A Degraded Reconstruction Enhancement-based Method for Tiny Ship Detection in Remote Sensing Images with A New Large-scale Dataset

Jianqi Chen, Keyan Chen, Hao Chen, Zhengxia Zou and Zhenwei Shi*, *Member, IEEE*

***Abstract*—The rapid detection of ships within the wide sea area is essential for intelligence acquisition. Most modern deep learning-based ship detection methods focus on locating ships in high-resolution (HR) remote sensing (RS) images. Seldom efforts have been made on ship detection in medium-resolution (MR) RS images. An MR image covers a much wider area than an HR one of the same size, thus facilitating quick ship detection. To this end, we propose a tiny ship detection method namely, Degraded Reconstruction Enhancement Network (DRENet), for MR RS images. Different from previous methods that mainly focus on feature fusion strategies to improve the expression ability of the detector, we design an additional network branch, i.e., degraded reconstruction enhancer, to learn to regress an object-aware blurred version of the input image in the training phase. Our intuition is that the proposed reconstruction branch may guide the backbone to focus more on tiny ship targets instead of the vast background. Moreover, we incorporate a CRoss-stage Multi-head Attention module in the detector to further improve the feature discrimination by leveraging the self-attention mechanism. To fill the gap of lacking a large-scale MR ship detection dataset, we introduce Levir-Ship, which contains 3876 GF-1/GF-6 multi-spectral images and over 3K tiny ship instances. Experiments on Levir-Ship validate the effectiveness and efficiency of the proposed method. Our method achieves 82.4 AP with 85 FPS, which outperforms many state-of-the-art ship detection methods. Our code and dataset are available at https://github.com/WindVChen/DRENet.**

***Index Terms*—Convolutional neural network (CNN), ship detection, deep learning, optical image, remote sensing.**

I. Introduction

SHIP detection in optical remote sensing (RS) images refers to locating ships in RS images and giving their positions and sizes automatically. As the main carrier of sea transportation, ship plays a very important role in the military and civilian fields [1]. The accurate and fast detection of ships in RS images has been a hot research topic.

With the rapid development of RS technology in recent years, research on ship detection in optical images is more active because of its content-rich and expression-intuitive features [2]–[6]. Traditional ship detection methods often require complicated hand-craft features extraction, and cannot adapt to changing environments [7], [8]. Nowadays, deep learning (DL) techniques, especially convolutional neural networks (CNN), have been widely applied in RS image ship detection. Compared with traditional methods, DL-based methods can learn robust multi-level features and classifiers in an end-to-end manner. Inspired by the great success of DL-based detectors in the field of computer vision (CV), in recent years, considerable works have introduced DL techniques into RS ship detection tasks [9]–[17], which confirmed better detection performance in terms of accuracy and stability than traditional counterparts.

The existing DL-based ship detection methods are mostly modified based on the object detectors in CV. Recent advances in optical RS ship detection include fusing the ship shape priors [18], extracting features from ship prow and stern [12], and applying the orientation information of ships [19]. However, most of these methods focus on high-resolution (HR) images, thus benefiting from the rich textures and clear edges.

Despite the great success in ship detection in HR RS images [19]–[21], seldom efforts have been made in ship detection in relatively lower resolution RS images, e.g., medium-resolution (MR, about $16m$/pixel) RS images.[1] Considering the urgent need for monitoring and early warning over a wide area of sea in practical applications [22], [23], ship detection in MR images is critically important. We argue that MR RS images are more suitable than HR ones for quick ship detection in vast sea areas, because an MR image can cover a much wider area than an HR image of the same size. For instance, to detect ships in a fixed sea area, supposed we need 1 hour in $16m$/pixel MR images (GF-1), then it means we need 256 hours, about 10 days, to finish the same task in $1m$/pixel HR images (SkySat-1). The time cost is apparently unbearable, especially in such a big data era.

Ship detection in real-world MR RS images has several challenges. First, ship detection in MR RS images can suffer from scarce textures and hazy edges of ships. For example, in a GF-1 image ($16m$/pixel), a ship may only occupy 20 pixels. Second, in the real scenario, an RS image may be covered by massive fractus clouds. The complex imaging conditions may

The work was supported in part by the National Key Research and Development Program of China under the Grant 2019YFC1510905, in part by the National Natural Science Foundation of China under the Grant 62125102, and in part by the Beijing Natural Science Foundation under the Grant 4192034. (Corresponding author: Zhenwei Shi)

Jianqi Chen, Keyan Chen, Hao Chen and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.

Zhengxia Zou is with Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China.

[1]Please refer to the website (https://doc.arcgis.com/en/imagery/workflows/resources/managing-medium-resolution-satellite-data.htm) for detailed definitions of high resolution and medium resolution.

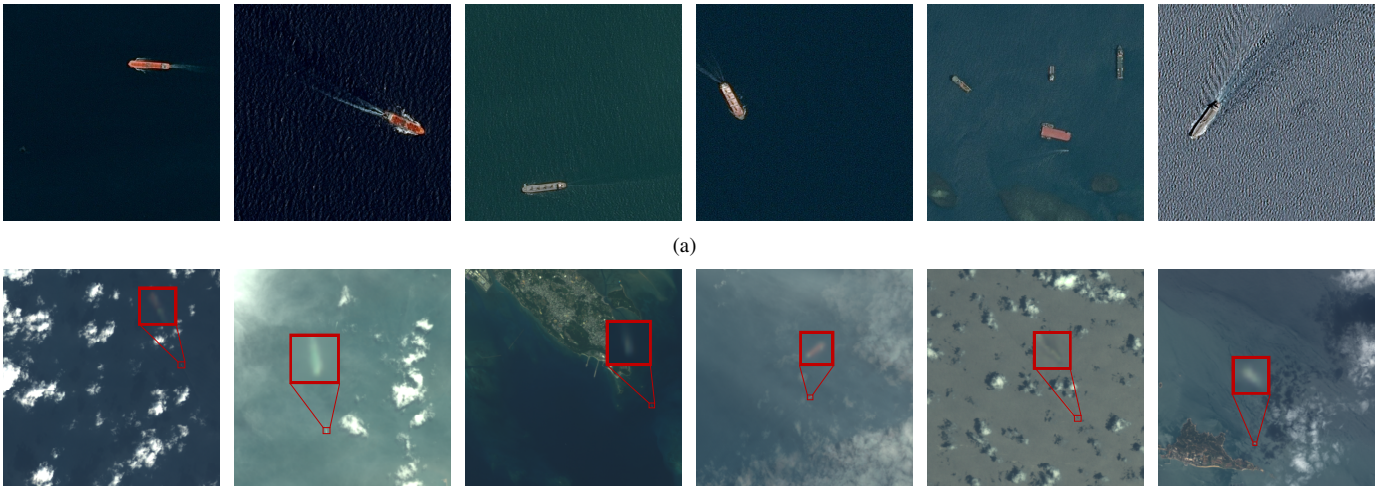


Fig. 1. Samples of 512×512 size from different datasets. (a): Samples from Airbus Ship Detection Competition dataset[2] with a spatial resolution of $2m$ and higher. (b): Samples from Levir-Ship dataset with a spatial resolution of $16m$.

also induce many false positives in a conventional detector. Fig. 1 gives an intuitive comparison of HR images and MR ones. The ships in MR images are difficult to be recognized (zoomed for best view) and also hard to be distinguished from fractus clouds. Most detection networks focus on improving the feature expression ability of the backbone or the neck by feature fusion [11], [24] and cross-stage connection [21], [25]. We argue that such feature enhancing methods are struggling to extract discriminative features of tiny ship targets.

To solve the above problems, we propose a degraded reconstruction enhancement ship detection network called DRENet - Tiny Ship Detection Based on Degraded Reconstruction Enhancement in Remote Sensing Images, which contains an efficient backbone to extract robust feature representations, a degraded reconstruction enhancer to help the backbone learn to distinguish ships from irrelevant backgrounds, and an object detector to locate ships. The pipeline is depicted in Fig. 2. Different from previous methods that mainly focus on feature fusion strategies [11], [24] to improve the expression ability of the detector, we design an additional network branch, i.e., degraded reconstruction enhancer, to learn to regress an object-aware blurred version of the input image in the training phase. In the enhancer, we design an image processing operation called "Selective Degradation" to blur the background. Our intuition is that the proposed reconstruction branch may guide the backbone to focus more on tiny ship targets instead of the vast background (e.g., fractus clouds). Please note that our proposed enhancer is only used in the training phase and is computing-free in the inferencing phase. Moreover, we incorporate a CRoss-stage Multi-head Attention (CRMA) module in the detector to further improve the feature discrimination by leveraging the self-attention mechanism.

To the most of our knowledge, there is still no public dataset for MR RS image ship detection. Therefore, we propose a dataset named "Levir-Ship" to promote our research. Levir-Ship contains 3876 images of 512×512 pixels collected from GaoFen-1 and GaoFen-6 satellites with the spatial resolution of $16m/$pixel. We conduct experiments on Levir-Ship and the results show the effectiveness and efficiency of our method.

The contribution of our work can be summarized as follows:

- We propose an effective method (DRENet) for efficient tiny ship detection. A degraded reconstruction enhancer is leveraged to guide the backbone to focus more on the target instead of the background, and the cross-stage multi-head attention is introduced to further improve the discrimination ability of the detector.
- We introduce the Levir-Ship detection dataset, consisting of 3876 GF-1/GF-6 images (each size of 512×512 pixels with the spatial resolution of $16m/$pixel) and more than 3K tiny ship instances.
- Extensive experiments on the Levir-Ship validate the effectiveness and efficiency of the proposed method. Our method achieves 82.4 AP with 85 FPS, and surpasses several state-of-the-art ship detection methods.

The rest of the paper are organized as follows. Related work is introduced in section II. In section III, we give a detailed description of our method. Section IV describes Levir-Ship dataset and the comparison with other ship detection datasets. Experimental results are reported in section V, and the conclusion is drawn in section VI.

II. Related Work

Ship detection has been extensively studied for decades. Generally, traditional methods perform a multi-stage coarse-to-fine process to obtain detection results [28]–[32]. The detection process can often be divided into two steps - region proposal and region refining, and make use of hand-craft features. Although achieving some good results in certain scenarios, these methods are vulnerable to complex situations, together with the limitation of heavy reliance on prior and complex operating procedures.

In recent years, benefiting from the development of hardware and big data, many effective DL object detection methods

[2]https://www.kaggle.com/c/airbus-ship-detection/overview.

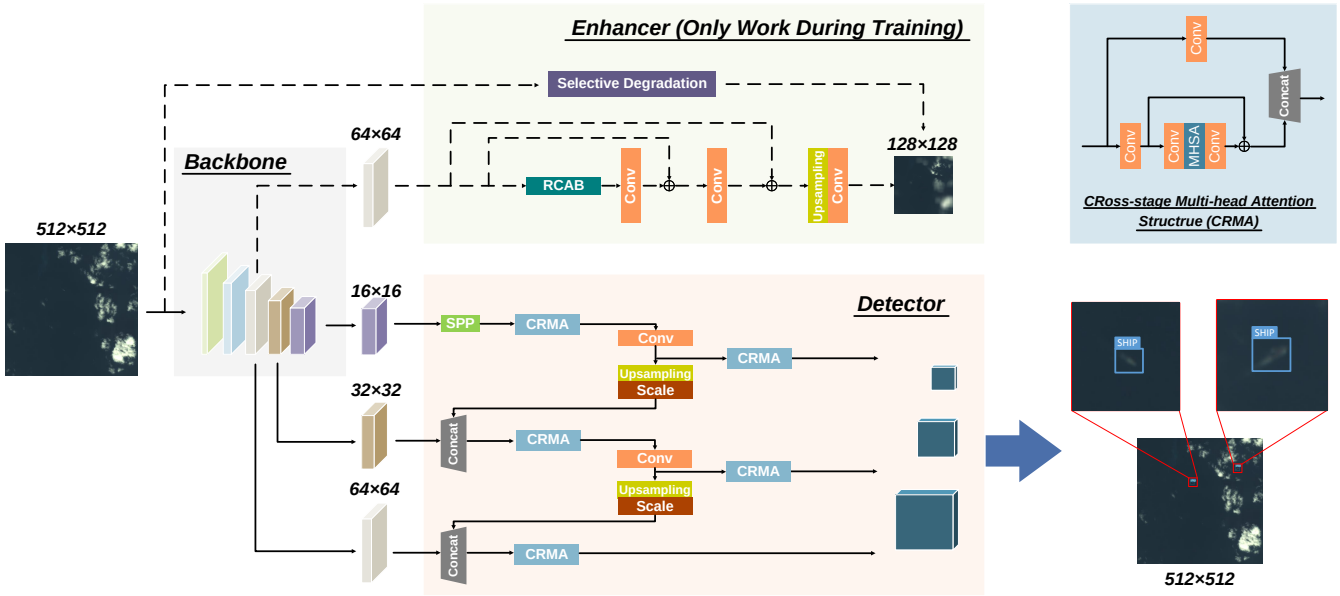


Fig. 2. The overall structure of DRENet. a) Backbone: A lightweight backbone to enable robust features extraction efficiently. b) Enhancer: The proposed degraded reconstruction enhancer to guide the backbone to extract many discriminative features by regressing degraded images. The degraded images come from the proposed "Selective Degradation" operation. c) Detector: The detector to display the final detection results. We introduce into the detector a CRoss-stage Multi-head Attention (CRMA) module, which can gain a large receptive field to locate ships accurately while saving much computation. (For details of RCAB and SPP, please refer to [26] and [27]. The implementations of upsampling layers are aligned with [26] and [27], with nearest-neighbor interpolation in the detector while sub-pixel convolution in the enhancer.)

have been proposed [33]. They are more flexible, unified, and strong to deal with the task, and can be divided into two-stage methods and one-stage methods. Two-stage methods are mainly based on the RCNN series [34]–[37]. They often play the idea of obtaining candidate bounding boxes first, and then classifying and regressing the results from the candidate. Comparing with two-stage methods, one-stage methods headed by Yolo [27], [38]–[41], SSD [42], RetinaNet [43], EfficientDet [44] directly give the detection results. They obtain a higher detection speed expending a little accuracy in general. However, both two-stage and one-stage methods rely on the setting of anchor boxes, which needs much prior knowledge and constrains the methods' generalization. Thus, a series of anchor-free methods are proposed, such as CornerNet [45], CenterNet [46], FCOS [47], etc, which prevent anchor boxes initialization and also achieve accurate detection results.

Motivated by the above approaches, many effective RS image object detection methods have emerged in recent years [48]–[53]. For example, Wang *et al*. [49] proposed a one-stage method, Zhang *et al*. [50] proposed an anchor-free method, and Huang *et al*. [52] proposed an object-adaptation label assignment, which achieve many good results. While more specific to ship target, many methods have made great contributions to ship detection in RS images with DL technology. Li *et al*. [11] proposed a hierarchical selective filtering layer structure based on Faster-RCNN adapting to different scales of ships. Tian *et al*. [12] added atrous spatial pyramid pooling on the basis of Mask-RCNN, and used dense layer jump connections to use feature information in the backbone effectively achieving high detection accuracy. Chen *et al*. [24] introduced dilated convolution and attention mechanism in YOLOv3, and proposed ImYOLOv3. Also, researchers introduce prior information to get high performance. Zhang *et al*. [18] pre-designed the structure of the candidate object according to the ship structure, and then inputted the proposals into CNN to detect the ship. Wu *et al*. [20] used the prominence of the ship's head to detect the ship by positioning the ship's bow. Tang *et al*. [54] extracted the ship by using the difference in hue, saturation, and value between the ship and the surrounding. These methods often focus on HR images and utilize the features extracted by clear edges and rich textures to achieve high performance. However, when tackling the tiny ship detection problem, they may perform badly. In the paper, we propose DRENet to address tiny ship detection.

III. Methodology

The DRENet, shown in Fig. 2, consists of an efficient backbone, a Degraded Reconstruction Enhancer (DRE), and a ship detector. We follow YOLOv5s (a small version of YOLOv5) [27] as the backbone to efficiently extract multi-scale image features. The degraded reconstruction enhancer is designed to guide the backbone to acquire many discriminative features by regressing an object-aware blurred version of the input image in the training phase. The supervision of the enhancer comes from a "Selective Degradation" operation we propose. In the detector, we introduce a CRoss-stage Multi-head Attention (CRMA) module to better locate ships within the input image. Details will be introduced as follows.

Algorithm 1 Selective Degradation for obtaining DRE supervision information

Input: I, the input image
Input: G, all the ground-truth boxes set in I
Define $F(\cdot)$, an increasing function to adjust the mean kernel size
Define $Center(\cdot)$, a function to get the box's central coordinate
Define $Distance(\cdot,\cdot)$, a function to compute the Euclidean Distance between two inputs
Define $EvenInt(\cdot)$, a function to return the nearest even number that is less than the input
Define $Resize(\cdot)$, a function to resize the image by nearest neighbor interpolation
Output: $\hat{I}$, the label of I in DRE

for $i = 0; i < H; i++$ **do**
  for $j = 0; j < W; j++$ **do**
    $k \leftarrow +\infty$
    for each $g \in G$ **do**
      $c \leftarrow Center(g)$
      $t \leftarrow Distance((i,j), c)$
      if $t < k$ **then**
        $k = t$
      end if
    end for
    $k = F(k)$
    $k \leftarrow EvenInt(k)$
    $\hat{I}(i,j) = \frac{1}{(k+1)^2} \sum_{m=i+\frac{k}{2}, n=j+\frac{k}{2}}^{m=i-\frac{k}{2}, n=j-\frac{k}{2}} I(m,n)$
  end for
end for
$\hat{I} = Resize(\hat{I})$
return $\hat{I}$

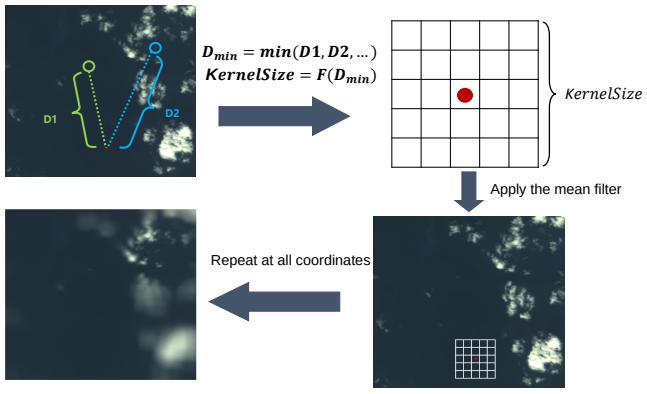


Fig. 3. The specific operation of Selective Degradation: calculate the mean filter kernel size at each point. $D1$ and $D2$ represent the distance from the current pixel position to each ship target, while $F(\cdot)$ is an increasing function to adjust the mean kernel size, where the output, i.e. the mean kernel size will increase as D_{min} increases.

A. Degraded Reconstruction Enhancer

The degraded reconstruction Enhancer (DRE) utilizes the features extracted by the backbone to reconstruct the degraded input image, which could guide the backbone to pay more attention to ships, keep more information about the objects, and omit the complex background.

1) Supervision of DRE: In this part, we propose a method named "Selective Degradation" to generate labels for the enhancer. Selective Degradation, with the nature of filtering, performs mean filters of different sizes on different positions in original images. Details of Selective Degradation are illustrated in Algorithm 1. Fig. 3 depicts how to get the mean kernel size of a certain position in an image. By Selective Degradation, we can get a "pseudo saliency map" in which ships would be kept completely, and the background is processed to be fuzzy. Through the training of DRE, the backbone will pay more attention to the tiny ships instead of the textures, details, or the fractus clouds of the background.

2) Design of DRE: As is shown in Fig. 2, we take feature maps from the backbone as the input of the enhancer. Here we just utilize the feature maps with the shape of 64×64 to reconstruct the degraded label with a shape of 128×128, leveraging the performance and convenience. We refer to the existing super-resolution method [26] to design the structure of DRE, but there are some differences. In [26], as the goal is to improve the quality of the super-resolution image, the network is actually complex and deep. However, instead of reconstructing perfect Selective Degradation labels, our goal is to acquire a discriminative feature representation of the backbone. Therefore, we make a relatively shallow enhancer network, with just a single RCAB (Residual Channel Attention Block, a component module in [26]) used, to concentrate more on the training of the backbone. Based on the mechanism of "Selective Degradation", the labels of the enhancer focus on the ship target and omit the complex background. Thus, the enhancer provides a tendency to obtain robust feature representations in the backbone, and classify ships more exactly.

Compared with the existing super-resolution methods [26], [55], [56], the proposed DRE focuses more on improving the detection results of tiny ships. The supervision of DRE is an object-aware blurred version of the input image (different from the clear image of existing super-resolution methods), which can guide the backbone to pay more attention to ships and be more robust against the disturbances from complex backgrounds. Further, different from the existing super-resolution methods that use labels of higher image resolution, our proposed DRE reconstructs labels of lower image resolution (128×128) than the input image (512×512), which saves much computation and memory. The structure of DRE is also much simpler and shallower than that of the existing super-resolution methods.

B. Efficient Detector

To achieve efficient ship detection, our detector is based on YOLOv5s. We have made some modifications for the tiny ship detection task to further improve efficiency.

In order to spread the rich spatial information of the low-level features to the deep-level features at a short and fast path, YOLOv5 adopts the PAN [57] structure. However, when assigning anchors to each output layer in tiny ship object

detection, most of the bounding boxes fall on the shallow detection layers while few on the deep detection layers. It is actually a time-costing operation but retrieving few benefits to enrich the spatial information of the deep detection layer. In DRENet, we remove the PAN path to improve the detection speed.

Considering different output layers of FPN [58] are assigned with an imbalanced number of training samples, we add two "Scale Layers" behind the two Upsampling layers in the detector (depicted in Fig. 2), following the work of [59]. The Scale Layer is actually just a single learnable parameter that can balance the backpropagated gradients from different output layers in the bottom-up pathway layers of FPN. Supposed there are two layers C1, and C2 in the bottom-up pathway from shallow to deep, and two layers P1, and P2 in the top-down pathway corresponded. In the original FPN, $P1 = C1 + upsample(P2)$, after we add the scale layer, it turns to be $P1 = C1 + \alpha \times upsample(P2)$, where α is a learnable parameter.

To achieve a full-image receptive field, we propose a CRoss-stage Multi-head Attention (CRMA) module, and replace the CSP-like layers in the bottleneck of YOLOv5s. The CSP-like layers in YOLOv5 are designed based on CSPNet [60]. The structure splits the input in half from the channel dimension, then passes the two parts into two branches (one branch with many convolution operations while the other just an identical mapping), and at last concatenates the two branches' outputs as the final output. In our work, we argue the many convolutions in the branch are low cost-effective and replace them with the Multi-Head Self-Attention (MHSA) layer designed in [61], which can enlarge the network's receptive field much while keeping low parameters and complexity, and also more suitable for vision task (with relative positional encodings used and Q, K, V in 2D style) compared with the original version in [62]. By fusing CSPNet's idea and the MHSA layer, we construct the CRMA module. More details of CRMA are shown in Fig. 2 and TABLE I. Also note that we replace all the five CSP-like layers with our CRMA in YOLOv5s's bottleneck, which can gain much reduction of computation and complexity of the network while keeping good performance.

TABLE I
DETAILS OF CRMA CONFIGURATION. CBS REPRESENTS OPERATIONS OF CONVOLUTION, BATCH NORMALIZATION, AND SILU ACTIVATION FUNCTION. MHSA DENOTES THE MULTI-HEAD SELF-ATTENTION LAYER.

Input: 16×16×512				
Pathway	Layer	Filters	Size	Output
1	CBS	256	1×1	16×16×256
	CBS	128	1×1	16×16×128
	MHSA	—	—	16×16×128
	CBS	256	1×1	16×16×256
	Residual	—	—	16×16×256
2	CBS	256	1×1	16×16×256
Output	Concat	—	—	16×16×512

C. Loss Function

The loss function of DRENet is composed of two parts, including the enhancer loss and the detector loss.

1) Enhancer Loss: We apply mean square error (MSE) to form the reconstruction error of the enhancer. The loss function is formulated as follows:

$$Loss_{enhancer} = \frac{\sum(y^* - y)^2}{N} \tag{1}$$

where y^* represents the pixel value of the reconstruction result, y represents the pixel value of the "Selective Degradation" label, and N represents the number of total pixels.

2) Detector Loss: The detector loss function contains regression loss and classification loss. We conduct typical loss in object detection task - CIoULoss [63] for regression error, and Binary Cross-Entropy Loss for classification. In DRENet, they are defined as:

$$loss_{reg} = 1 - CIoU \tag{2}$$

$$CIoU = IOU - \frac{\rho^2(b, b^{gt})}{c^2} - \frac{\nu^2}{1 - IOU + \nu} \tag{3}$$

$$\nu = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{4}$$

where b and b^{gt} denote the central points of predicted box and target box, $\rho(\cdot)$ denotes the Euclidean distance, c is the diagonal length of the smallest enclosing box covering the two boxes, ν measures the consistency of aspect ratio, and w^{gt}, h^{gt}, w, h respectively represent the width and height of two boxes.

$$loss_{cls} = -\frac{\sum(y_n \times \ln y_n^* + (1 - y_n) \times ln(1 - y_n^*))}{n} \tag{5}$$

where n represents the number of all anchor boxes, y_n^* denotes the probability of the n-th predicted bounding box. y_n is the label corresponding to the predicted bounding box, which is formulated by $CIoU$. Finally, we design the detector loss with a balanced factor of 0.05.

$$Loss_{detector} = loss_{reg} + 0.05 \times loss_{cls} \tag{6}$$

3) DRENet Loss: The training of the enhancer and detector jointly may occur unbalance between the two branches, thus leading to sub-optimal. How to weigh the two branches has a great impact on the performance of ship detection. Motivated by the work of Kendall *et al*. [64], we design a method to balance the training preference of the enhancer and detector automatically, which is achieved by two learnable weight coefficients. Finally, the DRENet loss is defined as:

$$Loss = \frac{1}{2a^2}Loss_{enhancer} + \frac{1}{2b^2}Loss_{detector} + \ln a + \ln b \tag{7}$$

where a and b are two coefficients automatically learned by the network. In training, a and b in the first two items of $Loss$ tend to be large values, while in the last two, they tend to be small. They are regulated by each other. This design can help the training progress smooth and steady.

D. Implementation Details

Our backbone follows the design of YOLOv5s. There are 5 stages in the backbone, each with downsampling by 2, thus we have different feature maps, 256×256, 128×128, 64×64, 32×32 and 16×16. Taking the features of 64×64, 32×32, and 16×16 as detector's input, we obtain the probability, position, and shape of the ship by inference. As for the enhancer, we take the 64×64 feature maps for reconstructing a degraded label with a shape of 128×128.

We adopt Stochastic Gradient Descent (SGD) optimizer with 0.99 momentum and 0.0005 weight decay. The learning rate is initially set to 0.01 and decays in a Cosine annealing strategy until 500 epochs. The batch size is 16. While for the optimization of the two weight coefficients a and b in loss, we adopt Adaptive moment estimation (Adam) optimizer with learning rate set to 0.01. Notably, the enhancer and detector are jointly optimized in the training phase, but in the testing phase, the enhancer can be removed and only the detector is used.

IV. LEVIR-SHIP

In recent decades, there are many effective object detection datasets in optical remote sensing.

NWPU-VHR-10 [65] contains 800 images in 10 categories. However, there are only 57 images including total 302 ships, which is not enough to train a DL network and suffers from overfitting. In addition, the spatial resolution of the dataset is 0.5-2m.

HRSC2016 [66] is a single-class dataset for ship detection. The images are all from 6 well-known ports. There are 1070 images, and 2976 objects in total. The spatial resolution of the dataset is 0.4-2m.

DIOR [67] is a relatively large dataset proposed in recent years. It contains 20 categories, 23463 images, and 192472 instances. The ship category has 2702 images and 62400 objects in total. However, the ship is also under a high spatial resolution.

HRRSD [68] is another large-scale RS dataset recently proposed. There are 13 categories, 21761 images, and 55740 instances. It contains 2165 images with 3886 ship objects, but the resolution of 0.15-1.2m is still too high.

Most ship detection datasets focus on high spatial resolution, such as 0.3m, 0.5m, and 2m [65]–[68]. Few datasets concentrate on tiny ship objects with a lower spatial resolution. In our work, we propose a tiny ship detection dataset named Levir-Ship under the spatial resolution of 16m.

Images in Levir-Ship are captured from multispectral cameras of GaoFen-1 and GaoFen-6 satellites. We only use the R, G, and B bands. 85 scenes have been collected in the dataset with pixel resolutions between 10000×10000 and 50000×20000. We crop the original images to finally get 1973 positive samples and 1903 negative samples with the size of 512×512. As shown in Fig. 4, we plot the distribution of width/height of ships in the Levir-Ship dataset. We can observe that the ship pixel size in Levir-Ship is almost below 20×20, and centralizes at around 10×10. The ship in Levir-Ship is relatively small compared to the vast background, which brings many challenges to the detection network.

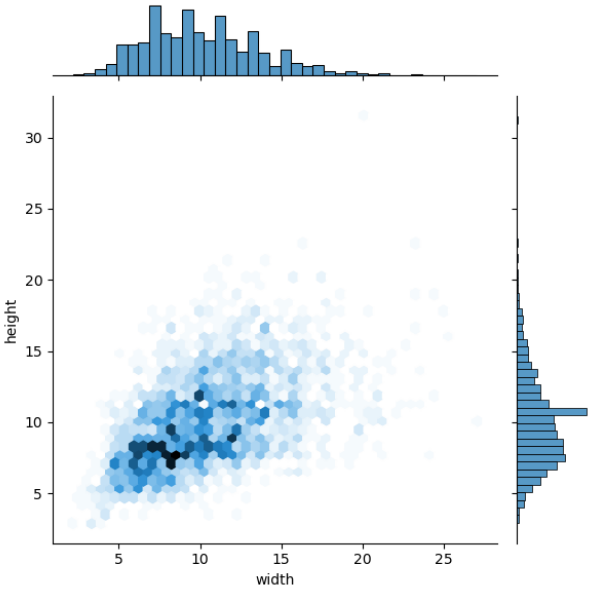


Fig. 4. Ship target size statistics in LEVIR-SHIP. We use the smallest enclosing rectangle of the actual ship target as a measure.

TABLE II shows the comparison between Levir-Ship and several existing RS object detection datasets. The ship in Levir-Ship is surrounded by complex background, and is hard to recognize, compared with those ships in HR images. In addition, due to the influence of different time, different photographers, and different locations, the images show different conditions, which brings many challenges to the task.

From Fig. 5, we can see that the background is complex, and ships in the images are hard to find even zoomed in several times. Meanwhile, since Levir-Ship covers a variety of environmental conditions, DL network can obtain better generalization, stability, and sufficiency.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conduct experimental comparative analyses of the proposed method. Our experiment is based on Levir-Ship, and we randomly divide the dataset into training set, validation set, and test set (details shown in TABLE III). All experiments are implemented on an NVIDIA Tesla V100 GPU.

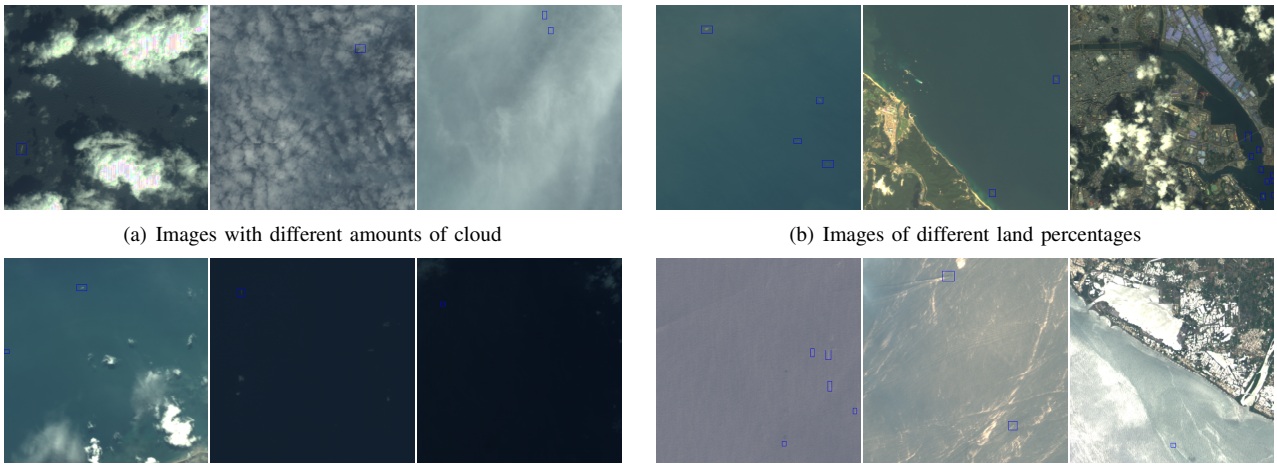
A. Evaluation Metrics

We evaluate the effectiveness and efficiency of the proposed method by detection accuracy and model complexity. The metrics are Average Precision (AP), Floating Point Operations (FLOPs), Model Parameters, and Frames Per Second (FPS). Details are described as follows.

To evaluate detection accuracy, We use AP_{50} (prediction is a true positive sample when the IoU is larger than 0.5) as the evaluation metric to measure the detection accuracy of the model. In calculating AP value, we conduct the method illustrated in the COCO dataset (in the Precision-Recall curve, according to classification confidence, 101 points are sampled

TABLE II
COMPARISON BETWEEN THE PROPOSED LEVIR-SHIP DATASET AND FOUR PUBLICLY AVAILABLE OBJECT DETECTION DATASETS. BB NUMBER DENOTES THE NUMBER OF BOUNDING BOXES. FOR DATASETS CONTAINING MULTIPLE CATEGORIES, ONLY INFORMATION ABOUT THE SHIP CATEGORY IS LISTED HERE.

Dataset	Images Number	BB Number	Source	Resolution	Year
NWPU VHR-10	57	302	Google Earth	0.5-2m	2014
HRSC2016	1070	2976	Google Earth	0.4-2m	2016
DIOR	2702	62400	Google Earth	0.5-30m	2018
HRRSD	2165	3886	Google Earth & Baidu Earth	0.15-1.2m	2019
LEVIR-SHIP (Ours)	3876	3219	GaoFen-1 & GaoFen-6	16m	2021



(a) Images with different amounts of cloud

(b) Images of different land percentages

(c) Images of different light intensity

(d) Images of different sea area characteristics

Fig. 5. A partial display of the data contained in LEVIR-SHIP under different situations that may be encountered in the process of ship detection. The blue bounding boxes are the ground truth.

TABLE III
THE DIVISION OF TRAINING SET, VALIDATION SET, AND TEST SET

	Image number	Target number
Training set	2320	2002
Validation set	788	665
Test set	788	552

at equal intervals, and then the Precision values corresponding to these points are accumulated and averaged to obtain the final AP value) to ensure that the calculation of the Precision-Recall curve area is more accurate. We use FLOPs, Parameters, and FPS to reflect the detection speed.

B. State-of-the-art Comparison

We compare DRENet with other state-of-the-art networks in ship detection on Levir-Ship dataset, including YOLOv3 [40], YOLOv5s [27], Retinanet [43], SSD [42], FasterRCNN [36], EfficientDet [44], FCOS [47], CenterNet [46], HSFNet [11], ImYOLOv3 [24], DFR and RFE structure proposed by Tian *et al*. [12]. The last three methods are proposed for ship detection task. We have also tried to compare with some traditional methods such as SVDNet [28] and PCANet [69] (configured as [28]), but unfortunately found these methods failed as the ship targets only cover a small area of the image. Unless otherwise specified, we train all networks for 500 epochs, using batch size 16.

TABLE IV shows the comparative results of networks. It can be seen that compared with other methods, our proposed method achieves the highest detection accuracy and real-time detection speed. DRENet achieves a healthy 1.5 point AP gap with the closest competitor, EfficientDet-D2 [44], while being much faster (85 vs.21 FPS). Compared with the fastest YOLOv5s [27], DRENet is a minor slower but much more accurate (82.4 vs.75.6 AP). Furthermore, the Params and FLOPs of DRENet are small. These can demonstrate the excellent detection performance of DRENet on the tiny ship detection task.

Fig. 6 gives some detection results between DRENet and other methods. Blue boxes represent true positive detections. Red boxes denote false negative detections. Yellow boxes indicate false positive samples. From the top to the bottom in the figure, the complexity of the background gradually rises. The ship in the image behaves very small, and the distinguishable features are not clear enough. Meanwhile, some ships are surrounded by fractus clouds, leading to a challenge in detection. We can observe that our method achieves better detection

TABLE IV
COMPARISON WITH SOME WELL-KNOWN DETECTION METHODS AND STATE-OF-THE-ART SHIP DETECTION METHODS. THE BOLD REPRESENTS THE OPTIMAL METRIC.

Methods	Params(Inference)	FLOPs(Inference)	AP	FPS
YOLOv3	61.52M	99.2G	69.9	61
YOLOv5s	7.05M	10.4G	75.6	**95**
Retinanet (ResNet50)	36.33M	104.4G	74.9	12
SSD (VGG16)	24.39M	175.2G	52.6	25
FasterRCNN (VGG16)	136.70M	299.2G	70.8	10
EfficientDet-D0	**3.84M**	**4.6G**	71.3	32
EfficientDet-D2	8.01M	20.0G	80.9	21
FCOS (ResNet50)	5.92M	51.8G	75.5	37
CenterNet (Hourglass-104)	191.24M	584.6G	77.7	25
HSFNet	157.59M	538.1G	73.6	7
ImYOLOv3	62.86M	101.9G	72.6	51
MaskRCNN (ResNet50)+DFR+RFE	24.99M	237.8G	76.2	6
DRENet (ours)	4.79M	8.3G	**82.4**	85

results than other methods. DRENet can detect more true positive ships and miss fewer targets. Additionally, the false positive detections can be well avoided, compared with other method detection results shown in the figure. For example, as for 3rd-row images, no methods mentioned can find the ship behind the cloud except for DRENet. The comparable results show DRENet has higher applicability and robustness for the tiny ship object detection task.

To further analyze the performance of our method under different background situations, we roughly divide the test set into further five subsets: "clam sea", "thin cloud", "thick cloud", "strong wave" and "fractus cloud", according to different background situations. To better illustrate the difference between these situations, for each subset we select an example image, displayed in Fig. 7. TABLE V shows the comparisons of DRENet with other methods in different situations, from which we can observe that the subsets "thick cloud", "strong wave" and "fractus cloud" are more difficult than the other two, as the performance of other methods has dropped a lot in these scenes. From the results, we can see that our method is more robust to complex backgrounds like "fractus cloud" than other methods while also keeping good performance in the simple scene like "calm sea".

C. Controlled Experiment

1) Degradation Function F: We compare the degradation effect on the original input image when taking different function $F(\cdot)$ to adjust the mean kernel size, including linear function, logarithmic function, and exponential function. As shown in Fig. 8, it can be seen that the exponential function conducts a more obvious degradation effect and can better meet the need of the enhancer to highlight the ship and blur the background. In order to keep the details of the ship and minimize the effect of the background, we apply the exponential function to calculate the size of the mean filter kernel at each pixel position. The steps are given in Algorithm 1.

2) Design of Enhancer : The degraded reconstruction enhancer is used for extracting more distinguishable basic feature representation between ship target and background. We conduct experiments on the different structures and supervisions of the enhancer.

Different Enhancer Structure. Here, we make comparisons on different inner structures of the enhancer to select the most suitable enhancer structure for the ship detection task. We evaluate the performance by AP and network complexity. TABLE VI shows the results in detail. "n-RCAB" represents the number of RCAB included in our enhancer network. "Upsampling" denotes different upsampling multiples, corresponding to reconstructing Enhancer labels of different resolutions. It can be seen from the results that when 1 RCAB structure and upsampling 2 times are carried out, the performance is the best. When the enhancer becomes more complex, the training of the enhancer focus more on high-level layers, instead of the basic layers of the backbone. This may illustrate why the performance declines as the structure becomes weighted. Meanwhile, when the upsampling factor gets larger, reconstruction of the degraded image becomes more difficult, which leads to the deterioration of the enhancer. Then the trained enhancer cannot achieve the proper feature representation the detector needs.

Different Supervision Form. We compare different label generation strategies in the enhancer to acquire better supervision for the reconstruction branch. As shown in TABLE VII, different generation strategies have different impacts on detection accuracy. The enhancer structure we take is 1 RCAB and upsampling 2 times. From the table, we can make some observations. First, direct down-sampling can improve the detection performance compared with only the detector without the enhancer. This may be because the backbone retains more spatial information by fitting the label, which

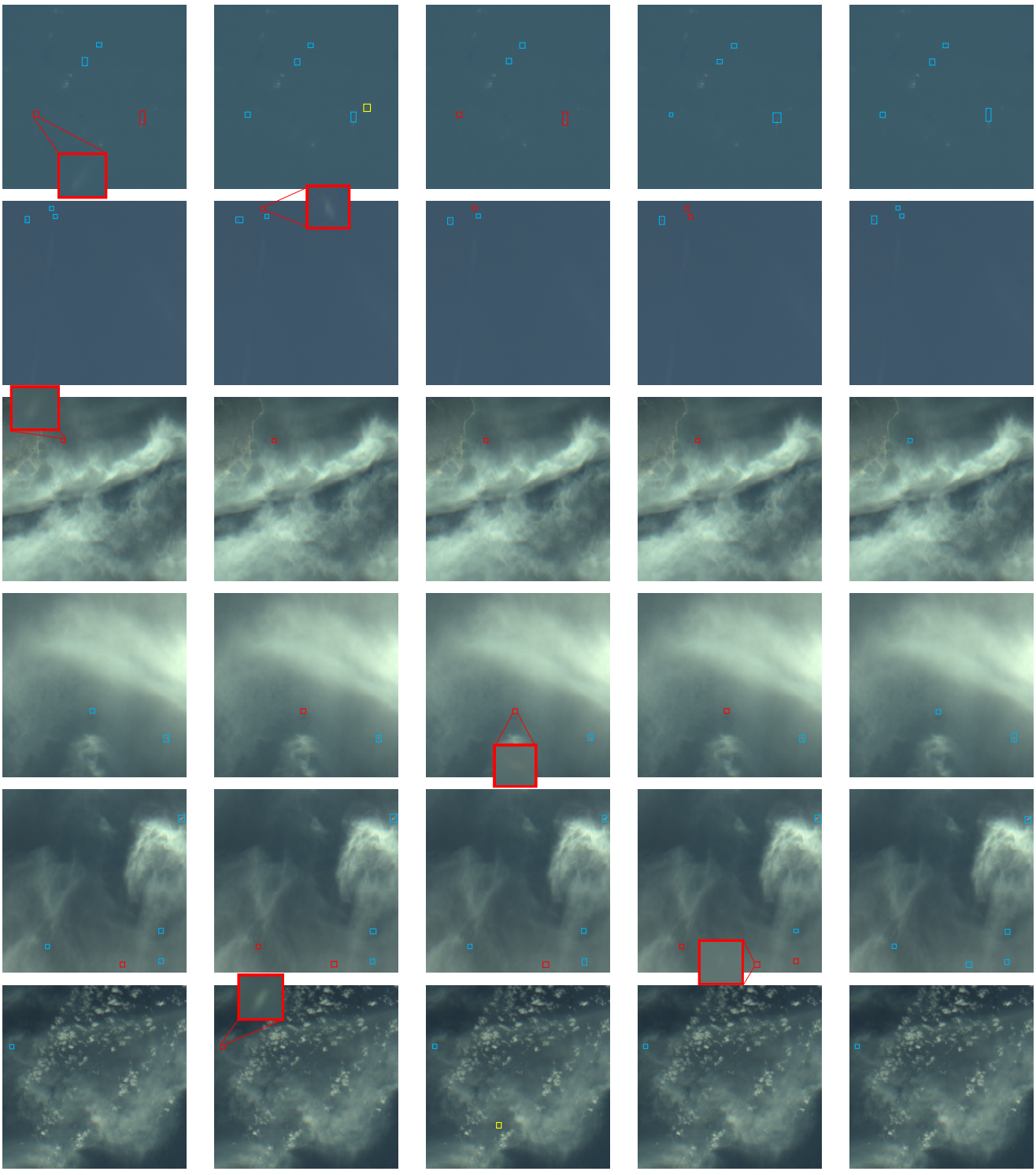
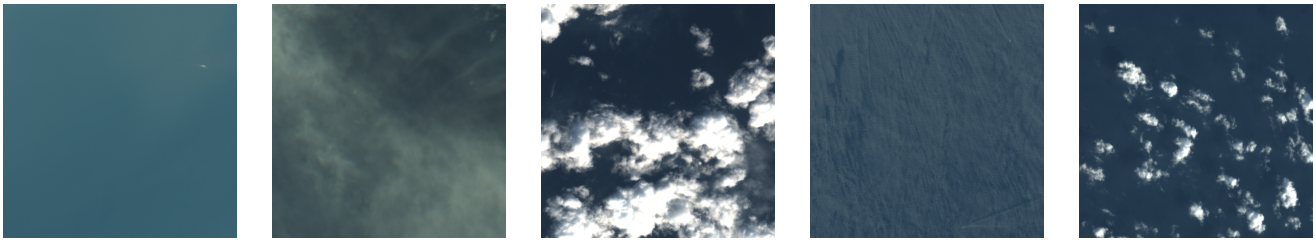


Fig. 6. Comparisons of the detection results by different methods. (a) YOLOv5s. (b) EfficientDet-D2. (c) ImYOLOv3. (d) Method [12]. (e) Ours. The blue box represents the real target detected, the red box represents the missed target and the yellow box represents false alarms. To facilitate observation, we randomly enlarge some missed ship targets.

makes more contributions to the detection of tiny ships. Second, we can also see the sudden-blur method reduces the detection accuracy, which is mainly because a ship in Levir-Ship dataset only occupies a small area, and the training of the network will suffer from a relatively small loss. Thus, the network learning deviates from the original direction and cannot focus on the details of the tiny ship while suppressing the background. Third, by using "Selective Degradation", i.e. the continuity-blur in TABLE VII to generate the label, the detection performance has been significantly improved.



(a) (b) (c) (d) (e)

Fig. 7. Images as examples for each background situation. (a) calm sea. (b) thin cloud. (c) thick cloud. (d) strong wave. (e) fractus cloud.

TABLE V
THE COMPARISONS OF DRENET WITH OTHER METHODS IN DIFFERENT SITUATIONS, WHERE "NUMBER" DENOTES IMAGES' NUMBER OF EACH SITUATION. THE HIGHEST AP VALUE IS MARKED IN BOLD.

Situations	Number	YOLOv5s	EfficientDet-D2	ImYOLOv3	Method [12]	DRENet (ours)
calm sea	262	76.8	**83.0**	75.9	78.4	82.1
thin cloud	238	84.3	83.9	83.7	82.4	**87.3**
thick cloud	60	60.5	78.9	56.6	70.1	**86.8**
strong wave	101	73.4	73.3	62.8	71.8	**82.8**
fractus cloud	127	72.1	74.7	61.8	64.4	**76.5**

TABLE VI
COMPARISON OF DIFFERENT NETWORK COMPLEXITY IN THE ENHANCER. "TRAIN" REPRESENTS THE TRAINING PROCESS WHERE WE KEEP THE DEGRADED RECONSTRUCTION ENHANCER. THE HIGHEST AP VALUE IS MARKED IN BOLD.

Structure	n-RCAB	Upsampling	AP	Params(Train)	FLOPs(Train)
YOLOv5s	—	—	75.6	7.05M	10.4G
YOLOv5s+Degraded reconstruction enhancer	1	×2	**76.8**	8.25M	20.3G
	1	×4	76.2	8.84M	40.0G
	2	×2	76.6	8.55M	22.7G
	2	×8	75.6	9.73M	121.2G

TABLE VII
COMPARISON OF DIFFERENT LABELS IN THE ENHANCER. "SUDDEN-BLUR" MEANS TO UNIFORMLY BLUR THE AREA OUTSIDE THE TARGET WITHOUT OPERATING ON THE AREA AROUND THE TARGET, WHILE "CONTINUITY-BLUR" REPRESENTS THE SELECTIVE DEGRADATION WITH THE EXPONENTIAL FUNCTION APPLIED AS $F(\cdot)$. THE HIGHEST AP VALUE IS MARKED IN BOLD.

Label generation method	AP
Original (YOLOv5s)	75.6
Down-sampling	76.6
Sudden-blur+Down-sampling	75.2
Continuity-blur(λ=1.01)+Down-sampling	76.8
Continuity-blur(λ=1.03)+Down-sampling	**78.7**
Continuity-blur(λ=1.05)+Down-sampling	77.1

In addition, we take different parameters λ in degradation function F, and finally select 1.03 because of its relatively great performance.

3) Structure of Detector: We have made some improvements to the original YOLOv5s prototype to achieve better performance in tiny ship detection. To verify the effect of different modifications, we conduct a series of ablation study on the detector. Results are shown in TABLE VIII. We can get some observations from the table. First, by adding "Scale Layer", a minor improvement appears without increasing the complexity. It may be because the Scale Layer alleviates the problem of unbalanced training of different layers in FPN to a certain extent. Second, after replacing the CSP bottleneck with the CRMA module we propose, the complexity is significantly reduced, and the detection AP has risen by 4.2. We believe this is due to the greater receptive field which leads to richer features that benefit tiny ship object detection. It verifies the effectiveness of the designed CRMA structure. We further make a trade-off of whether to remove PAN. From TABLE IX, we can observe the number of network parameters and operations be greatly reduced at the little expense of detection accuracy when the PAN structure is removed. Thus, we remove PAN in our final model.

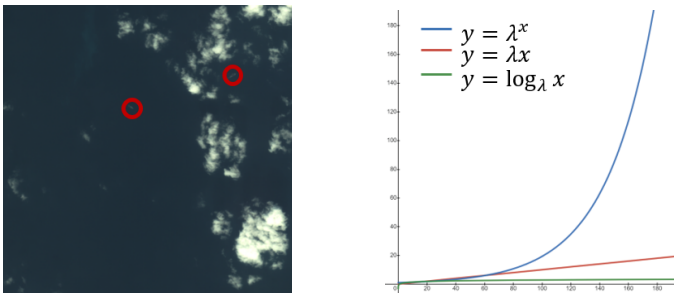
Compared with the YOLOv5s prototype, the amount of parameters and FLOPs has been reduced by nearly -32% and -19% respectively, while AP has increased 4.7 points. With faster speed, smaller model, and higher accuracy, the improved detector is more suitable for detecting ships in MR RS images.

4) Network Overall Setting: In this part, we will conduct experiments on loss function and training strategy.

Loss Function. DRENet is composed of an enhancer and

TABLE VIII
THE ABLATION STUDY OF THE MODIFICATIONS IN THE DETECTOR. THE OPTIMAL VALUES ARE MARKED IN BOLD.

Network description	Scale	CRMA	AP	Params	FLOPs
Baseline (YOLOv5s)	✗	✗	75.6	7.05M	10.4G
Add Scale	✓	✗	76.3	7.05M	10.4G
Replace CSP bottleneck with CRMA	✗	✓	79.8	5.69M	9.1G
Add Scale and CRMA	✓	✓	**80.8**	**5.69M**	**9.1G**



(a) Origin images (b) Different function $F(\cdot)$

(c) $y = log_\lambda x$ (d) $y = \lambda x$ (e) $y = \lambda^x$

Fig. 8. The degradation effect when takes a logarithmic function, linear function, and exponential function to calculate the mean filter kernel size at each position. The red circles in the images are the ships. λ is a constant of functions, which is set to ensure the ship areas are kept clear . x is the shortest distance to ship targets and y denotes the kernel size.

TABLE IX
THE TRADE-OFF BETWEEN KEEPING PAN AND REMOVING PAN. W DENOTES KEEPING, WHILE W/O DENOTES REMOVING.

	AP	Params	FLOPs
w	**80.8**	5.69M	9.1G
w/o	80.3	**4.79M**	**8.3G**

detector, and the two parts need to be trained jointly. The total loss contains the supervision of the detection label and degraded reconstruction label. Including the automatically balanced learning method mentioned in DRENet Loss (see Section III-C3), we also design a fixed balanced factor to verify the performance. The formula with a fixed balanced factor is indicated as:

$$loss = loss_{detector} + \alpha \times loss_{enhancer} \tag{8}$$

Where α is a fixed factor to balance the detector loss and enhancer loss. Comparative results are shown in TABLE X. It indicates that the automatic weight learning method can bring a significant improvement. Meanwhile, in the method with a fixed balanced factor, as the factor α gets larger from 0.1 to 10, AP also increases, which again reveals the effectiveness of the enhancer.

TABLE X
COMPARISON OF DIFFERENT CONFIGURATION METHODS OF THE LOSS FUNCTION. α DENOTES THE WEIGHT OF THE SUM OF THE TWO BRANCH LOSS FUNCTIONS. THE HIGHEST AP VALUE IS MARKED IN BOLD.

Combination description	weight α	AP
Fixed weight summation	100	80.9
	10	81.6
	1	81
	0.1	80.6
Automatic weight summation	—	**82.4**

TABLE XI
COMPARISON OF DIFFERENT TRAINING METHODS. PRE-TRAINING REFERS TO TRAINING THE ENHANCER FIRST, AND THEN JOINT TRAINING. THE HIGHEST AP VALUE IS MARKED IN BOLD.

Training method	Pretrained epoches	AP
Joint train without pre-training	—	**82.4**
Joint train with pre-training	100	81.9
	300	81.4
	500	80.5
	1000	77.5

TABLE XII
GENERALIZATION OF DRENET ON THE DOWNSAMPLED VERSION OF EXISTING HR DATASET.

Method	Enhancer	Efficient Detector	AP
YOLOv5s	-	-	51.5
DRENet	+	-	52.8
	-	+	50.9
	+	+	51.6

Training Strategy. TABLE XI shows the results of different training methods. Our goal is to seek better performance on ship detection, and the enhancer is just an auxiliary branch. So we try to pre-train the enhancer and then perform joint training. However, the experimental result shows that the performance of pre-training is worse than that of direct joint training, and as the pre-training epoch increases, AP gradually

declines. We believe that this phenomenon is mainly because the two branches complement each other. Since the enhancer can be regarded as a target location plus target segmentation, it is more difficult to directly train the enhancer, and network learning may appear to be deteriorated and divergent. While the detector can provide targets' position guidance for the enhancer, so that it can have a better learning effect. Therefore, the pre-training of the enhancer will result in poor detection accuracy, and will intensify with the number of epochs.

5) Generalization on Existing High-Resolution Dataset: We also analyze the generalization of our method on the existing HR dataset, that is, implementing experiments on the downsampled version of HR images.

Here we choose the HRRSD dataset to implement the experiments. Considering the size and resolution (0.15-1.2m) of the images in HRRSD, we first reshape all the images into 64×64, and then do ablation experiments to reveal the performance of DRENet.

From the results in TABLE XII, it appears that when only our efficient detector applied, the AP value drops a bit (-0.6), and we argue it is mainly caused by the size of the input. As the image size is only 64×64, after being processed by the network's backbone, the features consumed by CRMA will be at an even smaller size, thus the large-receptive-field characteristic of CRMA cannot gain more advantages than the convolution operations in YOLOv5s's CSP-like structure. Therefore, as we remove PAN structure in our efficient detector, the performance then deteriorates, which is also verified in TABLE IX.

While, we can also observe from the results that our enhancer continues displaying a big boost. When we only apply the enhancer, AP increases +1.3 points, and when we combine both the enhancer and detector, our DRENet achieves a near performance (AP +0.1) compared with baseline YOLOv5s. These results can further demonstrate the effectiveness of our enhancer. Considering the computation and complexity saved in our DRENet (demonstrated in TABLE IV), DRENet actually achieves good performance even in the downsampled version of the existing high-resolution dataset.

VI. Conclusion

In this paper, we propose a tiny ship detection dataset - "Levir-Ship", and an efficient tiny ship detection network - DRENet towards the real-world tiny ship detection task. Levir-Ship contains 3876 images from GaoFen-1 and GaoFen-6 satellites with multispectral cameras of 16m spatial resolution. DRENet performs two parts, including an enhancer to obtain more distinguishable features in the backbone, and a detector to carry efficient detection on tiny ships. The enhancer is inference-free, so it would not introduce extra inference cost on application. Experiments on Levir-Ship verify that our proposed DRENet has higher detection accuracy and ensures real-time detection speed at the same time compared with other state-of-the-art detection methods.

References

[1] H. He, Y. Lin, F. Chen, H.-M. Tai, and Z. Yin, "Inshore ship detection in remote sensing images via weighted pose voting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3091–3107, 2017.

[2] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2015.

[3] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1665–1669, 2017.

[4] F. Yang, Q. Xu, and B. Li, "Ship detection from optical satellite images based on saliency segmentation and structure-lbp feature," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 602–606, 2017.

[5] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sensing*, vol. 13, no. 21, p. 4441, 2021.

[6] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[7] J. Xu, X. Sun, D. Zhang, and K. Fu, "Automatic detection of inshore ships in high-resolution remote sensing images using robust invariant generalized hough transform," *IEEE geoscience and remote sensing letters*, vol. 11, no. 12, pp. 2070–2074, 2014.

[8] S. Li, Z. Zhou, B. Wang, and F. Wu, "A novel inshore ship detection via ship head classification and body boundary determination," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1920–1924, 2016.

[9] S. Zhang, R. Wu, K. Xu, J. Wang, and W. Sun, "R-cnn-based ship detection from high resolution remote sensing imagery," *Remote Sensing*, vol. 11, no. 6, p. 631, 2019.

[10] L. He, S. Yi, X. Mu, and L. Zhang, "Ship detection method based on gabor filter and fast rcnn model in satellite images of sea," in *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, 2019, pp. 1–7.

[11] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7147–7161, 2018.

[12] L. Tian, Y. Cao, B. He, Y. Zhang, C. He, and D. Li, "Image enhancement driven by object characteristics and dense feature reuse network for ship target detection in remote sensing imagery," *Remote Sensing*, vol. 13, no. 7, p. 1327, 2021.

[13] A. Van Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," *arXiv preprint arXiv:1805.09512*, 2018.

[14] Y. Wang, L. Wang, Y. Jiang, and T. Li, "Detection of self-build data set based on yolov4 network," in *2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*. IEEE, 2020, pp. 640–642.

[15] Y. Wang, C. Wang, and H. Zhang, "Combining a single shot multibox detector with transfer learning for ship detection using sentinel-1 sar images," *Remote sensing letters*, vol. 9, no. 8, pp. 780–788, 2018.

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

[17] Y. Gui, X. Li, and L. Xue, "A multilayer fusion light-head detector for sar ship detection," *Sensors*, vol. 19, no. 5, p. 1124, 2019.

[18] R. Zhang, J. Yao, K. Zhang, C. Feng, and J. Zhang, "S-cnn-based ship detection from high-resolution remote sensing images." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 41, 2016.

[19] L. Li, Z. Zhou, B. Wang, L. Miao, and H. Zong, "A novel cnn-based method for accurate ship detection in hr optical remote sensing images via rotated bounding box," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 686–699, 2020.

[20] F. Wu, Z. Zhou, B. Wang, and J. Ma, "Inshore ship detection based on convolutional neural network in optical satellite images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 4005–4015, 2018.

[21] W. Liu, L. Ma, and H. Chen, "Arbitrary-oriented ship detection framework in optical remote-sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 6, pp. 937–941, 2018.

[22] Y. You, Z. Li, B. Ran, J. Cao, S. Lv, and F. Liu, "Broad area target search system for ship detection via deep convolutional neural network," *Remote Sensing*, vol. 11, no. 17, p. 1965, 2019.

[23] Z. Song, H. Sui, and L. Hua, "A hierarchical object detection method in large-scale optical remote sensing satellite imagery using saliency detection and cnn," *International Journal of Remote*

Sensing, vol. 42, no. 8, pp. 2827–2847, 2021. [Online]. Available: https://doi.org/10.1080/01431161.2020.1826059

[24] L. Chen, W. Shi, and D. Deng, "Improved yolov3 based on attention mechanism for fast and accurate ship detection in optical remote sensing images," *Remote Sensing*, vol. 13, no. 4, p. 660, 2021.

[25] L. Chen, W. Shi, C. Fan, L. Zou, and D. Deng, "A novel coarse-to-fine method of ship detection in optical remote sensing images based on a deep residual dense network," *Remote Sensing*, vol. 12, no. 19, p. 3115, 2020.

[26] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[27] G. Jocher, A. Stoken, J. Borovec, NanoCode012, A. Chaurasia, TaoXie, L. Changyu, A. V, Laughing, tkianai, yxNONG, A. Hogan, lorenzomammana, AlexWang1900, J. Hajek, L. Diaconu, Marc, Y. Kwon, oleg, wanghaoyang0106, Y. Defretin, A. Lohia, ml5ah, B. Milanko, B. Fineran, D. Khromov, D. Yiwei, Doug, Durgesh, and F. Ingham, "ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," Apr. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4679653

[28] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with svd networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5832–5845, 2016.

[29] G. Mattyus, "Near real-time automatic vessel detection on optical satellite images," in *ISPRS Hannover Workshop*. ISPRS Archives, 2013, pp. 233–237.

[30] J. Han, P. Zhou, D. Zhang, G. Cheng, L. Guo, Z. Liu, S. Bu, and J. Wu, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 89, pp. 37–48, 2014.

[31] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Transactions on geoscience and remote sensing*, vol. 48, no. 9, pp. 3446–3456, 2010.

[32] Z. Shi, X. Yu, Z. Jiang, and B. Li, "Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 4511–4523, 2013.

[33] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[35] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[37] 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.

[38] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[39] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[40] Redmon, Joseph and Farhadi, Ali, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[41] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[42] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[43] 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.

[44] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[45] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

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

[47] 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.

[48] Z. Zou and Z. Shi, "Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1100–1111, 2017.

[49] G. Wang, Y. Zhuang, H. Chen, X. Liu, T. Zhang, L. Li, S. Dong, and Q. Sang, "Fsod-net: Full-scale object detection from optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[50] T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, "Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[51] Z. Huang, W. Li, X.-G. Xia, X. Wu, Z. Cai, and R. Tao, "A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2021.

[52] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A general gaussian heatmap label assignment for arbitrary-oriented object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 1895–1910, 2022.

[53] Z. Huang, W. Li, X.-G. Xia, H. Wang, F. Jie, and R. Tao, "Lo-det: Lightweight oriented object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[54] G. Tang, S. Liu, I. Fujino, C. Claramunt, Y. Wang, and S. Men, "H-yolo: a single-shot ship detection approach based on region of interest preselected network," *Remote Sensing*, vol. 12, no. 24, p. 4192, 2020.

[55] Lei, Sen and Shi, Zhenwei and Zou, Zhengxia, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1243–1247, 2017.

[56] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3633–3643, 2019.

[57] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[58] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[59] Y. Gong, X. Yu, Y. Ding, X. Peng, J. Zhao, and Z. Han, "Effective fusion factor in fpn for tiny object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1160–1168.

[60] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[61] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 519–16 529.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[63] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 993–13 000, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6999

[64] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[65] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 98, pp. 119–132, 2014.

[66] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines,"

in *International conference on pattern recognition applications and methods*, vol. 2. SCITEPRESS, 2017, pp. 324–331.

[67] 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 Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[68] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5535–5548, 2019.

[69] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

Jianqi Chen received his B.S. degree from the Image Processing Center, School of Astronautics, Beihang University in 2021. He is currently pursuing his M.S. degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include deep learning, object detection and artificial intelligence safety.

Keyan Chen received the B.S. degree from the School of Astronautics, Beihang University, Beijing, China, in 2019. He is currently working toward the M.S. degree in the Image Processing Center, School of Astronautics, Beihang University. His research interests include image processing, machine learning and pattern recognition.

Hao Chen received his B.S. degree from the Image Processing Center School of Astronautics, Beihang University in 2017. He is currently pursuing his doctorate degree in the Image Processing Center, School of Astronautics, Beihang University.His research interests include machine learning, deep learning and semantic segmentation.

Zhengxia Zou received his BS degree and his Ph.D. degree from Beihang University in 2013 and 2018. He is currently an Associate Professor at the School of Astronautics, Beihang University. During 2018-2021, he was a postdoc research fellow at the University of Michigan, Ann Arbor. His research interests include computer vision and related problems in remote sensing. He has published more than 20 peer-reviewed papers in top-tier journals and conferences, including TPAMI, TIP, TGRS, CVPR, ICCV, AAAI. His research was featured in more than 30 global tech media and was adopted by a number of application platforms with over 50 million users worldwide. His personal website is https://zhengxiazou.github.io/.

Zhenwei Shi (Member, IEEE) received the Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, from 2013 to 2014. He is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or coauthored over 200 scientific papers in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters, the IEEE Conference on Computer Vision and Pattern Recognition, and the IEEE International Conference on Computer Vision. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi also serves as an Editor for the Pattern Recognition, the ISPRS Journal of Photogrammetry and Remote Sensing, the Infrared Physics and Technology, and so on. His personal website is http://levir.buaa.edu.cn/.