

A Deformable Spatial Attention Mechanism-Based Method and a Benchmark for Dock Detection

Yuhong Tu , Yan Song , Beibei Li, Qiqi Zhu , Songxue Cui , and Haitian Zhu

Abstract—Dock is a significant site in the shipbuilding industry. The detection of docks contributes to many important fields. With the abundant methods and datasets, the deep learning-based object detection in remote sensing images has received wide attention. However, there is no dataset that includes the dock class. This article first proposes a dock dataset to build a benchmark and advance dock detection research. Further, object detection of docks using existing methods cannot yield convincing results due to the characteristics of docks. To meet the challenges in dock detection, a novel deformable spatial attention module (DSAM) is proposed to enhance the feature representation and localization of docks. Based on the DSAM, a novel network architecture is proposed to perform accurate and efficient dock detection. The ablation and comparison experiments reveal that the proposed methods are accurate and effective, which are superior to the existing methods.

Index Terms—Dataset, deep learning, deformable convolution, dock, object detection, spatial attention.

I. INTRODUCTION

SHIPBUILDING industry plays a key role in the national defense security [1], transportation [2], marine development [3], and national economy [4] all over the world. Shipyards are

Manuscript received 20 October 2022; revised 15 January 2023 and 2 March 2023; accepted 4 April 2023. Date of publication 10 April 2023; date of current version 20 April 2023. This work was supported in part by the Second Batch of Undergraduate Teaching Project in 2020 UAV Photogrammetry 3D Modeling Practice Innovation Teaching Research under Grant ZL202053, in part by the Integration and Application Demonstration in the Marine Field under Grant 2020010004, in part by the Jiangsu Provincial Marine Science and Technology Innovation Project, Research and Application Demonstration of Remote Sensing Monitoring Method for Jiangsu Coastal Zone Resources with Multi-source Remote Sensing Data, under Grant JSZRHYKJ202207, and in part by the Automated Identifying of Environment Changes Using Satellite Time-Series, Dragon 5 Cooperation 2020–2024, under Grant 57971. (Corresponding author: Yan Song.)

Yuhong Tu, Beibei Li, and Qiqi Zhu are with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, China (e-mail: tuyuhong@cug.edu.cn; 1137165964@qq.com; zhuqq@cug.edu.cn).

Yan Song is with the School of Geography and Information Engineering, the School of Public Administration, and the Key Laboratory of Geological Survey and Evaluation of Ministry of Education, China University of Geosciences, Wuhan 430074, China, and also with the National Engineering Research Center of Geographic Information System, Wuhan 430078, China (e-mail: songyan@cug.edu.cn).

Songxue Cui is with the National Satellite Ocean Application Service, Beijing 100081, China (e-mail: cuisongxue@163.com).

Haitian Zhu is with the National Satellite Ocean Application Service, Beijing 100081, China, and also with the Key Laboratory of Space Ocean Remote Sensing and Application, Ministry of Natural Resources, Beijing 100081, China (e-mail: zht@mail.nsoas.org.cn).

Dock dataset can be downloaded in https://drive.google.com/drive/folders/1_a01tOQtM0_aanoHXGPSkuqRRqWJaCjk?usp=share_link.

Digital Object Identifier 10.1109/JSTARS.2023.3265700

located near the seashore or water shore, utilized to construct new ships and repair old ones, exerting a crucial effect in the shipbuilding industry [5], [6]. The most complex process in the shipbuilding is carried out on the docks [7]. Therefore, the dock is the most critical component of a shipyard. From the perspective of industrial production, different scales of docks can be utilized to produce ships with various throughputs. So, it can directly and effectively reflect the production capacity of each shipyard by detecting the docks. From the perspective of land resource utilization, timely detection of long-term vacant docks and timely adjustment of the land use type are conducive to improving the land resource utilization in coastal areas. From the perspective of ecological environment, it will release a lot of waste water, exhaust gas, and dust in welding, painting, and outfitting during ship construction, which directly affect the ecological environment of the coastal area around the docks. The detection of docks contributes to the protection of ecological environment in the coastal areas.

At present, the manual statistics are employed for detection of docks in the traditional methods, which are subject to lack of macroscopic, inefficiency, and mass labor cost. Satellite remote sensing (RS) can observe objects from a macroscopic perspective, which can overcome the disadvantages of traditional methods. However, to the best of our knowledge, there are few researches on the detection of docks based on the RS images (RSIs) worldwide. For instance, Firat et al. [8] proposed an approach for end-to-end object detection by leveraging large amount of unlabeled data and a single-layer convolutional sparse autoencoders, which were evaluated on the dry docks. Combining multisource optical RSIs and deep learning methods, this article aims to achieve macroscopic and efficient detection of docks with low labor cost.

Object detection has been widely applied in RSIs and received extensive attention in recent years [9]. So far, there have been several public geospatial object detection datasets for RSIs available. For instance, the TAS dataset for automotive detection of aerial images [10], the SZTAKI-INRIA dataset for building detection [11], the RSOD with four classes [12], [13], the UCAS-AOD dataset for vehicles and planes [70], the NWPU VHR-10 dataset containing 10 categories [14], [39], the HRSC2016 dataset for ship detection [15], the DOTA dataset [16] including 15 categories and the DIOR benchmark in optical RSIs [18] containing 20 categories. Especially, the DOTA dataset has been expanded from DOTAv1.0 to DOTAv1.5 and DOTAv2.0 [17] in recent years. However, there is no dock class in all the published datasets, which definitely causes a challenge for deep

learning-based detection of docks. To build a benchmark for dock detection, a novel dataset for deep learning-based dock detection is proposed in this article first.

For the traditional object detection of images in natural scenes, the current deep learning-based methods can be divided into two main streams. The first is the two-stage method, which detects objects with the two-stage convolutional neural network (CNN) methods. The best-known two-stage methods include R-CNN [19], Fast R-CNN [20], Faster R-CNN [21], Mask R-CNN [22], R-FCN [23], and Cascade R-CNN [24]. The second refers to detecting the objects with the single-stage methods which do not need the proposals generated by an extra network like the two-stage ones. Therefore, it presents a faster and simpler architecture [25]. The best-known single-stage methods include YOLO series [26], [27], [28], [29], [30], SSD [31], RetinaNet [32], LADet [33], and so on.

In comparison with the object detection in natural scenes, that in RSIs faces more challenges [34] such as rotation, efficiency, and accuracy, which have been intensively studied. In the past few years, there are many researches focusing on the two-stage methods such as adopting the R-CNN architecture to detect various geospatial objects in RSIs [35], [36], [37], [38], [39]. Cheng et al. [39] proposed a rotation-invariant CNN model for multiclass geospatial object detection, which merges a novel rotation-invariant layer to the R-CNN. However, the methods based on R-CNN are time consuming. To further improve the accuracy and efficiency of object detection in RSIs, the faster R-CNN-based object detection methods for RSIs are developed and applied [40], [41], [42], [43], [44], [45]. For example, Li et al. [43] proposed a rotation-insensitive region proposal network (RPN) network by introducing a multiangle anchor frame into the RPN in the Faster R-CNN framework, which can effectively address the rotational variations of geospatial objects. Tang et al. [44] developed a hyperregion proposal network for vehicle detection and used hard negative sample mining to further improve the accuracy. The most of the above methods are based on the horizontal bounding boxes (HBBs). With the DOTA and DIOR benchmarks released, the object detection methods in RSIs have gradually developed toward the oriented bounding boxes (OBBs) based ones [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], the baselines of which are not restricted to the two-stage methods. Notwithstanding the OBBs can match the real forms of geospatial objects, it only can be implemented smoothly with more extra parameters, such as the orientation of objects, requiring more samples for training. Due to the limited number of dock samples in the proposed dock dataset, this article focuses on the detection of docks based on HBBs. Therefore, the relatively traditional Faster R-CNN is adapted as the baseline in the article because of its universality, stability, and reliability for object detection in RSIs with HBBs annotation.

With the rapid development of deep learning-based methods, attention mechanism has arguably become one of the most important concepts [56]. Attention not only indicates where to focus, but also improves the representation of interests [57]. The attention in object detection can be categorized into two types: spatial attention and channel attention. The spatial attention allows networks to learn the positions that should be focused

on from the spatial-wise [56]. For instance, Mnih et al. [58] presented a spatial attention model as a single RNN, which takes a glimpse window as its input and selects the next location to focus on using the internal state of the network, and generates control signals in a dynamic environment. The channel attention allows networks to learn the focuses from the channel-wise [56]. For example, Hu et al. [59] proposed a squeeze-and-excitation (SE) network that adaptively recalibrates the channel-wise feature responses by explicitly modeling the interdependencies among different channels. In the SE network, the channel-wise attention can be obtained based on the global average-pooled features. To comprehensively consider the spatial-wise and channel-wise, Chen et al. [60] proposed a novel spatial and channel-wise attention CNN which merges spatial and channel attentions in a CNN. Woo et al. [57] developed a convolutional block attention module (CBAM), comprising the concatenated channel and spatial attention modules (SAMs), which improves the performance while keeping the overhead small. Unlike CBAM, the BAM [61] arranges the channel and SAMs in parallel, which can learn what and where to focus or suppress efficiently through two separate pathways and can refine the intermediate features effectively.

Located in shipyards, the docks are utilized for ship construction and repair. Due to the long period of ship construction, the materials stacked on the same dock often change, the manifestations of docks appear differently over time. Moreover, the docks exhibit flexible orientations and various spatial distributions. Accordingly, the detection of docks in RSIs is more challenging. Nowadays, the mainstream object detection baseline is the backbone combined with feature pyramid network (FPN) [62], which can enable the multilevel features to the subsequent architecture. However, the faster R-CNN with FPN fails to consider the orientations and the features of real dock forms within the HBBs, as they are usually mixed with the features of background in HBBs. Meanwhile, the HBBs tend to overlap each other in the case of dense distribution, so that the network recognizes multiple docks as one. In addition, the docks are easily confused with other geospatial objects in the shipyard, even the background.

To make up the above deficiencies, a novel deformable spatial attention module (DSAM) is proposed based on the CBAM [57] and the deformable convolution [63], [64]. In the proposed DSAM, the extra deformable convolution layers are merged with the traditional SAM to learn the irregular location offsets which can match the real dock forms. Based on the location offsets, the network can focus on the informative features of the real dock forms. Therefore, with the DSAM, the dock features can be better represented and the overlap with omission caused by the HBBs annotation in the case of dense distribution can be relieved. Further, a novel network architecture is proposed based on the DSAM and the faster R-CNN with FPN to improve the accuracy in the detection of docks, which enhances the localization and feature representation of docks. Finally, the effectiveness of proposed methods is proved experimentally, and the accuracy is confirmed to be better than that of the existing methods.

The rest of this article is organized as follows. In Section II, the dock dataset is introduced and analyzed. In Section III,

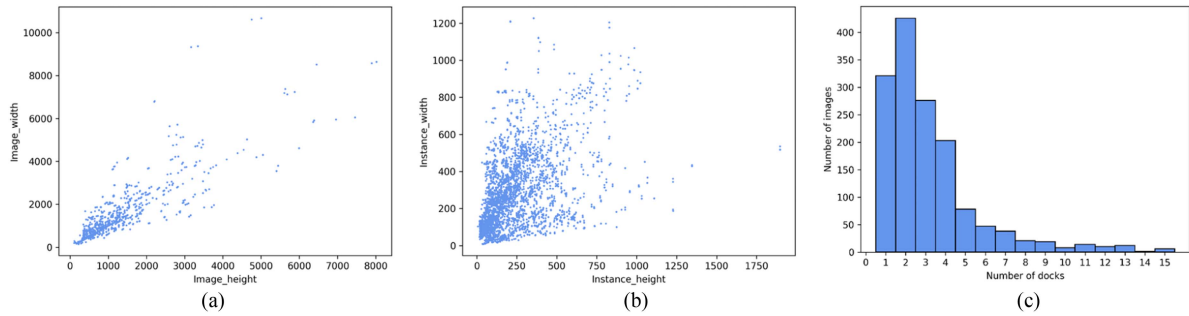


Fig. 1. Dock dataset statistics. (a) Image size statistics. (b) Instance size statistics. (c) Dock density statistics.

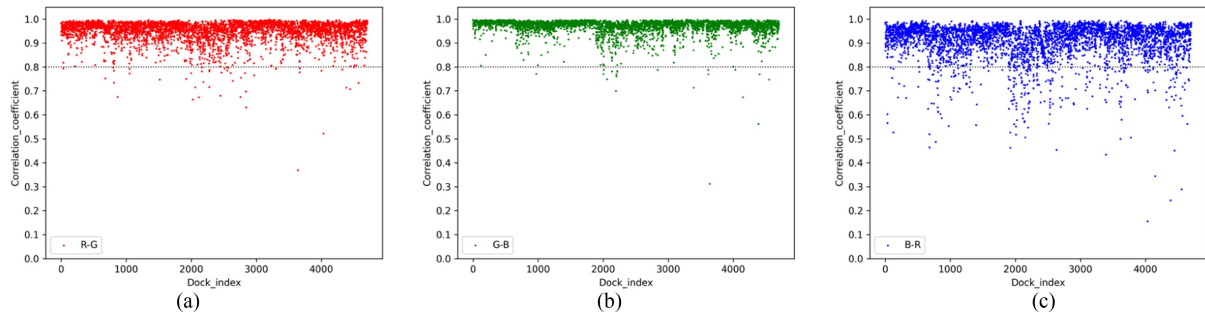


Fig. 2. Correlation coefficients within all instances. (a) Red-Green. (b) Green-Blue. (c) Blue-Red.

the DSAM and a novel network architecture are proposed. In Section IV, the comprehensive experiments are conducted to verify the accuracy and efficiency of the proposed methods. Finally, Section V concludes this article.

II. DOCK DATASET

First, a novel dock dataset including 1480 images and 4704 docks is proposed in this article. Due to the limited docks, there are RSIs in the dataset for the same dock from different times and sources. The dataset is collected from Google Earth (Google Inc.), map world (MW) [73], and ZY-3 [74]. The sources of the RSIs include 315 ZY-3 images, 575 Google Earth images, and 590 MW images and the corresponding resolutions are 2 m, 0.5 m, and 0.5 m, respectively.

In the dock dataset, all the RSIs that include three RGB visible bands cover the complete area of each shipyard, and the docks are labeled by HBBs in the format of VOC annotation [65]. Some typical RSIs labeled by HBBs in the dock dataset are given in Table I, in which the docks are labeled by the green HBBs. Further, the image size, the instance size (HBB size), and the dock density are illustrated in Fig. 1 while the ranges of data are given in Table II.

Due to the long period of ship construction, the same dock observed by different RSIs represents differently, such as being stacked by materials or just being vacant. Moreover, different scales of docks feature with variable characteristics. For instance, there are normally large guide rails and gantry cranes in the large-scale docks, which manifest their distinctly

interpretable characteristics. Nevertheless, the characteristics of diminutive docks are indistinctive because of the limited scale. Different locations of shipyards make the docks located inland or adjacent to the water, so the orientations of docks are unfixed. Therefore, it can be concluded that the docks are abundant in intraclass diversity. In addition, docks are densely distributed in some cases as shown in shipyard 5. The above discussion confirms that detection of docks is more challenging than that of other objects.








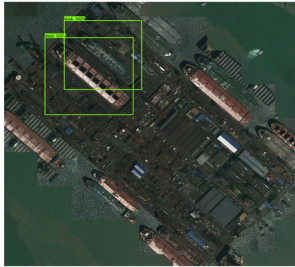







Due to the unique representation forms of docks in RSIs, the correlation coefficients between each band of a dock instance are calculated to further reflect the characteristics of the docks (see Fig. 2). It can be calculated with the flattened pixel vectors of different bands obtained by flattening the dock instance pixels within HBBs in images. The distributions of correlation coefficients which are larger than 0.9 and 0.8 are given in Table III.

Table III reveals the correlations between each band are high. A novel SAM and a novel network architecture for dock detection are proposed in the next section, which can overcome the above challenges to improve the detection accuracy.

III. METHODOLOGY

With the capability to address the multiscale feature representation, the FPN [62] has been merged into the mainstream network structure in recent years. Not only the FPN, but also the attention mechanisms enable the network to focus more on the significant features in images, which also enhance the feature

TABLE I
SOME SAMPLES IN THE DOCK DATASET

| Index | GE | MW | ZY-3 |
|-------|---|---|---|
| 1 |  |  |  |
| 2 |  |  |  |
| 3 |  |  |  |
| 4 |  |  |  |
| 5 |  |  |  |

representation capability of the network. The CBAM [57] is one of the most popular attention mechanisms, which consists of the channel attention module (CAM) and the SAM to comprehensively focus on the significant features from spatial-wise and channel-wise.

Although the extra network structures enhance the expression of features, the existing methods are not well adapted for the dock detection because of the dock characteristics. Specifically, the CBAM is usually merged with the backbone to enhance the feature representation for regions of interest. And, the SAM

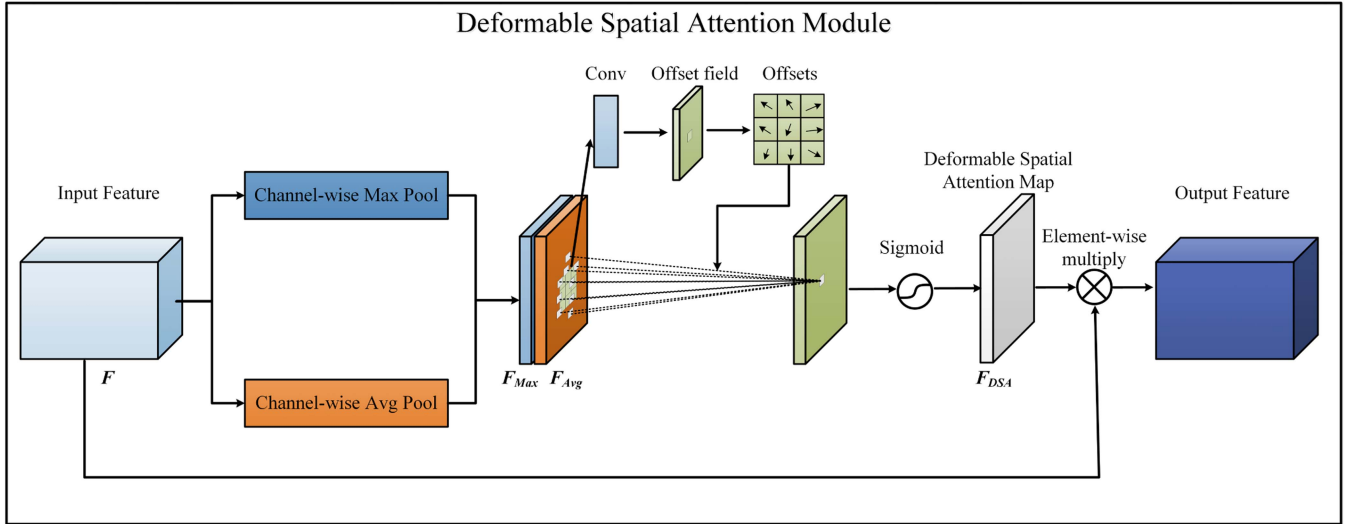


Fig. 3. Schematic diagram of deformable spatial attention module.

TABLE II
RANGES OF DATA IN THE DOCK DATASET

| Range type | Min size (pixels) | Max size (pixels) |
|---------------|-------------------|-------------------|
| Image size | 104*200 | 8019*8639 |
| Instance size | 38*8 | 985*1066 |

TABLE III
DISTRIBUTION OF CORRELATION COEFFICIENTS

| Bands correlation coefficient | >0.9 | >0.8 |
|-------------------------------|------|------|
| Red-Green | 90% | 99% |
| Green-Blue | 97% | 99% |
| Blue-Red | 69% | 93% |

in CBAM enables the network to focus on the informative parts from the spatial-wise. Nevertheless, the receptive field of conventional convolution in SAM is commonly a square area, which leads the network to focus on the part mismatched the real dock forms. Moreover, the orientations of docks vary widely within HBBs. Consequently, the SAM cannot enable the network to focus on the features of real dock forms well within limited square areas, which unduly merges the features of dock and background together. On the other hand, the CAM in CBAM enables the network to focus on the informative channels among the multi channels of the features from the channel-wise. And yet, the correlations between each visible band of the docks are strong based on the results in Table III. Excessive channel attention brings more parameters to the model, which may result in overfitting and reducing the final accuracy in test set. Thus, this article only focuses on the spatial attention which can improve the detection accuracy of the docks. Further, the network structure combined with attention module is critical. Owing to the multiblock structure of the backbone, the sizes of the feature maps are different. Therefore, where to merge the

attention module in the network will directly affect the ability of the network to localize and detect docks.

To address the above issues and improve the feature expression and the location accuracy of docks, a novel DSAM and a novel network structure are proposed in this article. Inspired by the deformable convolution [63], [64], the DSAM (see Fig. 3) is proposed, in which the traditional convolution of SAM is replaced with the deformable convolution. The deformable convolution is able to learn the irregular offsets with additional convolutional layers, which enables the network to focus on the features of irregular dock forms. The principle of DSAM can be expressed as

$$\begin{cases} F_{Avg} = \text{AvgPool}(F) \\ F_{Max} = \text{MaxPool}(F) \end{cases} \quad (1)$$

$$F_{DSA} = \sigma(f_{DC}^{N \times N}([F_{Avg}; F_{Max}])) \quad (2)$$

$$f_{DC}^{N \times N}(p) = \sum_{n=1}^{N \times N} w_n * x(p + p_n + \Delta p_n) * \Delta m_n. \quad (3)$$

In (1), F represents the input feature of DSAM, F_{Avg} and F_{Max} are the output of the sibling average pooling and max pooling, respectively. Then, the two outputs are aggregated and input to the deformable convolution layer, which is calculated in (2). The σ represents the sigmoid function while N represents the sampling kernel size. F_{DSA} is the deformable spatial attention map, which represents the spatial-wise weighting feature map. The deformable convolution operation can be calculated by (3), where Δp_n and Δm_n are the learnable offset and modulation scalar for the n th location, respectively. The value range of Δm_n is $[0, 1]$, while Δp_n is a real number with unconstrained range. The w_n and p_n denote the weight and prespecified offset for the n th location, respectively. The sampling is on the irregular and offset locations $x(p + p_n + \Delta p_n)$ which values are usually

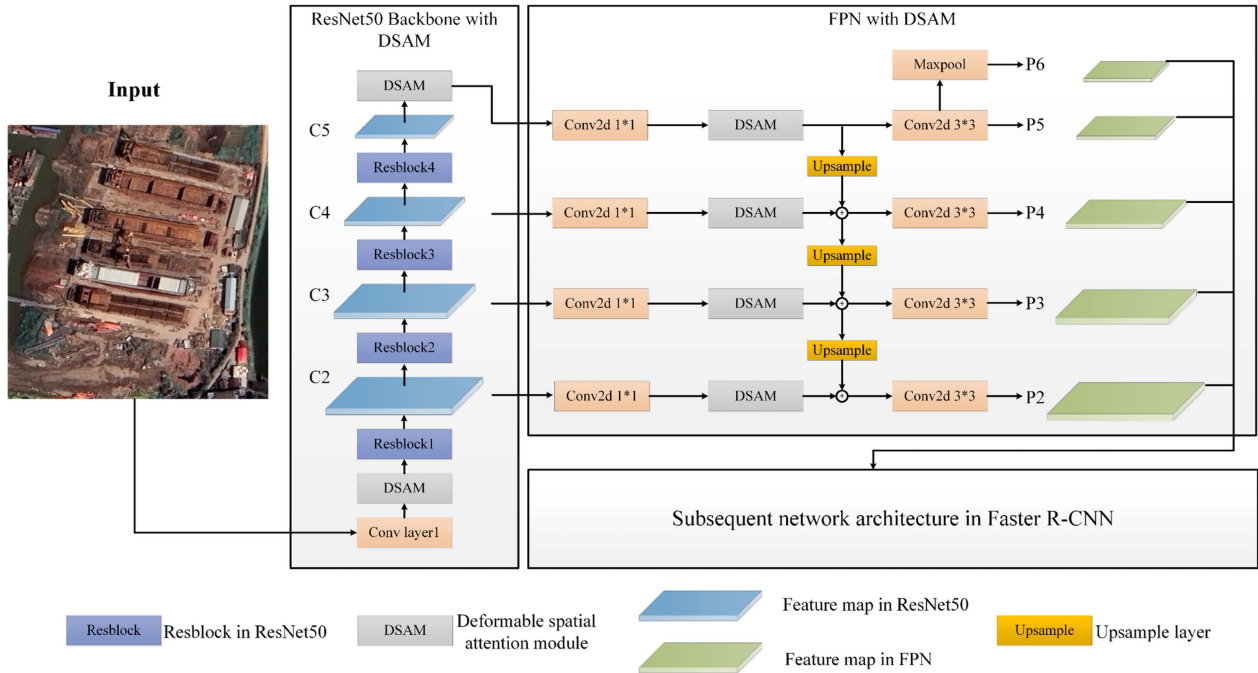


Fig. 4. Schematic diagram of the proposed network architecture.

fractional and can be calculated by the bilinear interpolation. Finally, F_{DSA} and F are element-wise multiplied to produce the spatially weighted feature (the output feature in Fig. 3) which is input to the subsequent structures.

In the DSAM, multichannel information can be first compressed by the sibling channel-wise operations. Then, the aggregated feature $[F_{Max}, F_{Avg}]$ is input to the deformable convolution layer to obtain the irregular directional offsets. Finally, the deformable spatial attention map is obtained to weight the input feature from the spatial-wise. With the DSAM, the network can learn to focus on the features in the irregular forms that approximately resemble the forms of docks. Therefore, the characteristics of docks can be better represented based on the DSAM, which are not restricted by the traditional convolutional reception field of squares.

Based on the proposed DSAM and the ResNet50 with FPN, a novel feature extraction network architecture is proposed, which is illustrated in Fig. 4. Because the ResNet50 is a bottom-up structure and the FPN is a top-down structure, the DSAM should be merged into the two structures, respectively. Concerned with the precise localization and feature representation, the DSAM is added into the ResNet50 where before the Resblock1 and after the Resblock4. Because the feature map which is the output of the Conv layer1 contains low-level semantic information and extensive location information, we add DSAM between the Conv layer1 and Resblock1 to mainly enhance the localization of docks. Further, the capability of features representation and classification of network should also be taken into account. The C5 feature contains extensive high-level semantic information. Hence, the DSAM is added after the Resblock4 to mainly enhance the feature representation of docks. With the addition of

TABLE IV
RESULTS OF DIFFERENT NETWORK ARCHITECTURES (FR REPRESENTS THE FASTER R-CNN [21], BACKBONE IS IN PARENTHESES)

| Network architecture | mAP |
|----------------------------------|--------------|
| FR (Proposed architecture) | 71.02 |
| FR (Without intermediate DSAM) | 68.23 |
| FR (ResNet50 with FPN, baseline) | 66.03 |
| FR (ResNet101 with FPN) | 65.89 |
| FR (Xception [71]) | 60.26 |
| FR (MobileNet [72]) | 59.72 |
| R-FCN [23] (ResNet50) | 63.47 |

DSAMs, the localization and features representation for docks of ResNet50 can be improved without excessive expansion. On the other hand, the FPN is the extra structure with ResNet50, which input features are the outputs of ResNet50. And, we set C2, C3, C4 and the output of the top DSAM in ResNet50 as the input to FPN. In the structure of FPN, the inner block (conv2d 1*1 in Fig. 4) unifies the dimensions of the input features to 256 dimensions. Moreover, there is an upsample layer between each FPN layer, which expands the spatial size of feature map. During upsampling, the spatial attention is significant for preserving high-level semantic information while accurately locating docks. Thus, the DSAM is added after each inner block in FPN to further reinforce the localization and feature representation of docks. With the addition of DSAM in the ResNet50 with FPN, the backbone can focus on the informative parts matched real dock forms and reinforce the dock features representation during the feature extraction stage.

TABLE V
RESULTS OF ABLATION EXPERIMENTS

| Baseline | | | Baseline + Different settings | | | |
|----------|--------------|--------------|-------------------------------|--------|--------|--------|
| DSAM | | ✓ | | | | ✓ |
| CAM | Faster R-CNN | | ✓ | | ✓ | ✓ |
| SAM | | | | ✓ | ✓ | |
| mAP | 66.03 | 71.02 | 67.11 | 68.45 | 69.13 | 66.78 |
| FPS | 10 | 8 | 8 | 8 | 7 | 6 |
| Param. | 41.35M | 41.41M | 41.39M | 41.36M | 41.39M | 41.45M |

TABLE VI
RESULTS OF DIFFERENT KERNEL SIZE IN DSAM

| Kernel Size | Param. | mAP |
|-------------|--------|--------------|
| 7*7 | 41.41M | 71.02 |
| 5*5 | 41.37M | 68.61 |
| 3*3 | 41.36M | 66.88 |

The bold value highlight the optimal detection result.

Finally, the multiscale features P2-P6 are input to the RPN and the subsequent processes are the same with the faster R-CNN.

IV. EXPERIMENTS AND RESULTS

The proposed dock dataset and methods are used to conduct experiments. The experiments are performed on one RTX3070 with 8G memory by PyTorch in Ubuntu. We use 70% of the dataset as training set and 30% as test set. To thoroughly evaluate the effectiveness of the proposed methods, we first perform extensive ablation experiments. Meanwhile, we verify the proposed methods outperform the existing methods. Further, we analyze the effect of kernel size hyperparameter on the final detection results. Finally, we compare the predicted results with the ground truth boxes in the test set. In particular, the mAP in the following results is calculated according to the evaluation indicator in the MS coco dataset [67], which is under the intersection over union of 0.5. And the pretrained weights [68] trained by MillionAID [69] are used in this article.

A. Comparison and Ablation Experiments

To verify the effectiveness of the proposed network architecture, experiments with different network architectures are conducted firstly. As analyzed in Section III, the localization and feature representation of docks should be concerned comprehensively. Thus, the top and bottom DSAMs in all ResNet50 and FPN are definitely required. We compare the cases with or without DSAM in the two intermediate layer of the FPN, and the results of other existing methods are given in Table IV.

With the results in Table IV, the optimal detection result can be obtained by the proposed network architecture, which is superior to the existing methods. Further, it is proved that the DSAM can further reinforce the feature representation and localization of docks during upsampling in FPN. The spatial attention is significant for preserving high-level semantic information with

accurately locating docks, which will directly affect the final detection results. Therefore, the proposed network architecture is reasonable, which focuses on the informative dock area during the whole bottom-up and then top-down feature extraction process based on the DSAM.



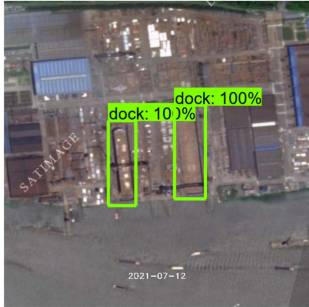







To further verify the effectiveness and superiority of the proposed methods, more ablation experiments are conducted based on the proposed network architecture. Different network architectures are compared in Table V, where also demonstrates the frames per second (FPS) and the parameters number of whole network (Param.).

As given in Table V, the mAP of the proposed architecture (only with DSAM) is optimal, which is greater 1.9% than the baseline with CBAM (CAM+SAM) and improves 4.9% than the baseline. However, the combination of DSAM and CAM (MCBAM) performs the worst in all improved architectures. It demonstrates excessive channel-wise attention yields negative results instead. As analyzed in Fig. 2, there are high correlations between the visible bands of the docks. In the MCBAM, the DSAM will be influenced by the excessive channel-wise attention. As a result, the DSAM cannot perform its original function and the detection accuracy is worse than the DSAM alone. The proposed DSAM focuses on the features within the approximate forms of the real docks, which comprehensively enhance the features representation and localization for docks from the spatial-wise. Hence, the proposed network architecture achieves the best accuracy.

B. Hyperparameter Experiments

The effectiveness of the proposed methods has been proved by the above experiments. Further, the comparison experiments of detailed hyperparameter are performed. Because of the different kernel size of the deformable convolution layer in DSAM, the receptive fields of the attention parts are diverse, which will affect the detection accuracy. The detection result is better when the network matches the real form of the docks more precisely. Thus, the kernel size of DSAM is a significant hyperparameter. In ResNet50, the bottom feature spatial dimension is large with a narrow channel dimension and contains abundant location information, while the top feature spatial dimension is small with a wide channel dimension and contains abundant high-level semantic information. We empirically set the kernel size of the top and the bottom DSAM in ResNet50 as 3 and 7. Further, considering the top-down structure of FPN and all the channel sizes of the input features are the same, the kernel sizes in FPN

TABLE VII
GROUND TRUTH BOUNDING BOXES AND THE PREDICTION RESULTS

| Index | Dock density | GT box | Prediction |
|-------|--------------|---|--|
| 1 | 1 |  |  |
| 2 | 2 |  |  |
| 3 | 2 |  |  |
| 4 | 8 |  |  |
| 5 | 9 |  |  |

are set the same empirically. We compare the effect of different kernel size of DSAM in FPN to the results which are given in Table VI.


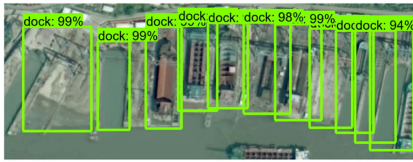
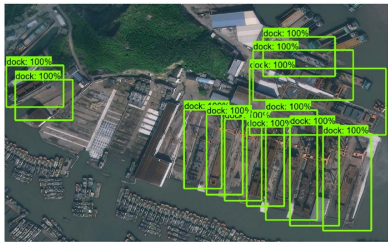
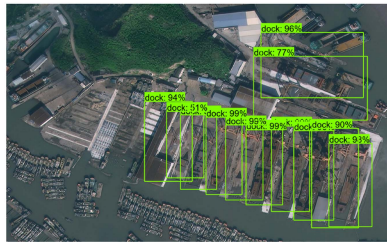




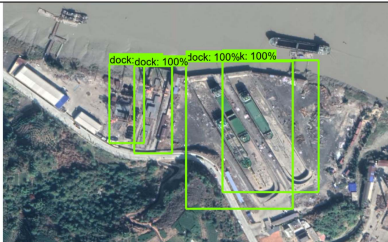
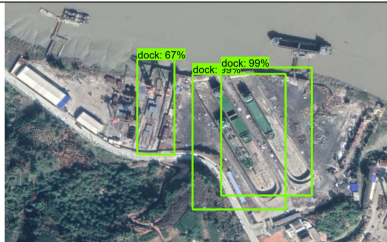
As given in Table VI, the optimal result can be obtained when the kernel size is set to 7. Due to the sizes of docks, small kernel size limits the feature representation and the localization of docks, which cannot match the real forms of the docks well.

Therefore, the kernel size of 7*7 is the optimal setting for the dock detection.

C. Detection Results

Based on the optimal parameters of the trained network, the test set is employed to compare the prediction results with the

TABLE VII
(CONTINUED)

| Index | Dock density | GT box | Prediction |
|-------|--------------|---|--|
| 6 | 11 |  |  |
| 7 | 12 |  |  |
| 8 | 2 |  |  |
| 9 | 2 |  |  |
| 10 | 4 |  |  |

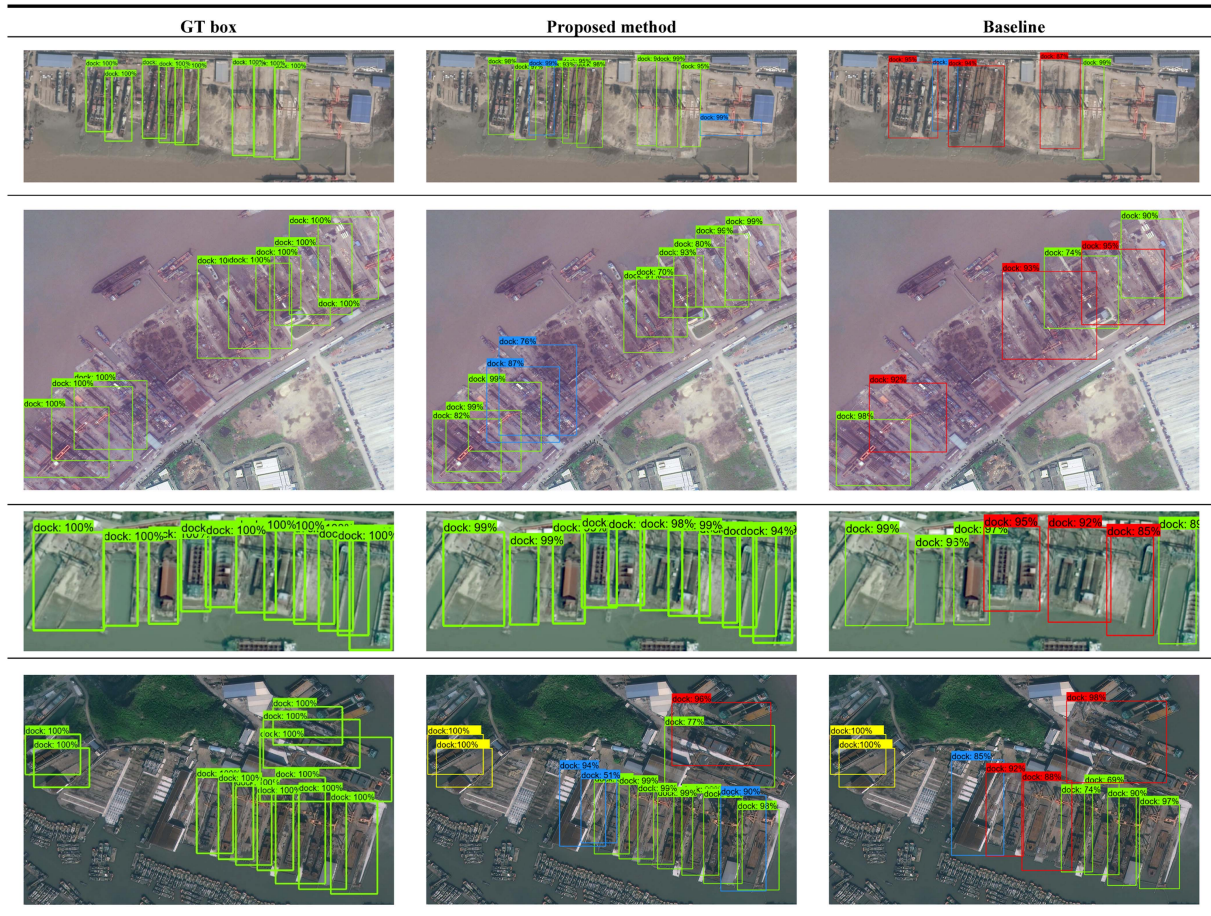
ground truth bounding boxes (GT boxes). A variety of representative cases from the test results are exhibited in Table VII, such as correct detection, missed detection, false detection, etc.

In Table VII, the detection results can be classified into various cases. Specifically, the detection results in shipyards 1 and 2 are correct, which means that the number and the area of docks are all basically correct. The detection results in shipyard 3 are the missed detection because two vacant docks are not detected,

which is possibly due to the insufficient distinctive features in the vacant docks. Further, there are few vacant docks in the dock dataset, which are not “learned” by the network. In shipyards 4–7, the detection results are acceptable in the case of dense distribution, which indicates that the proposed DSAM and network can detect the docks in dense distribution by focusing on the characteristics of each dock. The shipyard 8 is subject to false detection. The material area is incorrectly identified as a dock because of the similar characteristics of different objects,

TABLE VIII

SOME PREDICTION RESULTS IN THE CASE OF DENSE DISTRIBUTION (THE GREEN BOXES INDICATE THE CORRECT DETECTIONS. THE BLUE ONES INDICATE THE FALSE DETECTIONS. THE RED ONES INDICATE THE SYNCRETIC DETECTIONS. THE YELLOW ONES INDICATE THE MISSED DETECTIONS)



that is, the two types of objects both contain the materials for ship construction and gantry cranes. In shipyards 9 and 10, two adjacent docks are identified as one, which may be caused by the unobvious boundary information between docks. Moreover, some typical prediction results in the case of dense distribution are given in Table VIII. The proposed methods can identify the docks more accurately than the baseline in the case of dense distribution because the latter tends to identify the multiple neighboring docks as one. In summary, the proposed methods exhibit better detection results than the existing methods, with the mAP of 71.02%, although there are still some special cases of missed detections and false detections.

V. CONCLUSION

To perform accurate and efficient dock detection, a novel dock dataset is proposed first in this article, which includes 1480 images and 4704 docks labeled by HBBs. In order to overcome the challenges in dock detection, a novel DSAM is proposed to drive the network to learn features that match the actual form of the docks. Based on the DSAM, a novel feature extraction network is proposed. The proposed network comprehensively focuses on the feature representation and localization of docks, which merges the DSAM with the backbone together. Further,

the sufficient experiments prove that the proposed methods are accurate and efficient, which are superior to the existing methods. The mAP of the proposed methods reaches 71.02%. Finally, the predicted results are compared with the GT boxes, which demonstrates that the predicted results are compelling.

REFERENCES

- [1] Madrohim and L. Prakoso, "The total war strategy through the improvement of the role of national shipyard in supporting main weapon system of Indonesian navy," *J. Social Political Sci.*, vol. 4, no. 1, pp. 1–10, 2021.
- [2] S. Lee, B. Kim, M. Huh, S. Cho, S. Park, and D. Lee, "Mining transportation logs for understanding the after-assembly block manufacturing process in the shipbuilding industry," *Expert Syst. Appl.*, vol. 40, pp. 83–95, 2013.
- [3] K. Hossain, N. Zakaria, and M. Sarkar, "SWOT analysis of China shipbuilding industry by third eyes," *Procedia Eng.*, vol. 194, pp. 241–246, 2017.
- [4] G. Collins and M. Grubb, "A comprehensive survey of China's dynamic shipbuilding industry," 2008.
- [5] T. Rizwan et al., "Literature review on shipyard productivity in Indonesia," *Depik*, vol. 10, no. 1, pp. 1–10, 2021.
- [6] I. Syahtaria et al., "SWOT application on determining ship yard development strategies: A case study," *J. ASRO*, vol. 10, no. 3, pp. 37–41, 2019.
- [7] M. Liu, Z. Yang, and M. Meng, "Integrated modelling and simulation method for the project-type manufacturing process: Shipyard dock shop hoisting process," *IEEE Access*, vol. 9, pp. 89999–90010, 2021, doi: [10.1109/ACCESS.2021.3091470](https://doi.org/10.1109/ACCESS.2021.3091470).

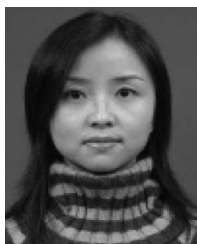
- [8] O. Firat, G. Can, and F. Vural, "Representation learning for contextual object and region detection in remote sensing," in *Proc. IEEE 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 3708–3713.
- [9] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605814, doi: [10.1109/TGRS.2021.3095186](https://doi.org/10.1109/TGRS.2021.3095186).
- [10] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 30–43.
- [11] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 33–50, Jan. 2012, doi: [10.1109/TPAMI.2011.94](https://doi.org/10.1109/TPAMI.2011.94).
- [12] Z. Xiao, Q. Liu, G. Tang, and X. Zhai, "Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images," *Int. J. Remote Sens.*, vol. 36, no. 2, pp. 618–644, 2015.
- [13] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017, doi: [10.1109/TGRS.2016.2645610](https://doi.org/10.1109/TGRS.2016.2645610).
- [14] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, 2016.
- [15] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016, doi: [10.1109/LGRS.2016.2565705](https://doi.org/10.1109/LGRS.2016.2565705).
- [16] G. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
- [17] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.
- [18] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, 2016, doi: [10.1109/TPAMI.2015.2437384](https://doi.org/10.1109/TPAMI.2015.2437384).
- [20] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 1137–1149, 2017, doi: [10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031).
- [22] K. He et al., "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
- [23] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 379–387.
- [24] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
- [25] J. Kaur and W. Singh, "Tools, techniques, datasets and application areas for object detection in an image: A review," in *Proc. Multimedia Tools Appl.*, 2022, vol. 81, pp. 38297–38351.
- [26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
- [27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.
- [28] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, vol. 1804, pp. 1–6.
- [29] A. Bochkovskiy et al., "Yolov4: Optimal speed and accuracy of object detection," 2020, [arXiv:2004.10934](https://arxiv.org/abs/2004.10934).
- [30] A. Kuznetsova, T. Maleva, and V. Soloviev, "Detecting apples in orchards using YOLOv3 and YOLOv5 in general and close-up images," in *Proc. Int. Symp. Neural Netw.*, 2020, vol. 149, pp. 233–243.
- [31] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, vol. 9905, pp. 21–37.
- [32] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.
- [33] J. Zhou, Y. Tian, W. Li, R. Wang, Z. Luan, and D. Qian, "LADet: A light-weight and adaptive network for multi-scale object detection," in *Proc. 11th Asian Conf. Mach. Learn.*, 2019, pp. 912–923.
- [34] P. Shamsolmoali, J. Chanussot, M. Zareapoor, H. Zhou, and J. Yang, "Multipatch feature pyramid network for weakly supervised object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021, doi: [10.1109/TGRS.2021.3106442](https://doi.org/10.1109/TGRS.2021.3106442).
- [35] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [36] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017, doi: [10.1109/TGRS.2016.2645610](https://doi.org/10.1109/TGRS.2016.2645610).
- [37] A. Salberg, "Detection of seals in remote sensing images using features extracted from deep convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1893–1896.
- [38] I. Sevo and A. Avramovic, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 740–744, May 2016, doi: [10.1109/LGRS.2016.2542358](https://doi.org/10.1109/LGRS.2016.2542358).
- [39] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016, doi: [10.1109/TGRS.2016.2601622](https://doi.org/10.1109/TGRS.2016.2601622).
- [40] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3652–3664, Aug. 2017, doi: [10.1109/JSTARS.2017.2694890](https://doi.org/10.1109/JSTARS.2017.2694890).
- [41] W. Guo, W. Yang, H. Zhang, and G. Hua, "Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network," *Remote Sens.*, vol. 10, 2018, Art. no. 131.
- [42] X. Han, Y. Zhong, and L. Zhang, "An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery," *Remote Sens.*, vol. 9, 2017, Art. no. 666.
- [43] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018, doi: [10.1109/TGRS.2017.2778300](https://doi.org/10.1109/TGRS.2017.2778300).
- [44] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, 2017, Art. no. 336.
- [45] Y. Zhong, X. Han, and L. Zhang, "Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 281–294, 2018.
- [46] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3500–3509.
- [47] J. Han, J. Ding, N. Xue, and G. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2785–2794.
- [48] X. Yang et al., "The KFIoU loss for rotated object detection," 2022, [arXiv:2201.12558](https://arxiv.org/abs/2201.12558).
- [49] J. Ding, N. Xue, Y. Long, G. S. Xia, and Q. Lu, "Learning ROI transformer for detecting oriented objects in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.
- [50] X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 18381–18394, 2021.
- [51] D. Liang et al., "Anchor retouching via model interaction for robust object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5619213, doi: [10.1109/TGRS.2021.3136350](https://doi.org/10.1109/TGRS.2021.3136350).
- [52] D. Wang et al., "Advancing plain vision transformer towards remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: [10.1109/TGRS.2022.3222818](https://doi.org/10.1109/TGRS.2022.3222818).
- [53] B. Zhong and K. Ao, "Single-stage rotation-decoupled detector for oriented object," *Remote Sens.*, vol. 12, 2020, Art. no. 3262.
- [54] Q. Ming, L. Miao, Z. Zhou, X. Yang, and Y. Dong, "Optimization for arbitrary-oriented object detection via representation invariance loss," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8021505, doi: [10.1109/LGRS.2021.3115110](https://doi.org/10.1109/LGRS.2021.3115110).
- [55] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, "Dense label encoding for boundary discontinuity free rotation detection," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15814–15824.

- [56] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.
- [57] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, vol. 11211, pp. 3–19.
- [58] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 2204–2212.
- [59] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [60] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6298–6306.
- [61] J. Park et al., "Bam: Bottleneck attention module," 2018, *arXiv:1807.06514*.
- [62] T. Lin et al., "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
- [63] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
- [64] X. Zhu et al., "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9308–9316.
- [65] M. Everingham and J. Winn, "The PASCAL visual object classes challenge 2012 (VOC2012) development kit," 2007, pp. 1–45.
- [66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [67] T. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [68] D. Wang et al., "An empirical study of remote sensing pre-training," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: [10.1109/TGRS.2022.3222818](https://doi.org/10.1109/TGRS.2022.3222818).
- [69] Y. Long et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-AID," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021, doi: [10.1109/JSTARS.2021.3070368](https://doi.org/10.1109/JSTARS.2021.3070368).
- [70] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 3735–3739.
- [71] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.
- [72] A. Howard et al., "Searching for mobilenetv3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.
- [73] "National platform for common geospatial information services," Ministry of Natural Resources of the People's Republic of China, Apr. 2023. Accessed: Apr. 2023. [Online]. Available: <https://www.tianditu.gov.cn/>
- [74] "China centre for resources satellite data and application," Apr. 2023. Accessed: Apr. 2023. [Online]. Available: <https://www.cresda.com/zgzywxxyzx/index.html>



Yuhong Tu received the B.S. degree in remote sensing science and technology from the School of Geography and Information Engineering from China University of Geosciences (Wuhan), Wuhan, China, in 2020. He is currently working toward the M.S. degree in photogrammetry and remote sensing with the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China.

His research interests include camera calibration and extrinsic calibration of combined sensor and deep learning-based object detection.



Yan Song received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2008, respectively.

She is currently an Associate Professor with the School of Geography and Information Engineering, China University of Geosciences (Wuhan). Her main research interests include photogrammetry and remote sensing, multitemporal remote sensing image change detection, remote sensing of islands, and coastal zones.



Beibei Li received the B.S. degree from the Henan Polytechnic University, Jiaozuo, China, in 2016. She is currently working toward the M.S. degree in resource and environment with the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China.

Her research interests include the extraction of remote sensing geographic information elements and remote sensing water depth inversion.



Qiqi Zhu received the B.S. degree in remote sensing science and technology from the China University of Mining and Technology, Xuzhou, China, in 2013, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2018.

She is currently an Associate Professor with the School of Geography and Information Engineering, China University of Geosciences (Wuhan). Her research interests are based on high-resolution, hyperspectral multi-source remote sensing data and multi-source geographic data in space, aviation and unmanned aerial vehicles.



Songxue Cui received the B.S. degree in marine technology from the Tianjin University of Science and Technology, Tianjin, China, in 2006, and the M.S. in cartography and geographical information system from the Ocean University of China, Qingdao, China, in 2009.

She is currently an Associate Professor with the National Satellite Ocean Application Service, Beijing, China. Her research interests include remote sensing of islands and coastal zones, satellite monitoring of marine disasters.



Haitian Zhu received the B.S. degree in information engineering from Wuhan University, Wuhan, China, in 2003, and the Ph.D. degree in signal and information processing from University of Chinese Academy of Sciences, Beijing, China, in 2015.

He is currently an Associate Professor with the Application Development Department, National Satellite Ocean Application Service, Beijing. His research interests include remote sensing of islands and coastal zones, and marine remote sensing application.