PistonNet: Object Separating From Background by Attention for Weakly Supervised Ship Detection

Yi Yang, Zongxu Pan, Yuxin Hu, and Chibiao Ding

***Abstract*—Object detection under weakly supervised learning is a challenging issue. In a remote sensing ship detection task, the weakly supervised learning method requires that the training set only has image-level class annotations. In the absence of location information, it is difficult to locate ships and extract features. Moreover, when the detector extracts more than one candidate region, the image-level annotation makes it difficult to determine whether the candidate regions are multiple ships or mixed with the detected background. To address these issues, this article analyzes the interaction between class information and location information, and proposes a weakly supervised detection method, i.e., PistonNet, based on data-efficient image transformers. PistonNet proposes an artificial point, which is inserted into the feature map. Artificial point can suppress the background and enhance the object by interfering in the weight distribution of object and background in self-attention calculation. PistonNet also proposes joint confidence probability to improve detection accuracy. Experiments on the GF1-LRSD and NWPU VHR-10 datasets show that the proposed method boosts detection performance effectively. The main improvements of PistonNet as well as the contributions of this article are threefold. First, PistonNet is a specific weakly supervised object detection method for single-class detection, which provides an innovative approach to classifying target and background. Second, PistonNet reaches the level of advanced supervised detectors on detection accuracy with fewer parameters. Finally, the objects' locations detected by PistonNet are obtained by segmenting regions on the heat map. PistonNet's background suppression ability makes it free from dependence on segmentation threshold.**

***Index Terms*—Artificial point, data-efficient image transformers (DeiT), remote sensing images, ship detection, weakly supervised learning.**

I. Introduction

In RECENT years, deep learning has a full application in object detection. Some widely used object detection methods [1]–[6] have been migrated and applied to various fields. Due to the excellent performance of deep learning in the field of object detection, some methods are applied to ship detection of remote sensing images, along with the corresponding optimization schemes for remote sensing field. [7]–[12] At present, the ability of data acquisition is greatly boosted to meet the requirements of deep learning. However, the efficiency of the dataset production is still limited by manual annotation. This asymmetry of time cost encourages the development of weakly supervised learning. Weakly supervised object detection (WSOD) simplifies the annotation from instance level to image level, which greatly reduced the labor cost of data production. Remote sensing datasets contain a large number of images of complex scenes, which greatly increases the difficulty of manual annotation. Therefore, the introduction of WSOD is of great significance in engineering application.

In 2016, Bilen and Vedaldi [13] designed an end-to-end weakly supervised deep detection network (WSDDN), which integrates the MIL [14] into the WSOD. MIL regards images as bags and object proposals as instances, then picks high-scored instances from the bags. WSDDN is a groundbreaking design, on which a host of research works has been proposed [15]–[19].

In the field of ship detection, WSOD still has the following defects. Under the restriction of only class annotation, WSOD could not achieve the learning of single-class object, while the supervised object detection can find the background region in the image through the object's bounding box (BBox) annotation to learn the difference. Existing WSOD methods applied to remote sensing images usually use multiclass datasets to ensure that each image involved in training contains at least one class of object [20]–[24]. Therefore, for weakly supervised training of ship data, it is necessary to make the background into a class separately. This data processing enables WSOD to be applied to the single-class ship detection task, but it also brings the following two problems.

1) First, compared with multiclass object detection, it is not the class of object contained in the image, but whether the image contains a ship that determines the classification difference in ship detection. WSOD methods based on MIL iteratively select the most contributing proposal as the pseudoinstance-level label, which guides the classification. Even if a background image is input, the detector will extract the candidate region to prove that the class of the image is "background". However, the background class is not determined by these candidate regions, but by the absence of ships in the image. In the images of the target class, background also exists, and these candidate regions representing background will lead to false classification.
2) Second, as two classes: a) background and b) target, do not have equal status. Ships have common features, but backgrounds do not. The supervised object detection methods can accurately extract object's features through

Manuscript received 22 March 2022; revised 23 May 2022; accepted 15 June 2022. Date of publication 21 June 2022; date of current version 7 July 2022. *(Corresponding author: Zongxu Pan)*

The authors are with the Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Aerospace Information Research Institute, School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yangyi183@mails.ucas.ac.cn; zxpan@mail.ie.ac.cn; yxhu@mail.ie.ac.cn; cbding@mail.ie.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3184637

BBox, while WSOD needs to learn features of all classes at the same time. Classification is the only source of the training loss of weakly supervised methods. Therefore, the classification difficulty of complex background will affect the target's classification and feature extraction. For example, in GF1-LRSD [25], which is an optical remote sensing dataset from GF-1 satellite, the average size of a ship is 10.9 pixels. In each image, the complex environment is the difficulty of detection. Under this condition, it is a challenge to achieve classification and localization by WSOD.

In 2021, TS-CAM [26] was proposed using WSOD, which is a non-end-to-end detector based on data-efficient image transformer (DeiT) [27]. TS-CAM combines the semantic-aware tokens with the semantic-agnostic attention map to make the output heat map of the network class- and position-sensitive. Objects are segmented on the heat map by setting a threshold. Compared with the end-to-end detector, the object segmentation method using heat map has the following two defects.

1) First, extraction of objects locations from feature maps directly depends on the setting of the threshold. With the change of threshold value, the result of segmentation also changes.
2) Second, the candidate regions lack matching confidence probabilities and the segmentation of the object depends on the pixel value in the heat map.

The distribution of pixel values is not strictly related to confidence probability. When the number of objects in the image is uncertain, the heat map cannot distinguish the candidate regions from multiple objects or mixed background. The precision and recall of the detector cannot be satisfactory at the same time.

However, the non-end-to-end design provides inspiration for solving the single-classification task of ship detection. The output of the detector is a heat map, which has not been converted into specific class information. It is expected that the target has candidate regions and the background does not, which can be achieved by interfering with the distribution pattern of pixels during heat map generation. The segmentation scheme based on the new distribution pattern can effectively solve the problem existing in the object extraction by the non-end-to-end detector.

Based on the abovementioned background research, this article proposes PistonNet, which designs artificial point to influence the interaction between class information and location information to suppress the background on the feature maps, and proposes joint confidence probability to improve detection accuracy. PistonNet has reached the following goals.

1) A specific WSOD method for single-class detection is proposed, which has engineering application significance.
2) The proposed method has a lightweight design and achieves the accuracy of advanced detectors using about 1/10 of the parameters.
3) The intervention of artificial point can effectively suppress the background during training. The combination of artificial point and joint confidence probability improves the detection accuracy of WSOD to multiple targets and makes the non-end-to-end detector free from dependence on segmentation threshold.

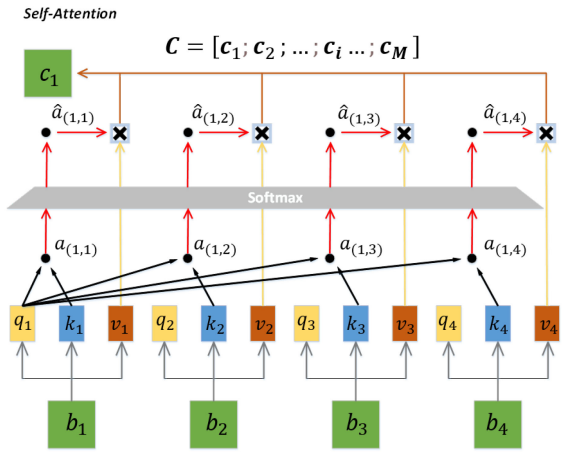


Fig. 1. Structure of self-attention layer. The generation of c_1 is shown in the figure, and the same goes for all the other vectors $(c_2, c_3, \ldots, c_M)$.

4) This article analyzes the mechanism of the interaction between class information and location information in WSOD and achieves the intervention in this process. This is instructive for the further study of weakly supervised learning.

II. Methods

A. Mechanism Analysis of WSOD Using DeiT

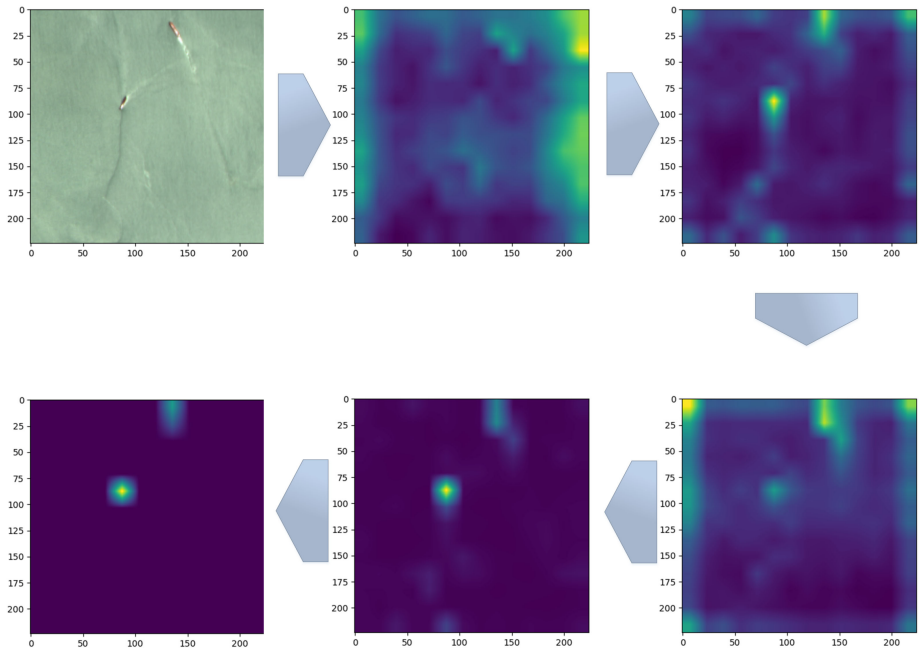
PistonNet uses DeiT as the backbone, which extracts features through cascading transformer blocks [28]. Within each transformer block, there is a self-attention layer and a multilayer perceptron (MLP) layer.

Before input to the backbone, an image of W × H resolution enters the convolution layer of N convolution kernels $W_{\text{conv}} \in \mathbb{R}^{P\times P}$ with step P to obtain the matrix $B_{\text{conv}} \in \mathbb{R}^{w\times h\times N}$, where $w = \frac{W}{P}$ and $h = \frac{H}{P}$. The matrix B_{conv} is reshaped to get $B \in \mathbb{R}^{M\times N}$, where $M = w \times h$. B contains vectors $b_i \in \mathbb{R}^{1\times N}, i = 1, 2, \ldots, M$. Each vector b_i concentrates information of a $P \times P$ patch at the corresponding position in the image.

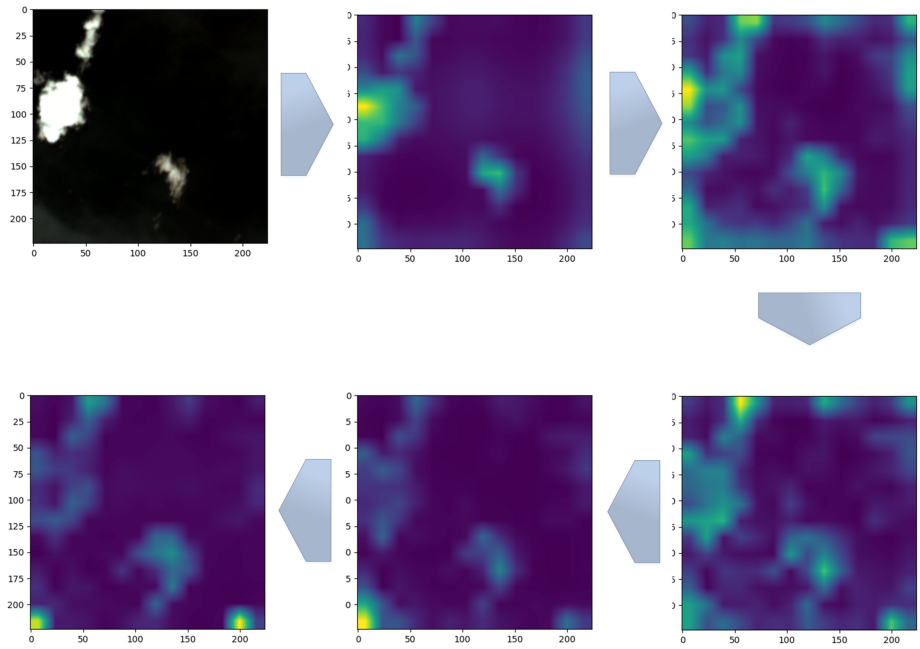
B is the input in the first transformer block and passes through the self-attention layer. The structure of the self-attention layer is shown in the Fig. 1. In the self-attention layer, the matrix B is multiplied by three transformation matrices (W_q, W_k, W_v) to obtain matrices Q, K, and V. Each group of $(q_i \in Q, k_i \in K, v_i \in V)$ corresponds to an input $b_i \in B$ in the self-attention layer. All the vectors q_i and k_j are conducted correlation calculation, gathering values $a_{(i,j)}$. All of $a_{(i,j)}$ forms the matrix $A \in \mathbb{R}^{M\times M}$. Each row vector $\text{row}_i = (a_{(i,1)}, a_{(i,2)}, \ldots, a_{(i,M)})$ of A is calculated by softmax to obtain $\hat{\text{row}}_i$, which forms the matrix $\hat{A}$. Vector c_i is gathered by summation of vector v_j, and $\hat{a}_{(i,j)}$ is the coefficient of summation $c_i = \sum_j \hat{a}_{(i,j)} v_j$.

The matrix $C = [c_1; c_2; \ldots; c_i; \ldots; c_M]$ enters MLP to obtain a matrix C_{out}, which is the final output of one transformer block and the input of the next transformer block with the same dimension as B. All the transformer blocks form a cascade.

In order to analyze the interaction mechanism between class information and location information in weakly supervised learning, in the detection head, the matrix C_{out} obtained from the



(a)



(b)

Fig. 2. Formation of location feature maps using DeiT only. As the transformer block deepens, the high weight in (a) gradually approaches the location of the ships. In (b), the weights are irregularly distributed, and parts of them are concentrated at the edge of the image. (a) Attention maps of ship. (b) Attention maps of background.

last transformer block is convolved and global average pooled to calculate classification loss. Using the trained model, this article extracts the matrix $\hat{A}$ of each transformer block. The matrix $\hat{A}$ is reshaped to $\hat{A}$_reshape $\in \mathbb{R}^{M\times w\times h}$, which is regarded as a $w \times h$ feature map with M channels. The feature map $\hat{A}$_heat $\in \mathbb{R}^{w\times h}$ of different channels is defined as location feature map. All the $\hat{A}$_heat are fused for visualization after histogram equalization.

To test the effect of WSOD using DeiT only, the target and the background images are input separately. Because of softmax calculation, points in $\hat{A}$_heat are distributed in the interval of (0,1) and the sum of their values is 1. These points are defined

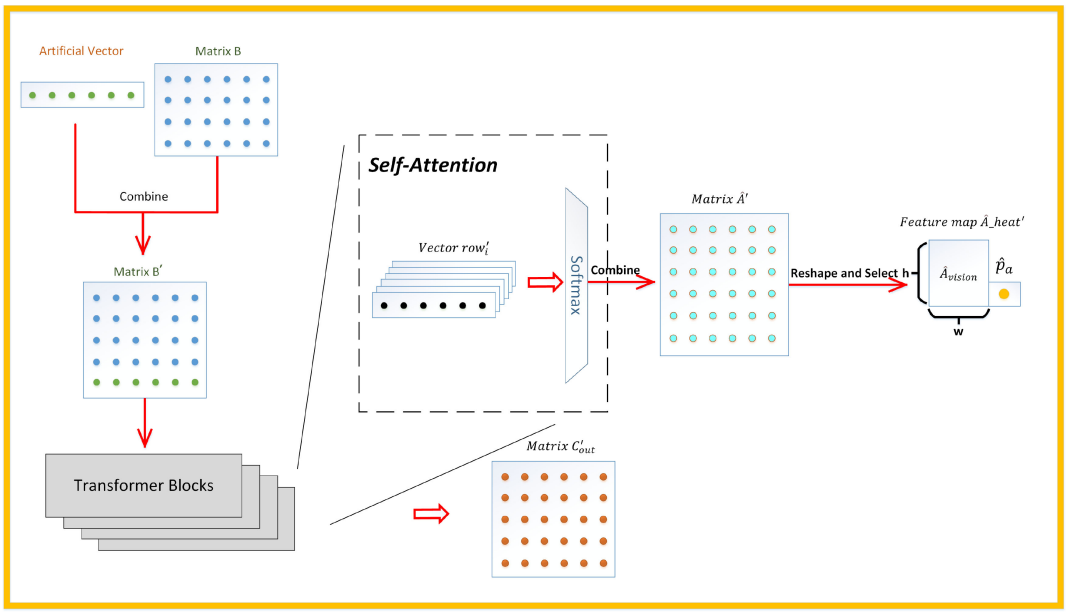


Fig. 3. Intervention of artificial point. Artificial point, unlike other points, does not represent information about a patch of the image. However, it participates in softmax with other points and adjusts the distribution of them.

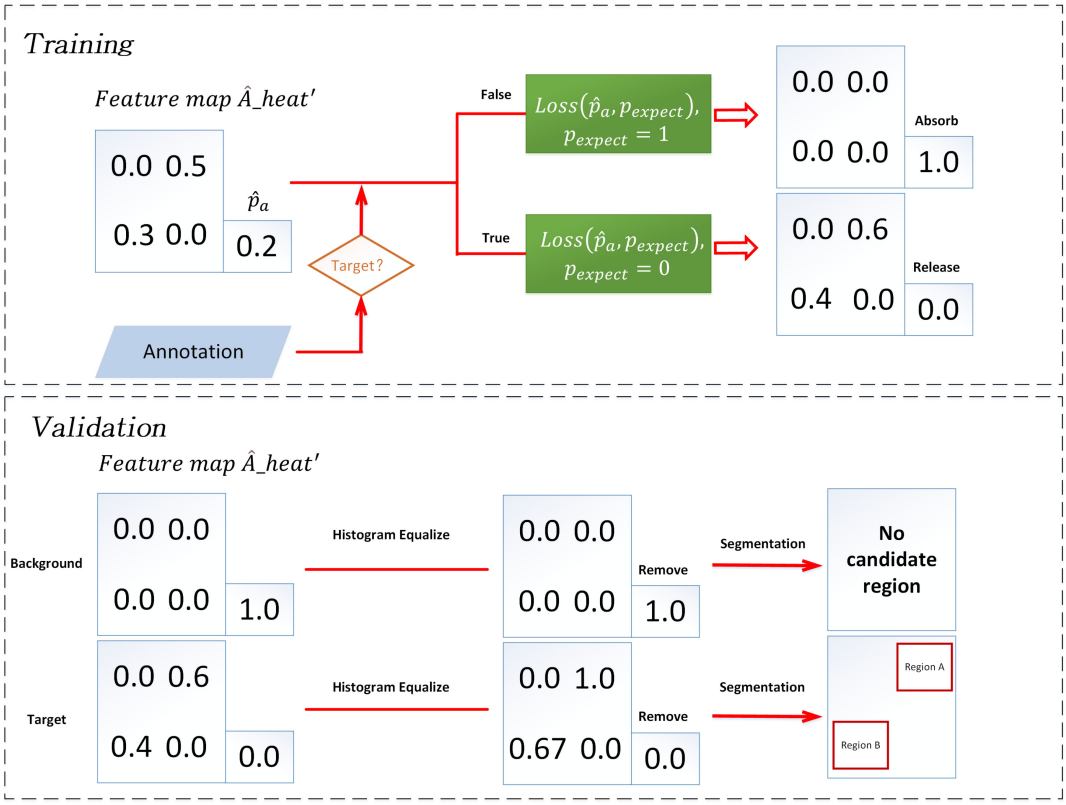


Fig. 4. Mechanism of piston. The values in the figure are for demonstration only; they show the desired control capabilities of the piston.

as feature points. As shown in Fig. 2(a), with the tranformer block deepens, feature points with high values gradually gather towards the objects' locations. The feature points in $\hat{A}$_heat have the property of weights and the source of $\hat{A}$_heat, $\hat{A}$ is multiplied by V in the self-Attention layers. It can be analyzed that the high weight value of objects' locations enables the network to classify correctly when participating in classification, which forms the basis of WSOD. $\hat{A}$_heat has the function of guiding the classification of class feature map C_{out}.

However, in Fig. 2(b), the performance of location feature map is not satisfactory. The distribution of feature points is disordered, and the position with the highest weight does not

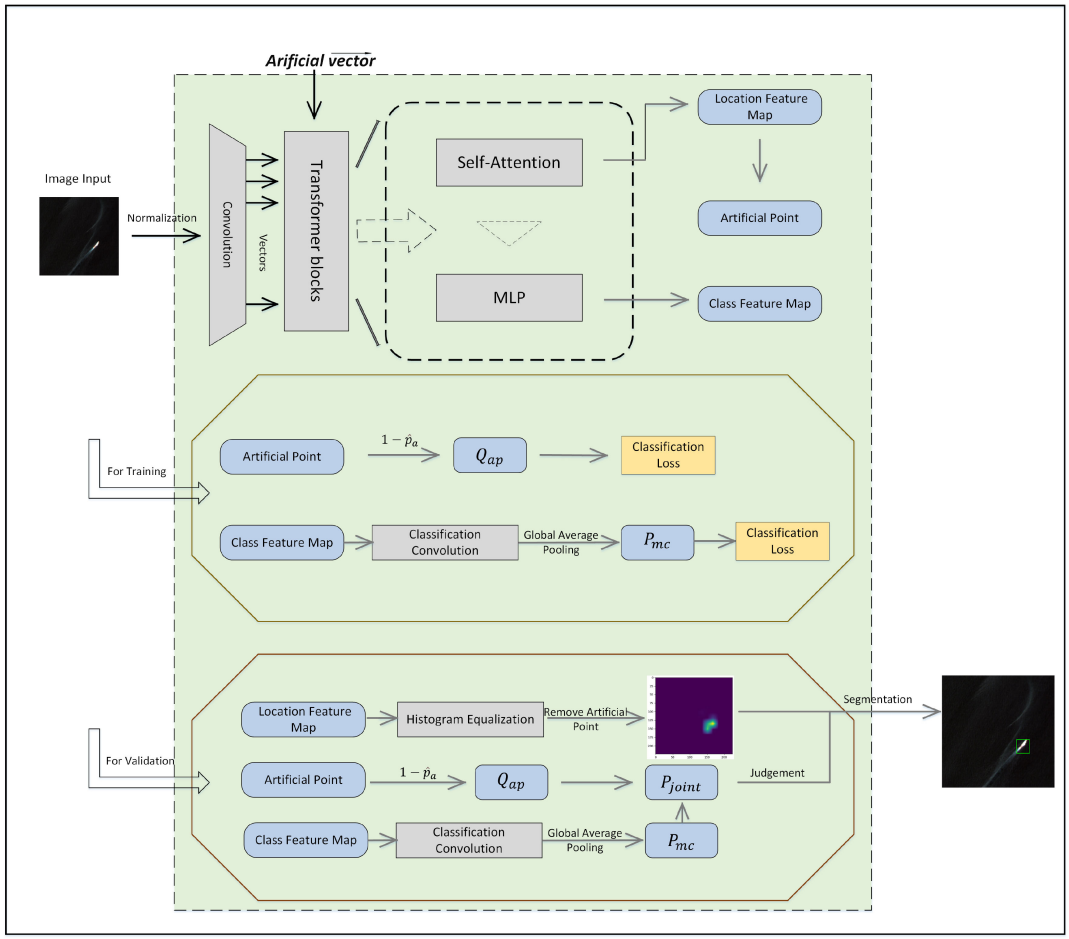


Fig. 5. Structure of PistonNet.

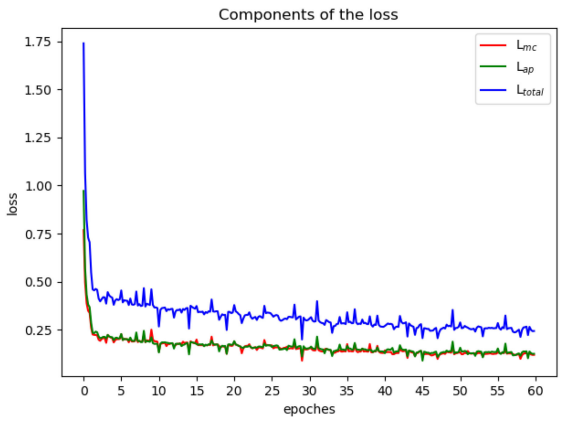


Fig. 6. Loss descent curve of PistonNet on the GF1-LRSD dataset. L_{ap} fluctuates more violently than L_{mc}, and this property is used in joint confidence probability designs.

contain any instance. There are the following two reasons for this result.

1) First, the weight of the location feature map is calculated by softmax. The sum of the weights is constant 1, regardless of whether the feature map contains objects or not. When the weight cannot be concentrated in the object's location, it will move randomly to a certain position.
2) Second, the visual heat map is generated by histogram equalization, so there must be highlighted areas on the heat map.

This issue has been discussed in Section I that candidate regions can also be generated in the background image. In the event of misclassification, these candidate regions become false detections. Moreover, when these candidate regions are mixed into the target's feature map, how to separate them is also a difficulty.

B. Intervention of Artificial Point

In order to solve this problem, mentioned in the previous section, the artificial point scheme is proposed in this article. The intervention of artificial point is shown in Fig. 3.

Before input to the transformer block, B is added with a row vector $s_{\text{art}} \in \mathbb{R}^{1\times N}$ initialized by a constant. The new input $B' \in \mathbb{R}^{(M+1))\times N}$ produces $C_{\text{out}}{}'$, row$'$, $\hat{A}'$, and $\hat{A}$_heat$'$. Accordingly, the lengths of all the vectors involved in transformer blocks are increased by 1. The dimension of the location feature map $\hat{A}$_heat$'$ output of the model is $w \times h + 1$. The extra point on the feature map is defined as the artificial

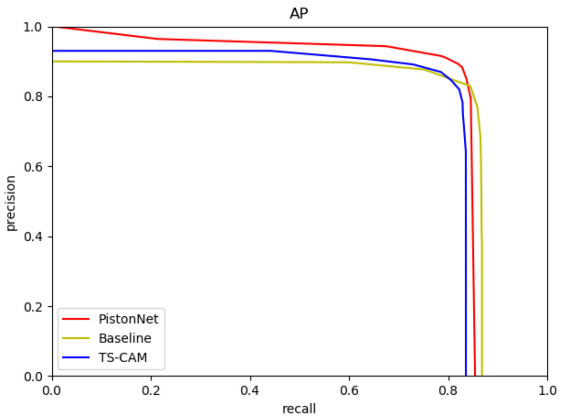


Fig. 7. Precision–recall curves on the GF1-LRSD dataset. The PistonNet has the best performance. When precision approaches 1, the curve of PistonNet shows a nonlinear rise, which will be explained in the section of joint confidence probability.

point $\hat{p_a}$. $\hat{A}$_heat$'$ consists of two parts: 1) $\hat{p_a}$ and 2) matrix $\hat{A}_{\text{vision}} \in \mathbb{R}^{w \times h}$.

The $\hat{p_a}$ extracted from the last transformer block calculates the classification loss with the class annotation. In the loss calculation of $\hat{p_a}$, when the input is the target image, the expected value of the class is set as 0; otherwise, it is 1.

The intervention of artificial point $\hat{p_a}$ acts as a piston. As shown in Fig. 4, in training, when the background image is input, the expected value of $\hat{p_a}$ is 1, which makes $\hat{p_a}$ absorbs all the weight in softmax calculation. When the target image is input, the expected value of $\hat{p_a}$ is 0, and all the weights are released into the $\hat{A}_{\text{vision}}$.

This "piston" mechanism effectively suppresses the background. For a background image, the weight sum of the $\hat{A}_{\text{vision}}$ is close to 0, so there is no candidate region. For a target image, the weight released by the piston will be concentrated near the objects, so it is not prone to cause false detection. It will be analyzed how does the piston suppresses the false detection in the following section.

In training of background image, the piston constrains each pixel's value of the background image to be close to 0. This constraint is also reflected in the training of the target image. Although target and background are divided into two classes in training, background actually exists in both classes of images. There is no significant difference between the backgrounds in the two classes of images. Therefore, the pixels in the background regions of the target image will also spontaneously approach 0. The piston uses the feature consistency of background regions in different classes to achieve background suppression.

During the validation process, the $\hat{A}$_heat$'$ is histogram equalized, and then the artificial point is removed. With background in the obtained heat map suppressed completely, the objects can be segmented by setting a very low threshold with an extremely low probability of false detection.

C. Joint Confidence Probability

It is worth noting that both the artificial point and the matrix C'_{out} output are class-sensitive. They are strongly correlated with the confidence probability of the image's class. This article defines the confidence probability from C'_{out} as P_{mc}, and the confidence probability from artificial point $\hat{p_a}$ as Q_{ap}, where $Q_{\text{ap}} = 1 - \hat{p_a}$. The joint confidence probability is proposed, as shown in the (1). The joint confidence probability is the weighted average of P_{mc} and Q_{ap}, and P_{mc} is also used as the weight. As the confidence probability increases, the proportion of Q_{ap} in P_{joint} increases. The advantages of this design will be analyzed in the following experiments:

$$P_{\text{joint}} = P_{mc} \times Q_{\text{ap}} + (1 - P_{mc}) \times P_{mc}. \tag{1}$$

D. Model Structure and Loss Calculation

Based on the artificial point, PistonNet is proposed in this article, and the whole structure of it is shown in Fig. 5. The input image is vectorized by convolution, and then combined with an artificial vector. In the cascaded transformer blocks, artificial point acquires class semantics and adjusts the distribution of the location feature map. The last transformer block has three outputs: 1) class confidence probability P_{mc}; 2) feature map $\hat{A}$_heat$'$; and 3) artificial point confidence probability Q_{ap}.

In training, P_{mc} and Q_{ap} calculate the classification loss, respectively. For P_{mc}, the classification loss uses focal loss [2], as shown in (2). P_{truth} is the ground truth of the input image, and γ and α are the balancing parameters.

$$L_{mc} = \begin{cases} -\alpha \left(1 - P_{mc}\right)^{\gamma} \log P_{mc}, & P_{\text{truth}} = 1 \\ -\left(1 - \alpha\right) P_{mc}^{\gamma} \log \left(1 - P_{mc}\right), & P_{\text{truth}} = 0 \end{cases}. \tag{2}$$

For Q_{ap}, the classification loss uses cross entropy loss as follows:

$$L_{\text{ap}} = \begin{cases} -\log Q_{\text{ap}}, & P_{\text{truth}} = 1 \\ -\log \left(1 - Q_{\text{ap}}\right), & P_{\text{truth}} = 0 \end{cases}. \tag{3}$$

The total loss is shown as follows, where β and λ are the balancing weights:

$$L_{\text{total}} = \beta L_{mc} + \lambda L_{\text{ap}}. \tag{4}$$

In validation, the images containing objects will be selected by P_{joint}. The precision–recall curve is drawn using P_{joint} as the threshold. Each corresponding feature map $\hat{A}$_heat$'$ is histogram equalized into the form of heat map, then the artificial point is removed. The heat map is segmented by a threshold and objects are extracted.

III. Experimental Results and Discussions

A. Dataset and Evaluation Metrics

The GF1-LRSD dataset [25] and the NWPU VHR-10 dataset [29] are used in the following experiments.

The GF1-LRSD dataset was collected from Gaofen-1 optical remote sensing satellite. GF1-LRSD contains 4406 512 × 512 images with a resolution of 16 m and 7172 labeled instances. Most objects in GF1-LRSD are smaller than 16 pixels, accounting for about 94% of the total objects. The average size of all objects in the dataset is 10.9 pixels. This dataset is divided into the training set and validation set by 4:1.

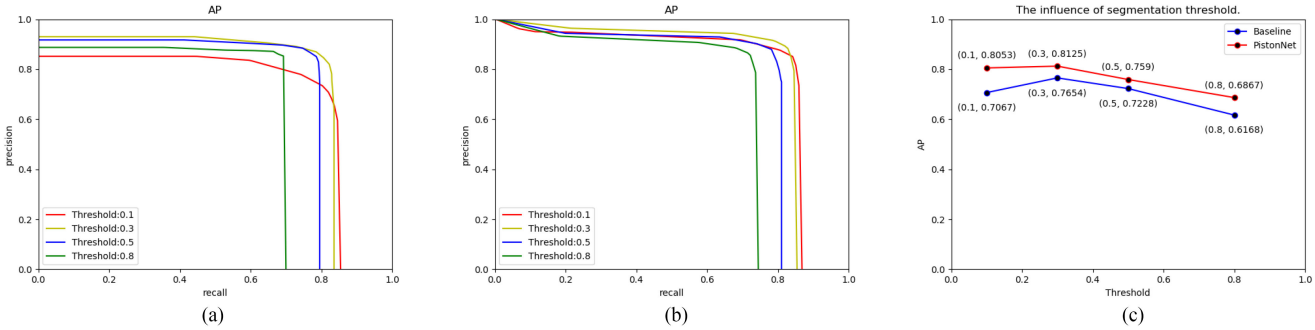


Fig. 8. Influence of segmentation threshold. PistonNet is more stable when the segmentation threshold is low, which means that the pixel value of the background is suppressed to close to 0. Therefore, PistonNet can extract as many candidate regions as possible without worrying about false detection. (a) Baseline. (b) PistonNet. (c) Comparison of methods' dependence on the threshold.

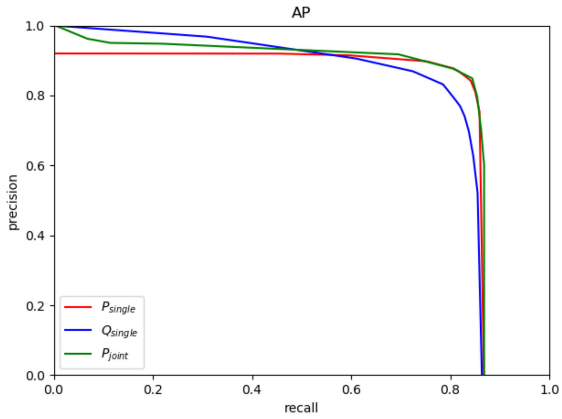


Fig. 9. Influence of joint confidence probability, which combines the advantages of the two probability prediction schemes and achieves the highest AP.

In order to make the training set meet the requirements of WSOD, images were cropped into 224 × 224 pieces. Pieces are divided into background and target, depending on whether they contain ships. The new training set contains a total of 13 363 224 × 224 images, including 5741 backgrounds and 7622 targets. In the validation process, each 512 × 512 image is cropped. The detector outputs the detection results of pieces and splices them back. In the validation set, it is regarded as a correct prediction that the center of the candidate region is within the BBox of the ship.

The NWPU VHR-10 dataset consists of a positive image set, including 650 images, and a negative image set, including 150 images, over ten object classes. The dataset is divided into the training set and the validation set by 4:1.

For the training set, the data processing method of [30] is referred to. Images (400 × 400 pixels) are cropped from the positive image set of the NWPU VHR-10 dataset. For these images, the target and background are reclassified and trained. In the validation process, we use the same method as for the GF1-LRSD dataset to input the cropped image (400 × 400 pixels) and splice the output. In the validation set of NWPU VHR-10, it is regarded as a correct prediction that the BBox of the candidate region overlaps more than 50% with the BBox of the ground truth.

TABLE I
AP COMPARISON OF DIFFERENT METHODS ON THE GF1-LRSD DATASET

Model	Baseline	TS-CAM	PistonNet
AP	0.7654	0.7704	**0.8125**

Bold figures indicate optimal result.

In this article, the average precision (AP) is used to evaluate the detector.

B. Experimental Environment

In this article, all the experiments are implemented under the Pytorch framework. The network is trained on 1 × Nvidia GeForce RTX 3090 and the batch size is set to 384. The number of epochs in training is 60, and the initial learning rate is 0.00005. The learning rate at epoch 30 and epoch 60 of the training process decays to $\frac{1}{5}$ of the current learning rate.

C. Ablation Experiments

PistonNet uses DeiT as the backbone to feature extraction, so DeiT is used as baseline in the following experiments. Loss curve in training is shown in Fig. 6. In order to evaluate the reliability of the proposed detector, the following experiments are conducted.

1) First, the influence of the artificial point on detection accuracy is tested. Precision–recall curves are shown in Fig. 7. The AP of PistonNet, baseline, and TS-CAM [26] is compared in the experiment. For the output heat map, PistonNet, baseline, and TS-CAM segment it with a threshold of 0.3.
   The heat map output from TS-CAM is obtained by coupling the feature maps of baseline. As given in Table I, TS-CAM has a 0.5% improvement in AP over baseline. As analyzed in Section I, multiclassification optimization cannot play a significant role in the target–background detection task. PistonNet is 4.71% higher than baseline, which proves the effectiveness of the proposed method.
2) Second, experiments will analyze the influence of segmentation threshold. At thresholds of 0.1, 0.3, 0.5, and 0.8, AP of baseline and PistonNet is calculated, respectively, and the result is shown in Fig. 8. The baseline

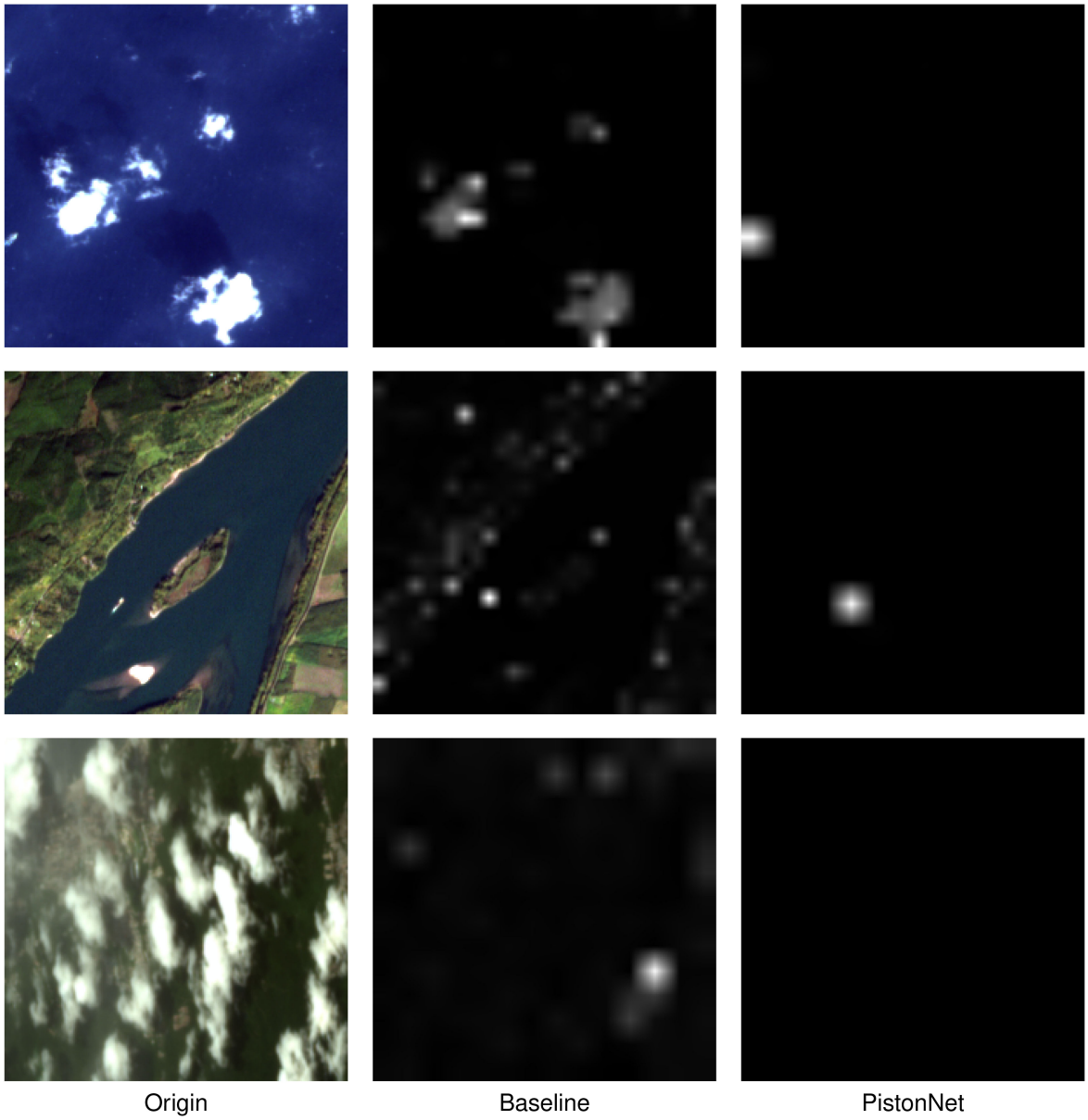


Fig. 10. Visual comparison of baseline and PistonNet. In the first two images, PistonNet shows better target extraction ability and background suppression ability. The third image illustrates that PistonNet has achieved the goal of this article; there is no candidate region in the background.

is highly dependent on the segmentation threshold, and AP fluctuates greatly as the threshold changes. Different from the baseline, AP of PistonNet at a low threshold (0.1) is not significantly lower than optimal AP. The AP–threshold curve of PistonNet generally shows a monotonically decreasing trend. Since that PistonNet has excellent background suppression ability, it can accurately extract objects under low threshold for segmentation, which makes it independent of threshold setting as a non-end-to-end detector.

3) Third, the effect of joint confidence probability is verified experimentally. The experiment compares the following three schemes.
   a) $P_{\text{single}} = P_{mc}$.
   b) $Q_{\text{single}} = Q_{\text{ap}}$.
   c) $P_{\text{joint}} = P_{mc} \times Q_{\text{ap}} + (1 - P_{mc}) \times P_{mc}$.

TABLE II
INFLUENCE OF JOINT CONFIDENCE PROBABILITY

Scheme	BEP	AP
P_{single}	0.8417	0.7878
Q_{single}	0.8029	0.7951
P_{joint}	0.8335	**0.8053**

P_{joint} shows the best overall performance. Bold figures indicate optimal result.

The detection accuracy of different schemes is given in Table II. Detector using joint confidence probability has the highest AP. As shown in Fig. 9, The curve of P_{single} is more stable with the highest value of break-even point (BEP), and the curve of

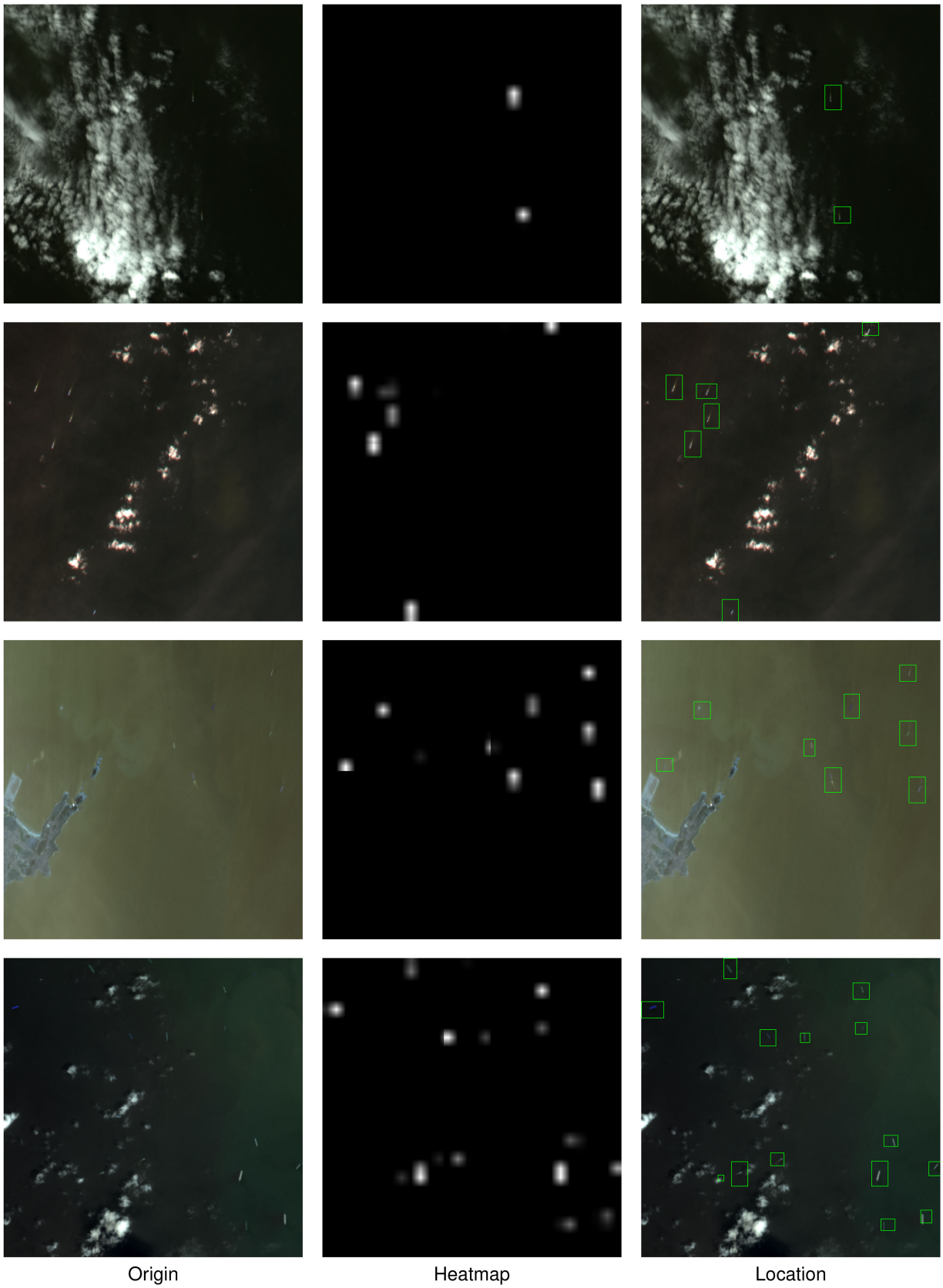


Fig. 11. Visualization of detection results from PistonNet on GF1-LRSD.

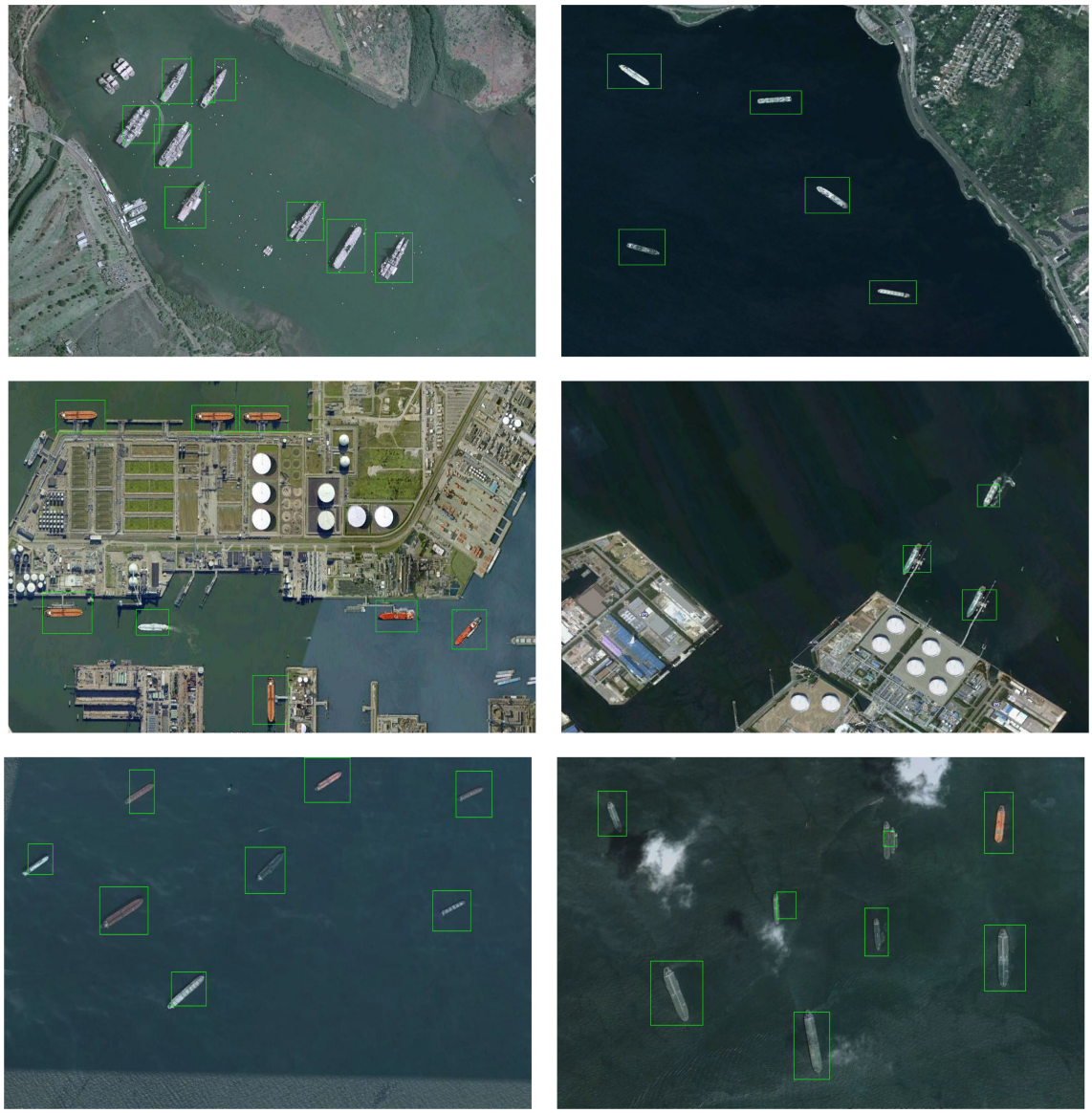


Fig. 12. Visualization of detection results from PistonNet on the NWPU VHR-10 dataset.

Q_{single} has a higher recall rate. P_{joint} inherits the advantages of the two curves and therefore performs better.

P_{mc} comes from class feature map C'_{out}, which is entirely designed to extract class semantics in training. Q_{ap} comes from the location feature map, and its function is to adjust the distribution of location information by class information. P_{mc} extracts class information in the class feature map, while Q_{ap} extracts class information in the location feature map. Therefore, the loss of Q_{ap} is harder to converge in training than P_{mc}. As shown in Fig. 6, the loss curve from Q_{ap} fluctuates more than the curve from P_{mc}.

Since it is difficult to extract class information in location feature map, Q_{ap} is hard to obtain the extreme value of the interval (0,1) in training. Although Q_{ap} is difficult to obtain a very high confidence probability; when Q_{ap} gets a high confidence probability, it is more reliable than P_{mc}.

When used as a confidence probability alone, P_{mc} is more reliable. So, P_{mc} is chosen as the weight of P_{joint}. The dominance of P_{mc} and Q_{ap} in the precision–recall curve gradually changes. As shown in Fig. 7, the AP curve of PistonNet shows a nonlinear rise at high precision.

4) Next, for the visual detection result, PistonNet has made significant improvements. The comparison with baseline is shown in Fig. 10. The following three scenarios are shown in this figure.

a) In the first image, the baseline outputs the correct classification, but the wrong location.

b) In the second image, the baseline outputs the correct classification and object's location, but with a lot of false detections.

c) In the third image, the baseline outputs the correct classification, but candidate regions are still given on the heat map of background, which is not expected.

Although these candidate regions are discarded after the baseline classifies the image as background, they impose an additional burden on the network in training to extract features that represent the image as "background". However, the class

TABLE III
OVERALL PERFORMANCE COMPARISONS ON THE GF1-LRSD DATASET

Model	Backbone	Weakly supervised	AP
Faster-RCNN [1]	ResNet	–	0.6049
YOLOv3 [3]	DarkNet	–	0.6782
SCRDet [7]	ResNet	–	0.6929
SSD512 [4]	VGG16	–	0.7228
R^3Det [10]	ResNet	–	0.7304
ATSS [31]	ResNet	–	0.7384
FCOS [6]	ResNet	–	0.7416
RetinaNet [2]	ResNet	–	0.7900
LR-TSDet [25]	ResNet	–	**0.8387**
TS-CAM [26]	DeiT	✓	0.7704
PistonNet	DeiT	✓	**0.8125**

Bold figures indicate optimal result.

TABLE IV
OVERALL PERFORMANCE COMPARISONS ON THE NWPU VHR-10 DATASET

Model	Backbone	Weakly supervised	AP
RICNN [32]	AlexNet CNN	–	0.7834
Faster-RCNN [1]	ResNet	–	0.8630
Fast-RCNN [33]	ResNet	–	0.9060
RICO [34]	ResNet	–	0.9080
WSDDN [13]	MIL	✓	0.4172
PCL [17]	MIL	✓	0.6376
OICR [18]	MIL	✓	0.6735
MELM [19]	MIL	✓	0.6930
TS-CAM [26]	DeiT	✓	0.7745
TCANet [22]	MIL	✓	0.7818
PistonNet	DeiT	✓	**0.8319**

Bold figures indicate optimal result.

TABLE V
VOLUMES COMPARISON OF DIFFERENT METHODS

Model	Backbone	Params (MB)
Faster-RCNN [1]	ResNet50	314.96
RetinaNet [2]	ResNet50	236.24
YOLOv3 [3]	DarkNet53	469.84
SSD512 [4]	VGG16	183.36
FCOS [6]	ResNet50	200.56
PistonNet	DeiT	21.04

of the background simply depends on the fact that it does not contain ships. The background images do not have features or candidate regions that represent the class of them. PistonNet does not have this problem and is able to focus entirely on learning ship features. Its improvement is reflected in both precision and visualization.

In contrast, PistonNet gives satisfactory detection results for different scenarios, which results from its effective suppression of background and accurate retention of object.

5) Finally, this article has compared PistonNet with other models. The overall comparison performance is given in Tables III and IV. PistonNet has exceeded some of the supervised detectors, and its detection accuracy is advanced. Moreover, PistonNet also has an excellent lightweight design. Table V compares the parameter volumes of PistonNet and other detectors. PistonNet has reduced the use of parameters by 88.47% compared with the detector with the fewest parameters. The number of parameters in PistonNet is 7.52% of the average number of parameters in these methods. PistonNet uses DeiT as backbone and designs 12 cascading transformer blocks to extract features. Compared with convolutional neural networks, PistonNet has fewer network layers and channels. As PistonNet adopts the method of segmenting objects from feature maps, the network does not need to train a detection head to further process feature maps. The lightweight design of PistonNet is determined by the advantages of ViT and weakly supervised learning.

To further show the performance of PistonNet, the detection results are shown in Figs. 11 and 12.

IV. CONCLUSION

This article proposed a weakly supervised detector called PistonNet for ship detection using optical remote sensing images. On the GF1-LRSD dataset and the NWPU VHR-10 dataset, PistonNet reached the advanced level of the supervised detectors. Experiments showed that PistonNet has the following contributions.

1) A specific WSOD method for single-class detection was proposed. PistonNet effectively suppressed the background through the piston mechanism, which can provide a solid baseline for related work.
2) PistonNet has a lightweight design and achieved the accuracy of advanced detectors. Full extensibility enables PistonNet to meet a variety of task requirements.

3) This article analyzed the generation and interaction of class semantics and location semantics theoretically, and experiments showed that the proposed methods can optimize this process.

In summary, in terms of the methodology, PistonNet is innovative in WSOD of target–background. With respect to the detection result, the proposed method is obviously superior to most supervised detectors in the remote sensing ship detection task. This further proves the effectiveness of the proposed method and the theoretical significance of this article.

At present, there is still a gap in accuracy between WSOD methods and supervised methods. On the one hand, it is more difficult to transform the objects' features into location parameters by the weakly supervised method. On the other hand, for dense objects, the weakly supervised method is difficult to accurately detect them. In this article, the PistonNet was designed for background suppression, and the multitarget detection was successfully achieved in the objective-background detection task. This method can also be extended to multiclass detection tasks. By using different classes as negative samples of each other, the network is still able to find the background to suppress and thus accurately detect the object. That is what we are dedicated to in the future.

REFERENCES

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

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

[3] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 1804–2767.

[4] W. Liu *et al.*, *SSD: Single Shot Multibox Detector*. Berlin, Germany: Springer, 2016.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[6] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2020, pp. 9626–9635.

[7] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8231–8240.

[8] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[9] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2844–2853.

[10] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3163–3171.

[11] Y. Yang, Z. Pan, Y. Hu, and C. Ding, "CPS-Det: An anchor-free based rotation detector for ship detection," *Remote Sens.*, vol. 13, no. 11, 2021, Art. no. 2208.

[12] J. Wang, W. Yang, H. Guo, R. Zhang, and G. S. Xia, "Tiny object detection in aerial images," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 3791–3798.

[13] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2846–2854.

[14] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1-2, pp. 31–71, 1997.

[15] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3512–3520.

[16] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4294–4302.

[17] P. Tang *et al.*, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[18] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3059–3067.

[19] F. Wan, P. Wei, J. Jiao, Z. Han, and Q. Ye, "Min-entropy latent model for weakly supervised object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1297–1306.

[20] 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 Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[21] 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, 2022.

[22] X. Feng, J. Han, X. Yao, and G. Cheng, "TCANet: Triple context-aware network for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6946–6955, Aug. 2021.

[23] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 675–685, Jan. 2021.

[24] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8002–8012, Nov. 2020.

[25] J. Wu, Z. Pan, B. Lei, and Y. Hu, "LR-TSDet: Towards tiny ship detection in low-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3890.

[26] W. Gao *et al.*, "TS-CAM: Token semantic coupled attention map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2886–2875.

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[28] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

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

[30] 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.

[31] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9756–9765.

[32] 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.

[33] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

Yi Yang received the B.S. degree in communication engineering from Tsinghua University, Beijing, China, in 2018. He is currently working toward the Ph.D. degree in signal and information processing with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

His research interests include remote sensing, object detection, and deep learning.

Zongxu Pan received the B.Eng. degree in electronic and information engineering from the Harbin Institute of Technology, Harbin, China, in 2010, and the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2015.

He is currently an Associate Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. He has authored or coauthored more than 50 papers in peer-reviewed journals and conference proceedings. His research interests include deep learning in remote sensing application, such as target detection and recognition in optical remote sensing and synthetic aperture radar images, and super-resolution reconstruction of remote sensing images.

Yuxin Hu received B.S. degree in communication engineering from Mongolian University, Hohhot, China, in 2002, and the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2007.

He is currently a Professor with the Key Laboratory of Technology, Geo-Spatial Information Processing and Application System, Aerospace Information Research Institute, Beijing. His research interests include synthetic aperture radar, data processing, and implementation of SAR processing system.

Chibiao Ding received the Ph.D. degree in electronic engineering from Beihang University, Beijing, China, in 1997.

Since 1997, he has been working with the Institute of Electronics, Chinese Academy of Sciences, Beijing. He is currently a Research Fellow and Vice Director with Aerospace Information Research Institute, Beijing. His research interests include advanced SAR system, signal processing technology, information systems, and SAR 3-D vision.