CFC-Net: A Critical Feature Capturing Network for Arbitrary-Oriented Object Detection in Remote-Sensing Images

Qi Ming, Lingjuan Miao, Zhiqiang Zhou, *Member, IEEE*, and Yunpeng Dong

***Abstract*— Object detection in optical remote-sensing images is an important and challenging task. In recent years, the methods based on convolutional neural networks (CNNs) have made good progress. However, due to the large variation in object scale, aspect ratio, as well as the arbitrary orientation, the detection performance is difficult to be further improved. In this article, we discuss the role of discriminative features in object detection, and then propose a critical feature capturing network (CFC-Net) to improve detection accuracy from three aspects: building powerful feature representation, refining preset anchors, and optimizing label assignment. Specifically, we first decouple the classification and regression features, and then construct robust critical features adapted to the respective tasks of classification and regression through the polarization attention module (PAM). With the extracted discriminative regression features, the rotation anchor refinement module (R-ARM) performs localization refinement on preset horizontal anchors to obtain superior rotation anchors. Next, the dynamic anchor learning (DAL) strategy is given to adaptively select high-quality anchors based on their ability to capture critical features. The proposed framework creates more powerful semantic representations for objects in remote-sensing images and achieves high-performance real-time object detection. Experimental results on three remote-sensing datasets including HRSC2016, DOTA, and UCAS-AOD show that our method achieves superior detection performance compared with many state-of-the-art approaches. Code and models are available at https://github.com/ming71/CFC-Net.**

***Index Terms*— Convolutional neural networks (CNNs), critical features, deep learning, object detection.**

I. Introduction

OBJECT detection in optical remote-sensing images is a vital computer vision technique which aims at classifying and locating objects in remote-sensing images. It is widely used in crop monitoring, resource exploration, environmental monitoring, military reconnaissance, etc. With the explosive growth of available remote-sensing data, identifying objects of interest from massive amounts of remote-sensing imagery has gradually become a challenging task. Most of the traditional methods use handcrafted features to identify objects [1]–[5]. Although much progress has been made, there are still problems such as low efficiency, insufficient robustness, and poor performance.

In recent years, the development of convolution neural networks (CNNs) has greatly improved the performance of object detection. Most CNN-based detection frameworks first extract features through convolution operation, and then preset a series of prior boxes (anchors) on the feature maps. Subsequently, classification and regression are performed on these anchors to obtain the bounding boxes of objects. The powerful ability to automatically extract features of CNN makes it possible to achieve promising object detection on massive images. Currently, the CNN-based models have been widely used in the object detection in remote-sensing images, such as road detection [6], vehicle detection [7], [8], airport detection [9], and ship detection [10], [11].

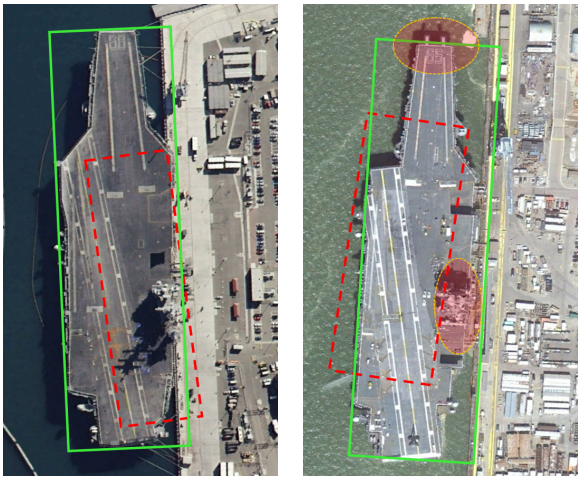
Although CNN-based approaches have made good progress, they are often directly derived from generic object detection frameworks. It is difficult for these methods to detect objects with a wide variety of scales, aspect ratios, and orientations in remote-sensing images. For example, the orientation of objects varies greatly in remote-sensing imagery, while the mainstream generic detectors utilize predefined horizontal anchors to predict these rotated ground-truth (GT) boxes. The horizontal boxes often contain a lot of background which may mislead the detection. There are some approaches that use rotated anchors to locate arbitrary-oriented objects [12]–[19], but it is hard for rotation anchors to achieve good spatial alignment with GT boxes, and they cannot ensure to provide sufficiently good semantic information for classification and regression.

Some recent researches address the above problems by designing more powerful feature representations [17], [18], [20]–[23]. However, they only focus on a certain type of characteristics of remote-sensing targets, such as rotation invariant features [20], [21] and scale sensitive features [22], [23]. They cannot automatically extract and utilize more complex and discriminative features. Another commonly used method is to manually set a large number of anchors covering different aspect ratios, scales, and orientations to achieve better spatial alignment with targets. In this way, sufficient high-quality anchors can be obtained and better performance can be achieved. Excessive preset anchors, however, bring about three problems: 1) most anchors are backgrounds that cannot be used for bounding box regression, which leads to severely

Manuscript received December 30, 2020; revised May 27, 2021; accepted June 25, 2021. *(Corresponding author: Zhiqiang Zhou.)*

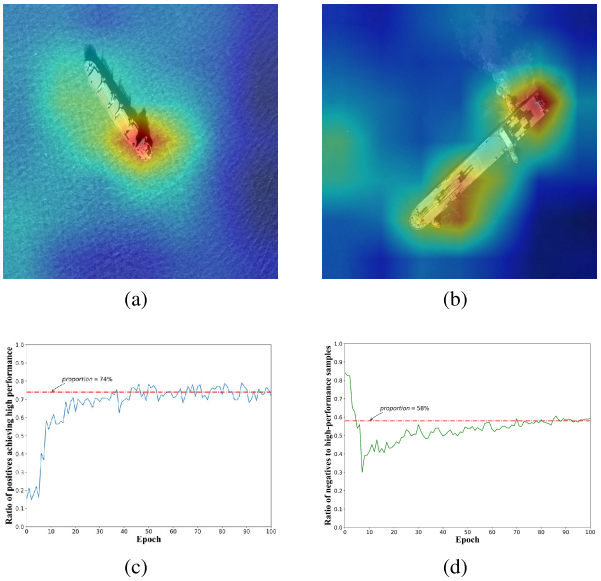
The authors are with the School of Automation, Beijing Institute of Technology, Beijing 100081, China (e-mail: chaser.ming@gmail.com; miaolingjuan@bit.edu.cn; zhzhzhou@bit.edu.cn; bitdyp@gmail.com).

Digital Object Identifier 10.1109/TGRS.2021.3095186



(a) (b)

Fig. 1. Illustration of the role of critical features in classification task. Predicted bounding boxes (green) are regressed from predefined anchor boxes (red). The ground truth classes of (a) and (b) are marked as A and B, respectively, while the predicted object categories are all A. Only the anchors that capture the critical features required to identify the object (such as island and bow here) can achieve the correct classification.



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

Fig. 2. Analysis of the importance of critical features. (a) and (b) Discriminative feature activation map in object detection. (c) Proportion of positive samples with high-quality detections among all positives. (d) Proportion of high-quality detections that regressed from negatives.

redundant calculation; 2) the parameters of the prior anchors need to be careful manually set, otherwise they would not obtain good alignment with GT boxes; and 3) there are a large number of low-quality negative samples in the excessive laid anchors which are not conducive to network convergence. The above-mentioned issues lead to the fact that densely preset anchors are still unable to effectively handle the difficulties of remote-sensing object detection.

To figure out how the complex variabilities of remote-sensing objects make it difficult to achieve high-performance detection, in this article we introduce the essential concept named critical features, referring to the discriminative features required for accurate classification or localization. Taking the classification task as an example, most anchor-based detectors treat the anchors in Fig. 1(a) and (b) as positive samples, due to that the Intersection-over-Union (IoU) between these anchors and GT boxes is higher than 0.5. But the anchor in Fig. 1(b) does not capture the discriminative features of the island and bow which are critical for precise ship classification. Although this anchor achieves accurate localization, it leads to incorrect classification results, thereby degrading detection performance. Furthermore, by visualizing the features extracted by CNN, it is found that the critical features for classification and regression are not always evenly distributed on the object, but may be on local areas such as the bow and stern [see Fig. 2(a) and (b)]. The preset anchors need to capture these critical features to achieve accurate detection. This is similar to the conclusion of some previous work [10], [24]. However, the mainstream rotation detectors tend to select anchors with high IoU with GT boxes as positives, but ignore high-quality anchors that contain critical features, which eventually leads to the unstable training process and poor performance. The statistics given in Fig. 2(c) supports this viewpoint. It can be seen that only 74% of positive anchors can achieve high-quality detection (with output IoU larger than 0.5) after regression, which indicates that even the positive anchors still cannot guarantee precise localization. We attribute this phenomenon to the fact that some of the selected positives do not capture the critical features required by the regression task. Besides, as shown in Fig. 2(d), surprisingly more than half of the anchors (about 58% in this case) that achieve accurate detection are regressed from the samples that are recognized as negatives. It means that a large number of negative anchors capture the critical features well but have not been effectively utilized at all. The inconsistency between the training sample division and the regression results will further lead to a gap between the classification scores and localization accuracy in the detection. Based on the above observations, we conclude that one of the key issues for object detection in remote-sensing imagery is whether the anchors can capture the critical features of the objects.

In this article, based on the viewpoint discussed above, the critical feature capturing network (CFC-Net) is proposed to achieve high-performance object detection in optical remote-sensing imagery. Specifically, CFC-Net first uses a well-designed polarization attention module (PAM) to generate different feature pyramids for classification and regression tasks, and then we can obtain task-specific critical features that are more discriminative and easy-to-capture. Next, the rotation anchor refinement module (R-ARM) refines the preset horizontal anchors to better capture the regression critical features to obtain high-quality rotation anchors. Finally, in the training process, the dynamic anchor learning (DAL) strategy is adopted to select the high-quality anchors that capture critical features as positives to ensure superior detection performance after training. Due to the proper construction and utilization of critical features, CFC-Net achieves the state-of-the-art detection performance using only one anchor, which

makes it became a both high-performance and memory-saving method. The code is available to facilitate future research.

The contributions of this article are summarized as follows.

1) We point out the existence of critical features through experiments, and interpret common challenges for object detection in remote-sensing imagery from this perspective. A novel object detection framework CFC-Net is then proposed to capture the critical features to achieve superior detection performance.
2) Polarized attention is proposed to construct task-specific critical features. Decoupled critical features provide more useful semantic information for individual tasks, which is beneficial to accurate classification and regression.
3) The dynamic anchor selection (DAS) strategy selects high-quality anchors that capture the critical regression features to bridge the inconsistency between classification and regression, and thus greatly improves the performance of detection.

The rest of this article is organized as follows. Section II introduces the related work of object detection. Section III elaborates on the proposed method. Section IV provides the experimental results and analysis. Finally, conclusions are drawn in Section V.

II. Related Work

Object detection in remote-sensing images has a wide range of application scenarios and has been receiving extensive attention in recent years. Most of the early traditional methods use handcraft features to detect remote-sensing objects, such as shape and texture features [1], [4], [5], scale-invariant features [2], and saliency [3]. For instance, Zhu *et al.* [4] achieves accurate ship detection based on shape and texture features. Eikvil *et al.* [5] utilizes spatial geometric properties and gray level features for vehicle detection in satellite images. These approaches have achieved satisfactory performance for specific scenes, but their low efficiency and poor generalization make it hard to detect objects in complex scenarios.

Recently, with the great success of CNNs, generic object detection has been strongly promoted. Mainstream CNN-based object detection methods can be classified into two categories: one-stage detectors and two-stage detectors. The two-stage detectors first generate a series of proposals, and then perform classification and regression on these regions to obtain the detection results [25]–[27]. These algorithms usually have high accuracy but slow inference speed. The one-stage detectors, such as the YOLO series [28]–[30] and SSD [31], directly conduct classification and regression on the prior anchors without region proposal generation. Compared with the two-stage detectors, one-stage methods have relatively low accuracy, but are faster and can achieve real-time object detection.

Deep learning methods have been widely used in object detection in remote-sensing images. A series of CNN-based approaches have been proposed and achieved good performance. However, some methods are directly developed from the generic object detection framework [22], [32], which detect objects with horizontal bounding box. It is hard for the horizontal box to distinguish densely arranged remote-sensing targets and is prone to misdetection. To solve this problem, some studies introduced an additional orientation dimension to achieve the oriented object detection [12]–[14]. For example, Liu *et al.* [12] integrates the angle regression into the YOLOv2 [29] to detect rotated ships. R^2PN [13] detects rotated ships by generating oblique region of interest (RoI). RR-CNN [14] uses the rotated RoI pooling layer, which makes the RoI feature better aligned with the orientation of the object to ensure accurate detection. However, to have a higher overlap with the rotated objects, these methods preset densely arranged rotation anchors. Most of the anchors have no intersection with the targets, which brings a lot of redundant computation and the severe imbalance problem. Some work alleviates the issue by setting fewer anchors but still maintaining detection performance [15], [33]. RoI Transformer [15] adopts horizontal anchors to learn the rotated RoI through spatial transformation, so that a few horizontal anchors can work well for oriented object detection. R^3Det [33] achieves state-of-the-art performance through cascade regression, in which feature alignment is performed on horizontal anchors. Despite the success of these methods, it is still difficult for horizontal anchors to accurately detect the rotation objects and the number of preset anchors is still large. Different from the previous work, our CFC-Net uses only one anchor for faster inference and achieves high-quality rotation object detection.

There are also some methods trying to construct better feature representation to alleviate the difficulty caused by large scale, shape, and orientation variations [16], [18]–[21], [23], [34], [35]. For instance, Li *et al.* [19] proposed a local-contextual feature fusion model to build powerful joint representations for object detection in remote-sensing images. ORN [21] performs feature extraction through the rotated convolution kernel to achieve rotation invariance. RICNN [20] optimizes the feature representation by learning a rotation-invariant layer. FMSSD [23] aggregates the context information in different scales to cope with the multi-scale objects in large-scale remote-sensing imagery. Li *et al.* [16] proposed a shape-adaptive pooling to extract the features of the ships with various aspect ratios, and then multilevel features are incorporated to generate a compact feature representation for ship detection. RRD [17] observes that shared features degrade performance due to the incompatibility of the classification and regression tasks. To solve the problem, the rotation-invariant and rotation-sensitive features are constructed for classification and regression tasks, respectively. The current work only pays attention to a certain aspect of the object characteristics, and cannot comprehensively cover the discriminative features required for object detection. According to the analysis in Section I, we believe that the detection performance largely depends on whether the prior anchors effectively capture these critical features, not limited to the rotation-invariant features or scale-invariant features. Therefore, the clear and powerful critical feature representation is very important for object detection. The proposed CFC-Net extracts and utilizes task-sensitive critical features for classification and regression tasks respectively, so that the

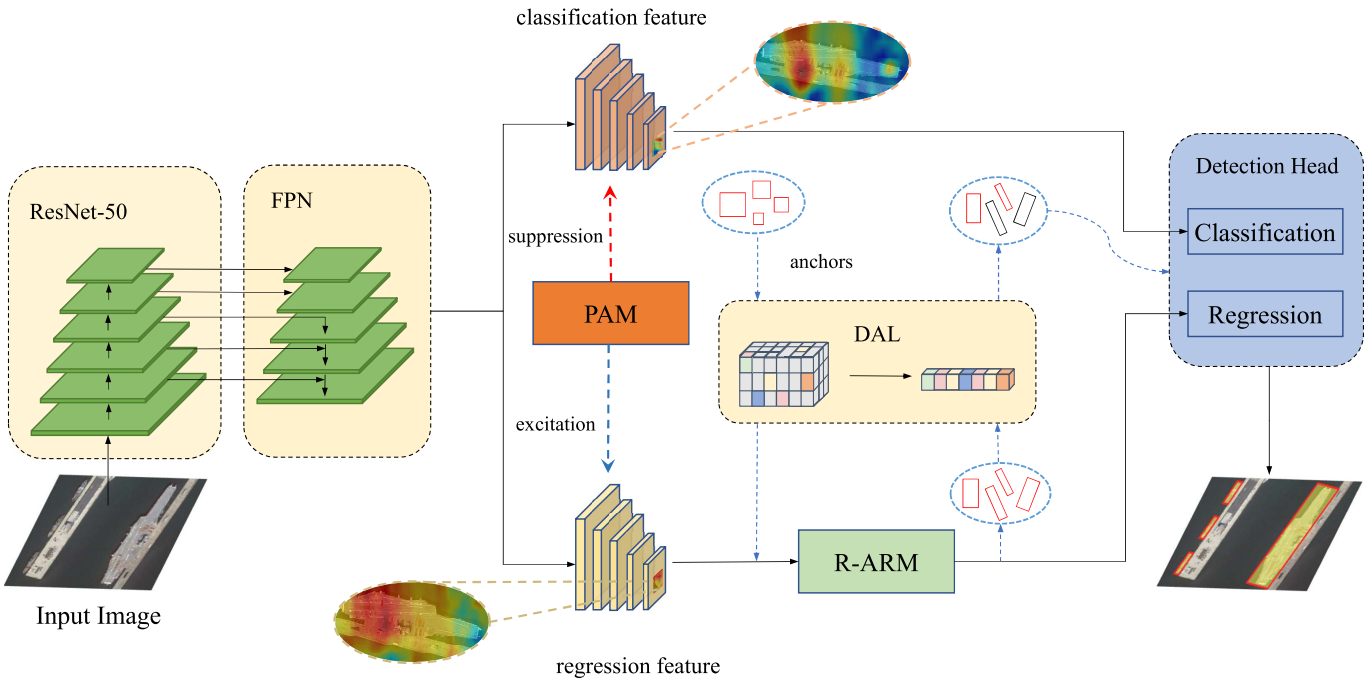


Fig. 3. Framework of the proposed CFC-Net.

detector can obtains substantial performance improvements from the more discriminative critical feature representation.

III. Proposed Method

The overall structure of CFC-Net is shown in Fig. 3. We uses ResNet-50 as the backbone network. First, we build multi-scale feature pyramids through feature pyramid network (FPN) [36], and then the decoupled features that are sensitive to classification and regression are generated through the proposed PAM. Subsequently, anchor refinement is conducted via R-ARM to obtain the high-quality rotation candidates based on the critical regression features. Finally, DAL strategy dynamically selects anchors that capture critical features for regression. In this way, the inconsistency between classification and regression can be alleviated and the detection performance can be improved. The details of the proposed CFC-Net are elaborated below.

A. Polarization Attention Module

In most object detection frameworks, both classification and regression rely on the shared features. However, as mentioned in [17] and [37], the shared features degrade performance owing to the incompatibility between the two tasks. For example, the classification branch is supposed to have the rotation invariance for different angles, while the regression branch of detectors should be sensitive to the change of the angle so as to achieve accurate orientation prediction. Therefore, rotation-invariant features are beneficial to classification task, but it is not conducive to bounding box regression.

We propose PAM to avoid the feature interference between different tasks and effectively extract the task-specific critical features. The overall structure of PAM is shown in Fig. 4. First, we build separate feature pyramids for different tasks, which is called dual FPN. Next, a well-designed polarization attention mechanism is applied to obtain the enhanced feature representation. Through the polarization function, different branches generate the discriminative features required for respective tasks. Specifically, for classification, we tend to select high-response global features to reduce noise interference. For regression, we pay more attention to the features of object boundaries and suppress the influence of irrelevant high activation regions.

Given the input feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$, we construct task-sensitive features as follows:

$$\begin{aligned}\mathbf{M} &= \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{M_s}(\mathbf{F}) \\ \mathbf{F}' &= \mathbf{M} + \psi(\sigma(\mathbf{M})) \odot \mathbf{F} + \mathbf{F}\end{aligned} \tag{1}$$

where $\otimes$ and $\odot$ represent tensor product and element-wise multiplication, respectively. σ denotes sigmoid function. First, we extract channel-wise attention map $\mathbf{M_c}$ and spatial attention map $\mathbf{M_s}$ from input features through convolution operations. The purpose of channel attention is to extract the channel-wise relationship of the feature maps. The weight of each channel is extracted by global average pooling and fully connected layers as follows:

$$\mathbf{M_c}(\mathbf{F}) = \sigma\left(\mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F_{gap}}\right)\right)\right) \tag{2}$$

where $\mathbf{F_{gap}}$ is obtained from input feature $\mathbf{F}$ via global average pooling, $\mathbf{W_0} \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W_1} \in \mathbb{R}^{C\times C/r}$ are the weights of the fully connected layers.

Besides, spatial attention is used to model the dependencies between pixels of the input image. The formula is as follows:

$$\mathbf{M_s}(\mathbf{F}) = \sigma\left(c^{3\times3}\left(\text{cat}\left(\left(c^{3\times3}, c_d^{1\times3}, c_d^{3\times1}, c_d^{3\times3}\right)(\mathbf{F})\right)\right)\right) \tag{3}$$

in which $c^{3\times3}$ represents convolution of 3×3 filters. $c_d^{1\times3}, c_d^{3\times1}, c_d^{3\times3}$ respectively denote dilated convolution of

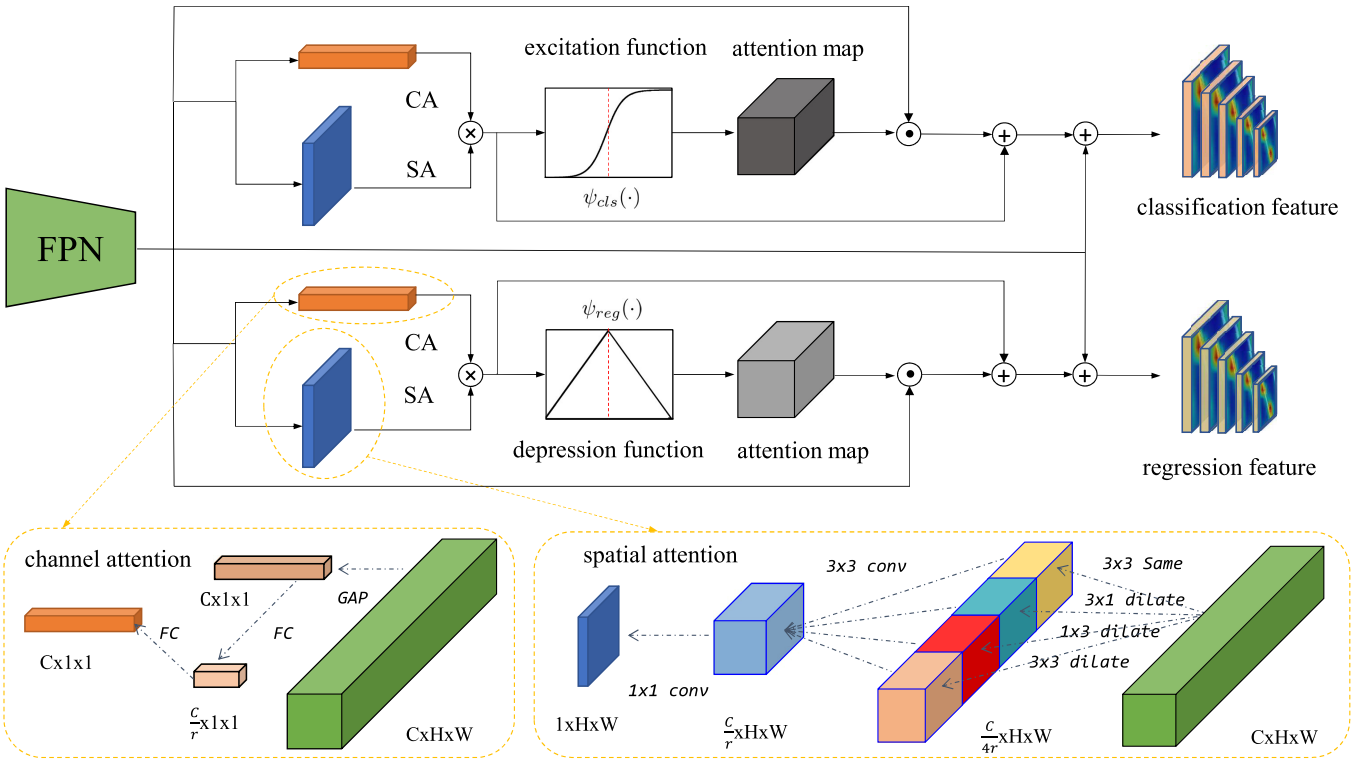


Fig. 4. Illustration of the PAM module. SA denotes spatial attention and CA represents channel-wise attention.

different kernel sizes, and their dilation rates are 2. Cat denotes concatenation of features. Dilated convolution is adopted here to expand the receptive field of the convolution kernels. At the same time, convolution kernels with different aspect ratios are used to better detect slender objects such as ships and bridges.

Next, the attention response map $\mathbf{M}$ for a specific task is obtained by multiplying the two attention maps. On this basis, we build the powerful task-sensitive critical feature representation through the task-specific polarization function $\psi(\cdot)$. For classification, the features are expected to pay more attention to the high-response part on feature maps, and ignore the part of less important clues which may be used for localization or may bring interference noise. We use the excitation function as follows:

$$\psi_{\text{cls}}(x) = \frac{1}{1 + e^{-\eta(x-0.5)}} \tag{4}$$

where η is the modulation factor used to control the intensity of feature activation (set to 15 in our experiment). The high-response area of critical classification features is enough to achieve accurate classification, there is no need to pursue too much information. Consequently, the effect of high-response critical classification features is excited, while irrelevant features with attention weight less than 0.5 are suppressed. In this way, the classifier is able to pay less attention to the difficult-to-classify areas and reduce the risk of overfitting and misjudgment.

Meanwhile, for the regression branch, the critical features are often scattered on the edges of object. We expect that the feature maps focus on as many visual clues as possible for object localization, such as object contours and contextual information. To this end, we use the following depression function to process the input features:

$$\psi_{\text{reg}}(x) = \begin{cases} x, & \text{if } x < 0.5 \\ 1 - x, & \text{otherwise.} \end{cases} \tag{5}$$

Different from the classification task, a strong response to a patch of the object edge is not conducive to locating the entire object. In (5), the depression function suppresses the area with the high response in the regression feature, which enforces the model to seek potential visual clues to achieve accurate localization. The curves of polarization function $\psi(\cdot)$ are shown in Fig. 4.

Finally, the polarization attention weighted features are combined with the original features to extract the enhanced critical features. As described in (1), the attention weighted features, the input features $\mathbf{F}$, and the attention response map $\mathbf{M}$ are merged by element-wise summation to obtain powerful feature representations for accurate object detection. The proposed PAM greatly improves detection performance via optimizing the representation of critical features. The explainable visualization results are shown in Fig. 5. It can be seen that PAM effectively extracts the critical features required for different tasks. For example, the extracted regression critical features are evenly distributed on the object, which is helpful to identify the object boundary and accurately localize the target. The classification critical features are concentrated more on the most recognizable part of an object to avoid interference from other parts of the object, and thus the classification results will be more accurate.

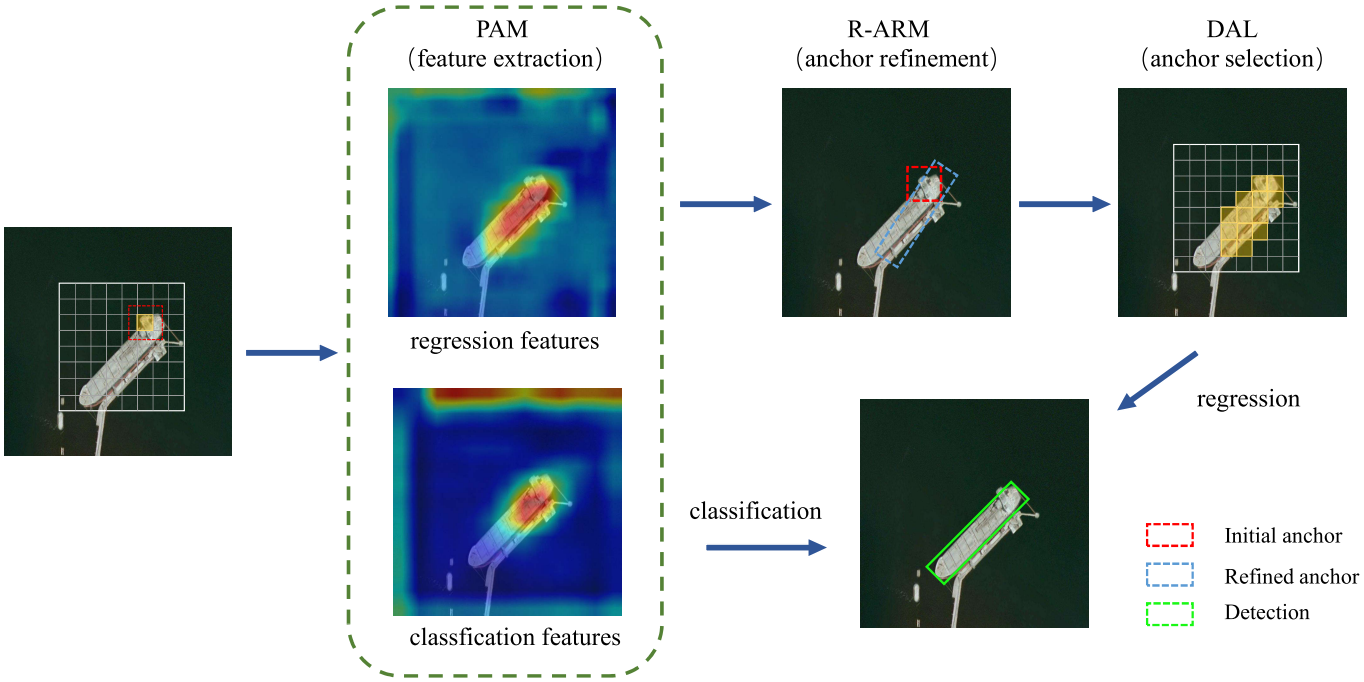


Fig. 5. Illustration of function of the proposed modules in the detection pipeline. The yellow area represents the center of the high-quality anchors.

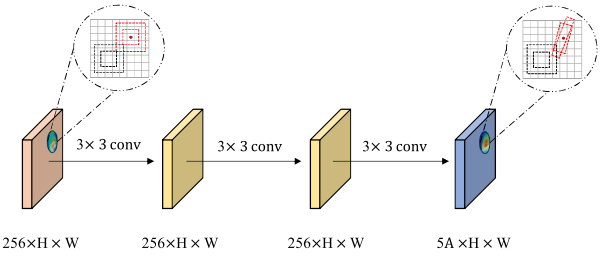


Fig. 6. Illustration of the R-ARM module. A denotes the number of anchors preset at each position of feature map, which is set to 1 in CFC-Net. Red boxes are the positive samples for anchor refinement.

B. Rotation Anchor Refinement Module

In the existing anchor-based object detectors, classification and regression are performed on densely preset anchors. It is difficult to achieve alignment between anchors and rotation objects owing to the large variation in the scale and orientation of the remote-sensing objects. To solve this problem, we proposed a R-ARM to generate high-quality candidates based on critical regression features with less reliance on the prior geometric knowledge of anchors. Given the regression-sensitive feature maps extracted by PAM, R-ARM refines the initial anchors to obtain the rotated anchors that better align with the critical regression features. The regions of these high-quality anchors capture the discriminative and semantic features of the object boundary, which helps to achieve accurate localization.

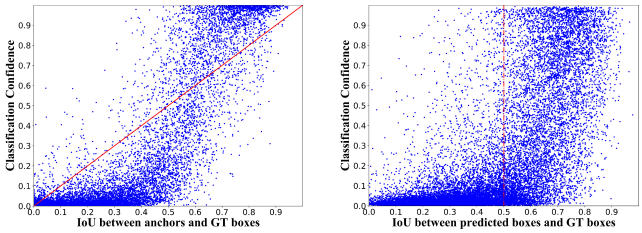
The architecture of R-ARM is shown in Fig. 6. It is stacked by three convolutional layers with 3×3 convolution kernels. The first two convolutional layers contains 256 filters in each layer to extract the anchor refining features from the input regression features. The last one predicts the objects at each location on the feature maps. We preset A initial horizontal anchors at each position of the feature map, which are represented as $(x^a, y^a, w^a, h^a, \theta^a)$. (x^a, y^a) are the center coordinates, and w^a, h^a denote the width and height of the anchors, respectively. $\theta^a = 0$ for the preset horizontal anchors. R-ARM regresses the angle θ and the box offsets of the prior anchors to get the rotation candidates which are expressed as (x, y, w, h, θ). R-ARM enables anchors to generate refined rotated boxes that are well aligned with the GT objects, and would help to capture more critical features for subsequent detection layers. Specifically, we predict offsets $\boldsymbol{t^r} = (t_x, t_y, t_w, t_h, t_\theta)$ for anchor refinement, which are represented as follows:

$$\begin{aligned} t_x^r &= \left(x - x^a\right)/w^a, \quad t_y^r = \left(y - y^a\right)/h^a \\ t_w^r &= \log\left(w/w^a\right), \quad t_h^r = \log\left(h/h^a\right) \\ t_\theta^r &= \tan\left(\theta - \theta^a\right) \end{aligned} \tag{6}$$

where x and x^a are for the refined box and anchor respectively (likewise for y, w, h, θ).

In CFC-Net, we set $A = 1$, which means that only one initial anchor is used. Therefore, we do not need to carefully set the hyperparameters of angle, aspect ratio, and scale for anchors such as the current anchor-based methods owing to the R-ARM. It is noted also that we do not integrate classification prediction in R-ARM such as some cascade regression approaches [33], [38]. This is due to the following considerations.

1) Classification in the refining stage is not accurate enough. As a result, the model may mistakenly exclude the potential high-quality candidates, leading to a poor recall of detections.
2) As mentioned in Section I, there is a gap between classification and regression. The high classification



(a) (b)

Fig. 7. Analysis of the classification and regression capabilities of anchors that use input IoU for label assignment. (a) Before regression. (b) After regression.

score does not guarantee accurate localization. Therefore, the training sample selection based on classification confidence in anchor refinement will degrade the detection performance.

Compared with previous one-stage detectors, CFC-Net needs fewer predefined anchors, but achieves better detection performance with the R-ARM. As illustrated in Fig. 5, guided by the critical regression features generated by PAM, the initial square anchor produces a more accurate rotated candidate via R-ARM. The refined anchor aligns well with the high-response region that captures critical features, which provides an effective semantic prior for subsequent localization.

C. Dynamic Anchor Learning

In the previous sections, we have introduced the critical feature extraction structure and high-quality anchor generation in CFC-Net. However, the misalignment between classification and regression tasks still exists, that is, the high classification scores cannot guarantee the accurate localization of the detections. This issue has been widely discussed in many studies [39]–[42]. Some of the work attributed it to the regression uncertainty [40], [42], which reveals that the localization results obtained by the regression are not completely credible. We believe that the gap between classification and regression mainly comes from unreasonable training sample selection [43], and further solve this problem from the perspective of critical features.

Current detectors usually select positive anchors for training according to the IoU between anchors and GT boxes. For simplicity, we denote the IoU between anchors and GT boxes as $\mathrm{IoU_{in}}$, while the IoU between the predicted boxes and GT boxes as $\mathrm{IoU_{out}}$. The selected positive anchors are supposed to have good semantic information which is conducive to object localization. As shown in Fig. 7(a), there is a positive correlation between the classification score and the $\mathrm{IoU_{in}}$. However, the high $\mathrm{IoU_{in}}$ does not guarantee good localization potential of the anchors, and as shown in Fig. 7(b), there is only a weak correlation between the classification confidence and localization capability of predicted boxes. We suggest that one of the main causes is that the samples selected according to the $\mathrm{IoU_{in}}$ do not align well with the critical features of the objects.

To resolve the above problems, a DAL method is adopted to select samples with strong critical feature capturing ability in the training phase. DAL consists of two parts: DAS and matching-sensitive loss (MSL). The rest of this section will elaborate on the implementation of the two strategies.

First, we adopt a new standard called matching degree to guide training sample division. It is defined as follows:

$$md = \alpha \cdot \mathrm{IoU_{in}} + (1-\alpha) \cdot \mathrm{IoU_{out}} - u^{\gamma} \tag{7}$$

in which $\mathrm{IoU_{in}}$ and $\mathrm{IoU_{out}}$ are the IoUs between the anchor box and the GT box before and after regression, respectively. α and γ are hyperparameters used to weight the influence of different items. u is the penalty term used to suppress the uncertainty during the regression process. The matching degree combines the prior information of spatial alignment, critical feature alignment ability, and regression uncertainty of the anchor to measure its localization capacity. Specifically, for a predefined anchor and its assigned GT box, $\mathrm{IoU_{in}}$ denotes the initial spatial alignment ability, while $\mathrm{IoU_{out}}$ indicates the critical feature alignment ability. Intuitively, higher $\mathrm{IoU_{out}}$ means that the anchor better captures critical regression features and has a stronger localization potential. However, actually, this indicator is unreliable due to the regression uncertainty. For example, the high-quality anchors with high $\mathrm{IoU_{in}}$ but low $\mathrm{IoU_{out}}$ may be mistakenly judged as negative samples [43] in the early stages of training.

Therefore, in (7) we further introduce the penalty term u to alleviate the influence from regression uncertainty. It is defined as follows:

$$u = |\mathrm{IoU_{in}} - \mathrm{IoU_{out}}|. \tag{8}$$

The change of IoU after regression indicates the probability of incorrect anchor assessment, and we use it to measure regression uncertainty. Uncertainty suppression item u imposes the penalty on samples with excessive IoU change after regression to ensure a reasonable training sample selection. We will confirm in the experimental part that the suppression of uncertainty during regression is the key to take advantage of the critical features.

With the evaluation of the matching degree, we can conduct better training sample selection. We first calculate the matching degree between all anchors and GT boxes in the images, and then candidates with matching degree higher than a certain threshold (set to 0.6 in our experiment) are selected as positive samples, while the rest are negatives. Next, for objects that are not assigned with any positives, the candidate with the highest matching degree would be selected as a positive sample.

The matching degree measures the ability of feature alignment. Therefore, the division of positive and negative samples is more reasonable, it would alleviate the misalignment between the classification and regression. It can be seen from Fig. 5 that DAL dynamically selects anchors that capture the critical regression features for bounding box regression. These high-quality candidates obtain accurate localization performance after the regression by alleviating the misalignment between classification and regression tasks.

We further integrate matching degree into the training process to construct a MSL for high-precision object detection.

The classification loss is as follows:

$$L_{\text{cls}} = \frac{1}{N_n}\sum_{i\in\psi_n} FL\left(p_i, p_i^*\right) + \frac{1}{N_p}\sum_{j\in\psi_p}(w_j+1)\cdot FL\left(p_j, p_j^*\right) \tag{9}$$

in which N_n and N_p inidcates the number of all negative and positive anchors, respectively. ψ_n and ψ_p respectively represent negative and positive samples. $FL(\cdot)$ is focal loss defined as RetinaNet [44]. p^* is the classification label for anchor ($p^* = 1$ if it is positive, while $p^* = 0$ otherwise). w_j represents the weighting factor, which is utilized to distinguish positive candidates with different localization ability. For a given target g, we first calculate its matching degrees (denoted by md) with all preset anchors. Then we select the matching degrees of positives (denoted by $\boldsymbol{md}_{\text{pos}}$, and $\boldsymbol{md}_{\text{pos}} \subseteq \boldsymbol{md}$). Assuming that the maximum value of $\boldsymbol{md}_{\text{pos}}$ is $md_{\max}$, we define a compensation value Δmd as follows:

$$\Delta md = 1 - md_{\max}. \tag{10}$$

Subsequently, Δmd is added to the matching degree of all positive candidates to obtain the weighting factor

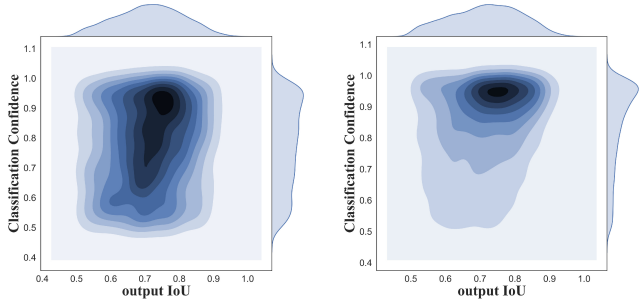
$$\boldsymbol{w} = \boldsymbol{md}_{\text{pos}} + \Delta md. \tag{11}$$

The weighting factor improves the contribution of the positive samples to the loss during training. In this way, the classification branch can discriminate anchors with different capabilities to capture critical features. Compared with the commonly used method that treats all positive anchors equally, this discriminative approach helps to distinguish positive samples of different localization ability. By introducing the localization information of anchors into the classification loss, the classifier can output more reliable classification confidence to select the detections with good localization, thereby bridging the gap between classification and regression.

Since matching degree measures the localization ability of anchors, it can be further used to promote high-quality localization. The matching-sensitive regression loss is defined as follows:

$$L_{\text{reg}} = \frac{1}{N_p}\sum_{j\in\psi_p} w_j \cdot L_{\text{smooth}_{L_1}}\left(\boldsymbol{t}_j, \boldsymbol{t}_j^*\right) \tag{12}$$

where $L_{\text{smooth}_{L_1}}$ represents the smooth-L_1 loss [26]. $\boldsymbol{t}$ and $\boldsymbol{t}^*$ are offsets for the predicted boxes and target boxes, respectively. The weighted regression loss adaptively pays more attention to the samples with high localization potential rather than good initial spatial alignment, therefore, high-precision detection would be achieved after training. It can be seen from Fig. 8(a) that the detectors trained with normal smooth-L_1 loss shows a weak correlation between the classification score and localization ability of the detections, which causes the predictions selected by the classification confidence to be unreliable. After training with a MSL function, as shown in Fig. 8(b), the model outputs the detections with both better localization performance and higher classification confidence, facilitating the selection of high-quality detection based on the classification score. The above analysis confirms the effectiveness of the MSL.



(a) (b)

Fig. 8. Correlation between the output IoU and classification score with and without MSL. (a) Without MSL. (b) With MSL.

DAS strategy and MSL can also be employed to the anchor refinement stage, and thus the multitask loss for CFC-Net is defined as follows:

$$L = L_{\text{cls}}(p, p^*) + \lambda_1 L_{\text{ref}}\left(\boldsymbol{t}^{\boldsymbol{r}}, \boldsymbol{t}^*\right) + \lambda_2 L_{\text{reg}}\left(\boldsymbol{t}, \boldsymbol{t}^*\right) \tag{13}$$

where $L_{\text{cls}}(p, p^*)$, $L_{\text{ref}}(\boldsymbol{t}^{\boldsymbol{r}}, \boldsymbol{t}^*)$, and $L_{\text{reg}}(\boldsymbol{t}, \boldsymbol{t}^*)$ are the classification loss, anchor refinement loss, and regression loss, respectively. $\boldsymbol{t}^{\boldsymbol{r}}$, $\boldsymbol{t}$ denotes the predicted offsets of refined anchors and detection boxes, respectively. $\boldsymbol{t}^*$ represents the offsets of GT boxes. These loss items are balanced via parameters λ_1 and λ_2 (we set $\lambda_1 = \lambda_2 = 0.5$ in our experiments).

IV. Experiments

A. Datasets

Experiments are conducted on three public remote-sensing datasets: HRSC2016, DOTA, and UCAS-AOD. The GT boxes in these datasets are annotated with oriented bounding box.

HRSC2016 [45] is a high resolution remote-sensing ship detection dataset with a total of 1061 images. The image sizes range from 300 × 300 to 1500 × 900. The entire dataset is divided into training set, validation set, and test set, including 436, 181, and 444 images, respectively. The images are resized to two scales of 416 × 416 and 800 × 800 in our experiments.

DOTA [46] is the largest publicly available dataset for oriented object detection in remote-sensing images. DOTA includes 2806 aerial images with 1 88 282 annotated instances. There are 15 categories in total, including plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). It is noted that images in DOTA are too large, we crop the original images into 800 × 800 patches with the stride 200 for training and testing.

UCAS-AOD [47] is an aerial aircraft and car detection dataset, which contains 1510 images collected from Google Earth. It includes 1000 planes images and 510 cars images in total. Since there is no official division of this dataset. we randomly divide it into training set, validation set, and test set as 5:2:3. All images in UCAS-AOD are resized to 800 × 800 in the experiments.

TABLE I
EFFECTS OF EACH COMPONENT OF CFC-NET

	Different Variants				
with PAM?	×	✓	×	✓	✓
with DAL?	×	×	✓	✓	✓
with R-ARM?	×	×	×	×	✓
mAP(HRSC2016)	70.5	76.2	78.7	83.8	**86.3**
mAP(DOTA)	68.1	69.2	69.5	70.5	**72.0**

TABLE II
ABLATION STUDY OF THE PROPOSED PAM

	Different Variants			
+ dual FPN	×	✓	✓	✓
+ attention	×	×	✓	✓
+ polarization function	×	×	×	✓
mAP	70.5	72.1	74.9	**76.2**

TABLE III
EVALUATION OF η IN PAM

η	-	5	15	50	150
mAP	75.49	75.41	**76.28**	76.06	74.85

B. Implementation Details

The backbone of our CFC-Net is ResNet-50 [48]. The model is pre-trained on the ImageNet and fine-tuned on remote-sensing image datasets. We utilize the feature pyramid of P_3, P_4, P_5, P_6, P_7 to detect multi-scale objects. For each position of the feature maps, only one anchor is set to regress the nearby objects. We use random flipping, rotation, and HSV jittering for data augmentation. We take matching degree threshold of positives to be 0.4 for the refinement stage, while 0.6 for detection stage for high-quality detections.

The mean Average Precision (mAP) defined in PASCAL VOC object detection challenge [49] is used as the evaluation metric for all experiments. For a fair comparison with other methods, experiments on all dataset use the mAP metric defined in PASCAL VOC 2007 challenge. Most of our ablation studies are conducted on the HRSC2016 dataset since remote-sensing ships often have a large aspect ratio and scale variation, which are major challenges for object detection in optical remote-sensing images. In the ablation studies on HRSC2016 dataset, all images are scaled to 416×416 without data augmentation. Ablation experiments on DOTA dataset are also provided to further prove the effectiveness of our method.

We train the model with the batch size set to 8 on RTX 2080Ti GPU. The network is trained with Adam optimizer. The learning rate is set to $1e-4$ and is divided by 10 at each decay step. The total iterations of HRSC2016, UCAS-AOD, and DOTA are 10, 5, and 40 k, respectively.

C. Ablation Study

1) Evaluation of Different Components: We conduct componentwise experiments on HRSC2016 and DOTA datasets to verify the contribution of the proposed components. The experimental results are shown in Table I. Since only one anchor is preset, it is difficult to capture the critical features required to identify the objects, so the baseline model only achieves the mAP of 70.5% on HRSC2016 and 68.1% on DOTA dataset. The detection performance is increased by 5.7% with PAM module on HRSC2016. It indicates that the PAM effectively constructs more powerful feature representations, so that even one preset anchor can make good use of critical features to achieve accurate detection. The performance of the model is improved by 8.2% on HRSC2016 with DAL, due to its ability of selecting high-quality anchors with good critical feature alignment in the learning process. The simultaneous use of DAL and PAM achieves a mAP of 83.8%, indicating that the two methods do not conflict and can effectively improve the detection performance. The proposed R-ARM refines the horizontal anchors to obtain high-quality rotated anchors. It further improves the performance by 2.5%. Finally, CFC-Net reaches the mAP of 86.3% and 72.0% on HRSC2016 and DOTA respectively, proving the effectiveness of our framework.

2) Evaluation of PAM: To verify the effect of the proposed PAM, we conducted comparative experiments on HRSC2016 dataset. The results are shown in Table II. Using dual FPN to extract independent features for classification and regression branches, the detection performance is improved by 1.6% compared with the baseline model. However, dual FPN does not fully extract the critical features for specific tasks. When we adopt the attention mechanism based on dual FPN, a further improvement of 2.8% is achieved. It indicates that the attention mechanism enables the features of different branches to better respond to the discriminative parts of the objects. Through the polarization function, the discriminative parts of the critical classification features are strengthened, while the high response regions in the critical regression features are suppressed to find more clues to further improve localization results. The improvement of 1.3% based on the attention-based model confirms our viewpoint. These experiments prove that the proposed components of PAM can effectively improve the detection performance.

We further conducted experiments to search for suitable η for PAM, and the experimental results are shown in Table III. When the η is small ($\eta = 5$), it can be seen from Fig. 9 that $\psi_{\text{cls}}(x)$ tends to be linear, the activation is weak. The performance is close to the baseline (75.41% compared to 75.49%). As the η increases, the activation effect gradually increases. In this case, the activation function helps the PAM to capture the key features for classification and thus it improves the mAP by 0.79% with η set to 15. However, when η is very large (for example, $\eta = 150$), $\psi_{\text{cls}}(x)$ tends to be a step function. The $\psi_{\text{cls}}(x)$ in this case directly suppresses the features of the low response. This will cause the neglect of potentially critical features and lead to unstable training. Therefore, the performance of the model with $\eta = 150$ in PAM is even 0.64% lower than the baseline.

Some visualization results are shown in Fig. 10. It can be seen that the heatmaps generated by PAM accurately respond

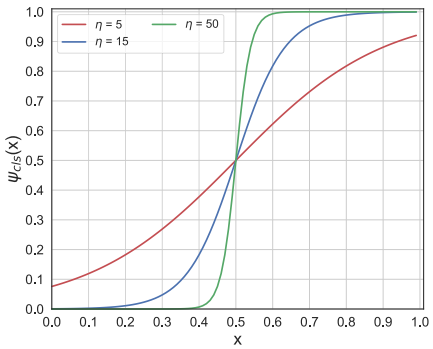


Fig. 9. Diagram of $\psi_{\text{cls}}(x)$ with different η.

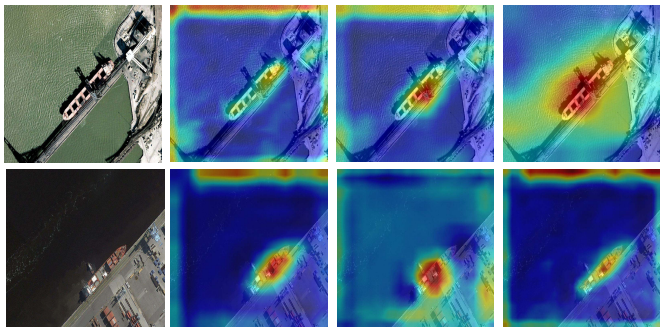


Fig. 10. Visualization results of critical features for classification and regression tasks. Compared with shared features, decoupled features can better focus on the critical features of their respective tasks. (a) Images. (b) Shared features. (c) Cls. features. (d) Reg. features.

TABLE IV
ABLATION STUDY OF DAL

	Different Variants			
with Input IoU?	✓	✓	✓	✓
with Output IoU?	×	✓	✓	✓
Uncertainty Supression?	×	×	✓	✓
Matching Sensitive Loss?	×	×	×	✓
mAP	70.5	71.3	76.2	**78.7**

to the area of task-sensitive critical features. The discriminative areas required for classification are often concentrated in the local part of objects, such as the stern and bow of ships. Meanwhile, the clues required for regression are more likely to be distributed on the edge of the objects or the contextual information.

3) Evaluation of DAL: We conduct componentwise experiments to verify the contribution of the DAL. The experimental results are shown in Table IV, in which input IoU, output IoU and regression uncertainty are denoted by the three terms in (7), respectively. For the variants with output IoU, α is set to 0.8 for stable training, and the detection performance slightly increases from 70.5% to 71.3%. It indicates that using output IoU alone is insignificant for training sample selection. With the suppression of regression uncertainty, the prior space alignment and posterior critical feature alignment would work together to dramatically improve the performance by 5.7%

TABLE V
ABLATION STUDY OF THE PROPOSED R-ARM

refinement stages	0	1	2
mAP	83.8	**86.3**	84.5

TABLE VI
ANALYSIS OF INFLUENCE OF DIFFERENT HYPERPARAMETERS. "-" MEANS THAT THE MODEL DOES NOT CONVERGE WITH THIS SETTING

γ	α	mAP	γ	α	mAP	γ	α	mAP
2	0.1	65.3	3	0.1	-	4	0.1	-
	0.2	76.2		0.2	69.8		0.2	47.4
	0.3	72.7		0.3	78.5		0.3	76.9
	0.5	70.0		0.5	74.4		0.5	**78.7**
	0.7	71.7		0.7	69.4		0.7	77.3
	0.9	43.9		0.9	69.1		0.9	72.1

compared with the baseline. Furthermore, matching degree guided loss function effectively distinguishes anchors with differential localization capability. The model using the matching sensitivity loss function achieves the mAP of 78.7%,

4) Evaluation of R-ARM: Based on DAL and PAM, we further conduct experiments to verify the effect of the proposed R-ARM and explore the influence of the number of refinement stages. For the model without R-ARM, we set the matching degree threshold of positives to 0.4. For the one-stage refinement model, the thresholds of the refinement stage and the detection stage are set to 0.4 and 0.6, respectively. The thresholds are set to 0.4, 0.6, and 0.8 for two-stage refinement modules. As shown in Table V, with one-stage R-ARM, the performance is increased by 2.5%, since the refined proposals provide high-quality samples which are better aligned with critical features of objects. However, adopting two-stage R-ARM drops the performance by 1.8% compared with the one-stage R-ARM. We suggest that as the threshold increases in detection stage, the number of positives decreases sharply, leading to insufficient positive samples and a serious imbalance between positives and negatives. Thus we use one stage R-ARM in CFC-Net.

5) Hyper-Parameters: To find suitable hyperparameter settings, we conduct parameter sensitivity experiments, and the results are shown in Table VI. As the α is reduced appropriately, the effect of feature alignment increases, and the mAP increases. For example, on condition that γ is equal to 4, as α decreases from 0.9 to 0.5, the mAP increases from 72.1% to 78.7%. It indicates that under the premise of uncertainty suppression, the feature alignment represented by the IoU_{out} is conducive to selecting anchors with high localization capabilities. However, when α is extremely small, the performance drops sharply (such as $\gamma = 4$), because the anchors selected by the dominant output IoU may contain too many false-positive samples. In this case, prior space alignment can help alleviate this problem and make anchor selection more stable. In addition, as γ decreases, the ability to suppress disturbance samples is stronger, but it may also

TABLE VII
COMPARISONS WITH DIFFERENT METHODS ON HRSC2016 DATASET

Methods	Backbone	Size	NA	mAP
Two-stage:				
R^2CNN [50]	ResNet101	800×800	21	73.1
RC1&RC2 [45]	VGG16	-	-	75.7
RRPN [51]	ResNet101	800×800	54	79.1
R^2PN [13]	VGG16	-	24	79.6
RoI Trans. [15]	ResNet101	512×800	5	86.2
Gliding Vertex [52]	ResNet101	512×800	5	88.2
Single-stage:				
RRD [17]	VGG16	384×384	13	84.3
BBAVector [53]	ResNet101	608×608	1	88.6
R^3Det [33]	ResNet101	800×800	21	89.3
R-RetinaNet [44]	ResNet101	800×800	121	89.2
GRS-Det [54]	ResNet101	800×800	1	89.6
CFC-Net	ResNet50	416×416	1	86.3
CFC-Net (aug)	ResNet50	800×800	1	88.6
CFC-Net (aug)	ResNet101	800×800	1	89.5
CFC-Net (aug + ms)	ResNet101	800×800	1	**89.7**

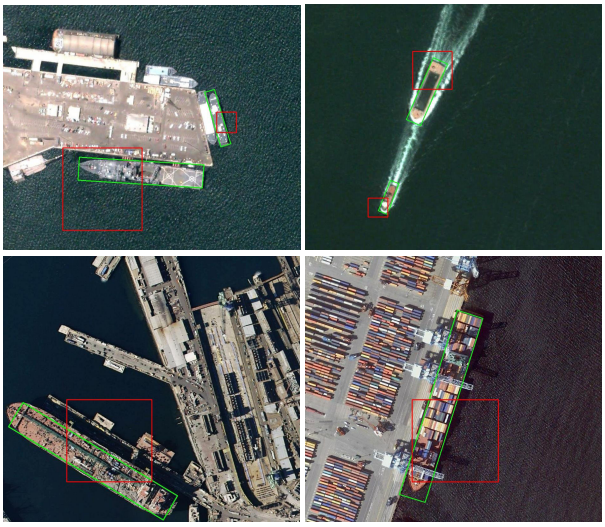


Fig. 11. Detection results on HRSC2016 dataset with our method. The red boxes and green boxes indicate the anchor boxes and detection results respectively.

suppress the mining of potential positives, resulting in performance degradation.

D. Main Results and Analysis

1) Results on HRSC2016: HRSC2016 contains lots of remote-sensing oriented ships with large aspect ratios, scales and orientations. Our method achieves competitive performances on HRSC2016 dataset. As shown in Table VII, "aug" represents using data augmentation, "ms" denotes multi-scale training and testing, and NA is the number of preset anchors at each location of feature maps. The proposed CFC-Net achieves the mAP of 86.3% when input images are rescaled to 416 × 416 without data augmentation, which is comparable to many previous advanced methods. With the input image resized to 800 × 800 and data augmentation adopted, our method reaches the mAP of 88.6%, which is better than many recent methods. With multi-scale training and testing, our method further achieves state-of-the-art performance on HRSC2016 dataset among the compared methods, reaching the mAP of 89.7%.

It is worth mentioning that our approach uses only one horizontal anchor at each position of feature maps, but outperforms the frameworks with a large number of anchors. These results show that it is unnecessary to preset a large number of rotated anchors for oriented object detection. Instead, the more important thing is to select high-quality anchors and capture the critical features for object recognization. For instance, the anchors in Fig. 11 have low IoUs with targets in the images and will be regarded as negatives in most detectors. However, they have a strong potential for accurate localization. CFC-Net effectively utilizes these anchors to achieve accurate prediction.

Moreover, our model is a single-stage detector, and uses the feature maps of $P_3 - P_7$. Compared with the $P_2 - P_6$ for two-stage detectors, the total amount of positions that need to set anchor is fewer, so the inference speed is faster. With the input image resized to 800 × 800, our model reaches 28 FPS on RTX 2080 Ti GPU. Besides, our method is similar to anchor-free methods that also set one anchor at each location on the feature maps, such as BBAVector [53], GRS-Det [54]. CFC-Net outperforms BBAVector and GRS-Det by 0.3% and 0.1%, respectively. Our inference speed is also faster than these anchor-free methods (12 FPS for BBAVector and 14 FPS for GRS-Det, while 28 FPS for CFC-Net).

2) Results on DOTA: We compare the proposed approach with other state-of-the-art methods on DOTA dataset. As shown in Table VIII, we achieve the mAP of 73.50%, which reaches the best performance among the compared methods. Some detection results on DOTA are shown in Fig. 12. It can be seen from the illustration that even though only one anchor is used, our CFC-Net still accurately detects densely arranged small objects (such as ships, small vehicles, and large vehicles in the third row). In addition, the proposed detector achieve accurate detection on objects with various scales. Take the second one (from the left) in the second row for example, CFC-Net outputs the precise detections of both large-scale roundabout and small vehicles at different scales. Besides, as shown in the third figure and the fifth figure in the first row, our method uses a few square anchors to detect objects with very large aspect ratios (such as bridges and harbors here), These detections indicate that it is not essential for preset anchors to have a good spatial alignment with the objects, while the key is to effectively identify and capture the critical features of the objects. The matching degree measures the critical feature capturing ability of anchors, and on this basis, the DAL strategy performs a more reasonable selection of training samples to achieve high-quality detection.

3) Results on UCAS-AOD: Experimental results in Table IX show that our CFC-Net achieves the best performance among the compared detectors, reaching the mAP of 89.49%. Note

TABLE VIII
PERFORMANCE EVALUATION OF OBB TASK ON DOTA DATASET

Methods	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
FR-O [46]	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.40	52.52	46.69	44.80	46.30	52.93
R-DFPN [55]	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
R^2CNN [50]	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN [51]	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
ICN [56]	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.20	53.64	62.90	67.02	64.17	50.23	68.16
RoI Trans. [15]	88.64	78.52	43.44	**75.92**	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
CAD-Net [34]	87.80	**82.40**	49.40	73.50	71.10	63.50	76.70	**90.90**	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
DRN [57]	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	**68.60**	52.50	70.70
O^2-DNet [58]	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
R^3Det [33]	89.54	81.99	48.46	62.52	70.48	74.29	77.54	90.80	81.39	83.54	61.97	59.82	65.44	67.46	60.05	71.69
SCRDet [59]	**89.98**	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	**86.86**	**65.02**	**66.68**	66.25	68.24	**65.21**	72.61
CFC-Net (ours)	89.08	80.41	**52.41**	70.02	**76.28**	**78.11**	**87.21**	90.89	84.47	85.64	60.51	61.52	**67.82**	68.02	50.09	**73.50**

Fig. 12. Visualization of detection results on DOTA dataset with our method.

TABLE IX
DETECTION RESULTS ON UCAS-AOD DATASET

Methods	car	airplane	mAP
YOLOv3 [30]	74.63	89.52	82.08
R-RetinaNet [44]	84.64	**90.51**	87.57
FR-O [46]	86.87	89.86	88.36
RoI Trans. [15]	88.02	90.02	89.02
CFC-Net	**89.29**	88.69	**89.49**

that the original YOLOv3 [30] and RetinaNet [44] are proposed for generic object detection, and the objects are annotated with horizontal bounding box. To make a fair comparison, we introduce an additional angle dimension and perform angle prediction to achieve rotation object detection. Our method outperforms the other compared single-stage detectors, and even better than some advanced two-stage detectors, such as RoI Transformer [15]. Besides, the detection performance of small vehicles is excellent, which indicates that our method is also robust to densely arranged small objects.

V. CONCLUSION

In this article, we introduce the concept of critical features and prove its importance for high-performance object detection through observations and experiments. On this basis, a CFC-Net is proposed to optimize the one-stage detector

from three aspects: feature representation, anchor refinement, and training sample selection. Specifically, decoupled classification and regression critical features are extracted through the polarization attention mechanism module based on dual FPN. Next, the rotation anchor refinement is performed on one preset anchor to obtain high-quality rotation anchors, which can be better aligned with critical features. Finally, matching degree is adopted to measure the ability of anchors to capture critical features, so as to select positive candidates with high localization potential. As a result, the inconsistency between classification and regression is alleviated and high-quality detection performance can be achieved. Extensive experiments on three remote-sensing datasets verify the superiority of the proposed method.

REFERENCES

[1] Y. Li, X. Sun, H. Wang, H. Sun, and X. Li, "Automatic target detection in high-resolution remote sensing images using a contour-based spatial model," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 5, pp. 886–890, Sep. 2012.

[2] J. Han *et al.*, "Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding," *ISPRS J. Photogramm. Remote Sens.*, vol. 89, pp. 37–48, Mar. 2014.

[3] 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.

[4] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[5] L. Eikvil, L. Aurdal, and H. Koren, "Classification-based vehicle detection in high-resolution satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 64, no. 1, pp. 65–72, Jan. 2009.

[6] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[7] H. Ji, Z. Gao, T. Mei, and B. Ramesh, "Vehicle detection in remote sensing images leveraging on simultaneous super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 676–680, Apr. 2020.

[8] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[9] N. Liu, Z. Cao, Z. Cui, Y. Pi, and S. Dang, "Multi-layer abstraction saliency for airport detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9820–9831, Dec. 2019.

[10] F. Wu, Z. Zhou, B. Wang, and J. Ma, "Inshore ship detection based on convolutional neural network in optical satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4005–4015, Nov. 2018.

[11] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[12] W. Liu, L. Ma, and H. Chen, "Arbitrary-oriented ship detection framework in optical remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 6, pp. 937–941, Jun. 2018.

[13] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1745–1749, Nov. 2018.

[14] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based CNN for ship detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 900–904.

[15] 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.*, Jun. 2019, pp. 2849–2858.

[16] L. Li, Z. Zhou, B. Wang, L. Miao, and H. Zong, "A novel CNN-based method for accurate ship detection in HR optical remote sensing images via rotated bounding box," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 686–699, Jan. 2021.

[17] M. Liao, Z. Zhu, B. Shi, G.-S. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5909–5918.

[18] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 294–308, 2020.

[19] 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.

[20] 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.

[21] Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Oriented response networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 519–528.

[22] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 3–22, Nov. 2018.

[23] 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.

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

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

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

[27] 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.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[29] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.

[30] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

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

[32] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[33] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," 2019, *arXiv:1908.05612*. [Online]. Available: http://arxiv.org/abs/1908.05612

[34] 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.

[35] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[36] 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.*, Jul. 2017, pp. 2117–2125.

[37] G. Song, Y. Liu, and X. Wang, "Revisiting the sibling head in object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11563–11572.

[38] Z. Tian, W. Wang, R. Zhan, Z. He, J. Zhang, and Z. Zhuang, "Cascaded detection framework based on a novel backbone network and feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3480–3491, Sep. 2019.

[39] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2888–2897.

[40] J. Choi, D. Chun, H. Kim, and H.-J. Lee, "Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 502–511.

[41] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 784–799.

[42] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for LiDAR 3D vehicle detection," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3266–3273.

[43] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," 2020, *arXiv:2012.04150*. [Online]. Available: http://arxiv.org/abs/2012.04150

[44] 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.*, Oct. 2017, pp. 2980–2988.

[45] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[46] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[47] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3735–3739.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[49] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[50] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: http://arxiv.org/abs/1706.09579

[51] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[52] 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.

[53] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2021, pp. 2150–2159.

[54] X. Zhang, G. Wang, P. Zhu, T. Zhang, C. Li, and L. Jiao, "GRS-Det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3518–3531, Apr. 2021.

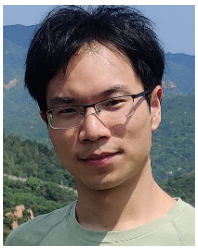
[55] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, Jan. 2018.

[56] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. Asian Conf. Comput. Vis.* Perth, WA, Australia: Springer, 2018, pp. 150–165.

[57] X. Pan *et al.*, "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11207–11216.

[58] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.

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



Qi Ming received the B.S. degree in automation from the School of Automation, Beijing Institute of Technology (BIT), Beijing, China, in 2018, where he is pursuing the Ph.D. degree in navigation, guidance, and control.

His research interests include computer vision, object detection, and remote-sensing image analysis.

Lingjuan Miao received the B.S. and M.S. degrees in control theory and engineering from the Harbin Institute of Technology, Harbin, China, in 1986 and 1989, respectively, and the Ph.D. degree in control theory and engineering from the China Academy of Launching Vehicle Technology, Beijing, China, in 2001.

Since 1992, she has been with the Beijing Institute of Technology, Beijing, as a Lecturer, since 1996 as an Associate Professor, and since 2001 as a Professor. Her research interests include GPS, inertial navigation systems, INS/GPS integrated navigation, and multisensor fusion technique.

Zhiqiang Zhou (Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control science and engineering from the School of Automation, Beijing Institute of Technology (BIT), Beijing, China, in 2004 and 2009, respectively.

From 2009 to 2012, he was a Post-Doctoral Researcher with the Institute of Automation, Chinese Academy of Sciences, Beijing. He is an Associate Professor with the School of Automation, BIT. His research interests include information fusion, pattern recognition, digital image processing, and vision-based navigation.

Yunpeng Dong received the B.S. degree in automation from Hebei University, Baoding, China, in 2020. He is pursuing the M.S. degree with the School of Automation, Beijing Institute of Technology, Beijing, China.

His research interests include object detection and FPGA for deep learning.