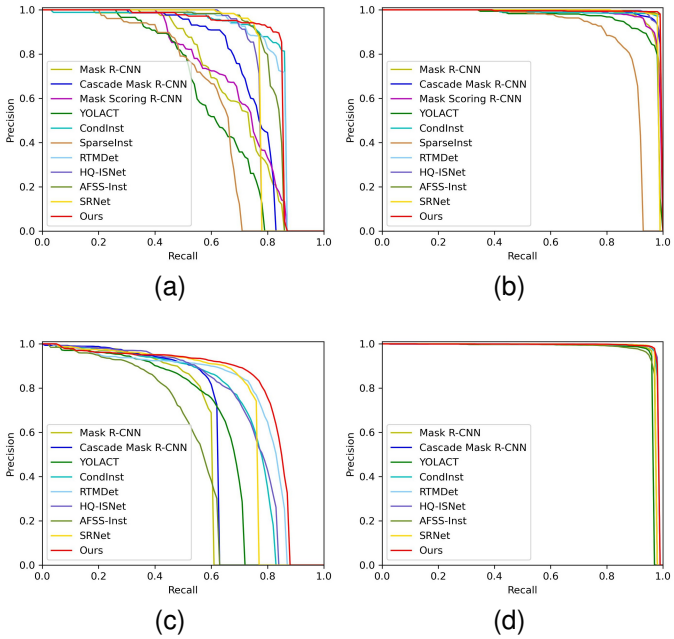

Fig. 7. PR curves of different methods on PSeg-SSDD and HRSID in inshore and offshore scenes. (a) PR curves for inshore scene on PSeg-SSDD. (b) PR curves for offshore scene on PSeg-SSDD. (c) PR curves for inshore scene on HRSID. (d) PR curves for offshore scene on HRSID.

The PR curves are illustrated in Fig. 7 to comprehensively show the instance segmentation performance of different methods on PSeg-SSDD and HRSID in inshore and offshore scenes. The PR curves are presented in Fig. 7 to provide a comprehensive illustration of the instance segmentation performance of different methods on PSeg-SSDD and HRSID in inshore and offshore scenes. Fig. 7(a) and (b) illustrate the PR curves on PSeg-SSDD. Fig. 7(a) reveals that when the recall is between 0.3 and 0.8, the precision of Yolact and SparseInst decreases rapidly. In contrast, our method, CondInst, and RTMDet maintain the highest precision within this range. Additionally, when the recall exceeds 0.8, our method consistently ensures higher precision compared to the others. Fig. 7(b) indicates that when the recall surpasses 0.7, the precision of the comparative methods starts to exhibit significant differences. SparseInst experiences the fastest decrease in its PR curve, whereas our method's PR curve consistently remains in the upper right corner, demonstrating that our method offers optimal instance

TABLE II
INSTANCE SEGMENTATION PERFORMANCE OF DIFFERENT METHODS ON PSEG-SSDD AND HRSID

Dataset	Model	Flops(G)	Params(M)	Scene	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
PSeg-SSDD [8]	Mask R-CNN [18]	114.73	43.97	inshore	0.419	0.688	0.474	0.474	0.298	0.142
				offshore	0.644	0.979	0.786	0.647	0.645	**0.700**
	Cascade Mask R-CNN [19]	245.49	76.80	inshore	0.451	0.739	0.522	0.495	0.353	0.500
				offshore	0.646	0.985	0.791	0.643	0.668	**0.700**
	Mask Scoring R-CNN [20]	153.77	60.01	inshore	0.425	0.696	0.482	0.484	0.295	0.217
				offshore	0.652	0.976	0.810	0.649	0.661	0.600
	Yolact [73]	47.67	34.73	inshore	0.353	0.610	0.404	0.399	0.258	0.041
				offshore	0.614	0.958	0.752	0.616	0.624	**0.700**
	CondInst [22]	209.66	33.98	inshore	0.421	0.830	0.365	0.392	0.490	0.900
				offshore	0.580	0.973	0.682	0.542	0.720	0.600
	SparseInst [74]	91.39	31.64	inshore	0.313	0.598	0.332	0.339	0.251	0.800
				offshore	0.582	0.882	0.731	0.528	**0.765**	0.600
	RTMDet [23]	106.63	57.31	inshore	0.415	0.830	0.360	0.342	0.603	0.400
				offshore	0.603	0.981	0.738	0.571	0.717	**0.700**
	HQ-ISNet [11]	273.15	82.98	inshore	0.455	0.759	0.529	0.453	0.474	0.400
				offshore	0.666	0.986	0.820	0.655	0.705	0.283
	AFSS-Inst [26]	16.15	8.2	inshore	0.455	0.815	0.477	0.452	0.477	0.367
				offshore	0.659	0.986	0.827	0.634	0.751	0.600
	SRNet [76]	162.80	45.44	inshore	0.442	0.766	0.477	0.440	0.453	0.600
				offshore	0.661	0.979	0.828	0.640	0.737	**0.700**
	Ours	54.24	21.48	inshore	0.498	0.830	0.536	0.452	0.664	0.900
				offshore	**0.671**	**0.987**	**0.843**	**0.656**	0.764	0.600
HRSID [33]	Mask R-CNN [18]	262.42	43.97	inshore	0.340	0.565	0.383	0.310	0.549	0.334
				offshore	0.653	0.958	0.834	0.645	0.725	0.500
	Cascade Mask R-CNN [19]	398.54	76.80	inshore	0.356	0.591	0.401	0.326	0.556	0.336
				offshore	0.674	0.968	0.857	0.667	0.743	0.413
	Yolact [73]	208.47	34.73	inshore	0.272	0.618	0.187	0.258	0.419	0.096
				offshore	0.586	0.956	0.710	0.580	0.683	0.356
	CondInst [22]	342.36	33.98	inshore	0.320	0.716	0.242	0.300	0.503	0.135
				offshore	0.607	0.976	0.748	0.597	0.703	**0.613**
	RTMDet [23]	272.97	57.31	inshore	0.404	0.759	0.387	0.393	0.530	0.101
				offshore	0.674	**0.977**	0.837	0.668	0.749	0.336
	HQ-ISNet [11]	481.36	82.98	inshore	0.396	0.725	0.410	0.390	0.497	0.095
				offshore	0.679	0.975	**0.871**	0.665	0.738	0.468
	AFSS-Inst [26]	64.58	8.2	inshore	0.282	0.518	0.280	0.280	0.335	0.031
				offshore	0.652	0.964	0.834	0.673	0.611	0.103
	SRNet [76]	651.20	45.44	inshore	0.395	0.716	0.406	0.386	0.529	0.082
				offshore	0.677	0.969	0.860	**0.676**	0.747	0.223
	Ours	263.21	23.15	inshore	0.425	0.791	0.399	0.410	0.563	0.201
				offshore	**0.680**	**0.977**	0.851	0.674	**0.751**	0.479

Bold items denote the optimal offshore values in the columns, the underlined items represent the optimal inshore values in the columns.

segmentation accuracy. The PR curves on HRSID are depicted in Fig. 7(c) and (d). Similarly, the PR curves of our method are situated in the upper right corner in both inshore and offshore scenes, signifying that our method maintains higher precision at the same recall levels. In summary, these results confirm the superior performance of our method.

Fig. 8 and Fig. 9 demonstrate the visualization results of instance segmentation by different methods on PSeg-SSDD and HRSID. As shown in the first and second rows of Fig. 8 and third row of Fig.9, our method is superior to compared method for densely arranged ship detection under complex background interference. It intuitively illustrates the effectiveness of the Gaussian center point extraction method. Moreover, the masks segmented by our method are smoother, and under- and over-segmentation occur less frequently, which illustrates the effectiveness of CAC and feature fusion method with FD. The third row in Fig.8 and first row in Fig.9 show the segmentation results for small ships under interference from inshore objects. Compared with other methods, our method do not occur false alarm and can accurately locate and segment ships under strong interference conditions. The visualization results fully demonstrate the advantages of our method.

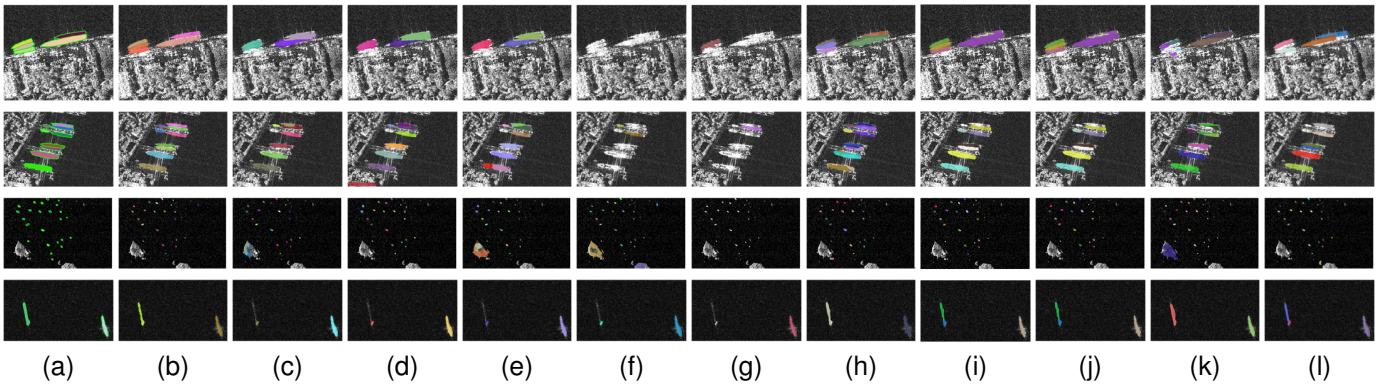

Fig. 8. Instance segmentation results of different methods on PSeg-SSDD. (a) Ground truth. (b) Our method. (c) Mask R-CNN. (d) Cascade Mask R-CNN. (e) Mask Scoring R-CNN. (f) Yolact. (g) CondInst. (h)SparseInst. (i) RTMDet. (j) HQ-ISNet. (k) AFSS-Inst. (l) SRNet.

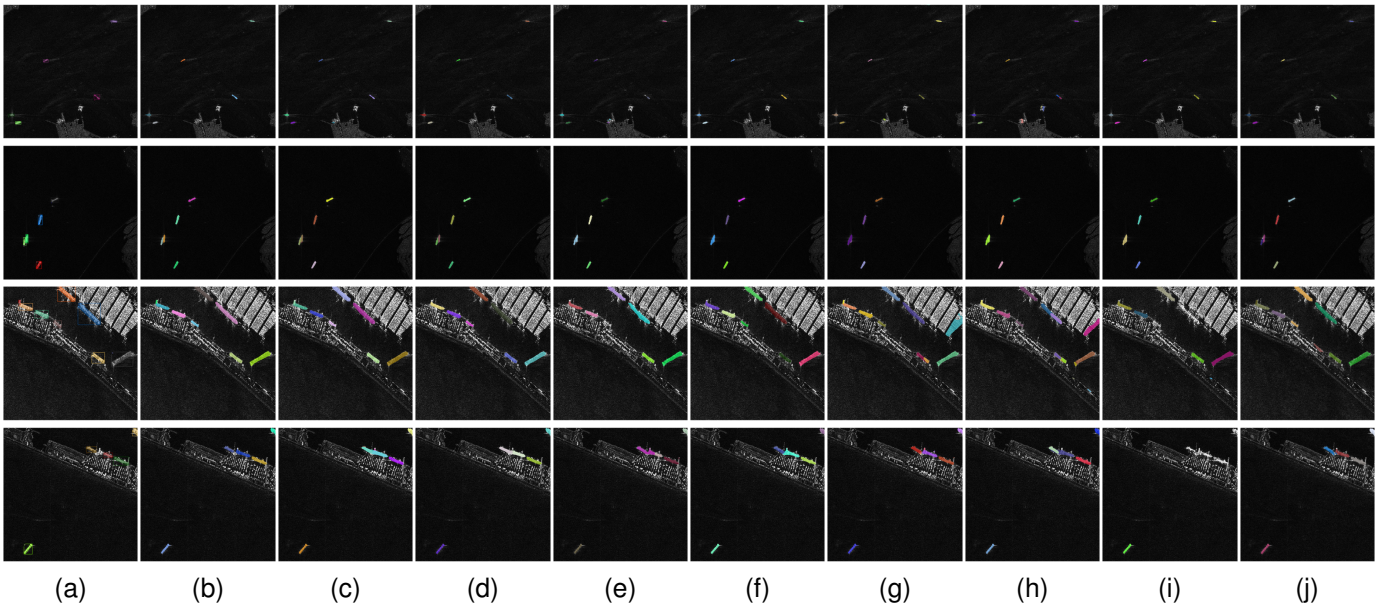

Fig. 9. Instance segmentation results of different methods on HRSID. (a) Ground truth. (b) Our method. (c) Mask R-CNN. (d) Cascade Mask R-CNN. (e) Yolact. (f) CondInst. (g) RTMDet. (h) HQ-ISNet. (i) AFSS-Inst. (j) SRNet.

In order to more thoroughly verify the effectiveness of our method, we further conduct instance segmentation experiments on large-scene image ALOS-2 from HRSID [33], which contains 18001×14804 pixels. Six representative regions, including inshore (slices 1, 2, 3 and 4) and offshore (slices 5 and 6) scenes, are chosen from the large-scene image ALOS-2. Fig. 10 illustrates the instance segmentation results, including where the slices are cropped. It can be seen from the visualization results that our method can effectively detect multi-scale targets in both inshore and offshore scenes although there are still some false alarms in complex inshore scene (pink mask in slice 2). In general, our method demonstrate excellent instance segmentation performance and effectiveness for the case of both the slice of the two datasets mentioned above and the large-scene image from HRSID.

D. Ablation Experiments

To illustrate the effectiveness of each part proposed in this paper, and quantitatively judge their improvement, we conduce ablation experiments on PSeg-SSDD for the Gaussian center point prediction method, the FD module and the CAC loss.

Table III shows the results of Gaussian center point prediction method ablation experiment. We use CondInst as the baseline, and then use FCOS and Gaussian heatmap for center point prediction. By comparing the model parameters, the computation complexity and instance segmentation accuracy, we can get the following conclusions.

1) The single-scale Gaussian center point prediction method can greatly reduce the model computation and parameters compared with the multi-scale FCOS, which are 76.95% and 38.46% lower, respectively.
2) The Gaussian center point prediction method is able to improve the accuracy of instance segmentation. Compared to FCOS, the AP metrics of inshore and offshore

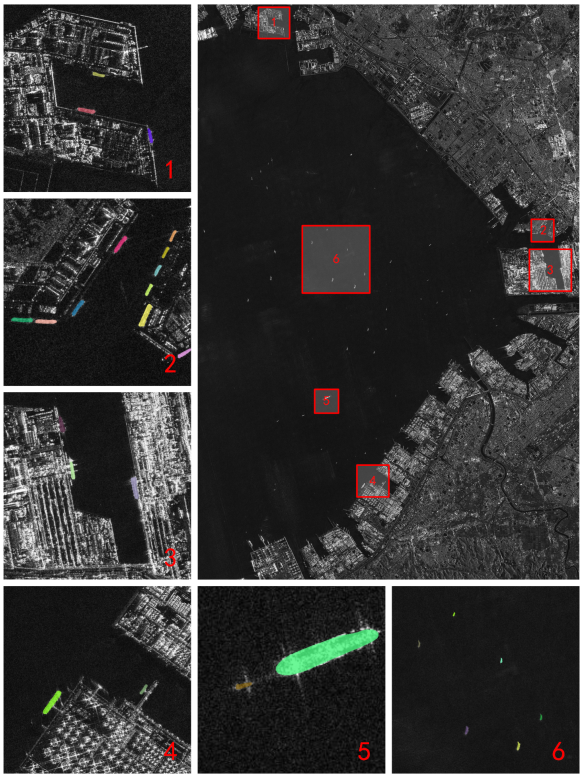

Fig. 10. Instance segmentation results on the large-scene image.

TABLE III
ABLATION EXPERIMENT ON GAUSSIAN CENTER POINT PREDICTION METHOD

Method	Flops(G)	Params(M)	Scene	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS	209.66	33.93	inshore	0.421	0.830	0.365	0.392	0.490	0.900
			offshore	0.580	**0.973**	0.682	0.542	0.720	**0.600**
Gaussian	48.32	20.88	inshore	0.459	0.849	0.462	0.452	0.543	0.900
			offshore	**0.647**	0.967	**0.804**	**0.626**	**0.744**	**0.600**

Bold items denote the optimal offshore values in the columns, the underlined items represent the optimal inshore values in the columns.

TABLE IV
ABLATION EXPERIMENT ON FD

FD	Flops(G)	Params(M)	Scene	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	54.23	20.97	inshore	0.472	0.826	0.491	0.442	0.592	0.495
			offshore	0.667	0.980	0.827	0.647	**0.773**	0.500
✓	54.24	21.48	inshore	0.498	0.830	0.536	0.452	0.664	0.900
			offshore	**0.671**	**0.987**	**0.843**	**0.656**	0.764	**0.600**

Bold items denote the optimal offshore values in the columns, the underlined items represent the optimal inshore values in the columns.

TABLE V
ABLATION EXPERIMENT ON CAC

Method	CAC	Scene	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Ours		inshore	0.459	0.840	0.468	0.447	0.567	0.900
		offshore	0.668	**0.987**	0.822	0.644	**0.778**	**0.600**
	✓	inshore	0.498	0.830	0.536	0.452	0.664	0.900
		offshore	**0.671**	**0.987**	**0.843**	**0.656**	0.764	**0.600**
Mask R-CNN		inshore	0.400	0.711	0.417	0.356	0.504	0.600
		offshore	0.630	0.909	0.807	0.597	0.752	**0.700**
	✓	inshore	0.438	0.754	0.469	0.386	0.539	0.400
		offshore	**0.639**	**0.919**	**0.820**	**0.609**	**0.758**	**0.700**
RTMDet		inshore	0.406	0.791	0.365	0.361	0.525	0.350
		offshore	0.603	**0.988**	0.738	0.571	0.717	0.700
	✓	inshore	0.415	0.830	0.360	0.342	0.603	0.400
		offshore	**0.619**	0.985	**0.773**	**0.588**	**0.720**	**0.712**

Bold items denote the optimal offshore values in the columns, the underlined items represent the optimal inshore values in the columns.

scenes have been improved by 9.03% and 11.55%, respectively.

3) At a high IoU threshold of 0.75, the Gaussian center point prediction method has higher segmentation accuracy compared to FCOS. It indicates that the method of encoding ship targets as elliptical Gaussian distributions is consistent with the physical properties of the ship and reduces the learning difficulty of the network.

Table IV shows the ablation experimental results of the FD. From this table, we can draw the following conclusions.

1) FD only brings a computational and parameter cost of 0.01 GFlops and 0.51 MParams, which is a lightweight network structure.
2) The adoption of FD bring about an improvement in instance segmentation accuracy of 5.51% for inshore scenes and 0.60% for offshore scenes.
3) FD has the most significant improvement in segmentation accuracy for inshore scenes. This indicates that FD effectively disperses the gradient flow of the classification and localization tasks and helps the feature maps generated by both tasks to better achieve the focus on the regions of interest for the task.

Table V shows the results of the CAC ablation experiment. Using our method, Mask R-CNN and RTMDet as baselines, we compare the accuracy of instance segmentation with and without CAC. According to the experimental results, we can get the following conclusions.

1) CAC can improve the segmentation accuracy in inshore and offshore scenes. Under the guidance of CAC, our method, Mask R-CNN and RTMDet improve the AP of both scenes by up to 3.8% and 1.6% respectively.
2) At an IoU threshold of 0.75, the instance segmentation accuracy of three methods has greatly improved. This indicates that CAC can help the network to sense the differences in the centroid location and the orientation of the mask during training, which makes the mask smoother and more consistent with the direction of the ship.
3) CAC is effective for both the two-stage algorithm Mask R-CNN and the single-stage algorithm RTMDet, proving the robustness of CAC.

E. Loss function Weights Selection

The method proposed in our article contains three branches, each of which have a corresponding loss function. The overall loss function contains three corresponding loss function weights, i.e., α_1, α_2 and α_3. α_3 defaults to 1. In order to determine the values of α_1 and α_2, we carry out loss function weights selection experiment. We vary the weights in steps of multiples of 10 and choose four different sets of weights to compare the performance of our method under different weights. Table VI shows the results of the loss function weight comparison experiment, from which we can get the following conclusions.

1) When α_1 is 10 and α_2 is 1, our method has the best segmentation accuracy in the inshore and offshore scenes, with APs of 0.498 and 0.671, respectively, which is higher than the results of other comparison experiments.
2) When the ratio of α_1 to α_2 rises or falls from 10, both cause the segmentation accuracy to deteriorate.

TABLE VI
INSTANCE SEGMENTATION PERFORMANCE UNDER DIFFERENT LOSS FUNCTION WEIGHT

α_1	α_2	Scene	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
1	10	inshore	0.451	0.816	0.459	0.445	0.524	0.595
		offshore	0.670	0.977	**0.851**	0.651	**0.771**	**0.600**
1	1	inshore	0.452	0.819	0.461	0.459	0.515	0.900
		offshore	0.668	0.984	0.842	0.651	0.764	0.600
10	1	inshore	0.498	0.830	0.536	0.452	0.664	0.900
		offshore	**0.671**	**0.987**	0.843	**0.656**	0.764	**0.600**
100	1	inshore	0.458	0.821	0.467	0.451	0.550	0.560
		offshore	0.666	0.984	0.841	0.653	0.763	**0.600**

Bold items denote the optimal offshore values in the columns, the underlined items represent the optimal inshore values in the columns.

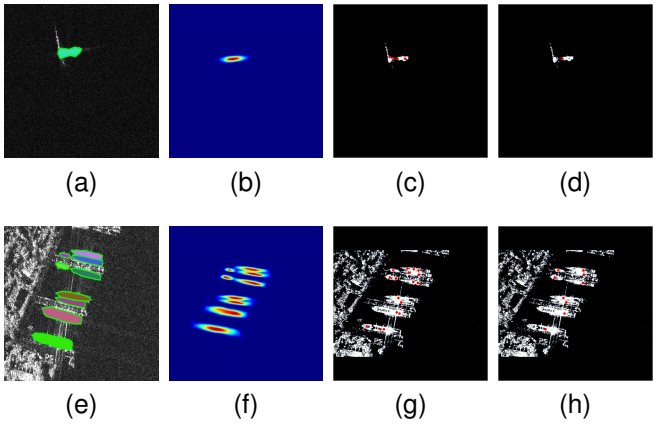

Fig. 11. Visualization of center point prediction. The images from left to right are ground truth, Gaussian heatmap, center point proposal, and center point prediction result, respectively. The red dots in (c), (d), (g), (h) represent the center points.

From the experimental results in Table VI, we can determine that setting α_1 and α_2 to 10 and 1, respectively, ensures the best segmentation accuracy of our method.

F. Visualization Experiments

To verify the effectiveness of our method, this experiment visualises the three methods proposed in this paper and analyses their contribution to instance segmentation.

Fig. 11 visualizes the center point prediction process. The images from left to right are ground truth, Gaussian heatmap, center point proposal, and center point prediction result, respectively. The images in the first row represent the offshore scene with scattered interference. As can be seen from Fig. 11(b), encoding ship instances as elliptic Gaussian distributions can effectively suppress scattering interference. By finding the local peak points in Fig. 11(b), the center point proposal in Fig. 11(c) is obtained. Since the Gaussian distribution of network learning is not smooth, there are multiple local peak points on a single instance. Through the Gaussian center point de-redundancy method proposed in this paper, the similarity between center points can be effectively assessed. As shown in Fig. 11 (d), our method retains the center point with highest heat value while suppressing the redundant center points. The images in the second row represent the inshore scene with dense arrangement of ships. As shown in Fig. 11(h), our method is able to discriminate between the center points of the same instance and those of different instances. Even when multiple instances are in close proximity to each other, our method is able to efficiently distinguish the center points of adjacent instances and suppress redundant center points.

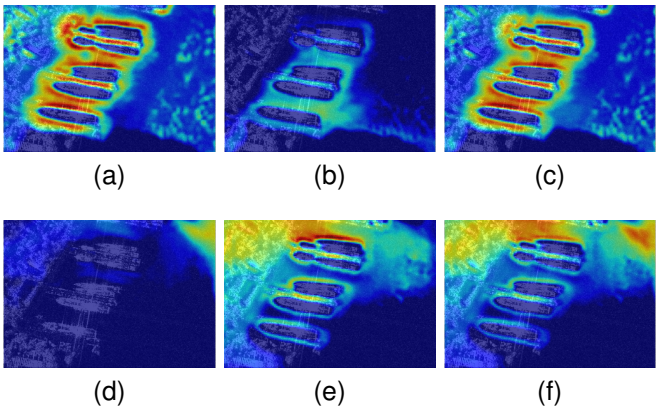

Fig. 12. Visualization of feature fusion in classification and localization branches without and after FD, respectively. (a), (b), (c) are classification branch feature map, localization branch feature map and mask feature map obtained by feature fusion without FD. Most of the ship edge features are derived from (a), illustrating the reversal of the primary and secondary relationships in the feature fusion process. (d), (e), (f) are feature maps after FD corresponding to (a), (b), (c). Ship edge features are mostly derived from (e), illustrating that the localization branch takes a dominant role in feature fusion, the problem of reversing the tasks of the classification and localization branches is solved.

Fig. 12 visualizes the feature fusion process. Without FD, Fig. 12(a), (b), (c) visualize the feature fusion process between the classification branch and the localization branch. Fig. 12(a) shows the feature map of classification branch, which contains massive information about ship's contour. Fig. 12(b) visualizes the feature map of localization branch, which provides less contour information compared with classification branch. Fig. 12(c) illustrates the mask feature map obtained by feature fusion, which is the sum of Fig. 12(a) and Fig. 12(b). From first row of Fig. 12, it can be seen that the contour information of the mask feature map is mostly provided by the classification branch, and the localization branch plays an auxiliary role, which is a reversal of the primary and secondary relationship.

According to the task definition of the branch, the localization branch completes the mask segmentation, and focuses on edge positions of ships. The classification branch completes the target center point prediction, and focuses on central positions. The visualization in Fig. 12 first row shows that there is a deviation in the tasks learned by different branches. The classification branch is more concerned about the edge position than localization branch, which results in the classification branch being unable to focus on its own task. This deviation must be corrected.

Fig. 12(d), (e), (f) illustrate the feature fusion process after FD. As can be seen in Fig. 12(d), the visualization color at the ship edges is blue, indicating that the classification branch feature map contains a few contour information. Fig. 12(e) shows the localization branch feature map, which contains more ship contour information compared with Fig. 12(d). Fig. 12(f) is the mask feature map, containing complete contour information of ships. According to the second row of Fig. 12, most of the ship contour information comes from localization

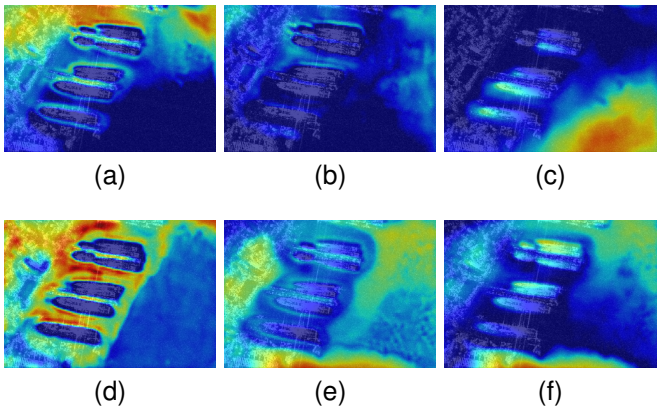

Fig. 13. Visualization of localization branch without CAC and with CAC, respectively. (a), (b), (c) are localization branch feature maps without CAC. (d), (e), (f) are localization branch feature maps with CAC. Compared to the first row, the feature maps in the second row have a more pronounced ship outline feature.

branch, which is consistent with the task definition of the branch.

Through above analysis, we can know that in the feature fusion process, classification branch can provide guidance information for localization branch. However, the mixing of gradient flows between different tasks branches will prevent the task branch from focusing on its native task, which causes the task branch to deviate from the learning direction. According to the structural design of our network, the classification branch should play an auxiliary role during the feature fusion process. This design not only conforms to the task definition of the branch, but also conforms to the principle of gradient backpropagation. During gradient backpropagation, smaller activation values mean smaller gradient flows. To reduce the effect of mixing of different branch gradient flows, classification branch as an auxiliary role is necessary during feature fusion. The FD module selects feature point from different spatial positions and channels for different task branches. When the corresponding weight of a feature point tends towards 0, it can suppress the gradient in backpropagation, while it tends towards 2, it can strengthen the gradient in backpropagation. The effectiveness of FD is verified through the visualization in Fig. 12.

Fig. 13 illustrates feature maps of localization branch without and with CAC guiding network training. The first row illustrates the feature maps without CAC guidance, from which it can be seen that the edges of the ship in feature maps are blurred. The second row illustrates the feature map with CAC guidance, which has clearer ship edge features compared to the feature map in the first row. Under the guidance of CAC, the network is able to perceive the boundaries of different instances more clearly, so that the dynamic mask head focuses on the corresponding instance.

V. Conclusion

In this article, we propose a BBox-free SAR ship instance segmentation method based on Gaussian heatmap. To tackle the issue of BBox-based NMS often omitting ships in densely arranged scenes, we introduce a center point prediction approach grounded in the mathematical properties of Gaussian distribution. This approach significantly enhances the accuracy of center point detection. To establish the relative positional relationship between pixels and target objects for the localization branch, we have devised a feature fusion module incorporating FD. FD effectively addresses the spatial location discrepancies in gradient flows during feature fusion. In order to guide the network in learning finer mask segmentation during training, we introduce the CAC. This addresses the limitation of the Dice Loss, which cannot perceive where the mask error occurs during instance segmentation. Our experiments, conducted on PSeg-SSDD and HRSID, demonstrate that our method proficiently extracts ship center points and precisely segments the ships. Ablation experiments underline the effectiveness of each component introduced in this paper. In the future, we plan to delve deeper into SAR ship instance segmentation algorithms to achieve higher accuracy while minimizing resource consumption.

References

[1] Z. Wu, B. Hou, and L. Jiao, "Multiscale cnn with autoencoder regularization joint contextual attention network for sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 2, pp. 1200–1213, 2020.

[2] W. Ao, F. Xu, Y. Li, and H. Wang, "Detection and discrimination of ship targets in complex background from spaceborne alos-2 sar images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 2, pp. 536–550, 2018.

[3] Z. Yue, F. Gao, Q. Xiong, J. Wang, T. Huang, E. Yang, and H. Zhou, "A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition," *Cognitive Computation*, vol. 13, pp. 795–806, 2021.

[4] G. Gao, "Statistical modeling of sar images: A survey," *Sensors*, vol. 10, no. 1, pp. 775–795, 2010.

[5] 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.

[6] L. Xu, H. Zhang, C. Wang, B. Zhang, and S. Tian, "Compact polarimetric sar ship detection with m-δ decomposition using visual attention model," *Remote Sensing*, vol. 8, no. 9, p. 751, 2016.

[7] D. J. Crisp, "The state-of-the-art in ship detection in synthetic aperture radar imagery," Defence Science and Technology Organisation Salisbury (Australia) Info ..., Tech. Rep., 2004.

[8] T. Zhang, X. Zhang, J. Li, X. Xu, B. Wang, X. Zhan, Y. Xu, X. Ke, T. Zeng, H. Su *et al.*, "Sar ship detection dataset (ssdd): Official release and comprehensive data analysis," *Remote Sensing*, vol. 13, no. 18, p. 3690, 2021.

[9] F. Gao, Y. Huo, J. Sun, T. Yu, A. Hussain, and H. Zhou, "Ellipse encoding for arbitrary-oriented sar ship detection based on dynamic key points," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–28, 2022.

[10] Z. Wu, B. Hou, B. Ren, Z. Ren, S. Wang, and L. Jiao, "A deep detection network based on interaction of instance segmentation and object detection for sar images," *Remote Sensing*, vol. 13, no. 13, p. 2582, 2021.

[11] H. Su, S. Wei, S. Liu, J. Liang, C. Wang, J. Shi, and X. Zhang, "Hq-isnet: High-quality instance segmentation for remote sensing imagery," *Remote Sensing*, vol. 12, no. 6, p. 989, 2020.

[12] F. Ma, X. Sun, F. Zhang, Y. Zhou, and H.-C. Li, "What catch your attention in sar images: Saliency detection based on soft-superpixel lacunarity cue," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[13] D. Xiang, Y. Xu, J. Cheng, Y. Xie, and D. Guan, "Progressive keypoint detection with dense siamese network for sar image registration," *IEEE Transactions on Aerospace and Electronic Systems*, 2023.

[14] D. Xiang, Y. Xie, J. Cheng, Y. Xu, H. Zhang, and Y. Zheng, "Optical and sar image registration based on feature decoupling network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[15] Y. Zhou, H. Liu, F. Ma, Z. Pan, and F. Zhang, "A sidelobe-aware small ship detection network for synthetic aperture radar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[16] F. Gao, L. Kong, R. Lang, J. Sun, J. Wang, A. Hussain, and H. Zhou, "Sar target incremental recognition based on features with strong separability," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.

[17] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer, 2020, pp. 677–694.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[19] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1483–1498, 2019.

[20] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6409–6418.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[22] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 282–298.

[23] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "Rtmdet: An empirical study of designing real-time object detectors," *arXiv preprint arXiv:2212.07784*, 2022.

[24] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[26] F. Gao, Y. Huo, J. Wang, A. Hussain, and H. Zhou, "Anchor-free sar ship instance segmentation with centroid-distance based loss," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 352–11 371, 2021.

[27] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 563–11 572.

[28] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, and Y. Fu, "Rethinking classification and localization for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 186–10 195.

[29] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[31] J. Zhang, M. Xing, G.-C. Sun, and N. Li, "Oriented gaussian function-based box boundary-aware vectors for oriented ship detection in multiresolution sar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[32] M. Zhao, X. Zhang, and A. Kaup, "Multitask learning for sar ship detection with gaussian-mask joint segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[33] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation," *Ieee Access*, vol. 8, pp. 120 234–120 254, 2020.

[34] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4974–4983.

[35] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9799–9808.

[36] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6910–6919.

[37] Z. Sun, C. Meng, J. Cheng, Z. Zhang, and S. Chang, "A multi-scale feature pyramid network for detection and instance segmentation of marine ships in sar images," *Remote Sensing*, vol. 14, no. 24, p. 6312, 2022.

[38] T. Zhang and X. Zhang, "A full-level context squeeze-and-excitation roi extractor for sar ship instance segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[39] T. Zhang and X. Zhang, "A mask attention interaction and scale enhancement network for sar ship instance segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[40] T. Zhang, X. Zhang, J. Li, and J. Shi, "Contextual squeeze-and-excitation mask r-cnn for sar ship instance segmentation," in *2022 IEEE Radar Conference (RadarConf22)*. IEEE, 2022, pp. 1–6.

[41] X. Ke, X. Zhang, and T. Zhang, "Gcbanet: A global context boundary-aware network for sar ship instance segmentation," *Remote Sensing*, vol. 14, no. 9, p. 2165, 2022.

[42] T. Zhang and X. Zhang, "Htc+ for sar ship instance segmentation," *Remote Sensing*, vol. 14, no. 10, p. 2395, 2022.

[43] S. Wei, X. Zeng, H. Zhang, Z. Zhou, J. Shi, and X. Zhang, "Lfg-net: Low-level feature guided network for precise ship instance segmentation in sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[44] Z. Shao, X. Zhang, S. Wei, J. Shi, X. Ke, X. Xu, X. Zhan, T. Zhang, and T. Zeng, "Scale in scale for sar ship instance segmentation," *Remote Sensing*, vol. 15, no. 3, p. 629, 2023.

[45] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 193–12 202.

[46] H. Ying, Z. Huang, S. Liu, T. Shao, and K. Zhou, "Embedmask: Embedding coupling for one-stage instance segmentation," *arXiv preprint arXiv:1912.01954*, 2019.

[47] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 906–13 915.

[48] M. Yasir, L. Zhan, S. Liu, J. Wan, M. S. Hossain, A. T. Isiacik Colak, M. Liu, Q. U. Islam, S. Raza Mehdi, and Q. Yang, "Instance segmentation ship detection based on improved yolov7 using complex background sar images," *Frontiers in Marine Science*, vol. 10, p. 1113669, 2023.

[49] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 649–665.

[50] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17 721–17 732, 2020.

[51] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[52] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1938–1958, 2021.

[53] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[54] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[55] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[56] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, "Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3205–3214.

[57] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.

[58] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3150–3158.

[59] M. Yi-de, L. Qing, and Q. Zhi-Bai, "Automated image segmentation using improved pcnn model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE, 2004, pp. 743–746.

[60] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE, 2020, pp. 1–7.
[61] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE access*, vol. 8, pp. 4806–4813, 2019.
[62] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach," *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, 2007.
[63] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.
[64] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 851–860.
[65] B. Prencipe, N. Altini, G. D. Cascarano, A. Brunetti, A. Guerriero, and V. Bevilacqua, "Focal dice loss-based v-net for liver segments classification," *Applied Sciences*, vol. 12, no. 7, p. 3247, 2022.
[66] L. Wang, C. Wang, Z. Sun, and S. Chen, "An improved dice loss for pneumothorax segmentation by mining the information of negative areas," *IEEE Access*, vol. 8, pp. 167 939–167 949, 2020.
[67] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2403–2412.
[68] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
[69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[70] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.
[71] Z. Tian, H. Chen, X. Wang, Y. Liu, and C. Shen, "Adelaidet: A toolbox for instance-level recognition tasks," 2019.
[72] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
[73] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.
[74] T. Cheng, X. Wang, S. Chen, W. Zhang, Q. Zhang, C. Huang, Z. Zhang, and W. Liu, "Sparse instance activation for real-time instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4433–4442.
[75] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1580–1589.
[76] X. Yang, Q. Zhang, Q. Dong, Z. Han, X. Luo, and D. Wei, "Ship instance segmentation based on rotated bounding boxes for sar images," *Remote Sensing*, vol. 15, no. 5, p. 1324, 2023.

Fei Gao received the B.S. degree in industrial electrical automation, and the M.S. degree in electromagnetic measurement technology and instrument from the Xi'an Petroleum Institute, Xi'an, China, in 1996 and 1999, respectively, and the Ph.D. degree in signal and information processing from the Beihang University, Beijing, China, in 2005.

He is currently a Professor with the School of Electronic and Information Engineering, Beihang University. His research interests include target detection and recognition, image processing, deep learning for applications in remote sensing.

Fengjun Zhong received the B.S. degree in electronic and information engineering from Beihang University, Beijing, China, in 2022, where he is currently pursuing the M.E. degree in information and communication engineering.

His research activities include target detection, instance segmentation, and remote sensing image processing.

Jinping Sun received the M.Sc. and Ph.D. degrees from Beihang University (BUAA), Beijing, China, in 1998 and 2001, respectively.

He is currently a Professor with the School of Electronic and Information Engineering, BUAA. His research interests include statistical signal processing, high-resolution radar signal processing, target tracking, image understanding, and robust beamforming.

Amir Hussain received the B.Eng. and Ph.D. degrees in electronic and electrical engineering from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively.

Following post-doctoral and senior academic positions at the University of the West of Scotland, Paisley, U.K., from 1996 to 1998, the University of Dundee, Dundee, U.K., from 1998 to 2000, and the University of Stirling, Stirling, U.K., from 2000 to 2018, respectively, he joined Edinburgh Napier University, Edinburgh, U.K., as the Founding Head of the Cognitive Big Data and Cybersecurity (CogBiD) Research Laboratory and the Centre for AI and Data Science. His research interests include cognitive computation, machine learning, and computer vision.

Huiyu Zhou received the B.Eng. degree in radio technology from the Huazhong University of Science and Technology, Wuhan, China, in 1990, the M.S. degree in biomedical engineering from the University of Dundee, Dundee, U.K., in 2002, and the Ph.D. degree in ratio technology, biomedical engineering, and computer vision from Heriot-Watt University, Edinburgh, U.K., in 2006.

He is currently a Professor with the School of Computing and Mathematical Sciences, University of Leicester, Leicester, U.K. His research interests include medical image processing, computer vision, intelligent systems, and data mining.