CANet: Centerness-Aware Network for Object Detection in Remote Sensing Images

Lukui Shi, Linyi Kuang, Xia Xu, Bin Pan, *Member, IEEE*, and Zhenwei Shi, *Member, IEEE*

***Abstract*— Recently, feature pyramid has been widely exploited in remote sensing detectors, which greatly alleviates the problem arising from scale variation across objects in remote sensing images. However, these object detectors with feature pyramid give insufficient consideration that objects in remote sensing images usually maintain symmetrical shape. To address this issue, we propose an anchor-free-based detector called Centerness-Aware Network (CANet), which could capture the symmetrical shape of objects in remote sensing images. The kernel structure of CANet is a new Centerness-Aware Model (CAM) that contains three components: Multiscale Centerness Descriptor (MSCD), Centerness Detection Head (CDH), and Feature Selective Module (FSM). Considering that symmetrical objects will maintain a rigid appearance around their center region, three components are integrated into the feature pyramid to extract and utilize the features around the center region. More precisely, the MSCD is embedded into the feature pyramid and highlights the center of current objects through the attention mechanism. Guided by the MSCD, the CDH could accurately capture the center of objects by per-pixel prediction. Furthermore, the FSM is connected to the CDH, which guides the CDH to adaptively select the optimal feature level from the pyramidal features. The selected feature level could describe the best semantic information around the center region, which helps the network progressively fit the symmetrical shape of remote sensing objects. Besides, we also design the hybrid loss function to effectively train CAM in the end-to-end way. The experiments show that our network is competitive with some state-of-the-art detection networks.**

***Index Terms*— Anchor-free, attention mechanism, remote sensing object detection.**

Manuscript received August 11, 2020; revised December 18, 2020 and March 17, 2021; accepted March 19, 2021. Date of publication April 2, 2021; date of current version December 16, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC1510900, in part by the National Natural Science Foundation of China under Grant 62001251 and Grant 62001252, in part by the Natural Science Foundation of Hebei Province of China under Grant F2019202062 and Grant F2020202008, and in part by the Science and Technology Program of Tianjin under Grant 18YFCZZC00060 and Grant 18ZXZNGX00100. *(Corresponding author: Bin Pan.)*

Lukui Shi and Linyi Kuang are with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China, and also with the Hebei Province Key Laboratory of Big Data Calculation, Tianjin 300401, China (e-mail: shilukui@scse.hebut.edu.cn; kuanglinyi@hotmail.com).

Xia Xu is with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: xuxia@nankai.edu.cn).

Bin Pan is with the School of Statistics and Data Science, Nankai University, Tianjin 300071, China, and also with the Key Laboratory of Pure Mathematics and Combinatorics, Ministry of Education, Tianjin 300071, China (e-mail: panbin@nankai.edu.cn).

Zhenwei Shi is with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3068970

I. Introduction

OBJECT detection in remote sensing images is a prerequisite for many practical applications, such as environmental monitoring, military investigation, and civilian application, which has attracted more and more attention [1]–[3]. With the advance of deep convolutional neural networks (CNNs) in recent years, many novel object detectors based on CNNs have been proposed [4]–[6]. However, different from natural scene images, remote sensing images are commonly captured from different high altitudes. Furthermore, the size of objects may have various scales due to the variation in shooting heights, in which case the multiscale of objects possibly brings obstacles to effectively detect objects [7].

To handle the scale variation of remote sensing objects, some networks generated high-quality fusion features by constructing and aggregating multilevel features, where feature pyramid network (FPN) [8] is regarded as a typical framework for detecting multiscale objects. The FPN is constructed as a top-down architecture with lateral connections to build multilevel feature maps [9]–[11]. The structure enables that high-level feature maps have more semantic information associated with large instances, and low-level feature maps reflect details associated with small instances. Taking FPN as the backbone for extracting features, many researchers have further developed novel methods that learn the characteristics of objects in remote sensing imagery. For example, some networks attempt to explore novel strategies that refine key information and incorporate semantics supplement, reconstructing more discriminative feature representations to enhance the detection capability. Wang *et al.* [12] embedded a nonlocal block and convolution block attention module (CBAM) into FPN to mine global information from the scene. By attaching instance-aware semantic information to pyramid feature representation, Xu *et al.* [13] proposed the hierarchical semantic propagation framework that alleviates the interference of complex background. For compensating the missing information of small objects caused by continuous down-sampling, some methods employed parallel dilated convolution layers with various rates to reconstruct diverse levels in feature pyramids and strengthen contextual clues for small objects [14]–[16]. To address the unbalance of multilayer features, methods proposed by literature [17]–[19] optimized feature extraction by integrating and refining all the feature levels. Besides, the literature [20] exploited asymmetric convolution blocks to strengthen rota-

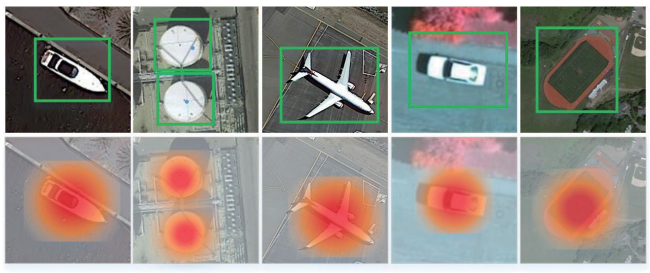


Fig. 1. First row shows several instances of remote sensing objects. The second row demonstrates that the remote sensing objects with symmetrical shape can maintain the distinctive appearance around their center region.

tional robustness of features, and further boost the FPN-based detector to adapt arbitrariness of the object arrangement.

Although these FPN-based approaches have been utilized auxiliary semantic information and achieved impressive performance, there still exists room for improvement. On the one hand, these FPN-based methods enumerate anchor boxes to densely cover the object location during training and testing. However, the anchor box mechanism is a heuristic rule depending on extra hyperparameters, in which case the performance of networks is sensitive to the predefined anchor setting schemes for specific data set [21]. On the other hand, many FPN-based methods generally implement anchor sampling via fixing the IoU threshold during training [22], which causes that detectors give an insufficient description on the shape characteristics of remote sensing objects. Therefore, it is necessary to eliminate these constraints caused by the anchor box mechanism for the promotion of FPN-based methods in the field of remote sensing.

Recently, anchor-free detectors have been proposed for object detection in remote sensing images, which can directly avoid the heuristic factor brought by the anchor boxes mechanism. These anchor-free detectors replace the traditional anchor box technology and adaptively focus on salient areas of objects with anchor-point or key-point detection way. The anchor-point detection encodes and decodes anchor-points to generate the proposal boxes around the sensitive area of the feature pyramid [23], [24]. Meanwhile, key-point detection detects the corresponding key points of objects, and groups these key points to generate the horizontal predicted boxes (PBs) [25], [26]. These anchor-free detectors have been explored more representative characteristics of remote sensing objects by combining with the novel strategy, i.e., self-attention mechanism and spatial shuffle-group enhance mechanism [27]–[30]. With flexible and robust detection methods based on multilevel feature structure, these novel anchor-free networks alleviate redundant heuristic factors related to anchor boxes, which achieves ideal results and shows great potential toward future trends.

Inspired by the superiority of the anchor-free detectors, we take FPN as the backbone to construct a new anchor-free detector for remote sensing images. Compared with objects in natural scene images, as shown in Fig. 1, remote sensing objects are obtained at a fixed bird-view perspective and have a fixed symmetrical shape. To our best knowledge, although the existing methods utilized various characteristics of objects (e.g., relevant global scene and dense object arrangement) [31]–[34], the symmetry as important prior knowledge for describing the shape of objects and still has not been achieved widespread exploitation. In our opinion, it may be potential to incorporate the symmetry information into appropriate representations to further enhance the detection capability. Actually, the symmetry as the property of remote sensing objects naturally facilitates the objects of interest in maintaining rigid appearance around their geometric center, therefore these features around the geometric center can serve as high-quality representations for remote sensing objects. At present, the visual attention mechanism has been proposed to extract the key region feature. For example, some methods [35]–[37] integrated spatial and channel attention blocks to capture significant areas, and the literature [38], [39] constructed recurrent neural network (RNN) to capture key sequence information. If an appropriate attention module combines with FPN to focus on the center of the object, FPN-based detectors can achieve the purpose of perceiving the symmetry of the object, and their performance may be improved in remote sensing images.

This article proposes Centerness-Aware Network (CANet) for object detection in remote sensing images, which is a single-stage anchor-free detector with a per-pixel prediction way. The core of CANet is a newly designed structure, the Centerness Aware Model (CAM), which aims at emphasizing the features around the center of the detected symmetrical objects. The CAM is divided into three components, namely multiscale centerness descriptor (MSCD), centerness detection head (CDH), and feature selective module (FSM). These submodules are implemented jointly to progressively guide the network to focus on the center of the symmetrical objects. Specifically, MSCD first implements the spatial attention mechanism into multilevel features to highlight the central region of objects. Then CDH is connected to each level of features, which is responsible for classification and location on each network branch. Furthermore, the CDH is guided by the FSM, and adaptively selects the training branch that reflects the best features around the geometric center of the object. Finally, the CAM is trained with a hybrid loss function by the end-to-end way. The contributions of this article can be summarized as follows:

1) We develop the novel CAM which facilitates the FPN-based detector to capture the symmetrical shape for remote sensing objects.
2) The proposed CAM can describe the centerness of symmetrical objects by highlighting, perceiving, and refining central semantics from hierarchical features in the remote sensing images.

The next chapters in this article are organized as follows. Section II describes the proposed network in detail. Section III further demonstrates the implementation details and experimental results. In Section IV, we summarize this article.

II. Methodology

This section mainly describes the details of our proposed network. We first introduce the overall structure of the network. Then the structure of the CAM in the network is described by Sections II-B–II-D, which contain MSCD, CDH,

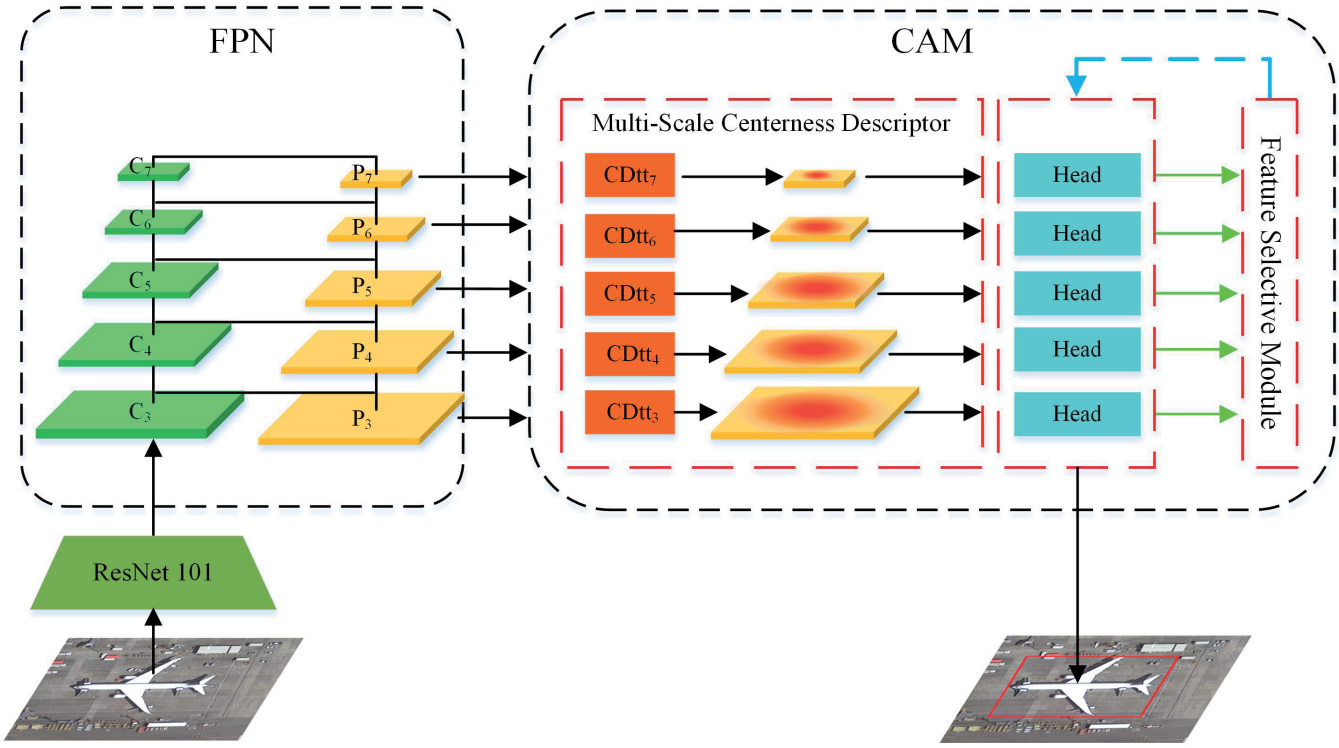


Fig. 2. Overall architecture of CANet. FPN is plugged into CAM that is decomposed into MSCD, CDH, and FSM. In the figure, the black line represents the data flow forward by the network. Each green line represents the loss of the head at each level of features, and the blue line represents the level of feature selected for the current training instance. The gradient of the loss at the selected feature level will participate in backpropagation. In brief, CDtt_i denotes the centerness descriptor submodule connecting the P_i in MSCD, and Head denotes CDH.

and FSM. Finally, we illustrate a hybrid loss that optimizes CAM.

A. Overall Architecture of CANet

From the perspective of the overall structure, as shown in Fig. 2, the network can be divided into ResNet101 backbone, FPN, and CAM, where CAM is further decomposed into MSCD, CDH, and FSM. The ResNet101 is adopted as the backbone of the overall architecture that extracts multilevel features from the original images, and these generated features are denoted as C_3, C_4, C_5, C_6 and C_7, respectively. Then C_3 to C_7 are connected with FPN to establish the top-down pathway and lateral connection, which operates feature fusion between all level features and finally generates discriminative pyramidal features P_3 to P_7. The P_i denotes ith level in the pyramid feature, which integrates with spatial details and global semantics from different level features.

The CAM is designed as the core structure in the network, and it is composed of three submodules. The first submodule, namely MSCD, gives weight to each pixel in P_3 to P_7 by implementing a visual attention mechanism, which enhances multilevel feature maps to generate salient regions around the center of detected objects. For enhancing feature maps on each scale, the CDH achieves the corresponding prediction in a pixel-by-pixel way. The CDH has two parallel fully convolutional branches, i.e., classification branch and localization branch. The classification branch predicts the category of objects at each point in the feature maps. Meanwhile, the localization branch predicts the 4-D vector (l, t, r, b) which describes the relative offsets from the current point to the four predicted box boundaries. During the training process, as the last submodule of CAM, the FSM adaptively selects the optimal training branch for better reflecting center semantics information, which progressively boosts the network to learn the appearance of the object center.

B. Multiscale Centerness Descriptor

The objects of interest (such as vehicles, ships) in remote sensing images generally maintain the symmetrical shape even if the surroundings around objects are varied and complicated. The symmetrical shape causes these objects to present rigid appearance and remain significantly distinguishable around their central region. To utilize the central area in the multiscale feature for the detection task, we design a visual attention network, namely MSCD, which is integrated into the FPN and guides the network to focus on the features around the center of the objects.

Fig. 3 illustrates the structure of MSCD in detail. To obtain pyramidal heatmaps for the multilevel features, the descriptor employs five groups of submodule centerness detection $(\text{CD})\text{tt}_i$ layers whose parameters are shared with each other, where each CDtt_i is constructed by a set of continuous fully convolutional layers and responsible for handling corresponding feature level P_i. The CDtt_i receives feature maps with specific levels and generates corresponding centerness heatmap. Each group of convolution layer contains four $1 \times 3 \times 3$ convolution (each followed by sigmoid activation function). After convolution layers and activation functions are continuously stacked,

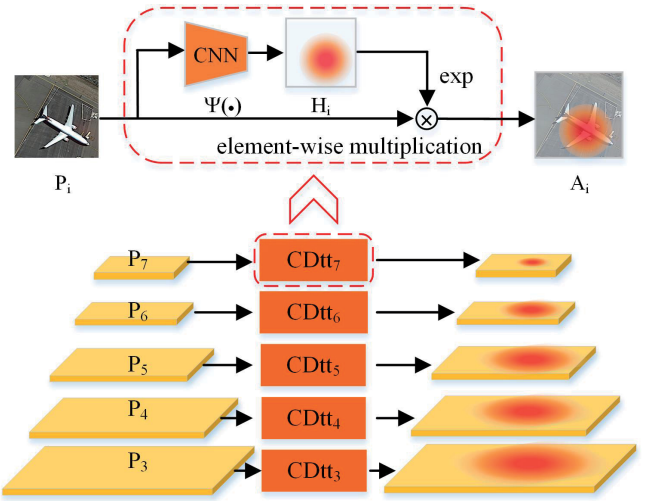


Fig. 3. Detailed structure of the MSCD. The feature pyramid is initially input into the MSCD, and then an enhanced feature pyramid is generated. In the figure, the CDtt_i denotes submodule in the MSCD, and the symbol $\bigotimes$ denotes element-wise multiplication.

the input P_i whose scale is $w \times h \times c$ will generate centerness heatmap with $w \times h \times 1$ size, and the value of centerness reflects the normalized distance from the positive pixel to the geometrical center of the current object. Finally, the channel of the heatmap is expanded to c dimension, and the generated heatmap is denoted as H_i whose scale is the same as P_i. The process from P_i to H_i can be defined as follows:

$$H_i = \Psi(P_i), \quad i = 3, 4, 5, 6, 7 \tag{1}$$

where $\Psi(\cdot)$ represents the mathematical function integrating convolution operations, sigmoid activation function, and dimension expansion operation, which converts the feature maps to the centerness heatmaps. The pyramidal centerness heatmaps execute the exponential operation, and then the exponential heatmaps are sequentially fused with the original feature maps through the dot product operation, where A_i denotes the enhanced feature map corresponding to P_i and its scale is $w \times h \times c$. The process is briefly described as follows:

$$A_i = P_i \bigotimes \exp(H_i), \quad i = 3, 4, 5, 6, 7 \tag{2}$$

where $\bigotimes$ represents element-wise product, and $\exp(\cdot)$ represents exponential operation. The fusion operation highlights the features around the center of objects. Furthermore, the fused feature maps will be input into the predicted branches for classification and localization task.

To train the fully convolutional network in MSCD, we first need to divide the sample area for each training instance. Conventionally, the ground truth (GT) of each training instance is labeled as a horizontal box and denoted with the format of $b = [x, y, w, h]$. For each horizontal box $[x, y, w, h]$, the (x, y) represents the center coordinate of the box, and (w, h) reflects the information about dimension of width and height, which is further assigned to each feature level on pyramidal features. When being attached to ith feature level, the horizontal box is recorded as $b_p^l = [x_p^l, y_p^l, w_p^l, h_p^l]$. Consistent with the sample division of the classification task in the

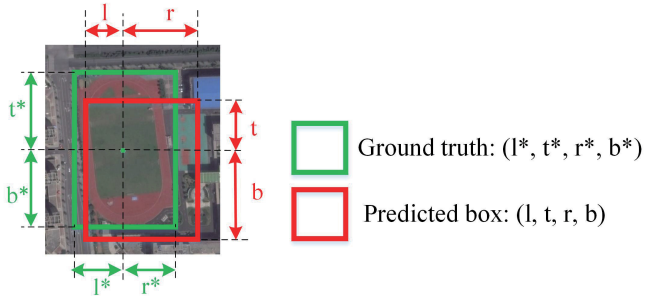


Fig. 4. Graphics illustrates the 4-D vector (l, t, r, b) and the vector how to encode the location of the predicted box and the ground truth box.

Fig. 5, b_p^l is divide into the effective region $[x_e^l, y_e^l, w_e^l, h_e^l]$ and the ignored region $[x_i^l, y_i^l, w_i^l, h_i^l]$. We set the fixed constants $\varepsilon_e = 0.3$ and $\varepsilon_i = 0.5$, so that $x_e^l = x_p^l$, $y_e^l = y_p^l$, $w_e^l = \varepsilon_e w_p^l$, $h_e^l = \varepsilon_e h_p^l$, $x_i^l = x_p^l$, $y_i^l = y_p^l$, $w_i^l = \varepsilon_i w_p^l$, $h_i^l = \varepsilon_i h_p^l$. The effective region b_e^l is taken as the positive sampling area, while the ignored region $(b_i^l - b_e^l)$ is not arranged as the training sample. In addition, the remaining area is defined as b_n^l in entire training image, and it is the negative sample area during training. To calculate the target values of the positive sample, (x, y, w, h) is converted into the 4-D vector (l, t, r, b), which is described as

$$\begin{aligned} l &= x - x_{\min}, \quad t = y - y_{\min} \\ r &= x_{\max} - x, \quad b = y_{\max} - y. \end{aligned} \tag{3}$$

The Fig. 4 illustrates the 4-D vectors (l, t, r, b) how to encode the predicted box for locating the position. The 4-D vector explicitly denotes the distance from a pixel to the left, right, upper, and lower boundaries of the ground truth box. Then the positive sample that backs onto the 4-D vector (l^*, t^*, r^*, b^*) is converted into the centerness value between 0 and 1, which is computed by

$$\text{centerness}^* = \sqrt{\frac{\min(l^*, r^*) \times \min(t^*, b^*)}{\max(l^*, r^*) \times \max(t^*, b^*)}}. \tag{4}$$

In addition, the target value of the negative sample is specified as 0. Because the pixel-level centerness reflects continuous predicted values between 0 and 1, the binary cross entropy (BCE) loss is adopted to train MSCD. Here the loss of the lth level feature is built upon the pixel-level, which is given by

$$\begin{aligned} L_{\text{CN}}^I(l) = -\frac{1}{N(S)} \sum_{i,j \epsilon S} &(\text{centerness}(l, i, j) \log(H(l, i, j)) \\ &+ (1 - \text{centerness}(l, i, j)) \log(1 - H(l, i, j))). \end{aligned} \tag{5}$$

In (5), the S denotes the training area where the pixel-level samples are located, which is calculated by $S = b_e^l \bigcup b_n^l$. The $N(S)$ denotes the number of pixel-level samples on the training area S. Besides, $\text{centerness}(l, i, j)$ and $H(l, i, j)$, respectively, denote the centerness target value and predicted value on the coordinate (i, j) in P_l. Supervised by the loss function from training samples, the MSCD acquires the ability to enhance the features around the center of objects.

C. Centerness Detection Head

The CDH connects with the enhanced feature maps after MSCD performs the operation, and perceives salient regions to effectively detect objects. As shown in Fig. 5, CDH contains parallel classification branch and location branch. Each classification branch consists of four groups of convolution, where the size and stride of each convolution kernel adopt $K \times 3 \times 3$ and 1, respectively. The feature maps by implementing each group of convolution operation are followed by the sigmoid activation function to ensure that the probability range is 0 to 1 for the classification task. The branch finally generates a group of feature maps stacked by K channels, where K represents the number of categories of the predicted objects. The predicted value of each pixel on the ith channel reflects the probability that the point on the original image belongs to the ith category object. On the other hand, the localization network is constructed by four convolution kernels with $4 \times 3 \times 3$ size, and the stride of the convolution is designated as 1. Different from the classification branch, each group of convolution is followed by relu activation function. In addition, each localization branch finally generates feature maps stacked by four channels. In the final prediction maps, the 4-D vectors on the predicted maps reflect the location information. Sequentially, these 4-D vectors are decoded as the prediction boxes that back onto input images. Finally, the redundant results are eliminated by nonmaximum suppression (NMS). During the prediction process, multilevel branches need to share parameters with each other.

Since pixels distributed on the sensitive area tend to make more accurate predicted results, we implement training losses by pixel-level on classification and location task, which will make each detection head perceive the salient area. For the classification loss, we still divide the training sample into the valid regions and the ignored regions according to the guidance in Fig. 5. The division ratio is consistent with the trained MSCD, so as to ensure that the ambiguous regions on the ground truth are excluded from the training samples, which further guarantees that CDH handles the pixel-level features around the center of the objects. However, based on the division, negative samples are classified into the vast majority during the training process. To overcome the imbalance of the positive and negative samples, we use the focal loss [40] to accelerate its convergence, which is defined as

$$L_{\mathrm{FL}}(p_t) = -\alpha(1-p_t)^{\gamma}\log(1-p_t) \tag{6}$$

where α and γ denote the weighting factor and modulating factor, respectively, and p_t denotes the probability that one pixel belongs to the true category. In the experiment, we set $\alpha = 0.25$ and $\gamma = 2$.

For the location task, we first introduce the concept of Generalized Intersection over Union (GIoU) [41], which is a metric for comparing predicted box with the ground truth. Its expression is as follows:

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{|C \setminus (\mathrm{PB} \cup \mathrm{GT})|}{|C|} \tag{7}$$

where IoU denotes the intersection over the union between the predicted box and the ground truth, C denotes the smallest horizontal box enclosing the current predicted box and the ground truth box, PB denotes the predicted box, and GT denotes the ground truth box. Note that the predicted box is derived from one anchor point, and the detailed decoding process is described in (3) and Fig. 4. The GIoU not only measures the intersection of coverage between two boxes, but also reflects the empty area between two boxes. In this article, the regression loss for location adopts $L_{\mathrm{GIoU}} = 1 - \mathrm{GIoU}$, which establishes function with GIoU for per-pixel on the predicted map. As is shown in Fig. 5, the division ratio of the positive sample for the loss is consistent with the classification task, while the ignored regions in the classification task are divided into negative samples in the location branch. When training the location branch, the loss simultaneously drives the area growth of the bounding boxes overlapping with the ground truth and reduces the empty area between two boxes. As the location of the predicted boxes gradually approaches ground truth boxes during training, anchor points around the center of the object are decoded as a series of predicted boxes. Compared with the boxes generated from other locations, these boxes are convenient to match the ground truth, which forces their losses to converge. With the assistance of anchor points around the center, the location branch of the network can perceive the center of the objects and completely generate the horizontal predicted boxes.

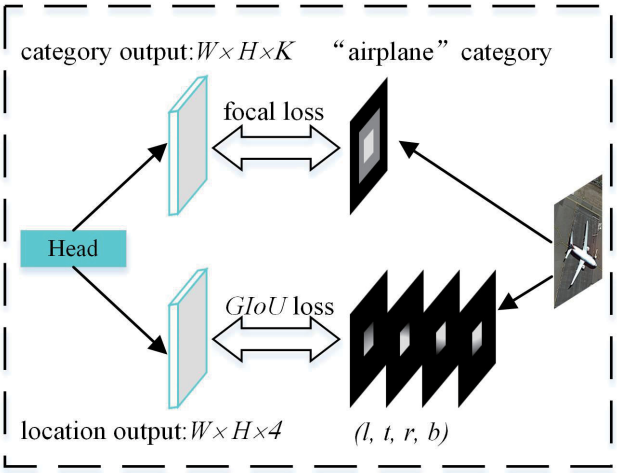


Fig. 5. Structure of CDH. The CDH performs classification and location, where focal loss and GIoU loss are used to train the corresponding task. In addition, the figure also illustrates the division area of training samples corresponding to the task. In the sample division, white area and black area indicate positive and negative sample areas, respectively, while gray area indicates ignored areas that is not used as training samples.

D. Feature Selective Module

As we know, the FPN-based networks generally include multiple predicted branches, where each branch is connected to a specific level of pyramid features. The higher levels contain more abstract semantic information that is suitable for detecting larger objects, while the lower levels are suitable for detecting smaller objects due to their detailed granularity [42]. To quantitatively make training instances match with feature levels, many traditional networks utilize heuristic feature selection to determine the feature level according to the size of

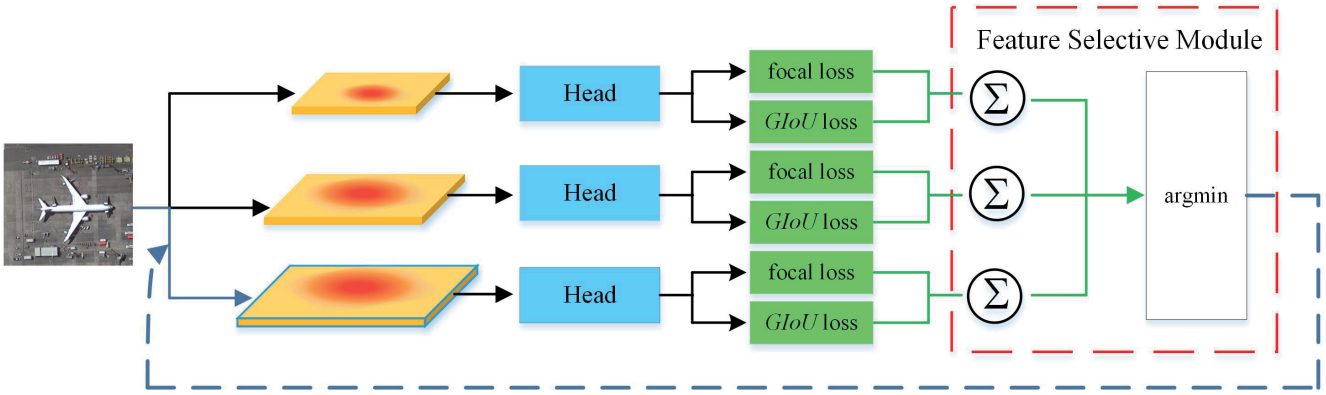


Fig. 6. Working process of FSM. The enhanced pyramidal features are used for classification and location by CDH, and then losses (i.e., focal loss and GIoU loss) at each level are generated. Based on the comparison of the sum of the losses at each layer, the feature level that generates the smallest sum of loss is selected for training. For simplicity, only three levels are shown in the figure, and Head denotes CDH.

the instance. For example, the FPN-based detectors assign the level of a feature according to the following predesigned formula [8]:

$$l = \lfloor l_0 + \log_2\left(\sqrt{wh}/224\right)\rfloor \tag{8}$$

where l is the level of a feature on the instance with $w \times h$, and $l_0 = 5$ that defines the parameter of the lowest level. Based on the heuristic rule, the size of the instance becomes decisive factor for selecting features during training multiple predicted branches, and the selected level of feature is used to describe the current object. However, when applying (8) for the network, even if two instances with the same category are similar in appearance around their center region, they may be accidentally assigned to different levels of features. In other words, the heuristic selection method only relies on the predefined rules to passively select features. As a result, it is possible for the network not to obtain the optimal level of feature containing the best semantics around their center.

To guide the network to adaptively select features for describing the center region of the objects, we introduce FSM that draws inspiration from online feature selection strategy in literature [21]. The FSM analyzes the multilevel semantic information that contains enhanced features of the central area, and further selects the optimal trained branch to assign the current object, without designing rules in advance. The process guarantees that the network can automatically obtain the optimal features around the center area of the object, which makes the network gradually have the ability to perceive the center of objects. Fig. 6 illustrates the working process of FSM. For simplicity, only three levels are shown in the figure. When an instance is the input into the CANet, the enhanced feature maps are first generated from MSCD. Then each CDH predicts the category and location on all feature levels, and the losses between the predicted result and the ground truth are calculated. Here, $L^I_{\mathrm{FL}}(l)$ and $L^I_{\mathrm{GIoU}}(l)$ denote the mean value of the classification loss and regression loss on all the valid regions of the ith feature level, respectively. The detailed calculation process is shown in (9) and (10)

$$L^I_{\mathrm{FL}}(l) = \frac{1}{N(S)} \sum_{i,j \epsilon S} L_{\mathrm{FL}}(l,i,j) \tag{9}$$

$$L^I_{\mathrm{GIoU}}(l) = \frac{1}{N\left(b^l_e\right)} \sum_{i,j \epsilon b^l_e} L_{\mathrm{GIoU}}(l,i,j) \tag{10}$$

where $N(b^e_l)$ denotes the number of pixels inside b^e_l, and S denotes the training area where the positive and negative samples are located. Besides, $L_{\mathrm{FL}}(l,i,j)$ and $L_{\mathrm{GIoU}}(l,i,j)$ denote the focal loss and GIoU loss on the coordinate (i,j) in P_l, respectively. The CDH generates the losses on all levels, and then the following formula is utilized to generate the optimal feature level l^* via comparing all sum of losses

$$l^* = \arg\min_l L^I_{\mathrm{FL}}(l) + L^I_{\mathrm{GIoU}}(l). \tag{11}$$

The optimal feature level is selected by FSM, and it contains the most suitable central feature for describing the center region of the current object. As the network is continuously trained by a large number of samples, it has the ability to infer the category and location of the object according to the center area of objects. During testing, the predicted branches corresponding to the most appropriate feature level naturally outputs higher confidence than other branches.

E. Hybrid Loss Function

Guided by FSM, the optimized level l^* of features is determined from multiple levels. To take advantage of the optimal level for training CAM, we joint three losses at feature level l^* into the following hybrid loss function, where three losses contain classification loss, location loss, and centerness loss. The hybrid loss is described by

$$L^I(l^*) = L^I_{\mathrm{FL}}(l^*) + L^I_{\mathrm{GIoU}}(l^*) + L^I_{\mathrm{CN}}(l^*). \tag{12}$$

Equation (12) means that only the loss corresponding to the l^* level will participate in backpropagation, thereby achieving the purpose of simultaneously optimizing the three parts in the hybrid loss. During the training process, MSCD is gradually optimized, and it further highlights the pixel-level features to generate the salient region. Then the salient area guides the network to generate high-quality predicted boxes around the center of the objects. The entire process of CANet is briefly summarized in Algorithm 1.

Algorithm 1 CANet

Input: Training data $\{I_0, \ldots, I_m\}$, testing data $\{T_0, \ldots, T_n\}$ pretraining weight W and hyperparameter;
Output: Trained network and predicted boxes;
Initialize the pretraining weight W for the network;
for $t = 0$ to m **do**
  Loss function: $L^I \leftarrow \phi$;
  Extract P_3, P_4, P_5, P_6, P_7 from I_t through FPN;
  for $l = 3$ to 7 **do**
    Generate heatmap H_l based on (1);
    Generate enhanced feature map A_l based on (2);
    Obtain predicted boxes on A_l through CDH;
    Calculate $L^I_{Fl}(l)$ and $L^I_{IoU}(l)$ on A_l based on (9) and (10);
    $L^I \longleftarrow L^I \bigcup (L^I_{Fl}(l) + L^I_{GIoU}(l))$;
  end for
  Select optimal feature level l^* based on (11);
  Generate hybrid loss on level l^* based on (12);
  Update weights W of network through hybrid loss;
end for
Generate bounding boxes from $\{T_0, \ldots, T_n\}$ through the trained network;
Select the predicted boxes through NMS;

III. Experiment

This section provides a concise description of experimental details, such as data sets, evaluation metrics, implementation details, and experimental results over the public remote sensing object detection data sets.

A. Data Sets and Evaluation Metrics

1) Data Sets: In the experiment, our proposed network is validated on NWPU very-high-resolution (VHR)-10 [43] and RSOD [44]. The NWPU VHR-10 is a publicly available remote sensing object detection data set released by Northwestern Polytechnical University. It contains 800 optical remote sensing images, of which 715 images are color images obtained from Google Earth, and the remaining 85 images are locally sharpened color infrared images. The categories of remote sensing objects are airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle.

RSOD is a remote sensing object detection data set released by Wuhan University in 2015. The data set contains 4993 airplane objects in 446 pictures, 191 playground objects in 189 images, 191 oil tank objects in 189 images, and 180 overpass objects in 176 images. In the comparative experiment, since two sets do not specify unified partition of training and testing sets, we randomly and, respectively, select 75% of the positive samples as a training set and the rest positive samples as a testing set.

2) Evaluation Metrics: In order to evaluate the performance of our proposed detection network, we use the average precision (AP) as the evaluation metrics. Generally, the detection results can be divided into four cases according to the coverage of the prediction box and ground truth box: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Then these four cases are combined with the following formula to calculate the precision rate (P) and recall rate (R). The $P-R$ curve is formed by connecting the P and R coordinate points under different confidence levels

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + FP} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{14}$$

The average accuracy (AP) is indicated through the graph area enclosed by the $P-R$ curve and the coordinate axis. The solution process can be described by the following formula:

$$\text{AP} = \int_0^1 P(r)dr = \sum_{k=1}^{N} P(k)\Delta R(k) \tag{15}$$

where r represents the recall rate, and $P(r)$ represents the precision rate under the r condition. The formula (15) implies that the value of the AP ranges from 0 to 1. The larger the value of AP in a certain category is, the better the detection effect of that category is. To evaluate the overall prediction performance of the network over multiple types of objects, we also average the APs of different categories, and finally obtain the mean average accuracy (mAP).

B. Implementation Details

The backbone of the network adopts RestNet101 pretrained on the ImageNet. By fine-tuned over the remote sensing data sets during our training phase, the backbone can extract features of studied images. The two studied data sets are trained on one GPU and their batch size was set to 1. Total training iterations for NWPU VHR-10 and RSOD are 20000 and 30000, which take around 2 and 3 hours, respectively. We adopt Adam as a training optimizer, and the learning rate is 0.00001. During the process of testing, if the overlap of the predicted box and the ground truth box exceeds 50%, it indicates that the target has been successfully and effectively detected. Such predict results are considered as TP in the experiment.

C. Results and Analysis

To evaluate the performance of the proposed network, the proposed network is made quantitatively in comparison with several state-of-the-art detection networks on the studied data sets. These compared methods are divided into remote sensing algorithms and natural scene algorithms. For remote sensing algorithms, the context-aware detection (CAD)-Net [45] (taking FPN as backbone), rotation-invariant CNN (RICNN) [46], and the network proposed by Li *et al.* [47] are implemented in comparative experiment. Besides, we also adopt a series of the natural scenes algorithms including four anchor-based detectors (Faster R-CNN [48], FPN [8], single shot multibox detector (SSD) [49], Cascade R-CNN [50]) and three anchor-free detectors (FoveaBox [51], guided anchoring (GA)-RetinaNet [52], fully convolutional one-stage object detection (FCOS) [24]). We select VGG16 as the network backbone of SSD, while the backbones of other natural scenes algorithms are designated as ResNet101, which

TABLE I
COMPARISON OF DIFFERENT METHODS ON NWPU VHR-10 (%)

method Class	CAD-Net ResNet101	RICNN ResNet101	Li et al. ZF CNN	Cascade R-CNN ResNet101	FoveaBox ResNet101	GA-RetinaNet ResNet101	FCOS ResNet101	CANet (ours) ResNet101
Airplane	97.00	88.71	99.70	99.54	99.48	99.99	99.99	**99.99**
Ship	77.90	78.34	**90.80**	88.53	83.60	84.28	85.21	85.99
Storage tank	95.60	86.33	90.61	95.98	96.83	97.92	96.94	**99.27**
Baseball diamond	93.60	89.09	92.91	94.46	95.14	96.53	**97.75**	97.28
Tennis court	87.60	42.33	90.29	94.02	85.65	96.98	95.80	**97.80**
Basketball court	87.10	56.85	80.13	**88.21**	84.37	85.12	80.34	84.77
Ground track field	99.60	87.72	90.81	97.16	95.54	95.34	**99.67**	98.38
Harbor	**100**	67.47	80.29	91.45	90.78	89.72	95.04	90.38
Bridge	86.20	62.31	68.53	82.30	81.67	81.32	81.82	**89.16**
Vehicle	89.90	72.01	87.14	90.25	**91.15**	91.85	88.92	90.25
mAP	91.50	73.11	87.12	92.19	91.42	91.91	92.14	**93.33**

TABLE II
COMPARISON OF DIFFERENT METHODS ON RSOD (%)

method Class	Faster R-CNN ResNet101	FPN ResNet101	SSD VGG16	Cascade R-CNN ResNet101	FoveaBox ResNet101	GA-RetinaNet ResNet101	FCOS ResNet101	CANet (ours) ResNet101
Aircraft	81.34	91.23	71.21	**92.20**	91.64	91.09	90.74	91.76
Playground	95.47	97.14	99.47	99.79	99.25	**99.82**	99.80	97.90
Oiltank	96.78	97.82	89.65	97.23	96.60	**98.38**	97.67	97.06
Overpass	85.45	88.40	85.81	87.45	86.81	83.05	86.43	**94.11**
mAP	89.76	93.65	86.53	94.17	94.02	92.91	93.68	**95.21**

ensures that the backbones in the comparison networks are as identical as possible. We also keep the data set usage consistent to make a fair comparison of each other, and their evaluation results are shown in Tables I and II.

Table I shows the mAP of our proposed network has reached 93.33% on NWPU VHR-10. The mAP value reflects that our network has the best experimental performance compared to other comparison networks. When analyzing the AP of specific categories in the NWPU VHR-10, it could conclude that the detection results from other methods are not ideal for ship (such as 77.90% in CAD-Net, 78.34% in RICNN) and vehicle (such as 72.01% in RICNN), while our network could obtain 85.99% and 89.16% for these challenging objects. In addition, compared to all other algorithms, our network gains the best experimental result in detecting oil tank, tennis court, and bridge categories. These AP values related to these objects reached 99.27%, 97.80%, and 89.16%, respectively, which is certainly slightly ahead of the experimental results of the other algorithms. On the other hand, as shown in Table II, the mAP of our method on the RSOD data set is 95.21%, which is also the highest among all networks. For the overpass category, our network can reach 94.11% AP value, which outperforms all other networks. Besides, our network reaches AP values of 91.76%, 97.90%, and 97.06% on aircraft, oil tank and overpass, respectively. These AP values are also relatively ideal experimental results compared to other networks.

Table III summarizes the computation time of seven different methods on NWPU VHR-10. The experiment reports the average running time per testing image by following the widely adopted scheme in literature [46]. As an advanced single-stage FPN-based detector, FCOS [24] has the lowest computation compared with other networks. The consumption time of CANet has only a 0.01 s gap with FCOS and outperforms five other networks, which demonstrates that our proposed method as a single-stage object detector still has relatively high efficiency. In our opinion, the competitive high efficiency of the proposed network benefits from a single-stage anchor-free structure avoiding high computational costs related to the anchor boxes. In summary, all comparative experiments demonstrate that CANet has superior performance in keeping speed/accuracy trade-off.

TABLE III
COMPUTATION TIME COMPARISON OF SEVEN METHODS

Methods	average running computation per image (seconds)
RICNN	8.47
Li et al.	2.89
Cascade R-CNN	0.12
FoveaBox	0.12
GA-RetinaNet	0.18
FCOS	0.09
CANet (ours)	0.10

D. Ablation Studies

In this section, we set up a series of ablation studies on the NWPU VHR-10. The ablation experiment verifies the effectiveness of the proposed network structure from two aspects. By gradually adding submodules to CAM and comparing the effects of different paired components, the ablation experiment is set to prove the effect of submodules in CAM. Then we also explore the optimal loss item combination to construct the hybrid loss function in the second group of ablation experiments. Before implementing experiments in different situations, we randomly select 70% positive samples of NWPU VHR-10 for training and the remaining positive samples for testing networks. To make a fair comparison with each other, we maintain the same hyper parameters when carrying out various experiments. These experiment arrangements and corresponding results are shown in Tables IV and V.

TABLE IV
ANALYSIS OF CAM IN ABLATION STUDIES (%)

FPN	+ MSCD	+FSM	+CDH	+anchor-based branches	mAP
✓				✓	87.85
✓	✓			✓	89.61
✓		✓		✓	88.73
✓			✓		90.25
✓		✓	✓		91.22
✓	✓	✓	✓		**93.15**

TABLE V
ANALYSIS OF HYBRID LOSS IN ABLATION STUDIES (%)

smooth L_1 loss	IoU loss	$GIoU$ loss	BCE loss	mAP
✓			✓	90.37
	✓		✓	92.82
		✓	✓	**93.15**
		✓		91.49

As shown in Table IV, each paired component shows the corresponding combination mode in the experiments. These networks take FPN as the backbone for exacting features, and their results verify the complementarity of the submodules of CAM. The first network only adopts anchor-based detection head as baseline, which has exactly the same structure as the RetinaNet network, and its mAP has reached 87.85%. Then we also, respectively, added MSCD and FSM to baseline, and their experimental results are slightly improved by 1.76% and 0.88%. The improvement of the experimental results shows that the two modules have effects on enhancing the performance of anchor-based detection. In order to prove the advantages of CDH, the combination of FPN+CDH is conducted in the fourth set of experiments, whose experimental result is 91.22% and higher than the combination in the first group. Meanwhile, we combine FSM+CDH components with FPN structure, and their overall performance reaches 91.22% mAP. Finally, we inserted the MSCD+FSM+CDH components (CAM) into the FPN-based network, and its mAP achieves 93.15%, which is significantly higher than the experimental results of other combinations. These experiments with different paired components not only demonstrate that the proposed MSCD, FSM, and CDH are actually complementary to each other, but also show that the CAM has significant advantages in enhancing FPN-based detection.

We also explored the effects of various items combinations on hybrid loss function, and the detailed results are shown in Table V. The attributes in the table represent various loss items, where smooth L_1 loss, IoU loss, and GIoU loss are implemented in the regression loss for location task, and BCE loss reflects the loss function related to proposed attention model. Since classification loss is implemented in each set of experiments, the focal loss is not shown in the table. The mAP in Table IV shows the experimental results under the different combination modes of various loss terms. In the first three sets of experiments, we separately organized smooth L_1 loss, IoU loss, and GIoU loss as regression loss for location task, and we ensure that Focal loss and BCE loss are adopted for classification and attention tasks. Experiments show that the effect of combining GIoU loss is 2.78% and 0.33% higher than smooth L_1 loss and IoU Loss, respectively. In addition, it is worth mentioning that we also experimented to exclude attention loss while retaining classification and regression loss on the overall performance of the network. In this case, the mAP has achieved 91.49%. This experimental result shows that the proposed network has a better performance when it is supervised by attention loss. The above results of ablation studies in the section reflect that CANet enhances the detection effect of objects by constructing various submodules and combining various effective losses.

E. Visualization Experiment

This section introduces visualization experiments to further illustrate the performance of the network in actual scenarios. Considering that CAM guides the network to focus on the center of the object by utilizing the attention mechanism, we further explore the effect of the attention mechanism on enhancing the features around the center of the detected objects. Fig. 7 is visualized as the activation response maps that before and after the attention mechanism is implemented in CAM, where each pair of maps are derived from the same feature level. In addition, we have also visualized the centerness heatmaps generated by attention mechanism, where the pixel values in the heatmaps reflect the weighted values for enhancing original feature maps. The brighter areas in the feature maps represent the activation response that the neural network can easily perceive, while the brighter areas in the centerness heatmaps represent a high centerness value that denotes high weighted values.

As shown in Fig. 7, the original image in the first row illustrates that tennis courts are densely distributed in remote sensing images, and thus the semantic information in the original feature map is ambiguous around their edge area. In this case, the detector based on pixel-by-pixel prediction is easily affected by ambiguous semantics around their edge area. However, guided by the centerness heatmap, the central features of the tennis court are enhanced, and the ambiguous features on the edges are suppressed. As a result, the boundaries between adjacent objects become clearer, which contributes to the bounding box to accurately cover the densely distributed objects. In the second row, we can observe that these harbors in the original image are frequently disturbed by the background (such as sea, houses), which causes the features of harbors to be not obvious in the whole image. Particularly, the noise in the background can easily bring about great interference to the pixel-by-pixel prediction detector. However, centerness heatmaps highlight features around the center of the indistinctive objects, which enables the network to focus on acquiring effective information in the harbor. Therefore, our network can accurately capture these challenging objects. The above activation response comparison results show that CAM can perceive symmetric remote sensing objects and generate high-quality feature maps.

To further demonstrate the predicted results of our proposed network, Figs. 8 and 9 show the predicted visualization results of the partial remote sensing objects on NWPU VHR-10 and RSOD, respectively. In Fig. 8, the first row shows some objects with insignificant features, and these objects are easily disturbed by the background. For example, the harbors are located

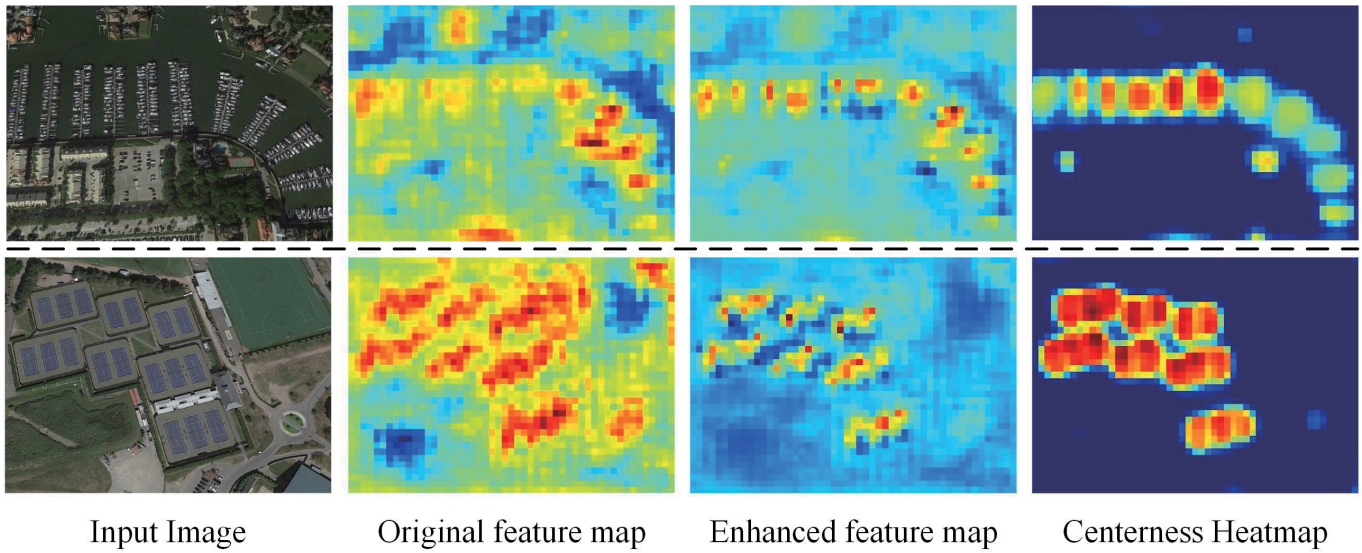


Fig. 7. Activation response maps for harbors and tennis courts. This set of images include input images, centerness heatmaps, original feature maps and enhanced feature maps.

Fig. 8. Results predicted by CANet on NWPU VHR-10, which contains the harbor, basketball court, baseball diamond, bridge, vehicle, playground, tennis court, ground track field, storage tank, airplane object categories.

Fig. 9. Results predicted by CANet on RSOD, which contains the airplane, overpass, playground, oil tank object categories.

on the river, and their appearance is easily disturbed by the small boats scattered in the water and shore. The appearance of the bridge over the river is similar to the road, and the basketball courts and baseball diamonds tend to confuse grass courts. The insignificant features and background influences of these examples bring obstacles for remote sensing detectors.

However, our network is guided by the attention mechanism to effectively capture features around the center of remote sensing objects, in which case central features with high discrimination exclude background influences. Meanwhile, objects in the second row are densely distributed in the images. For example, each tennis court in the second row is adjacent to other tennis courts, which causes the edge of these tennis courts to have ambiguous information. But our network can perceive the center of these detected objects and suppress ambiguous information, so as to accurately locate the tennis courts.

In Fig. 9, the first row contains a large number of tiny airplanes that tend to be missed due to their insufficient semantic information. However, the proposed network can accurately capture tiny airplanes and achieve ideal results. Based on the results in Section III, we analyze and summarize that CANet has the following two advantages in detecting these tiny airplanes compared with anchor-based methods. First, the anchor-based network generally adopts smooth L_1 loss to regress location during training network, while our network uses GIoU loss to locate the predicted boxes in the detection task. The theoretical derivation in literature [53] proves that the loss related to IoU is more suitable for small object locations. Second, CANet can obtain more training samples for small objects compared with anchor-based detection networks. The anchor-based detection networks select samples based on the coverage of the anchor box and the instance box. In this case, a small number of anchor boxes can cover the instance boxes, and thus anchor-based networks could not obtain sufficient samples for training tiny objects [54]. Compared with the traditional sampling method based on anchor box matching, the pixel-by-pixel sampling method could produce a relatively large number of samples for handling tiny objects, which supports that our network has sufficient ability to detect challenging objects with tiny scale. Then the second row shows that objects with large scale (such as overpasses and playgrounds) can also be successfully detected by our network. Therefore, Fig. 9 also shows that the network is suitable for detecting objects with cross-scale. In summary, these visualization experiments also prove that our proposed network has advantages in detecting symmetrical remote sensing objects.

IV. Conclusion

This article proposes CANet for object detection in remote sensing images, which is a single-stage anchor-free detector based on per-pixel prediction. In the network, the CAM is designed as the core structure and plugged into the pyramidal feature network. The novel model could capture symmetric characteristics of remote sensing objects by highlighting, perceiving, and refining central semantics from hierarchical features in the remote sensing images. The experiment shows that our proposed network can outperform some current state-of-the-art methods. Moreover, we conduct ablation experiments by constructing different paired components in CAM and combining various effective items for hybrid loss function. The results of ablation studies verify that the CAM can embed into the FPN-based single-stage detection network, and the flexible model can promote FPN to effectively describe symmetrical objects. In the future, we will further explore new methods that utilize symmetrical characteristics of objects to handle more detection tasks in complex situations. Specifically, based on the current model, we are considering how to design novel losses that directly describe the symmetry of objects, the purpose of which is to further help the network enhance the perception of symmetric remote sensing objects.

Acknowledgment

The authors would like to thank Cheng Gong from Northwestern Polytechnical University and Long Yang from Wuhan University for providing the NWPU VHR-10 and RSOD data sets in their study, respectively.

References

[1] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*. [Online]. Available: http://arxiv.org/abs/1905.05055

[2] Z. Shao, Y. Pan, C. Diao, and J. Cai, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 4062–4076, Jun. 2019.

[3] P. Ren, M. Xu, Y. Yu, F. Chen, X. Jiang, and E. Yang, "Energy minimization with one dot fuzzy initialization for marine oil spill segmentation," *IEEE J. Ocean. Eng.*, vol. 44, no. 4, pp. 1102–1115, Oct. 2019.

[4] X. Wang, Z. Shao, X. Zhou, and J. Liu, "A novel remote sensing image retrieval method based on visual salient point features," *Sensor Rev.*, vol. 34, no. 4, pp. 349–359, Aug. 2014.

[5] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with SVD networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5832–5845, Oct. 2016.

[6] Q. Wang, Z. Yuan, Q. Du, and X. Li, "GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 3–13, Jan. 2019.

[7] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8232–8241.

[8] 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. (CVPR)*, Jul. 2017, pp. 2117–2125.

[9] X. Ying *et al.*, "Multi-attention object detection model in remote sensing images based on multi-scale," *IEEE Access*, vol. 7, pp. 94508–94519, 2019.

[10] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2019.

[11] E. Liu, Y. Zheng, B. Pan, X. Xu, and Z. Shi, "DCL-net: Augmenting the capability of classification and localization for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 18, 2021, doi: 10.1109/TGRS.2020.3048384.

[12] J. Wang, Y. Wang, Y. Wu, K. Zhang, and Q. Wang, "FRPNet: A feature-reflowing pyramid network for object detection of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, early access, Dec. 8, 2020, doi: 10.1109/LGRS.2020.3040308.

[13] C. Xu, C. Li, Z. Cui, T. Zhang, and J. Yang, "Hierarchical semantic propagation for object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4353–4364, Jun. 2020.

[14] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.

[15] H. Qin, Y. Li, J. Lei, W. Xie, and Z. Wang, "A specially optimized one-stage network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 401–405, Mar. 2020.

[16] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[17] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2020.

[18] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[19] L. Zhang, Y. Wang, and Y. Huo, "Object detection in high-resolution remote sensing images based on a hard-example-mining network," *IEEE Trans. Geosci. Remote Sens.*, early access, 2020, doi: 10.1109/TGRS.2020.3038673.

[20] Y. Wu, K. Zhang, J. Wang, Y. Wang, Q. Wang, and Q. Li, "CDD-net: A context-driven detection network for multiclass object detection," *IEEE Geosci. Remote Sens. Lett.*, early access, Dec. 22, 2020, doi: 10.1109/LGRS.2020.3042465.

[21] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 840–849.

[22] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyond anchor-based object detector," 2019, *arXiv:1904.03797*. [Online]. Available: http://arxiv.org/abs/1904.03797

[23] C. Zhu, F. Chen, Z. Shen, and M. Savvides, "Soft anchor-point object detection," 2019, *arXiv:1911.12448*. [Online]. Available: http://arxiv.org/abs/1911.12448

[24] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[25] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[26] X. Zhou, D. Wang, and P. Krahenbuhl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: https://arxiv.org/abs/1904.07850

[27] X. Sun, Y. Liu, Z. Yan, P. Wang, W. Diao, and K. Fu, "SRAF-Net: Shape robust anchor-free network for garbage dumps in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, early access, Oct. 5, 2020, doi: 10.1109/TGRS.2020.3023928.

[28] Y. Lin, P. Feng, and J. Guan, "IENet: Interacting embranchment one stage anchor free detector for orientation aerial object detection," 2019, *arXiv:1912.00969*. [Online]. Available: http://arxiv.org/abs/1912.00969

[29] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 7, 2020, doi: 10.1109/TGRS.2020.3005151.

[30] J. Chen, F. Xie, Y. Lu, and Z. Jiang, "Finding arbitrary-oriented ships from remote sensing images using corner detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1712–1716, Oct. 2019.

[31] J. Zhang, C. Xie, X. Xu, Z. Shi, and B. Pan, "A contextual bidirectional enhancement method for remote sensing image object detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4518–4531, 2020.

[32] J. Lei, X. Luo, L. Fang, M. Wang, and Y. Gu, "Region-enhanced convolutional neural network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5693–5702, Aug. 2020.

[33] C. Tao, L. Mi, Y. Li, J. Qi, Y. Xiao, and J. Zhang, "Scene context-driven vehicle detection in high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7339–7351, Oct. 2019.

[34] 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.*, early access, Oct. 26, 2020, doi: 10.1109/TGRS.2020.3030990.

[35] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2020.

[36] M. Zhou, Z. Zou, Z. Shi, W.-J. Zeng, and J. Gui, "Local attention networks for occluded airplane detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 381–385, Mar. 2020.

[37] Q. Guo, H. Wang, and F. Xu, "Scattering enhanced attention pyramid network for aircraft detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, early access, Nov. 10, 2020, doi: 10.1109/TGRS.2020.3027762.

[38] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[39] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, Jul. 2019.

[40] 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. (ICCV)*, Oct. 2017, pp. 2999–3007.

[41] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[42] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 447–456.

[43] 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 J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[44] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[45] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[46] 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.

[47] 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.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[49] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[50] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[51] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[52] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2965–2974.

[53] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520.

[54] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019.

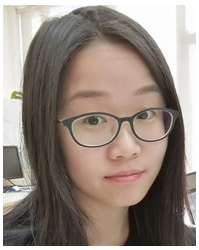
Lukui Shi received the B.S. degree in computer and application and the M.S. degree in computer application technology from the Hebei University of Technology, Tianjin, China, in 1996 and 1999, respectively, and the Ph.D. degree in computer application technology from Tianjin University, Tianjin, in 2006.

Since 2014, he has been a Professor with the School of Artificial Intelligence, Hebei University of Technology, China. He has authored two books and more than 20 articles. His research interests includes machine learning, lung sound recognition, and data digging.

Dr. Shi was a member of the Discrete Intelligent Computing Professional Committee of Chinese Association of Artificial Intelligence and a member of Visual Big Data Professional Committee of China Society of Image and Graphics.

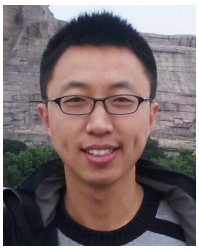
Linyi Kuang received the B.S. degree in software engineering from the Hebei University of Technology, Tianjin, China, in 2018, where he is pursuing the M.S. degree with the School of Artificial Intelligence.

His research interests include machine learning and intelligent computing.



Xia Xu received the B.S. and M.S. degrees from the School of Electrical Engineering, Yanshan University, Qinhuangdao, China, in 2012 and 2015, respectively, and the Ph.D. degree from the School of Astronautics, Beihang University, Beijing, China, in 2019.

She is working as an Assistant Professor with the College of Computer Science, Nankai University, Tianjin, China. Her research interests include hyperspectral unmixing, multiobjective optimization, and remote sensing image processing.



Bin Pan (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Astronautics, Beihang University, Beijing, China, in 2013 and 2019, respectively.

Since 2019, he has been an Associate Professor with the School of Statistics and Data Science, Nankai University, Tianjin, China. His research interests include machine learning, remote sensing image processing, and multiobjective optimization.

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

From 2005 to 2007, he was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China. From 2013 to 2014, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA. He is a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or coauthored more than 100 scientific articles in related journals and proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE Conference on Computer Vision and Pattern Recognition. His research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi received the Best Reviewer Awards for his service to the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) in 2017. He has been an Associate Editor for the *Infrared Physics and Technology* since 2016.