Not All Boxes Are Equal: Learning to Optimize Bounding Boxes With Discriminative Distributions in Optical Remote Sensing Images

Qi Ming, Lingjuan Miao, Zhiqiang Zhou, *Member, IEEE*, Nicolas Vercheval, and Aleksandra Pižurica, *Senior Member, IEEE*

***Abstract*— Detecting oriented objects in optical remote sensing images has been consistently challenging due to difficulties in bounding boxes' localization. The cascaded regression framework, widely used for high-quality bounding box refinement, has demonstrated effectiveness in this domain. However, our experiments reveal a discontinuity issue in bounding box optimization in cascaded regression framework. As a result, performance gain is not guaranteed across all stages in this framework. In this article, we propose a distribution discriminative detector (DDDet) to address the above issues and enhance the optimization of bounding boxes in oriented object detection. Specifically, a novel conditional anchor refinement framework (CARF) is designed to improve cascaded regression structure. CARF distinguishes bounding boxes with different distributions, adaptively optimizing them within the well-assigned regressors. Subsequently, the aligned convolution module (ACM) is integrated into each regressor, facilitating the continuous alignment between features and refined anchors. Furthermore, the geometry-guided training sample selection (GTSS) method is incorporated into CARF to assign labels based on object shape priors. Experimental results show that DDDet obtains state-of-the-art performance on mainstream datasets for oriented object detection in remote sensing image, which demonstrates the effectiveness of the proposed method. Our method surpasses many current single-stage detectors, two-stage detectors, and refine-stage detectors, achieving the mAP of 79.41% on the DOTA dataset and 44.15% on the FAIR1M dataset.**

***Index Terms*— Anchor refinement, bounding box regression, convolutional neural networks (CNNs), feature alignment, object detection.**

I. Introduction

RECENTLY, due to the rapid development of remote sensing technology, the available remote sensing images have increased dramatically. It is a very challenging and important task to efficiently identify objects from massive remote sensing images. In addition to directly serving object-oriented applications, object recognition in remote sensing scenes has a wide range of uses. For example, object-based contour information can be used as auxiliary prior and contextual information for remote sensing change detection [1], [2]. Some previous methods use handcrafted features to extract objects in images [3], [4], [5]. Since remote sensing images often have complex scenes and variable object distributions, these traditional methods suffer from low detection accuracy and slow inference speed.

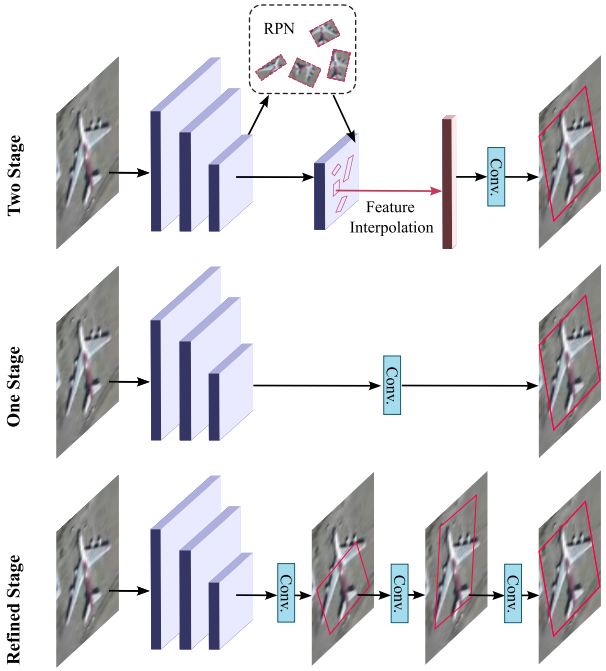


Fig. 1. Frameworks of different object detectors.

In the past decade, convolutional neural networks (CNNs) have achieved great success in the field of computer vision. The powerful local feature extraction capability of CNNs has achieved excellent performance in various downstream vision tasks such as image classification [6], object detection [7], [8], and image segmentation [9]. Some works also introduce CNN-based detectors into the object detection in remote sensing images and bring great breakthroughs [10], [11], [12], [13], [14], [15].

Currently, remote sensing object detectors are generally divided into two categories: one-stage detectors and two-stage detectors. As shown in Fig. 1, two-stage detectors first generate a series of region of interest (RoI) through a region proposal network (RPN). Then, more powerful features are extracted within RoIs for subsequent classification and

Manuscript received 23 January 2024; revised 7 March 2024 and 6 April 2024; accepted 25 April 2024. Date of publication 2 May 2024; date of current version 14 May 2024. *(Corresponding author: Zhiqiang Zhou.)*

Qi Ming, Lingjuan Miao, and Zhiqiang Zhou 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).

Nicolas Vercheval and Aleksandra Pižurica are with the Faculty of Engineering and Architecture, Ghent University, 9000 Ghent, Belgium (e-mail: nicolas.vercheval@ugent.be; aleksandra.pizurica@ugent.be).

Digital Object Identifier 10.1109/TGRS.2024.3396134

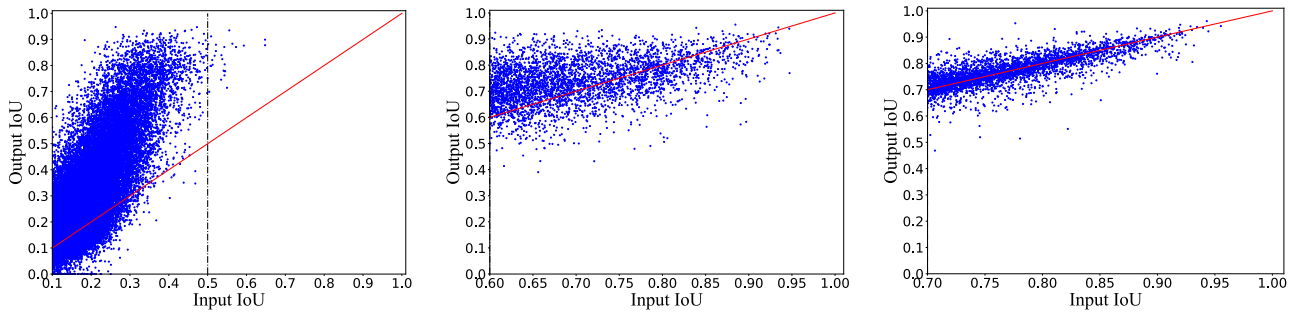


Fig. 2. Illustration of the IoU changes before and after the bounding boxes go through different refinement stages. Input IoU is the IoU between input anchor and the corresponding GT box. Output IoU denotes the IoU between output refined box and GT box. The visualization results show that in the later stages, many anchors suffer severe performance degradation after regression. (a) First stage. (b) Second stage. (c) Third stage.

regression [16], [17]. One-stage detectors directly treat object detection as a regression task and predict the objects in one step [18], [19], [20]. The feature interpolation adopted in two-stage detectors is helpful for robust feature extraction, but it is time-consuming. Therefore, two-stage detectors have better performance but slower running speed than one-stage detectors.

For object detection in remote sensing images, one-stage detectors have more potential and are more suitable for the following reasons: 1) sizes of remote sensing images are often very large, and therefore a fast detector is required to achieve efficient object detection and 2) the density of objects in remote sensing images is extremely uneven. Two-stage detectors need to select a fixed number of RoIs for prediction, which is inflexible. Too few RoIs may cause missed detection, while too many RoIs will cause excessive feature interpolation time and slower inference speed. Recently, there have also been some works that introduce the advantages of two-stage detectors into a one-stage detectors for good performance, which are called refined stage detectors [21], [22], [23], [24]. These methods use multiple regressions to continuously refine the initially preset bounding boxes (also known as anchors), achieving better detection accuracy at the cost of slightly reduced running speed. Different intersection-over-union (IoU) thresholds are adopted for different refinement stages to select high-quality positive samples for cascaded regression.

The refined stage detectors achieve superior performance in remote sensing object detection. However, we found that the bounding boxes were not well-optimized in each refinement stage. As shown in Fig. 2, we visualized the IoU changes in anchors before and after regression in a refined stage detector shown in Fig. 1. We used the common IoUs of 0.5, 0.6, and 0.7 as the thresholds to select positives for each refinement stage. In the first stage, the initial preset anchors are hard to align well with ground truth (GT) to achieve IoU higher than 0.5 [see Fig. 2(a)]. Therefore, we select the candidates with maximum IoUs as positives. Although these samples have low relatively input IoU, they achieve good localization results after anchor refinement. However, in the second stage, high-quality refined anchors from the first stage are not well-optimized. As shown in Fig. 2(b), many anchors even suffer from drop in accuracy after refinement. The performance degradation is even more exacerbated in the third stage. Almost half of the positive samples have worse localization performance after regression [see Fig. 2(c)]. The above experimental phenomena widely exist in the refined stage detectors, which limits the high-quality detections in remote sensing images.

More specifically, the performance bottlenecks in refined detectors can be attributed to the following aspects:

1) Imbalanced loss contribution across different stages. Training samples in earlier stages are of lower quality and larger prediction offsets, leading to higher losses. The overall loss contribution will be dominated by earlier stages, and the model tends to focus on samples in these stages and neglects to optimize high-quality samples in later stages.
2) Misalignment between IoU thresholds and bounding box distributions. Within each stage, low-quality samples slightly above the IoU threshold tend to incur larger regression losses, hindering more accurate predictions.
3) Imbalanced number of positives across stages. Since the IoU thresholds in the later stages are higher, positive samples in later stages are very rare and potentially cause the regressor to overfit.
4) Misalignment between anchors and features. In the cascaded regression framework, the feature extraction process does not adjust with continuous bounding box optimization.

In this article, we delve into the above problems and proposed a distribution discriminative detector (DDDet) to optimize bounding box regressions for better performance. Specifically, a novel conditional anchor refinement framework (CARF) is proposed to refine bounding boxes with different distributions separately. Then, aligned convolution layers are introduced to eliminate misalignment between features and anchor regions to extract more informative features. Furthermore, to fully use the potential high-quality samples, we proposed a geometry-guided training sample selection (GTSS) for high-quality training sample selection. Our DDDet achieves state-of-the-art performance on multiple publicly available remote sensing datasets, including HRSC2016 [25], UCAS-ADO [26], DOTA [27], and FAIR1M [28]. In summary, the contributions of this article are as follows.

1) We observed that the bounding boxes are not continuously optimized in current cascaded regression framework and give a deep analysis on this phenomenon with the experimental results. Then, a DDDet is designed

to solve the issues to achieve substantial performance gains.
2) We propose a novel CARF to improve cascaded regression structure. CARF distinguishes bounding boxes with different distributions, adaptively optimizing them within the well-assigned regressors.
3) A GTSS method is introduced into the CARF for better training sample selection. GTSS selects potential high-quality samples during the cascaded regression process, which helps achieve more accurate predictions.

The rest of this article is organized as follows. Section II introduces related work of object detection in remote sensing images. Section III elaborates on our analysis and method. Section IV shows experimental results and comparison with other methods. Finally, conclusions are drawn in Section V.

II. Related Work

A. Generic Object Detection

Object detection aims to detect objects from various classes in images or videos. The introduction of CNNs leads to significant improvements in detection accuracy [16], [17], [18], [19], [29], [30], [31]. Object detectors can be broadly classified into two categories: one-stage detectors and two-stage detectors. Two-stage detectors first generate a series of candidate regions, and then perform classification and bounding box regression on regions to detect objects. Representatives of this class include R-CNN [32] and its variants such as Fast R-CNN [16] and Faster R-CNN [17]. In contrast, one-stage detectors aim to directly classify and locate objects together in one regression step. Representative works include YOLO [18] and SSD [29]. Two-stage detectors generally perform better in accuracy but have slower inference speeds, while one-stage detectors are suited for real-time scenarios.

To bridge the gap between one-stage and two-stage detectors, cascaded refine-stage detectors are proposed to incorporate bounding box refinement process to improve detection accuracy without significantly compromising on speed. RefineDet [33] used an anchor refinement module (ARM) for coarse adjustment of anchors, followed by an object detection module (ODM) that precisely classifies objects and refines bounding box positions. R^3Det [21] introduced a feature refinement module that iteratively refines features at multiple scales and adjusts the anchor positions, leading to more accurate detection of objects. CFC-Net [22] adopted cascaded optimization to iteratively refine object localization by regressing boxes at multiple stages. Refine-stage detectors offer a promising compromise between the high speed of one-stage detectors and the high accuracy of two-stage detectors.

Recently, transformer-based methods have achieved tremendous success in the field of object detection. DETR [34] views the object detection task as a set prediction problem, predicting the set containing all the object boxes in one step. Swin-Transformer [35] combines the self-attention mechanism of transformers with the local perceptual abilities of CNNs for efficient detection results. On this basis, Bae [36] introduced deformable part region learning module to adjust flexibly in response to its geometric transformations.

B. Object Detection in Remote Sensing Images

Object detection in remote sensing images has been a popular research topic for several decades. The increasing demand for automatic object extraction from high-resolution remote sensing images has spurred the growth of object detection algorithms. With the advent of CNNs, the performance of object detection in remote sensing images has also been significantly improved [14], [15], [23], [37], [38]. Remote sensing images differ from natural images, and it is important to consider the unique challenges posed by remote sensing data.

First, remote sensing images are from a bird's-eye view, and the objects in the remote sensing images are often multioriented. Therefore, recent works usually adopt oriented bounding boxes (OBBs) to represent rotated objects in remote sensing images [11], [22], [37], [39], [40]. For example, Zand et al. [15] preset rotated anchors on the feature maps and introduce angle prediction to regress rotated objects. Second, regression of OBBs brings many optimization problems during training process. Yang et al. [37] pointed out that periodicity of angle would cause sudden increases in angle regression loss. Ming et al. [23] suggested that the angle representation redundancy leads to suboptimal angle regression. To resolve these issues, a series of works based on angle classification were proposed to avoid potential loss oscillation issue [11], [40]. Moreover, there are also methods that approximate OBB as Gaussian distributions for better optimization [41], [42].

Feature extraction is very important for multiorientation remote sensing object detection. Large selective kernel network (LSKNet) [38] dynamically adjusts its spatial receptive field to incorporate unique prior knowledge and long-range context for more accurate detection of objects in remote sensing scenarios. ReDet [14] encodes rotation equivariance and invariance by integrating rotation-equivariant networks for feature extraction to adaptively extract features based on RoI orientation, which improves orientation prediction.

To achieve high-accuracy detection performance, some methods used two-stage detectors for detecting objects in remote sensing images [14], [37]. However, two-stage methods need to set a large number of RoIs to ensure a high recall in remote sensing scenes, which greatly reduces inference speed. Therefore, one-stage detectors are very popular. Compared with one-stage detectors, refine-stage detectors introduce only a slight increase in inference overhead but can achieve higher detection accuracy. S^2ANet [13] aligned the rotated regional features for oriented objects and incrementally refined the detection of objects for precise localization. RDD [43] decoupled orientation from OBBs, and then performed multiple regressions on horizontal anchors to obtain better priors for accurate oriented object detection.

C. Label Assignment Strategy in Object Detection

The anchor-based methods densely preset a large number of prior anchors on the feature maps. Then, the samples whose IoU with the GT is larger than the set threshold (usually 0.5) will be selected as positives for regression. This process is also known as label assignment. Label assignment helps reduce the search space for object detection algorithms and improve the accuracy of object detection. Over the years, various label

assignment strategies have been proposed and used in object detection [10], [12], [22], [39], [44], [45], [46].

Some works suggested that the IoU between anchor and GT is not consistent with the corresponding prediction accuracy [22], [39]. For example, DAL [39] uses the IoU of samples before and after regression to comprehensively consider the localization potential of anchors. ATSS [45] dynamically selects high-quality positives based on statistical characteristics of training samples.

There are also some works that specially design label assignment strategies for remote sensing objects [10], [46]. For objects with large ratios (such as bridges and ships), it is hard for anchors to achieve good spatial alignment with them. SASM [46] designs dynamic IoU thresholds for objects of different shapes to adaptively select positives, thereby providing more samples to ensure sufficient learning process. EARL [47] introduces an elliptical distribution-aided approach to improve training sample selection method, focusing the model on high-quality samples. Yu et al. [48] introduced a soft label assignment mechanism to select arbitrary-oriented training samples, focusing on the most representative items for more stable training optimization. Zhang et al. [49] proposed a task-collaborated detector to incorporate classification and localization confidence into label assignment, and then anchors with precise and consistent predictions are selected as positives.

For two-stage detectors or refined stage detectors, different IoU thresholds are used for each regression stages [22], [33], [50]. Generally, the input proposals in later stages are more accurate, so higher IoU thresholds are applied to select high-quality positives for regression. Most of the existing label assignment strategies are studied within separate regressors. Unlike previous work, our framework takes into account the bounding box distribution between different stages to make a better design.

III. Methodology

In this section, we first dive into the bottleneck of cascaded regression framework. Then, a DDDet is proposed for high-quality object detection in remote sensing images. The modules in DDDet solve the above issues in cascaded regression models from various aspects, leading to stable and sustained performance improvements.

A. Bottleneck of Cascaded Regression Framework

The cascaded regression is a common framework used to improve the accuracy of object detectors. It aims to progressively refine the initial anchors through multiple regression steps. Object detectors based on cascaded regression have achieved state-of-the-art performance in many fields [21], [22], [33], [50]. However, as shown in Fig. 2, we found through experiments that cascaded regression in one-stage detectors does not continuously optimize all the bounding boxes. Even many high-quality boxes suffer performance degradation after regression. The above issues limit the performance of this framework, and the reasons can be attributed to the following factors.

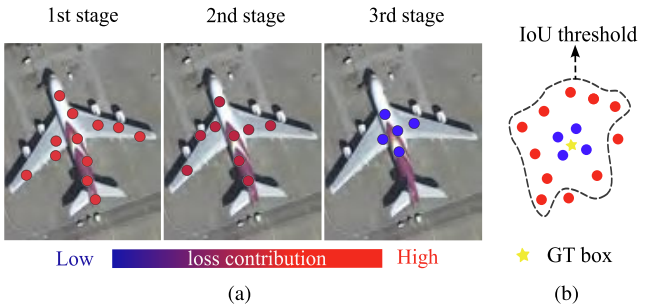


Fig. 3. Illustration of the imbalanced loss contribution of positive samples across stages (a) and within a certain stage (b). (a) Positives in earlier stages are relatively inaccurate and therefore dominate the regression loss. (b) In each stage, the losses of high-quality positives close to GT are relatively small, so the model tends to optimize low-quality positives far away from GT.

1) Imbalanced Loss Contribution Across Different Stages: In the cascaded regression framework, different IoU thresholds are typically set to select positives for the corresponding stages. Positive samples in later stages are more accurately located than those in earlier stages, therefore leading to smaller regression loss. In this case, the regression loss is dominated by weak samples with lower accuracy in earlier stages. And strong samples with more accurate localization are not further optimized [see Fig. 3(a)]. As a result, the detector tends to compensate for weak ones rather than optimize strong ones. Some previous works balanced the loss contribution by adjusting the weight for different stages [22], [50]. However, the distribution of bounding boxes is dynamically changing, and thus fixed loss weights are difficult to ensure balanced losses across stages. For instance, increasing the loss contribution in later stages by reweighting might result in poorer predictions in earlier stages [see Fig. 6(b)]. Consequently, insufficient high-quality positives may be propagated to later stages, which does not guarantee an overall performance gain.

2) Misalignment Between IoU Thresholds and Bounding Box Distributions: Imbalanced loss contributions also occur within each stage. As shown in Fig. 3(b), within a certain regression stage, low-quality positive samples near the IoU threshold often have larger loss contributions. As a result, the detector tends to optimize samples around the set IoU threshold, while high-qualities near the GT are not well-refined.

3) Imbalanced Number of Positives Across Stages: Different IoU thresholds lead to different numbers of available positive samples. A higher IoU threshold for later stage often results in fewer positives. Therefore, later stage is more likely to suffer from overfitting. In addition, the difference in the number of training samples also exacerbates the imbalanced loss contribution across different stages.

4) Misalignment Between Anchors and Features: Multiple steps of regression are based on the initial anchor features. Anchors are shifted relative to the initial positions to achieve localization results. However, initial features are used in all subsequent regression stages. The same features cannot provide more accurate semantic information to achieve further refinement.

B. Deformable Feature Alignment

Region-based feature extraction methods [16], [17], [22], [51] are very common in two-stage detectors, but they are also

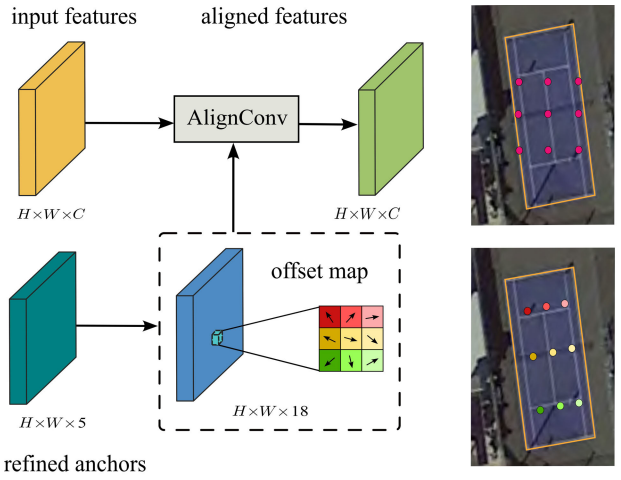


Fig. 4. Structure of ACM. The sampling points are shifted according to the anchor area to achieve alignment between features and bounding boxes.

time-consuming. One-stage detectors abandon these operations to greatly improve inference speed. However, they therefore suffer from performance degradation due to inaccurate features. In cascaded regression models, the misalignment between features and training samples is even more severe. As the bounding boxes are gradually optimized, they move further away from the initial locations and features. Detecting objects using the initial anchor features is obviously inaccurate in this case.

To address the problem, we adopt the aligned convolution (AlignConv) module in the cascaded regression framework. The overview of AlignConv is shown in Fig. 4. AlignConv aligns the sampling grid with the input feature map, which helps better capture fine-grained spatial information. The AlignConv consists of two steps: 1) using regular grid $\mathcal{R}$ to sample according to the anchor area on input feature map and 2) performing a weighted summation of sampled values by kernel weights $\mathbf{W}$.

For a standard convolution operation, we have

$$\mathbf{F}_{\text{out}}(\mathbf{p_0}) = \sum_{\mathbf{p_n}\in\mathcal{R}} \mathbf{W}(\mathbf{p_n}) \cdot \mathbf{F}_{\text{in}}(\mathbf{p_0} + \mathbf{p_n}) \tag{1}$$

in which $\mathbf{F}_{\text{in}}$ is the input feature map, and $\mathbf{F}_{\text{out}}$ is the corresponding output feature map defined on $\Omega = \{0, 1, \ldots, H-1\} \times \{0, 1, \ldots, W-1\}$. $\mathbf{p_n}$ is the location from sampling grid $\mathcal{R} = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}$. $\mathbf{p_0}$ denotes the location on the ouput features $\mathbf{F}_{\text{out}}$, and $\mathbf{p_0} \in \Omega$. $\mathbf{W}$ is the weight value.

In AlignConv, the offsets based on the anchor position will be applied to the regular grid $\mathcal{R}$ to determine the sampling position, and therefore, (1) is as follows:

$$\mathbf{F}_{\text{out}}(\mathbf{p_0}) = \sum_{\mathbf{p_n}\in\mathcal{R}} \mathbf{W}(\mathbf{p_n}) \cdot \mathbf{F}_{\text{in}}(\mathbf{p_0} + \mathbf{p_n} + \Delta\mathbf{p_n}) \tag{2}$$

where $\Delta\mathbf{p_n}$ is the localization offset in offset field $\mathcal{O}$. Given the anchor box at the current location $\mathbf{p_0}$ as (cx, cy, w, h, θ). The rotated sampling field is denoted as follows:

$$\mathcal{R}_{\text{rbox}} = \frac{1}{s}\left((cx, cy) + \frac{1}{k}(w, h) \cdot \mathbf{r}\right) \cdot \mathbf{R}^{\mathbf{T}}(\theta) \tag{3}$$

where s is the stride of filters for downsampling, and k denotes the kernel size of filters. $\mathbf{R}(\theta)$ is the rotation matrix denoted in [13]. Then, the offset field $\mathcal{O}$ is calculated as follows:

$$\mathcal{O} = \sum_{\mathbf{p_n}\in\mathcal{R}} (\mathcal{R}_{\text{rbox}} - \mathbf{p_0} - \mathbf{p_n}). \tag{4}$$

Equation (2) dynamically adjusts the sampling locations according to the input anchor box, which helps align the features within the anchor area. During the cascaded regression process, the position of input anchor is continuously refined. Aligned convolution extracts critical and effective features based on refined boxes, which bridges the gap between features and predicted boxes, achieving continuous and accurate optimization for bounding box regression.

The previous region-based feature alignment operations [16], [51] are based on bilinear interpolation, and they are very time-consuming. Large-scale remote sensing image interpretation tasks require a fast and efficient solution. AlignConv only calculates the offset of the sampling point to achieve feature alignment to improve performance, which ensures high-speed interpretation at the expense of relatively small computational overhead. As shown in Fig. 6(c), AlignConv achieves feature alignment, significantly enhancing the prediction accuracy of samples in each stage of the cascaded regression framework.

The application of aligned convolution has solved the misalignment between anchors and features in current refined stage detectors. By inserting aligned convolution in the cascaded bounding box regression framework, the regressors dynamically select the feature extraction area based on the optimized bounding box localization, therefore enhancing feature representation and improving model robustness and detection accuracy.

C. Conditional Anchor Refinement Framework

The cascaded regression framework has achieved significant success in object detection in remote sensing imagery [13], [21], [22]. However, current frameworks mostly directly apply multilevel regressors to refine detections for continuous performance gains. We suggest that these methods have not taken into account the matching issue between samples from different distributions and corresponding regressors.

For the tth regressor in the classic cascaded regression framework, the bounding box regression process is represented as follows:

$$\mathbf{b}_t = f_{t-1}(\mathbf{X}_{t-1}, \mathbf{b}_{t-1}) \tag{5}$$

where $\mathbf{b}_{t-1}$ represents anchor boxes input to the regressor, and $\mathbf{b}_t$ is the predicted bounding box after regression. $\mathbf{X}$ is the input feature, and f symbolizes the regression process of the regressor.

As discussed in Section III-A, there are many problems in the classic cascaded regression framework. To solve the above issues, we propose a CARF. As shown in Fig. 5, in our CARF, samples from different distributions (i.e., bounding boxes with different localization accuracy) are separately refined toward specific regressors. This ensures continuous optimization for input boxes from two perspectives: 1) the features fed into the regressors are well-aligned with the spatial positions of input boxes and 2) each regressor optimizes only for samples from a specific distribution, avoiding suboptimal refinement issues caused by imbalanced loss contributions as discussed

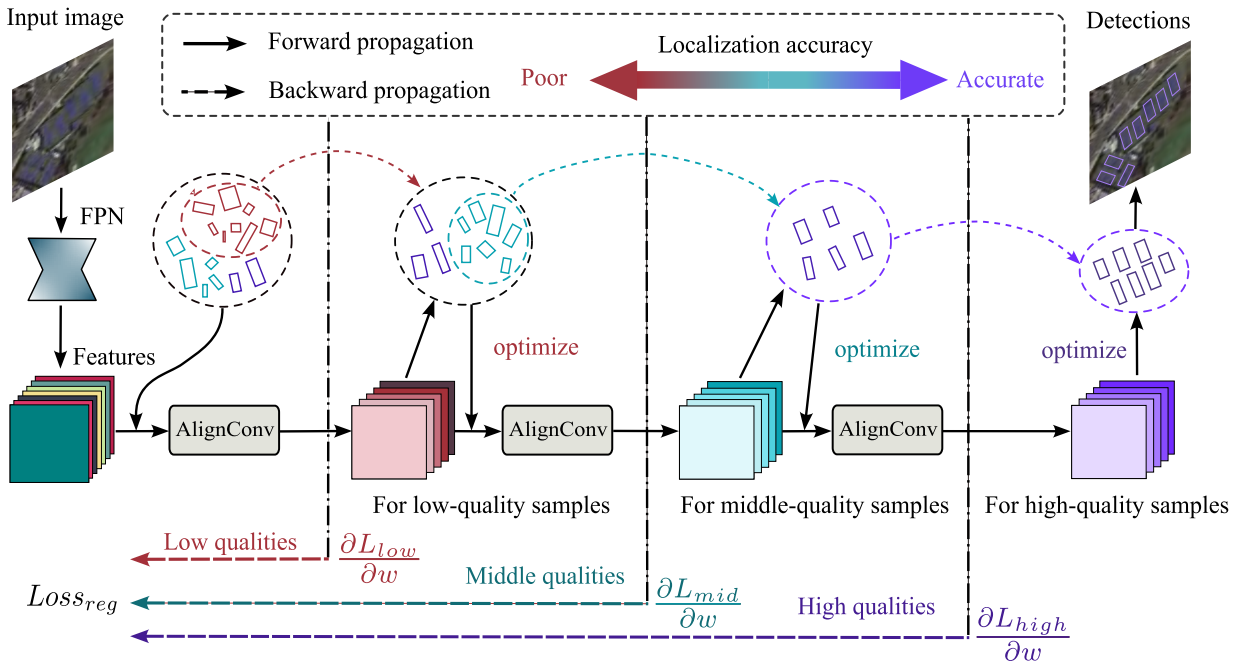


Fig. 5. Overview of the proposed method.

in Section III-A. In CARF, we achieve the matching between sample distributions and regressors by controlling the gradient flow during the backpropagation process, as indicated by the arrows in Fig. 5. Specifically, the bounding boxes regression process of the tth regressor in CARF is expressed as follows:

$$\mathbf{b}_t = \left\{\mathbf{b}_{t-1}^{+},\ f_{t-1}\left(\mathbf{X}_{t-1}, \mathbf{b}_{t-1}^{-}\right)\right\}. \tag{6}$$

Different from the classic cascaded regression framework, CARF divides the refined positive samples from the previous stage into two parts $\mathbf{b}^{+}$ and $\mathbf{b}^{-}$, which are defined as below

$$\begin{cases} \mathbf{b}_{t-1}^{+} = \left\{\mathbf{b}_{t-1} \mid \mathrm{IoU}(\mathbf{b}_{t-1}, \mathbf{g}) \geq T_{t-1}^{+}\right\} \\ \mathbf{b}_{t-1}^{-} = \left\{\mathbf{b}_{t-1} \mid T_{t-1}^{-} \leq \mathrm{IoU}(\mathbf{b}_{t-1}, \mathbf{g}) \leq T_{t-1}^{+}\right\} \end{cases} \tag{7}$$

where $\mathbf{g}$ represents the ground truth, and $\mathrm{IoU}(\cdot)$ is used to calculate the spatial overlap between two bounding boxes [52]. T^{-} and T^{+} are the IoU lower bound and IoU upper bound, respectively, used to further divide positive samples into $\mathbf{b}^{+}$ and $\mathbf{b}^{-}$. Among them, $\mathbf{b}^{-}$ consists of samples specifically optimized by the current stage regressor, while $\mathbf{b}^{+}$ includes high-quality samples with higher IoU, which are retained to participate in the regression in the next stage.

The CARF framework divides samples of different qualities, so that the regressor can better perform optimization for samples of specific quality [as illustrated in Fig. 6(d)]. As samples with different qualities are assigned to regressors for separate optimization, this achieves a decoupling between the loss function and sample distribution. In addition, we can simply adjust the loss contribution weights between regressors to balance the regression loss. It is evident that the appropriate selection of parameters for T^{-} and T^{+} determines the interval of sample division, thereby affecting the regression accuracy of the model. The following discussion will delve into the considerations regarding parameter settings and detailed label assignment strategy.

Most refined stage detectors suffer from unreasonable sample quality division, leading to imbalanced loss contribution

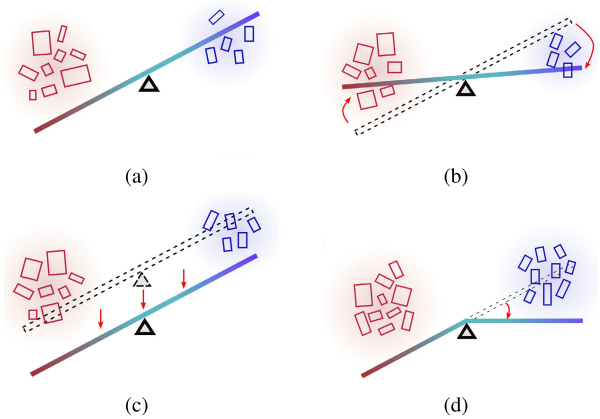


Fig. 6. Comparison of several strategies on samples of different quality during cascade regression process. (a) Classical cascaded regression framework. (b) Weight of loss contribution at different stages. (c) ACM and (d) proposed CARF. Red boxes represent low-quality positives, while blue boxes represent high-quality positives. Samples above the bar indicate those that can be effectively optimized, whereas those below the bar cannot achieve accurate predictions.

across different stages. Our CARF differentiates samples of different quality, optimizing samples with similar regression potential using the same regressor. Consequently, the model adjusts the loss contributions between different regressors, alleviating imbalanced loss contribution across stages. Moreover, dynamic sample quality interval division also resolves the misalignment between IoU thresholds and bounding box distributions, ensuring each regressor can effectively optimize all the assigned samples.

D. Geometry-Guided Training Sample Selection

Objects in remote sensing imagery often exhibit significant variations in scale and aspect ratio. Predefined anchors often struggle to effectively cover objects well, leading to insufficient positives for regression. To adaptively select high-quality

Algorithm 1 Geometric-Guided Training Sample Selection

Input: t is the number of regressors. T_0 is the initial threshold used to select positive samples. $\mathcal{O} = \{\mathbf{O_1}, \ldots, \mathbf{O_t}\}$ is a list containing the IoU between the sample and GT in each regression stage. $\mathcal{G}$ is the set of ground-truth boxes.

Output: $\mathcal{P} = \{\mathbf{P_1}, \ldots, \mathbf{P_t}\}$ contains positives for each regressor. $\mathcal{B}^+ = \{\mathbf{B_1^+}, \ldots, \mathbf{B_t^+}\}$ and $\mathcal{B}^- = \{\mathbf{B_1^-}, \ldots, \mathbf{B_t^-}\}$ are the divisions of $\mathcal{P}$, respectively, as defined in 7.

1: $\mathcal{T} = \{T_0\}$
2: **for** i in $[0, t-1]$ **do**
3: $v = \text{Mean}(\{\mathcal{O}[i] \mid \mathcal{O}[i] \geq T_0\})$
4: $T_i = 0.5 \cdot (\mathcal{T}[-1] + v)$, $\mathcal{T} = \mathcal{T} \cup \{T_i\}$ ▷ Obtain IoU thresholds in each stage;
5: **end for**
6: **for** i in $[0, t-1]$ **do**
7: **for** g in $\mathcal{G}$ **do**
8: $\mathbf{cd} = \sqrt{(\frac{\Delta cx}{w_g})^2 + (\frac{\Delta cy}{h_g})^2}$
9: $\mathcal{C} \Leftarrow \text{TopK}(\mathbf{cd}, k)$ ▷ Calculate shape guide center distance between GT and anchors, and then select the top-k samples as candidates;
10: **for** c in $\mathcal{C}$ **do**
11: $o = \text{IoU}(c, g)$
12: $r = \text{Max}\{\frac{h_g}{w_g}, \frac{w_g}{h_g}\}$, $f(r) = \frac{1}{1+\ln(r)}$
13: $T^+ = T[i+1] \cdot f(r)$, $T^- = T[i] \cdot f(r)$ ▷ Compute shape-guided IoU thresholds for each regressor;
14: **if** $T^- \leq o \leq T^+$ **then**
15: $\mathcal{B}^-[i] = \mathcal{B}^-[i] \cup \{c\}$
16: **else if** $o \geq T^+$ **then**
17: $\mathcal{B}^+[i] = \mathcal{B}^+[i] \cup \{c\}$
18: **end if**
19: **end for**
20: **end for**
21: **if** $i == 0$ **then**
22: $\mathcal{P}[i] = \mathcal{P}[i] \cup \mathcal{B}^-[i]$
23: **else**
24: $\mathcal{P}[i] = \mathcal{P}[i] \cup \mathcal{B}^-[i] \cup \mathcal{B}^+[i-1]$
25: **end if**
26: **end for**
27: **return** $\mathcal{P}$, $\mathcal{B}^-$, $\mathcal{B}^+$

samples, we propose a geometric-guided training sample selection (GTSS) strategy. GTSS incorporates geometric prior information of objects into the training sample selection process in two aspects: 1) IoU threshold and 2) label assignment. The pseudocode of the GTSS is shown in Algorithm 1.

Next, we will introduce the IoU threshold selection at different stages in CARF. As mentioned in Section III-C, we use two IoU thresholds, T^+ and T^-, to dynamically distinguish positives at different stages. Considering that the majority of anchors are negatives, we initially set a lower bound T_0 (set to 0.5) to filter out negatives. Subsequently, we compute the mean of potential high-quality samples at each stage as the partition criterion, as illustrated in line 3 in Algorithm 1. Based on this criterion, a series of IoU thresholds are calculated for different regressors, as shown in line 4 in Algorithm 1 and Fig. 7. The IoU upper bound T^+ of the previous stage serves as the IoU lower bound T^- for the current stage.

Label assignment is subsequently carried out based on the acquired IoU thresholds. Generally, a smaller distance between the anchor center and the object center is more likely to yield favorable spatial alignment effects [10], [45]. Therefore, we incorporate the center distance as prior information

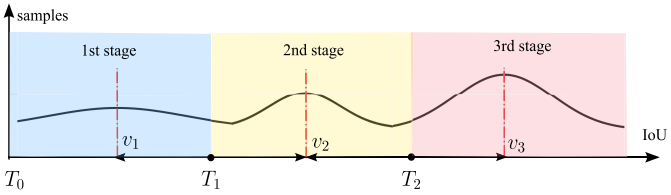
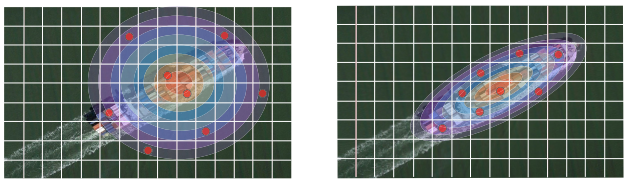


Fig. 7. Illustration of IoU thresholds' determination at different stages. T_0 is set to filter out negatives. v denotes the mean value of potential high-quality samples for each regressor.



(a) (b)

Fig. 8. Comparison of label assignment strategies based on center distance measurements (a) with shape guidance and (b) without shape guidance. Red points are anchor points in images. It shows that shape priors enhance the spatial overlap between anchor points and objects.

into the label assignment process. However, there are many objects with large aspect ratios in remote sensing images. For such objects, the traditional Euclidean distance metric may erroneously regard low-quality anchors with centers located outside the object boundaries as positives (as illustrated in Fig. 8). Hence, we adopt a shape-guided center distance metric (as described in Algorithm 1, line 8). This method dynamically selects anchors with centers within the GT region, ensuring a better coverage of the object region by the anchors.

During the training process, we initially compute the IoU thresholds for each stage and then select the top k samples based on shape-guided center distance. Subsequently, we modulate the IoU thresholds using the aspect ratio (as indicated in Algorithm 1, line 13) to yield the final IoU thresholds, which are used to determine positives at each regression stage.

The traditional topk-based label assignment strategies use the Euclidean distance between anchor points and the center of GT box for training sample selection. These methods treat anchor points at the same distance from the object's center as equally qualified samples. In this case, as shown in Fig. 8(a), many anchor points selected as positives actually do not fall on the object region. Our GTSS method introduces the shape prior of objects, adjusting the calculation of the center distance using the aspect ratio of the target. It can be seen from Fig. 8(b) that this approach places greater emphasis on anchor points near the object area, therefore selecting samples with higher spatial overlap with the object for accurate regression processes.

To address the issue of imbalanced number of positives across stages in the existing cascaded regression frameworks, GTSS is proposed to optimize the training samples among different regressors. First, by adopting shape-guided center distance for label assignment, the number of low-quality samples in earlier regressors is reduced, and samples with localization potential are selected for training. Then, a dynamic sample quality interval division strategy is designed to ensure a consistent distribution of samples across intervals. By balancing the number of positives across stages and selecting

high-quality samples, the detection performance is greatly improved.

E. Loss Function

We used RetinaNet [53] as the baseline and extended it with additional regression stages to achieve a cascaded RetinaNet. The multitask loss is represented as follows:

$$L = L_{\text{cls}} + \lambda \cdot L_{\text{reg}}. \tag{8}$$

In (8), λ is a hyperparameter to adjust the contribution of different losses. Focal loss [53] was used as the classification loss, as shown below

$$L_{\text{cls}} = -\sum_{i=1}^{t}\sum_{j=1}^{N}\left(1-\hat{p}_j\right)^{\gamma}\log\left(\hat{p}_j\right) \tag{9}$$

where

$$\hat{p}_j = \begin{cases} p_j, & \text{if } p_j^* = 1 \\ 1-p_j, & \text{otherwise.} \end{cases} \tag{10}$$

p_j^* is the GT label for classification, while p_i is the corresponding prediction. t is the number of regressors. N denotes the number of anchors.

As discussed in Section III-A, positive samples in earlier stages tend to exhibit larger deviations from the GT, leading to more substantial loss contributions. To address variations in loss contributions across different stages, we introduced IoU loss to build an IoU-balanced loss for regression process. The regression loss is presented below

$$L_{\text{reg}} = \sum_{i=1}^{t}\frac{1}{1-T_i^-}\sum_{j=1}^{N_p}\left(1-\text{IoU}\left(\mathbf{b},\mathbf{b}^*\right)\right) \tag{11}$$

where N_p is the number of positive samples. T^- is the IoU lower bound. $\mathbf{b} = (cx, cy, w, h, \theta)$ is the predicted box, and $\mathbf{b}^* = (cx_g, cy_g, w_g, h_g, \theta_g)$ is the corresponding GT box. IoU loss is scale-invariant and normalizes the deviation between anchors and GT, preventing the loss from being dominated by low-quality positives. Furthermore, we fine-tuned the loss contribution for each stage using the IoU lower bound. This adaptive adjustment further reduces the contribution of earlier stages to the loss, achieving a more balanced and stable training process.

IV. Experiments

In this section, we evaluate the performance of the proposed method on multiple public remote sensing datasets, including HRSC2016 [25], UCAS-AOD [26], DOTA [27], and FAIR1M [28]. Section IV-A will introduce the information of these datasets. Section IV-B describes the experimental setup and parameters. In Section IV-C, we compare the performance of the proposed method with the existing state-of-the-art models. Finally, in Section IV-D, ablation studies are conducted to verify the performance gains of the proposed framework.

A. Datasets

We explored diverse aerial image datasets in our extensive experiments, including HRSC2016 [25], UCAS-AOD [26], DOTA [27], and FAIR1M [28].

The HRSC2016 dataset [25] is collected for high-resolution remote sensing ship detection, including 1061 images with sizes ranging from 300 × 300 to 1500 × 900 pixels. It is partitioned into training, validation, and test sets containing 436, 181, and 444 images, respectively. UCAS-AOD [26] is a dataset for aerial plane and car recognition, comprising 1510 images (1000 for planes and 510 for cars).

DOTA [27] stands out as a substantial dataset with 2806 aerial images and 188 282 annotated instances across 15 categories, 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). Image sizes range from about 800 × 800 to 4,000 × 4000 pixels, and for our experiments, we cropped images into 800 × 800 patches with a stride of 200.

FAIR1M [28] is a benchmark for fine-grained object recognition in aerial imagery, boasting over 1 million instances and 15 000 images. Objects are annotated to 37 categories using OBBs, and image widths vary from 1000 to 10 000 pixels. The objects in FAIR1M include Boeing 737, Boeing 777, Boeing 747, Boeing 787, Airbus A320, Airbus A220, Airbus A330, Airbus A350, COMAC C919, COMAC ARJ21, other-airplane, passenger ship, motorboat, fishing boat, tugboat, engineering ship, liquid cargo ship, dry cargo ship, warship, other-ship, small car, bus, cargo truck, dump truck, van, trailer, tractor, truck tractor, excavator, other-vehicle, baseball field, BC, football field, TC, roundabout, intersection, and bridge.

B. Implementation Settings

In our experiments, the baseline model is the cascaded RetinaNet introduced in Section III-E. Ablation studies were specifically conducted and DOTA datasets. We use two refinement stages in both the baseline model and our framework. For the baseline model, IoU thresholds used for training sample selection are set to 0.4, 0.6, and 0.7 for three regression stages, respectively. For our proposed CARF framework, the initial IoU threshold for the first stage T_0 is set to 0.4. We set $\lambda = 1$ in (8) and $k = 20$ in GTSS for training sample selection.

In the comparison with the state-of-the-art methods on public remote sensing datasets, we strive to ensure consistency with widely adopted training strategies for a fair comparison. The total training iterations were set to 36 epochs for HRSC2016 and UCAS-AOD. For the larger-scale remote sensing datasets DOTA and FAIR1M, models were trained for 12 epochs. The SGD optimizer was used for training, with an initial learning rate set to 1×10^{-3} and the learning rate is divided by 10 at each decay step. Due to the large sizes of images in DOTA and FAIR1M, we adopted a cropping strategy, cropping the images into patches of 1024 × 1024 with a stride of 512. The images in HRSC2016 and UCAS-AOD are resized to 800 × 800 for training and testing. All the models are trained on four NVIDIA RTX 3090 GPUs, with a batch size set to 8.

C. Comparison With State-of-the-Art

1) Main Results on HRSC2016: We conducted a performance comparison between DDDet and the existing state-of-the-art methods on the HRSC2016 dataset, with the

TABLE I
COMPARISON WITH STATE-OF-THE-ARTS ON THE HRSC2016 DATASET

Methods	RRD [54]	RoI Trans. [55]	Gliding Vertex [56]	OPLD [57]	DAL [39]
mAP(%)	84.30	86.20	88.20	88.44	88.95
Methods	CFC-Net [22]	GWD [41]	TIOE-Det [11]	GCL [58]	DDDet (ours)
mAP (%)	89.70	89.85	90.16	90.19	**90.28**

TABLE II
COMPARISON WITH STATE-OF-THE-ARTS ON THE UCAS-AOD DATASET

Methods	FR-O [17]	RoI Trans. [55]	RIDet [23]	CFC-Net [22]
mAP(%)	88.36	89.02	89.23	89.49
Methods	TIOE-Det [11]	DAL [39]	S^2ANet [13]	DDDet (ours)
mAP (%)	89.49	89.87	89.99	**90.03**

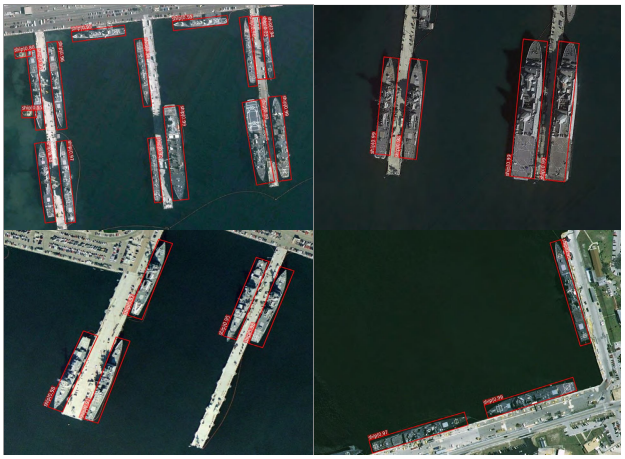


Fig. 9. Some detection results using our method on the HRSC2016 dataset.

experimental results presented in Table I. The HRSC2016 dataset comprises a substantial number of ships with large aspect ratios, which are difficult to detect accurately. DDDet achieves a notable mAP of 90.28% on this dataset. Some detection results are depicted in Fig. 9. It illustrates that DDDet achieves precise localization of objects with large aspect ratios. The GTSS strategy adopted in DDDet helps adaptively select high-quality samples based on object shape, which contributes to the accurate detection of slender ship targets.

2) Main Results on UCAS-AOD: The UCAS-AOD dataset includes lots of cars and airplanes. The object size in UCAS-AOD is small. Therefore, the feature map contains few effective features, which makes it hard to detect object accurately. As shown in Table II, our approach achieved the highest mAP of 90.03% among the compared methods. This success can be attributed to the iterative refinement of region features by the deformable alignment module during cascaded regression. The continual location optimization contributes to a robust feature representation, leading to superior performance in subsequent classification and regression tasks.

3) Main Results on FAIR1M: FAIR1M is a recent remote sensing dataset designed for fine-grained object recognition. In our study on FAIR1M, the experimental results are

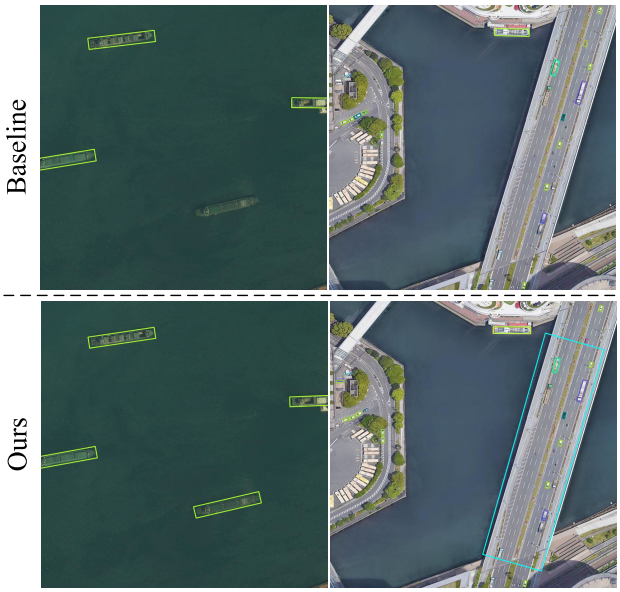


Fig. 10. Visualization of detections on the FAIR1M dataset.

presented in Table III. Our approach achieved the mAP of 44.15%, outperforming other comparative methods. Notably, FAIR1M poses challenges with numerous visually similar objects that are hard to distinguish. Visualization of detections is shown in Fig. 10. It is evident that DDDet accurately localizes objects and reports the correct categories of objects in FAIR1M. Our method selects higher quality samples, thereby achieving precise object detection results compared with the baseline model.

4) Main Results on DOTA: DOTA is currently the most widely used large-scale dataset for oriented object detection in remote sensing images. The objects within the DOTA dataset present notable challenges due to their extensive scale variations and changes in aspect ratios. The experimental results are presented in Table IV. Our proposed DDDet achieves the best performance with the mAP of 79.41%. Visualization of some detections is shown in Fig. 11. As can be found in Fig. 11, DDDet exhibits robust detection capabilities in diverse and intricate scenarios. Examples include densely arranged LVs (first row, first subplot), bridges with large aspect ratios (second row, fourth subplot), and varying-sized ships and ports (second row, second subplot).

D. Ablation Study

1) Evaluation of the Proposed Components in DDDet: First, we conducted experiments on the hyperparameter λ in the loss function. As shown in Table V, the best performance is achieved when $\lambda = 1.0$. The reasonable selection of different λ has little impact on performance, indicating that the loss function is relatively robust. Then, to validate the effectiveness of the proposed modules in DDDet, we conducted componentwise ablation experiments on the DOTA, HRSC2016, and UCAS-AOD datasets. The experimental results are presented in Table VI. The baseline model, a cascaded RetinaNet, achieved the mAP of 72.1% on the DOTA dataset. The aligned convolutional module optimizes feature representation, bridging the spatial gap between anchors and features, leading to a performance improvement of 0.3 points. Next, integrating the CARF into the cascaded regression process further

TABLE III
COMPARISON WITH OTHER METHODS ON THE FAIR1M DATASET

The items with red and blue colors indicate the best and second-best results of each column, respectively. * denotes using multi-scale training and testing.

Method	FCOS [59]	DAL [39]	RIDet [23]	FR-O [17]	CFC-Net [22]	TIOE-Det [11]	Gliding Vertex [56]	RoI Trans. [55]	DDDet (ours)
mAP(%)	23.70	29.00	31.58	33.70	34.31	35.16	35.86	**38.27**	**44.15**
Boeing 737	10.34	32.53	28.25	36.05	30.89	**37.62**	36.32	35.84	**38.98**
Boeing 747	43.54	74.39	80.62	85.19	**83.87**	**86.71**	82.61	82.74	83.83
Boeing 777	5.96	**13.14**	12.92	12.45	10.72	11.06	11.29	12.81	**13.01**
Boeing 787	13.67	39.91	45.28	45.35	38.60	**46.32**	**48.69**	43.90	40.10
C919	0.00	2.11	0.15	15.45	5.67	0.00	**24.48**	15.77	**22.11**
A220	11.71	41.32	39.89	**49.50**	42.44	48.75	**50.01**	48.68	48.73
A321	3.95	58.38	53.69	63.16	50.68	**68.49**	65.27	67.35	**68.84**
A330	15.03	44.59	62.80	65.89	55.13	**69.98**	**69.98**	65.56	63.09
A350	14.20	54.88	55.27	62.69	59.20	**78.19**	65.18	62.92	**73.20**
ARJ21	13.75	1.57	8.53	31.25	5.30	8.62	**33.24**	**33.60**	32.08
passenger ship	**10.65**	3.83	6.11	6.24	7.19	3.73	8.92	**15.20**	9.92
motorboat	46.21	53.04	55.20	44.37	**63.38**	58.45	52.04	58.04	**69.95**
fishing boat	**9.59**	5.71	5.49	3.71	8.72	5.12	5.11	9.37	**11.99**
tugboat	19.81	21.08	30.15	26.05	19.70	**30.51**	28.49	30.17	**38.75**
engineering ship	**13.24**	7.11	5.84	6.88	7.67	10.38	9.73	**10.87**	10.56
liquid cargo ship	12.92	12.05	17.21	9.50	**21.23**	5.56	15.67	19.28	**30.06**
dry cargo ship	**35.08**	28.41	29.58	17.78	30.54	18.71	26.75	33.02	**40.24**
warship	20.75	11.91	14.47	6.37	23.21	2.52	13.67	**24.90**	**40.64**
small car	42.56	48.05	52.73	51.44	62.43	**65.89**	49.53	57.73	**73.45**
bus	15.55	7.71	15.27	21.00	**34.50**	4.73	22.04	31.23	**45.48**
cargo truck	31.72	25.04	30.32	32.89	41.15	36.29	36.69	**42.46**	**52.51**
dump truck	23.90	22.82	29.50	40.04	42.18	41.31	39.52	**45.26**	**55.87**
van	34.59	43.26	45.01	45.96	51.65	**65.89**	43.65	54.49	**74.13**
trailer	12.14	2.48	3.82	7.82	11.41	0.53	11.65	**15.54**	**16.64**
tractor	1.07	1.03	0.05	**3.77**	1.69	0.18	2.90	3.55	**4.18**
excavator	7.90	5.06	5.03	9.28	10.26	9.83	12.49	**12.78**	**20.22**
truck tractor	1.09	0.55	0.53	1.71	0.71	0.10	**3.66**	2.59	**2.88**
basketball court	23.09	38.76	37.47	39.92	40.21	**50.23**	39.85	42.87	**53.14**
tennis court	74.76	75.37	77.78	76.97	79.41	**80.23**	76.98	78.40	**88.86**
football field	49.64	46.10	52.69	52.36	58.01	**60.70**	50.79	59.30	**63.25**
baseball field	82.90	84.66	85.63	87.56	84.34	**88.57**	86.85	86.60	**88.21**
intersection	55.14	44.06	51.41	57.11	51.98	**65.07**	58.59	58.18	**60.22**
roundabout	**26.46**	13.96	17.05	22.28	18.22	21.02	20.49	19.34	**24.79**
bridge	**22.79**	15.08	17.96	7.75	14.31	11.94	16.21	20.76	**33.85**

yielded a performance gain of 0.5 points. CARF optimally allocates samples of different qualities to respective regressors, continuously refining bounding boxes for better performance. On this basis, the GTSS strategy further improved performance by 0.3%. We propose that GTSS introduces object shape priors into the label assignment process, which helps mine more potential high-quality positives. Finally, the adoption of IoU-balanced loss achieved adaptive balanced loss contributions between different stages, leading to a performance improvement of 0.2%. Stable performance gains can also be achieved on HRSC2016 and UCAS-AOD, which proves the effectiveness of the proposed components.

Our method is based on a refine-stage detector; it introduces a minimal amount of parameters and computational overhead. As shown in Table VII, our framework only brings about 0.3% more parameters. When the size of input image is set to 800 × 800, our method leads to a decrease in inference speed by 1.1 frames/s. Overall, the computational cost introduced by the proposed method is very small. DAL [39] is an advanced one-stage detector, and RoI Transformer [55] is a two-stage detector. Our method is a refine-stage method, and therefore, the time cost of our method falls between those of one-stage and two-stage detectors.

2) Evaluation of Effect of CARF: DDDet incorporates CARF to address challenges in optimizing bounding boxes during the cascaded regression process. We further conducted experiments to demonstrate the applicability and robustness of CARF. The results are shown in Table VIII. We explored how CARF improves the performance of cascaded regression detectors with different numbers of refinement stages. The baseline is a basic RetinaNet with AP_{50} of 70.1%. Obviously, the model with two refinement stages achieved a significant performance improvement of 2.3 points. However, when optimizing bounding boxes with three regressors, the performance only increased by 0.2 points. It shows that the additional refinement stages brought limited performance gains. We attribute this to various issues discussed in Section III-A, which leads to performance bottlenecks in cascaded regression detectors. Following this, the introduction of CARF yielded notable performance gains of 1.3%–1.0% for two- and three-regressor

TABLE IV

PERFORMANCE COMPARISON WITH STATE-OF-THE-ARTS ON THE DOTA DATASET

. The items with red and blue colors indicate the best and second-best results of each column, respectively. 'Ms' means using multi-scale training and testing.

	Methods	Ms	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	$AP_{50}(\%)$
One-Stage	CFC-Net [22]	✓	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
	R^3Det [21]	✓	89.80	83.77	48.11	66.77	78.76	**83.27**	87.84	90.82	85.38	85.51	65.67	62.68	67.53	78.56	72.62	76.47
	DAL [39]		89.69	83.11	55.03	71.00	78.30	81.90	88.46	90.89	84.97	**87.46**	64.41	65.65	76.86	72.09	64.35	76.95
	SLA [12]	✓	88.33	84.67	48.78	73.34	77.47	77.82	86.53	90.72	**86.98**	86.43	58.86	68.27	74.10	73.09	69.30	76.36
	GWD [41]	✓	89.06	84.32	55.33	77.53	76.95	70.28	83.95	89.75	84.51	86.06	**73.47**	67.77	72.60	75.76	**74.17**	77.43
	RIDet [23]	✓	89.31	80.77	54.07	76.38	**79.81**	81.99	**89.13**	90.72	83.58	87.22	64.42	67.56	**78.08**	**79.17**	62.07	77.62
	KLD [42]	✓	88.91	85.23	53.64	**81.23**	78.20	76.99	84.58	89.50	86.84	86.38	**71.69**	68.06	75.95	72.23	**75.42**	78.32
	TIOE-Det [11]	✓	89.76	85.23	**56.32**	76.17	**80.17**	**85.58**	88.41	90.81	85.93	87.27	68.32	70.32	68.93	78.33	68.87	**78.69**
Multi-Stage	RoI Trans. [55]	✓	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 [60]	✓	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
	SCRDet [37]	✓	**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
	Gliding Vertex [56]		89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	**70.91**	72.94	70.86	57.32	75.02
	Mask OBB [61]	✓	89.56	**85.95**	54.21	72.90	76.52	74.16	85.63	89.85	83.81	86.48	54.89	69.64	73.94	69.06	63.32	75.33
	CSL [40]	✓	**90.25**	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
	OPLD [57]	✓	89.37	85.82	54.10	79.58	75.00	75.13	86.92	90.88	86.42	86.62	62.46	68.41	73.98	68.11	63.69	76.43
	AProNet [62]	✓	88.77	84.95	55.27	78.40	76.65	78.54	88.45	90.83	86.56	87.01	65.62	70.29	75.43	78.17	67.28	78.16
	DDDet(ours)	✓	89.55	**85.88**	**57.93**	**80.38**	79.56	83.16	**89.13**	90.85	86.94	**87.40**	68.51	**71.30**	**78.41**	**79.29**	62.84	**79.41**

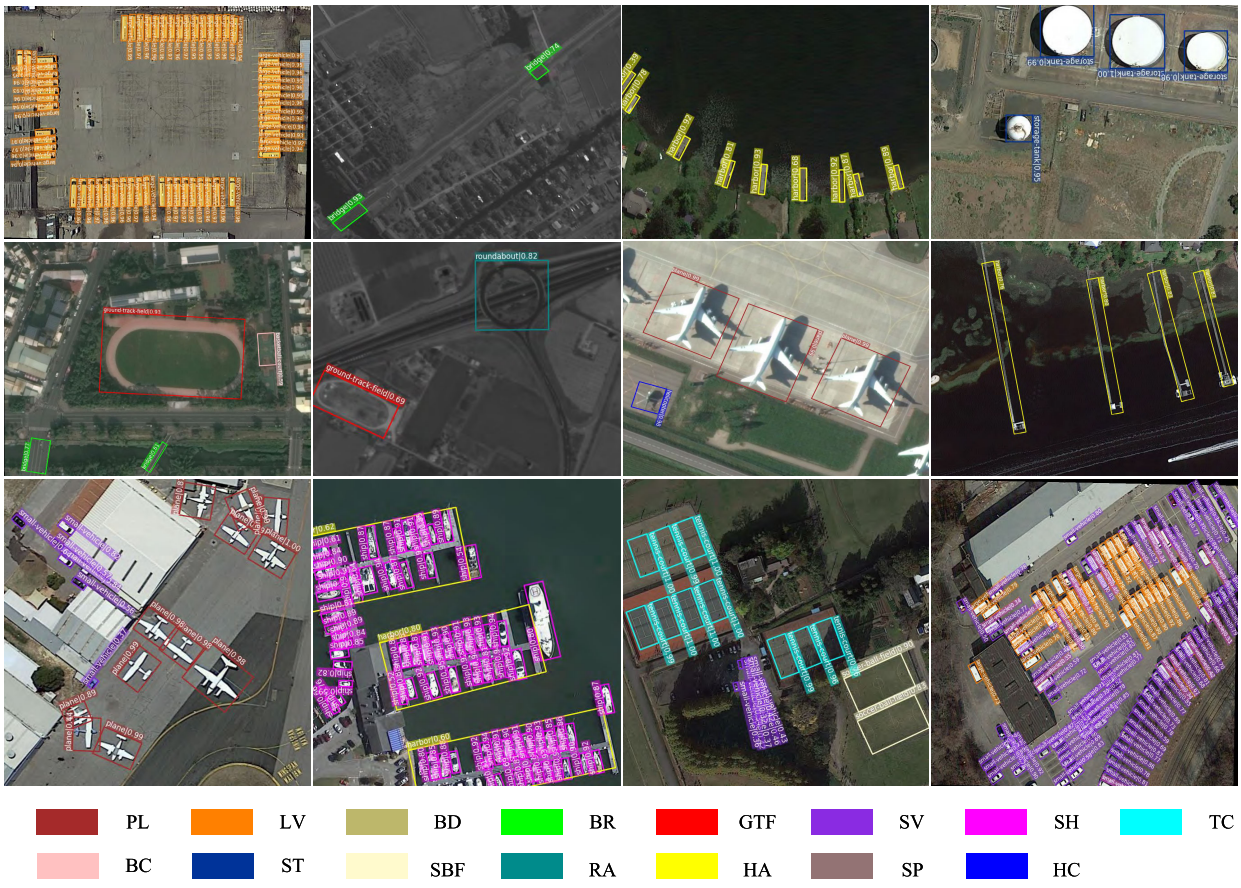


Fig. 11. Visualization results of our method on the DOTA dataset.

detectors, respectively. The experimental results demonstrate that our approach effectively enables the cascaded regression detectors to progress through more stages.

In addition, we introduced AP_{75} metric in our experiments to evaluate the performance of the proposed method in high-precision oriented object detection. Additional refinement stages could generate high-quality candidate boxes, thereby continuously improving the performance of high-quality detection results. Our method further enhances performance on this basis, significantly increasing AP_{75}. When the number of refinement stages is 3, although AP_{75} cannot be further optimized, our method can still achieve stable gains in

TABLE V
IMPACT OF λ IN THE LOSS FUNCTION ON DETECTION PERFORMANCE

λ	0.1	0.5	1.0	2.0	5.0
mAP(%)	71.6	72.9	**73.4**	73.3	72.8

TABLE VI
EFFECTS OF THE PROPOSED COMPONENTS ON REMOTE SENSING DATASET

	Different Models				
+ Aligned Conv.		✓	✓	✓	✓
+ CARF			✓	✓	✓
+ GTSS				✓	✓
+ IoU-balanced loss					✓
$mAP_{DOTA}(\%)$	72.1	72.4	72.9	73.2	**73.4**
$mAP_{HRSC2016}(\%)$	86.9	87.5	88.7	89.1	**89.8**
$mAP_{UCAS\text{-}AOD}(\%)$	88.1	88.3	89.1	89.5	**89.7**

TABLE VII
EVALUATION OF THE PROPOSED METHOD ON MODEL PARAMETERS AND INFERENCE SPEED

Method	DAL [39]	RoI Trans. [55]	Baseline	Baseline + Ours
Parameters(M)	36.1	49.6	36.8	36.9
FPS	18.3	9.1	13.6	12.5

TABLE VIII
EVALUATION OF CARF ON THE DOTA DATASET

Num Stages	1	2		3	
+ CARF	×	×	✓	×	✓
$AP_{50}(\%)$	69.8	72.1	**73.4**	72.3	73.3
$AP_{75}(\%)$	42.3	43.7	45.1	44.1	**45.8**

TABLE IX
ANALYSIS OF THE IMPACT OF PARAMETER IN GTSS

k	1	5	10	20	50
mAP(%)	63.5	71.2	73.1	**73.2**	73.0

high-precision detection by 0.7 points. Therefore, for different application scenarios, the appropriate number of refinement stages can be selected to meet the requirements of the actual task.

3) Evaluation of Settings of GTSS: As demonstrated in Table VI, GTSS has been proven to enhance the detection accuracy in CARF. Furthermore, we explored the influence of different values of k within GTSS. As shown in Table IX, the best performance is 73.2%, which is achieved when $k = 20$. Alternative values near $k = 20$ do not lead to significant performance drop. Therefore, we suggest that the choice of k is robust. However, a very small k, such as $k = 1$, leads to a sharp performance drop by 9.7 points. It proves that selecting only a few positives for training limits the utilization of potential high-quality samples, requiring a longer training schedule for model convergence. Conversely, adopting a large k would select many low-quality samples as positives. This makes it challenging for the model to discern effective features for object recognition, therefore resulting in performance degradation.

V. CONCLUSION

In this article, we identified a significant challenge in current cascaded regression framework—bounding boxes lack high-quality continuous optimization. Furthermore, this observation prompted an in-depth analysis of the phenomenon. To address these limitations and achieve substantial performance improvements, we introduced the DDDet. Specifically, the innovative CARF is proposed to uniquely address the challenge by distinguishing bounding boxes with different distributions, adaptively optimizing them within the well-assigned regressors. Next, the aligned convolution module (ACM) is incorporated into each regressor, enabling the continuous acquisition of high-quality features. Furthermore, we integrated the GTSS method into CARF for adaptive label assignment. Our DDDet achieves state-of-the-art performance on multiple mainstream oriented object detection datasets. The experimental results validate the superiority of the proposed approach.

However, the proposed framework still has some limitations. The continuous optimization of bounding box framework is designed to meet the demands of high-accuracy detection tasks, where the accuracy of object detection is more important than inference speed. Although our method has improved performance while ensuring inference speed, there is still room for improvement by sacrificing speed advantages to further increase detection accuracy. Compared with the bounding box optimization of the refine-stage detectors, the RoI-based bounding box regression in two-stage detectors, despite being time-consuming, can better align features to achieve performance improvements. Since RoI-based operations typically output a fixed number of candidate boxes, the distribution of the bounding boxes is different from those of detectors in the refine-stage detectors. Therefore, future work will explore the issue of uneven distribution of bounding boxes in the cascaded regression framework of two-stage detectors, and further improve high-quality detection performance.

REFERENCES

[1] Z. Lv, M. Zhang, W. Sun, J. A. Benediktsson, T. Lei, and N. Falco, "Spatial-contextual information utilization framework for land cover change detection with hyperspectral remote sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411911.

[2] Z. Lv, J. Liu, W. Sun, T. Lei, J. A. Benediktsson, and X. Jia, "Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411115.

[3] 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.

[4] 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.

[5] 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.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[7] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[8] 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, 2010.

[9] Q. Ming and X. Xiao, "Towards accurate medical image segmentation with gradient-optimized dice loss," *IEEE Signal Process. Lett.*, vol. 31, pp. 191–195, 2024.

[10] J. Song, L. Miao, Q. Ming, Z. Zhou, and Y. Dong, "Fine-grained object detection in remote sensing images via adaptive label assignment and refined-balanced feature pyramid network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 71–82, 2023.

[11] Q. Ming, L. Miao, Z. Zhou, J. Song, Y. Dong, and X. Yang, "Task interleaving and orientation estimation for high-precision oriented object detection in aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 196, pp. 241–255, Feb. 2023.

[12] Q. Ming, L. Miao, Z. Zhou, J. Song, and X. Yang, "Sparse label assignment for oriented object detection in aerial images," *Remote Sens.*, vol. 13, no. 14, p. 2664, Jul. 2021.

[13] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5602511.

[14] J. Han, J. Ding, N. Xue, and G. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2786–2795.

[15] M. Zand, A. Etemad, and M. Greenspan, "Oriented bounding boxes for small and freely rotated objects," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4701715.

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

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[18] 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. (CVPR)*, Jun. 2016, pp. 779–788.

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

[20] 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.

[21] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3163–3171.

[22] Q. Ming, L. Miao, Z. Zhou, and Y. Dong, "CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605814.

[23] Q. Ming, L. Miao, Z. Zhou, X. Yang, and Y. Dong, "Optimization for arbitrary-oriented object detection via representation invariance loss," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[24] 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.

[25] 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*, vol. 2, 2017, pp. 324–331.

[26] 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.

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

[28] X. Sun et al., "FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 116–130, Feb. 2022.

[29] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf.* Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[30] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[31] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 379–387.

[32] 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. (CVPR)*, Jun. 2014, pp. 580–587.

[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.

[34] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[35] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[36] S.-H. Bae, "Deformable part region learning and feature aggregation tree representation for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10817–10834, Sep. 2023, doi: 10.1109/TPAMI.2023.3268864.

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

[38] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16794–16805.

[39] Q. Ming, Z. Zhou, L. Miao, H. Zhang, and L. Li, "Dynamic anchor learning for arbitrary-oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2355–2363.

[40] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 677–694.

[41] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.

[42] X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback–Leibler divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18381–18394.

[43] B. Zhong and K. Ao, "Single-stage rotation-decoupled detector for oriented object," *Remote Sens.*, vol. 12, no. 19, p. 3262, Oct. 2020.

[44] K. Kim and H. S. Lee, "Probabilistic anchor assignment with IoU prediction for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Aug. 2020, pp. 355–371.

[45] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[46] L. Hou, K. Lu, J. Xue, and Y. Li, "Shape-adaptive selection and measurement for oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, 2022, pp. 923–932.

[47] J. Guan, M. Xie, Y. Lin, G. He, and P. Feng, "EARL: An elliptical distribution aided adaptive rotation label assignment for oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5619715.

[48] Y. Yu, X. Yang, J. Li, and X. Gao, "Object detection for aerial images with feature enhancement and soft label assignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624216.

[49] C. Zhang, B. Xiong, X. Li, and G. Kuang, "TCD: Task-collaborated detector for oriented objects in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4700714.

[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] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[52] 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.

[53] 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. 2980–2988.

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

[55] 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. (CVPR)*, Jun. 2019, pp. 2849–2858.
[56] 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.
[57] Q. Song, F. Yang, L. Yang, C. Liu, M. Hu, and L. Xia, "Learning point-guided localization for detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1084–1094, 2021.
[58] Q. Ming, L. Miao, Z. Zhou, J. Song, and A. Pizurica, "Gradient calibration loss for fast and accurate oriented bounding box regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611015.
[59] 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.
[60] 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.
[61] J. Wang, J. Ding, H. Guo, W. Cheng, T. Pan, and W. Yang, "Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images," *Remote Sens.*, vol. 11, no. 24, p. 2930, Dec. 2019.
[62] X. Zheng, W. Zhang, L. Huan, J. Gong, and H. Zhang, "AProNet: Detecting objects with precise orientation from aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 181, pp. 99–112, Nov. 2021.

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 currently 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 Harbin Institute of Technology, Harbin, China, in 1986 and 1989, respectively, and the Ph.D. degree in control theory and engineering from China Academy of Launching Vehicle Technology, Beijing, China, in 2001.

Since 1992, she has been with Beijing Institute of Technology, Beijing, first as a Lecture, since 1996 as an Associate Professor, and since 2001 as a Professor. Her main 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 currently an Associate Professor with the School of Automation, BIT. His research interests include information fusion, pattern recognition, digital image processing, and vision-based navigation.

Nicolas Vercheval received the B.S. and M.S. degrees in mathematics and applied mathematics from the University of Bologna, Bologna, Italy, in 2013 and 2017, respectively. He is currently pursuing the Ph.D. degree with Ghent University, Ghent, Belgium.

His research interests include hidden representation, generative models, and explainability of deep learning models.

Aleksandra Pižurica (Senior Member, IEEE) received the Diploma degree in electrical engineering from the University of Novi Sad, Novi Sad, Serbia, in 1994, the M.Sc. degree in telecommunications from the University of Belgrade, Belgrade, Serbia, in 1997, and the Ph.D. degree in engineering from Ghent University, Ghent, Belgium, in 2002.

She is currently a Professor of statistical image modeling with Ghent University. Her research interests include the area of signal and image processing and machine learning, including multiresolution statistical image models, Markov random field models, sparse coding, representation learning, and image and video reconstruction, restoration, and analysis.

Prof. Pižurica is a member of the EURASIP Technical Area Committee Signal and Data Analytics for Machine Learning. She received the Scientific Prize "de Boelpaepe" by the Royal Academy of Science, Letters and Fine Arts of Belgium for her contributions to statistical image modeling and applications to digital painting analysis, in 2015. She received numerous other recognitions for her work, among which as co-recipient of the David Hestenes Prize at AGACSE 2018 and the Best Paper Award of the IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion contest, in 2013 and 2014. She served as the TPC Co-Chair of the 30th EUSIPCO Conference (in 2022), an Europe Liaison for IEEE ICIP 2020 and ICIP 2024, the Plenary Co-Chair of EUSIPCO 2024, and elected as the TPC Co-Chair of IEEE ICIP 2026. She served as an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING from 2012 to 2016 and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2016 to 2019. She is a Senior Area Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING from 2016 to 2019 and since 2022.