Context-Based Oriented Object Detector for Small Objects in Remote Sensing Imagery

QUNYAN JIANG, JUYING DAI, TING RUI, FAMING SHAO, GUANLIN LU, AND JINKANG WANG
Department of Mechanical Engineering, College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China
Corresponding author: Juying Dai (dinajy2001@aeu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671470, and in part by the China National Key Research and Development Program under Grant 2016YFC0802904.

ABSTRACT Object detection in remote sensing imagery is a challenging task in the field of computer vision and has high research value. To improve the classification accuracy and positioning accuracy of object detection, we propose a new multi-scale oriented object detector suitable for small objects. Firstly, the feature fusion network based on information balance (IBFF) is proposed to reduce the reuse of different layers’ features from the backbone network and reduce the interference of redundant information based on the premise that the output features have sufficient information, and retain enough shallow detail information. Secondly, to efficiently utilize deep and shallow features, enhance important features, and reduce background noise interference, different attention-based context feature fusion modules (DACFF) are designed according to the characteristics of different feature fusion stages. Finally, an improved strategy of oriented bounding box regression is proposed to obtain the oriented bounding box with a simpler and more effective strategy. The proposed method was evaluated on two public remote sensing datasets, DOTA and HRSC2016, and their mAP values are 80.96% and 95.01%, respectively, which verified the effectiveness of the proposed algorithm.

INDEX TERMS Object detection, remote sensing imagery, feature fusion, attentional mechanism.

I. INTRODUCTION
As an important research direction of computer vision, optical remote sensing object detection is widely used in the fields of traffic guidance, military monitoring, and urban planning, among others. Optical remote sensing is an important basic technology and has important research significance. In recent years, although object detection algorithms in natural scenes have made great breakthroughs, due to the particularity of remote sensing imagery, the object detection algorithms in this field still face great challenges.

Remote sensing datasets mainly have the following characteristics [8], [9]: (1) Large direction change. Remote sensing imagery is taken from an aerial perspective, and the direction of the captured object is random, so the direction of the object is uncertain. Moreover, the traditional horizontal bounding box cannot tightly surround the object. (2) Small and dense objects. The numerous small objects in remote sensing imagery account for a large proportion of overall objects and have dense distribution. These objects interfere with each other, making it very easy to miss the detection of objects. (3) Objects with large size changes. The sizes of objects in different categories or within the same category vary greatly under different resolutions, resulting in large scale changes and great difficulty in detection. Because of the above characteristics, it is difficult to detect remote sensing images. Therefore, appropriate methods should be selected according to the characteristics of remote sensing datasets, and the traditional target detection method should be improved in order to apply it to object detection in remote sensing imagery.

Earlier optical object detection algorithms in remote sensing imagery are based on manual design features. Although these algorithms benefit from strong interpretability, they
have poor robustness, a limited scope of application, and low detection accuracy. Early detection algorithms cannot meet the needs of practical applications. In recent years, with the development of deep learning, excellent object detection networks continue to emerge [1], [2], [3], [4], [5], and the classification and positioning accuracy have been greatly improved. However, traditional object detection is often used to detect the horizontal angle of view, such as PASCAL VOC [6] and ImageNet [7]. Compared to remote sensing datasets, the objects contained in this kind of datasets are large in size, small in number, and feature fixed object orientation.

Because of the particularity of remote sensing detection objects, it is not good to use the commonly used object detection algorithms directly. In recent years, great progress has been made in performance improvements based on the RCNN framework [10], [11], [12], [13], [14], [15], [16], [17] in the research field of remote sensing object detection. These methods provide good detection performance by using the horizontal anchor as the region of interest (ROI) and then relying on region-based features for category recognition [10], [11], [15]. Non-maximum suppression (NMS) is used as a post-processing technology to eliminate redundant anchors and find the best position for object detection. Based on the horizontal anchor, using the NMS method often leads to the problem of missed detection. Therefore, researchers choose the oriented bounding box as the positioning bounding box for remote sensing object detection to make the object position more accurate and reduce the number of missed objects. However, current detection algorithms still have great limitations. In order to improve the ability of object detection in remote sensing imagery, the network layers of the detection model are often deepened. However, with the deepening of the convolution layer, the resolution decreases significantly, and the geometric details of the feature map will be difficult to carry. Especially for small objects, the geometric information in the deep features is seriously lost, and the semantic information is diluted and reduced by the surrounding environmental information. For this reason, the detection of small and dense objects in remote sensing imagery is still a great challenge. The contributions of this paper are as follows:

1) To improve the detection performance of small and dense objects, this paper proposes an oriented object detector named CIODet based on context information.

2) For small objects, a feature fusion network based on the information balance of different layers is proposed. The proposed feature fusion network makes the deep features and shallow features from the backbone network have the same utilization rate, to ensure that the features output from the feature fusion network have sufficient semantic information, while preserving the detail information of the shallow features as much as possible, so as to improve the detection accuracy of small objects.

3) For the chaotic background, an attention-based context feature fusion module is proposed. This module makes efficient use of the deep and shallow features, highlights the important features, reduces the interference of background noise, and makes full use of context information.

4) To make the prediction box realize directional encirclement, the strategy of oriented bounding box regression is improved, and the oriented bounding box is obtained with a simpler and more effective strategy that has obvious improvement effects on slender objects.

The rest of this paper is organized as follows. The relevant work is outlined in Section II. Section III describes the proposed algorithm in detail, and Section IV verifies the effectiveness of our proposed algorithm by comparing it with other methods. Section V provides a summary.

II. RELATED WORK

A. OBJECT DETECTION IN REMOTE SENSING IMAGERY

According to classification of the anchor, the oriented object detectors in remote sensing imagery can be divided into anchor-based detectors and anchor-free detectors [18], [19], [20], [21], [22]. This paper mainly focuses on anchor-based detectors. Anchor-based detectors can be divided into multi-stage detectors [23], [24], [25], [26], [27], [28], [29], [30], [31] and one-stage detectors [32], [33], [34], [35], [36], [37].

You Only Look Once [38] and Single Shot Multibox Detector [5] are classic one-stage algorithms. Although one-stage algorithms have advantages in speed, their accuracy is lower than that of two-stage algorithms. The region-based detection algorithm [3], [4] generates a group of more accurate proposals in the first stage and sends them to the RCNN network. In the second stage, the algorithm carries out classification and regression, which greatly improves the accuracy of object detection. Due to the particularity of remote sensing imagery objects, the effects when directly using these general algorithms are not good; thus, these algorithms need to be adjusted and improved. The ROI transformer [14] converts the horizontal region of interest from RPN output to the rotating region of interest. This strategy does not need to increase the number of anchors and can obtain an accurate rotating region of interest. Although the detection accuracy of this method has been improved, it still needs to be improved greatly, especially for small objects and very large objects. SCRDet [31] is designed as a sampling fusion network that integrates multi-layer features into effective anchor sampling to improve the detection sensitivity of small objects. The attention mechanism has a certain effect on the improvement of object detection, and it also has a good effect on the detection of remote sensing images when the attention mechanism is added to the feature extraction. CAD-Net [26] uses spatial attention to learn object collaboration, and integrates global context information in object detection. While improving detection accuracy, improving efficiency is also very important. \textit{R}D\textit{Det} [36] uses horizontal anchors in the first stage to obtain faster speeds and more proposals. In the refinement stage, this algorithm uses refined rotated anchors to adapt to dense scenes. It is found that rotation feature plays an important role in detection and classification. ReD\textit{Det} [24] is proposed to adaptively extract
rotation invariant features from equivariant features according to the direction of the ROI based on rotation-equivariant features to improve the accuracy of detection.

B. MULTISCALE FEATURE FUSION
In a deep neural network, the feature map extracted by the shallow network contains rich bottom information such as location and edge, while the deep feature map has high-level semantic information. To better extract the context information of features at different scales, we use multi-scale feature fusion. By complementing the information of deep features with shallow features, cross layer information flow can be obtained to improve the performance of the network [39], [40], [41], [42], [43], [44]. FPN [40] transfers deep semantic information to the bottom layer to supplement the shallow semantic information and thus obtain high-resolution and strong semantic features; this process has a good effect on the detection of small objects. PANet [41] adds strong positioning features from the shallow layer to deep layer based on FPN to further improve the detection accuracy of deep features. Based on FPN and PANet, NAS-FPN [45] seeks to find the best path of information flow among various multi-scale features and demonstrates the importance of repeatedly following top-down and bottom-up paths. Great progress has been made in the schemes for fusing multi-scale feature maps to obtain more informative features. In recent years, increasingly more researchers have included the idea of an attention mechanism [46], [47] to better highlight more important features. Inevitably, there will be local position mismatches in the feature map fusions of different layers. To solve this problem, researchers have explored a variety of methods to deal with the features before and after fusion [48], [49].

C. ORIENTED BOUNDING BOX REGRESSION
Conventional object detection generally uses a horizontal bounding box during positioning, but the detection objects in the remote sensing dataset have the characteristics of a small size, dense distribution, and random direction. Therefore, using the horizontal bounding box for detection can easily cause inaccurate object positioning and object loss. To solve this problem, oriented bounding box detection is introduced into the field of remote sensing to study the problem of oriented object detection. The oriented bounding box gives the detected object a more compact surrounding bounding box during detection to obtain accurate positioning. Most existing oriented object detectors are improved based on the five-parameter method and eight-parameter method. To detect objects in any direction, the authors in [8] regressed the coordinates of the four vertices of the oriented bounding box based on Faster R-CNN. An anchor with a rotation angle is proposed to convert the horizontal RoI into rotational RoI. At the same time, rotational RoI and rotational RoI learning are combined to avoid a large number of anchors. The oriented bounding box is often used for scene text detection, but it also has important reference value for other object detection. In the first stage of $R^2$CNN [51], the region proposal network still uses the horizontal bounding box to extract the region of interest. In the second stage, the oriented bounding box is regressed based on the horizontal candidate region to reduce memory consumption. The RoI transformer [14] then inserts an RROI learner between RPN and RCNN to convert the horizontal region generated by RPN into a rotating region to ensure high efficiency and low complexity. Although the RRoI learner can capture rotational features, it cannot give the generated feature map rotational invariance. Therefore, the authors in [24] used ReDet to add a rotation-equivariant network to the detector to extract the rotation-equivariant features and thus accurately predict the direction and reduce the size of the model. To avoid confusion about the sequential label points for oriented objects, the authors in [27] located the quadrilateral by learning the offset of four points on a non-rotating rectangle and used the quadrilateral to determine the position of the object.

III. MATERIALS AND METHODS
A. OVERALL ALGORITHM FRAMEWORK
The overall framework of the network is shown in Fig. 1. Firstly, the image is input into the backbone network (Resnet101) for feature extraction. Then, the features of the three layers from the conv3, conv4, and conv5 modules are extracted. We next send the three features into the IBFF network for cross-layer fusion to obtain three scale feature maps with richer positioning data and semantic information. Then, we use the RPN network to obtain the region of interest and expand the threshold of NMS to reduce the risk of missing dense objects [20]. Finally, according to the classification and regression results, the final oriented bounding box is generated using the improved oriented bounding box algorithm.

B. CROSS-LAYER FEATURE FUSION NETWORK
There are multi-scale objects in the imagery, which makes the information content of each pixel on the imagery significantly different. This situation makes the optimal network depth for dealing with small and large objects different. As the number of network layers increases, the edge and other detail information of the detected object will gradually lose, and the semantic information will gradually lose as the network depth exceeds a certain threshold. Compared with large objects, the loss of detail information of small objects is more severe with the increase of model levels, and the optimal level of semantic information extraction is also relatively shallow. Therefore, the previous work shows that the fusion of different layers of features can perform better in classification and regression. In order to fully retain the detail information of small objects and extract rich semantic information, a feature fusion network structure is designed, as shown in Fig. 1. In this way, the information of different layers is more evenly used in the process of feature fusion, so that each object retains enough semantic information and detail information as much as possible, which is conducive to the detection of small objects.
In the previous feature fusion scheme of FPN + PAN, the output feature maps of different layers suffer from the problem of information imbalance. As shown in Fig. 2, feature map $F_b1$ uses $F1$, $F2$, and $F3$ once each, while $F_b2$ and $F_b3$ reuse some of the feature maps output by the backbone network. The structure of FPN + PAN encoder and decoder makes the shallow information lack of effective processing, and the semantic information has undergone too many unnecessary fusion, so that the noise in the shallow information cannot be suppressed, while the deep semantic information is reduced due to dilution by the surrounding environmental information. We believe that undifferentiated fusion should be carried out for different layers of information, and key information should be used to make the fusion information more effective.

As shown in Fig. 1, firstly, we fuse the three features of the output of the backbone from the top-down path to obtain $F_t3$, $F_t2$, and $F_t1$. $F_t1$ combines the features of $F1$, $F2$, and $F3$ at the same time, and $F_t2$ combines the features of two layers, $F2$ and $F3$. $F_t3$ does not carry out feature fusion. To make $F_t2$ and $F_t3$ fuse the information of the three feature maps $F1$, $F2$, and $F3$ (shown in the green area), $F1$ is convoluted and fused with $F2$ to obtain $F_b2$. $F_b2$ is then convoluted and fused with $F_t3$ to obtain $F_c3$. After convolution, $F_b1$ is fused with $F_t2$ to obtain $F_c2$. Then, $F_c1$, $F_c2$, and $F_c3$ all contain the features of $F1$, $F2$, and $F3$ after one fusion to obtain the features with richer bottom information and semantic information. The output features of the IBFF network contain relatively balanced information, without excessive redundancy and with less interference noise, which is conducive to improving the performance of object detection.

**C. ATTENTION-BASED CONTEXT FEATURE FUSION MODULE**

The attention mechanism emphasizes or selects the important information contained in the processing object by redistributing the weight parameters. This mechanism also suppresses some irrelevant detail information and focuses attention on useful information. The shallow feature map has higher resolution and contains more geometric detail information, which is beneficial to small object detection. Deep feature map...
has low resolution, large receptive field and rich semantic information, which is also very important for small object detection. To make efficient use of deep and shallow features, highlight important features and reduce background noise interference, different attention-based context feature fusion modules are designed according to the characteristics of feature fusion in different stages. When features are fused via the top-down path, we emphasize the integration of the semantic information of deep features into shallow features and adopt the method of sharing the channel attention map of deep features for feature fusion. When features are fused via the bottom-up path, we emphasize the integration of the bottom information of shallow features into deep features and adopt the method of sharing the spatial attention map of shallow features for feature fusion. After conducting feature fusion on the two paths, the features include more abundant information. Since different features have different contribution levels to the detection results, we apply a newly designed cross-layer feature fusion module, as shown in Fig. 3 (c), to obtain the final output features. The calculation formula for channel attention is as follows:

\[ M^C(F) = \sigma(\text{MLP}(F^c_{\text{max}}) + \text{MLP}(F^c_{\text{avg}})) \]  

(1)

where \(\sigma\) represents the sigmoid function; MLP is the multilayer perceptron; and \(F^c_{\text{max}}\) and \(F^c_{\text{avg}}\) represent the feature maps after maximum pooling and average pooling, respectively, and generate two different sets of descriptive information. Then, the two features are sent to the shared network, and the elements are summed. After the sigmoid function, the channel attention map \(M_c(F) \in \mathbb{R}^{C \times 1 \times 1}\) is generated. The calculation formula of spatial attention is as follows:

\[ M^S(F) = \sigma(f^{7 \times 7}(\text{[F}^s_{\text{avg}}; F^s_{\text{max}}])) \]  

(2)

where \(f^{7 \times 7}\) represents the convolution operation with a filter size of \(7 \times 7\). \(F^s_{\text{avg}} \in \mathbb{R}^{1 \times H \times W}\), and \(F^s_{\text{max}} \in \mathbb{R}^{1 \times H \times W}\) represents the results obtained after average pooling and maximum pooling along the channel direction, respectively. After connecting them, effective feature descriptors are generated. After using convolution and the sigmoid function, the feature descriptor is generated as a spatial attention map that emphasizes or suppresses the location information.

**D. ORIENTED BOUNDING BOX REGRESSION**

Five-parameter representation and eight-parameter representation are common schemes of oriented bounding box parameter representation. Five-parameter representation suffers from a boundary problem caused by the periodicity of the angle, which will produce unnecessary losses, resulting in a poor effect of the detection bounding box. Due to the ambiguity of the starting and ending order of the four vertices in eight-parameter representation, the unique representation...
can be determined by manually setting the rules and starting point. However, this method also produces boundary problems and affects the prediction of the vertex position on the boundary. To avoid the above problems, this paper slides the vertices on the horizontal bounding box to each corresponding edge to accurately describe an oriented object. The introduced parameters include four length ratios and obliquity factors. The four length ratios represent the relative slip offset on each boundary of the quadrilateral. The obliquity factor is the ratio of the area of the oriented bounding box to its horizontal bounding box, and horizontal bounding box detection is selected for objects that are close to horizontal.

The idea of this method is intuitively depicted in Fig. 4. The green box represents the oriented bounding box $B_o$, and the orange box represents the horizontal bounding box $B_h$. We use $(x, y, c_1, c_2, c_3, c_4)$ to represent the oriented bounding box, and the horizontal rectangle can be represented by $(x, y, w, h)$. Here, $(x, y)$ represents the center point of the rectangle, and $(w, h)$ represents the width and height of the rectangle. The obliquity factors $r$ is the area ratio between $B_o$ and $B_h$, indicating the inclination degree of the object. Here, we adopt a more simple and efficient expression to improve the regression strategy:

$$
\begin{align*}
    r &= |O|/|B_h| \\
    \alpha_1 &= x_1 + \alpha_2 + \alpha_3 + \alpha_4 \\
    \alpha_{(1,3)} &= ||l_{(1,3)}||/w, \\
    \alpha_{(2,4)} &= ||l_{(2,4)}||/h,
\end{align*}
\tag{3}
$$

where $\alpha_i, i \in \{1, 2, 3, 4\}$ is an additional variable, $||l_i|| = ||c_i - c'_i||$ is the distance between $c_i$ and $c'_i$, and $l_i = (c_i, c'_i)$ represents the slip distance from $c'_i$ to $c_i$. It is worth noting that for horizontal objects, all $\alpha_i$ values are set to 1, where $|\cdot|$ is the cardinality, and we choose whether to use horizontal or directional detection as the final result according to the value of $r$. If the value is close to 1, we use the horizontal detection box, and if the value is close to 0, we use the oriented detection box. The expression of regression loss $L_{reg}$ is as follows:

$$
L_{reg} = \lambda_1 \times L_h + \lambda_2 \times L_\alpha + \lambda_3 \times L_r
\tag{5}
$$

$$
L_\alpha = \sum_{i=1}^{4} \text{smooth}_{L_i}(\alpha_i - \bar{\alpha}_i)
\tag{6}
$$

where $L_h$ is the loss of horizontal bounding box regression, which is the same as that in [4]; $L_\alpha$ is the loss of length ratio ($\alpha_1, \alpha_2, \alpha_3, \alpha_4$) regression; $L_r$ is the loss of obliquity factors $r$ regression; and $\lambda_1, \lambda_2, \lambda_3$ are super parameters that balance the importance of each loss term.

IV. EXPERIMENTS

In this section, we demonstrate the effectiveness of our method on DOTA and HRSC2016. Then, we compare our method with the state-of-the-art methods. Finally, we use ablation experiments to verify the effectiveness of the proposed methods.

A. DATASETS

**DOTA Dataset:** DOTA is a remote sensing imagery dataset from different sensors and platforms used for remote sensing object detection. This dataset includes 2806 images with sizes between 800 × 800 and 4000 × 4000, for a total of 188,282 instances. We use the training and validation sets for training and the test set for testing. Since the size of each image in DOTA is about 4000 × 4000 pixels, the original image is cropped to a size of 1024 × 1024 with a stride of 612. As shown in Fig. 5, statistics are provided on the sizes of objects in the training set of DOTA. This dataset presents significant challenges. As shown in Fig. 5(a), most objects in the dataset are small- and medium-size objects, with small objects accounting for 62.01% and medium objects accounting for 31.06%. The dataset also contains many slender objects. The model is trained through 180000 iterations, in which the learning rate starts from 0.0005 and decreases to 0.0001 and 0.00001 at 120000 iterations and 150000 iterations, respectively.

**HRSC2016 Dataset:** HRSC2016 is a challenging ship detection dataset with arbitrary orientation. The dataset is obtained from Google Earth, and all images are taken at six famous ports. The size of each image in the dataset ranges between 300 × 300 and 1500 × 900, with a total of 2,976 instances. As shown in Fig. 5(b), the dataset contains a large number of slender objects with a large aspect ratio. The training set and test set include 617 images and 444 images, respectively. For experiments on HRSC2016, we scale the image to 512 × 512 for training and testing. Here, the model is trained through 120000 iterations, in which the learning rate starts from 0.0005 and decreases to 0.001 and 0.0001 at 80000 iterations and 100000 iterations, respectively.

B. EXPERIMENTAL SETUP AND PARAMETER EVALUATION

The experiment is conducted under the PyTorch deep learning framework using a GeForce RTX 2080Ti GPU along with the random gradient descent method (SGD) to train the network. Here, the momentum parameter is 0.949, and the initial learning rate is 0.0005. In the training phase, hard negative mining is used, and random rotation, random clipping, and random
image distortion are used as data enhancement strategies to enrich the samples.

We mainly use the Average Precision (AP), mean Average Precision (mAP), and Precision-Recall (PR) curves to evaluate model performance. The PR curve describes the relationship between precision and recall. Precision and recall are defined as follows:

\[
\text{precision} = \frac{TP}{TP + FN} \quad (8)
\]
\[
\text{recall} = \frac{TP}{TP + FP} \quad (9)
\]

where, for TP, the predicted value is the same as the true value, and the predicted value is the positive sample; for FP, the predicted value is different from the true value, and the predicted value is a positive sample; and for FN, the predicted value is different from the true value, and the predicted value is a negative sample. AP is the average of the precision at different recall rates, and mAP represents the average of AP in all categories. The calculation formula is as follows:

\[
mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i \quad (10)
\]

C. COMPARISONS WITH STATE-OF-THE-ART METHODS

Results on the DOTA Dataset: As shown in Table 1, we arrange the physical objects according to their general physical dimensions from left to right (largest to smallest), with the largest physical objects represent in the leftmost column and the smallest physical objects in the rightmost column. Our proposed approach for DOTA is then compared with other state-of-the-art approaches. The mAP of the proposed network is found to be 80.96%, which is better than the mAP values of other advanced networks and demonstrates our approach’s effectiveness in remote sensing imagery object detection.

Compared with the state-of-the-art methods, our method can achieve the highest AP in five categories. For example, the AP of bridges is 63.94%, the AP of planes is 91.62%, the AP of ships is 89.03%, the AP of large vehicles is 88.69%, and the AP of small vehicles is 80.61%. These results show that our proposed method has good detection performance for some small, dense, and slender objects. The experimental results in Fig. 6 further qualitatively demonstrate that the proposed model offers good detection performance for such objects, highlighting the effectiveness of the proposed model.

However, a small-scale physical object does not correspond directly to a small-scale detection object. In short, when the spatial resolution of the visual sensor is low, an object still has the opportunity to obtain a larger pixel detection object and thus more complete information expression in the dataset image. In response to this problem, we conduct a performance analysis based on the information in Table 2. The results are shown in Table 2. The AP results using our method are 79.17%, 85.89%, and 72.14% for small, medium, and large objects, respectively. It can be clearly seen that in our method, small-scale objects feature significant improvements in detection performance, which further verifies the effectiveness of the strategy we proposed.

Results on the HRSC2016 Dataset: The objects in DOTA contain a large number of aircraft, storage tanks, and baseball diamonds, as well as other objects with small aspect ratios. These types of objects can achieve good bounding box effects, even if there is a certain angle deviation, by approaching the square bounding box. Thus, it is difficult to effectively
verify the effectiveness of our rotating oriented bounding box regression strategy. We further experiment with using the proposed method on HRSC2016 because HRSC2016 is full of ship objects, which have a distinct long and narrow visual appearance.

We believe that the proposed feature fusion network can fully fuse the edge information of shallow features and the semantic information of deep features to improve detection performance and further improve the positioning accuracy with our improved oriented bounding box regression strategy. As shown in Table 3, our method has an advantage when using the strategy based on an oriented bounding box, and the mAP is 95.01%.

In order to further verify the effectiveness of the proposed algorithm, we also compare the inference speed of the network on the dataset. When the image input size is 512×800, the inference speed of the model reaches 15.1 FPS. Compared with other two-stage models, the inference speed in this paper is faster, which verifies that the algorithm proposed in this paper also has a certain improvement in detection efficiency.

D. ABLATION STUDIES

1) ABLATION STUDIES OF IBFF
To evaluate the effectiveness of the proposed feature fusion network, we compare the performance of different feature fusion networks under the same backbone network. As shown

| TABLE 1. Comparison with state-of-the-art methods for the DOTA OBB task. |
|----------------|--------|--------|--------|--------|--------|--------|--------|--------|
| Method         | Backbone | BR     | HA     | GTF    | SBF    | BD     | RA     | BC     | SP     | ST     | PL     | SH     | HC     | LV     | SV     | mAP    |
|----------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| PloU [53]      | DLA-34 | 24.1   | 57.1   | 60.2   | 46.5   | 69.7   | 37.1   | 77.2   | 70.9   | 61.9   | 68.0   | 64.8   | 64.0   | 64.4   | 38.3   | 60.5   |
| P-RSDet [52]   | ResNet101 | 47.33  | 60.79  | 72.03  | 59.45  | 73.65  | 57.87  | 80.12  | 90.82  | 65.21  | 81.32  | 89.02  | 72.76  | 52.59  | 73.71  | 70.58  | 69.82  |
| O2-DRN [19]    | Hourglass104 | 47.33  | 58.21  | 61.21  | 60.93  | 82.14  | 60.17  | 82.23  | 90.76  | 66.98  | 81.36  | 89.31  | 78.62  | 61.03  | 74.03  | 71.32  | 71.04  |
| DRN [54]       | Hourglass104 | 47.22  | 69.30  | 64.10  | 57.65  | 82.34  | 61.93  | 86.18  | 90.57  | 69.63  | 84.89  | 89.71  | 85.84  | 58.48  | 74.43  | 76.22  | 73.23  |
| R^2 Det [36]   | ResNet152 | 51.11  | 68.10  | 65.62  | 65.10  | 80.81  | 57.18  | 84.89  | 90.83  | 68.98  | 84.42  | 89.24  | 78.32  | 60.88  | 76.03  | 70.67  | 72.81  |
| RSDet [27]     | ResNet152 | 54.7   | 67.0   | 69.9   | 65.2   | 83.9   | 69.2   | 88.0   | 91.2   | 70.2   | 85.6   | 90.0   | 75.4   | 64.6   | 79.6   | 70.6   | 75.0   |
| DA-Net* [32]   | ResNet101 | 53.28  | 76.16  | 69.55  | 65.03  | 82.41  | 65.70  | 84.76  | 90.68  | 73.37  | 86.33  | 89.7   | 89.04  | 58.86  | 79.54  | 78.24  | 76.11  |
| DCL [55]       | ResNet152 | 53.54  | 73.29  | 72.76  | 67.49  | 83.60  | 66.88  | 86.59  | 90.67  | 70.56  | 86.98  | 89.26  | 87.31  | 69.99  | 82.56  | 79.04  | 77.37  |

| TABLE 2. Experimental results of our proposed approach and the state-of-the-art methods on DOTA. |
|----------------|--------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Method         | AP_x  | AP_y   | AP_z   | AP_w   | AP_v   |
|----------------|--------|--------|--------|--------|--------|
| FR-O [8]       | 46.75  | 58.29  | 56.58  | 55.50  | 27.83  |
| R2CNN [51]     | 59.39  | 66.17  | 60.27  | 63.67  | 34.16  |
| RRPN [50]      | 60.33  | 67.16  | 60.22  | 63.95  | 33.39  |
| ICN [56]       | 66.45  | 75.62  | 67.71  | 71.50  | 46.58  |
| RAdet [57]     | 70.96  | 75.11  | 66.58  | 72.63  | 46.76  |
| RoI-Transformer [14] | 71.54  | 77.04  | 66.67  | 72.94  | 45.34  |
| FAdet [58]     | 74.26  | 80.74  | 70.26  | 76.83  | 50.63  |
| MResNet [29]   | 71.48  | 84.57  | 71.23  | 79.93  | 52.26  |
| SCRDet++ [31]  | 76.39  | 83.72  | 74.42  | 80.53  | 54.53  |
| ResDet [24]    | 76.42  | 85.89  | 76.07  | 82.98  | 53.98  |
| CLODet(ours)   | 79.17  | 85.24  | 72.14  | 84.07  | 61.34  |
in Table 4, we design the following ablation experiments: (1) Instead of using a feature fusion network, the output features of modules conv3, conv4, and conv5 in Fig. 1 are directly used as the inputs of the subsequent RPN and fully connected neural network; (2) only the top-down path feature fusion network is used; (3) the top-down path and bottom-up path feature fusion networks are used at the same time; (4) the proposed IBFF network is adopted, but the proposed DACFF module is not used; (5) IBFF and DACFF are used at the same time.

The experimental results are shown in the Table 4. The feature fusion network proposed in this paper offers good performance and has the highest mAP, which is 78.99%. Moreover, the AP values of most categories are improved. Specifically, the AP values of small objects are significantly improved, while those of large objects are less significantly improved. These results show that the proposed feature fusion network based on information balance can effectively improve the detection precision of small objects.

We compare Parameters, GFLOPs and FPS of the different feature fusion network, and the experimental results are shown in the Table 5. Compared with the common feature fusion networks of FPN+PAN, the proposed feature fusion network has higher detection efficiency. In the input image size of 1024*1024, the FPS of the IBFF method reaches 13.5, which is higher than FPN+PAN method. Compared with FPN+PAN, IBFF had fewer parameters, only 66.1M. It is
verified that the algorithm proposed in this paper has a certain improvement effect on detection efficiency.

2) ABLEATION STUDIES OF DACFF
To verify the effectiveness of the feature fusion module proposed in this paper, as shown in Table 6, ablation experiments are designed as follows: (1) The multi-scale (MS) structure and DACFF module are not used; (2) the multi-scale structure is not adopted, but the DACFF module is adopted; (3) the multi-scale structure is adopted but not the DACFF module; (4) the multi-scale structure and DACFF module are adopted at the same time.

The experimental results are shown in Table 6. Under the network with or without a multi-scale scheme, the feature fusion module we proposed can effectively improve the mAP of the detection network. The mAP of the multi-scale scheme and DACFF module is the highest—4.22% higher than that of the scheme without the multi-scale structure or DACFF module. The experimental results show that the mAP improvement achieved using our approach is not only due to the multi-scale structure. Although the multi-scale structure can improve a certain mAP, adding the feature fusion module proposed in this paper plays a great role. When the multi-scale structure is not adopted, the mAP when adding the three proposed feature fusion modules is found to be 1.82% higher than the mAP when not adding the modules. When using the multi-scale structure, the mAP when adding the three proposed feature fusion modules is 1.97% higher than that when not adding the feature fusion modules. As shown in Table 4, after adding the three proposed feature fusion modules, the detection results of each class are improved to varying degrees. The experimental results show that the proposed feature fusion module can effectively improve the detection precision of the network.

To further demonstrate the effectiveness of the proposed feature fusion module, we next compare the intermediate heatmaps with and without the DACFF module. Fig. 8 shows the comparison of heatmaps at different levels of the network before and after the introduction of the DACFF module. The first row shows the results for when the module is not introduced, while the second row shows the results after introduction of the module. The first to third columns present the visualization results of $F_c^1$, $F_c^2$, and $F_c^3$, respectively. As shown in the figure, with deepening of the level, the heatmap increasingly highlights the detected objects. We believe that this result is due to deepening of the level.
FIGURE 8. The heatmaps [59] of different layers. The first row presents the output heatmap without the DACFF module, and the second row is the output heatmap with the DACFF module.

FIGURE 9. Detection results for images with different rotation angles.

enriching the advanced semantic information, thus enabling the visualization to more effectively express the object information. Such image features can provide a data foundation for the algorithm to effectively distinguish between the
TABLE 7. The ablation analysis of strategies based on an oriented bounding box.

| Method     | mAP@.5 | mAP@.75 |
|------------|--------|---------|
| RBox reg.  | 73.49  | 34.18   |
| Vertex reg.| 68.12  | 44.52   |
| Our        | 80.96  | 62.20   |

FIGURE 10. Comparison of Rotated IoU under different expressions of the oriented bounding box.

object and background. Through visual comparison results of these two rows of depth feature maps, it can be clearly seen that after introduction of the DACFF module, the heatmap focuses more on the object itself and has an obvious effect on suppressing background noise.

3) ABLEATION STUDIES OF STRATEGY BASED ON ROTATING BOUNCING BOX

To verify the effectiveness of the improved oriented bounding box regression strategy in this paper, we compare the proposed method with two baseline methods using oriented bounding box representation (denoted by RBox Reg.) and quadrangle representation (denoted by Vertex Reg.). As shown in Table 7, our improved oriented bounding box strategy offers the highest mAP, with the mAP@.5 of 80.96% and the mAP@.75 of 62.20%. To further demonstrate the effectiveness of our improved strategy, we conducted the following qualitative analysis. As shown in Fig. 9, we rotate the pictures by 45°, 90°, 135°, and 180°, and determine the detection results of the rotated images with different schemes. RBox Reg. has a poor detection effect due to inaccurate angle regression. Because of the confusion in defining the vertex order in training, Vertex Reg. has difficulty in determining the objects oriented in some directions. As shown in Fig. 9, it can be qualitatively seen from the test results for the rotated images that even with a dense distribution and large aspect ratio, the improved method can accurately detect horizontal and rotated objects, and the detection effect is better than that of the other two strategies.

The basic mechanism of most object detection algorithms is to cover the object area through the bounding box and to capture the image within the bounding box and send the image to the classifier to judge the object attribute. Ideally, the bounding box will be tight around the object. However, that is often not the case. If the proportion of the object covered by the bounding box is small, or the background accounts for a large proportion in the bounding box, accurate classification of the classifier will be difficult. Therefore, it is important to improve the IoU of the algorithm for object detection. Fig. 10 shows that our improved approach provides varying degrees of improvement for different categories of objects. The improvement effect is obvious for slender objects, such as bridge, harbor, ground track field, ship, large vehicle, and small vehicle, as represented by solid lines (the other objects are represented by dotted lines in the figure). However, for objects with an aspect ratio close to 1, the improvement effect is small. Experimental results show that our improved oriented bounding box scheme has better detection performance, especially for slender objects.

V. CONCLUSION

This paper proposes oriented object detection based on context information. The proposed IBFF feature fusion network attaches importance to and retains the extraction and utilization of shallow feature information when fusing feature maps of different scales, enhances the edge of the object and improves the localization ability during detection, especially for the improvement of small object detection performance. In addition, the feature fusion network adopts the attention-based context feature fusion module, which can efficiently extract the important information contained in the deep and shallow feature maps according to the characteristics of different layers, reduce the interference of other noises, and further improve the detection performance of the model. We also improve the oriented bounding box regression strategy to further improve the positioning accuracy. Our model is evaluated on two public remote sensing imagery datasets, DOTA and HRSC2016, and the mAP values are 80.96% and 95.01%, respectively. The experimental results show the effectiveness of the proposed algorithm.

REFERENCES

[1] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.
[2] A. Bochkovskiy, W. Chien-Yao, and H. Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” Apr. 2020, arXiv:2004.10904.
[3] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, Dec. 2015, pp. 1440–1448.
[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” in Proc. 14th Eur. Conf. Comput. Vis. (ECCV), vol. 9905, Amsterdam, Netherlands, 2016, pp. 21–37.
[6] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes challenge: A retrospective,” Int. J. Comput. Vis., vol. 111, no. 1, pp. 98–136, Jan. 2015.
[7] 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 Beach, FL, USA, Jun. 2009, pp. 248–255.
[8] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, “DOTA: A large-scale dataset for object detection in aerial images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, Jun. 2018, pp. 3974–3983.

[9] Z. Liu, L. Yuan, L. Weng, and Y. Yang, “A high resolution optical satellite imaging dataset for ship recognition and some new baselines,” in Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods, Porto, Portugal, 2017, pp. 324–331.

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

[11] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, “Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 8, pp. 3652–3664, Aug. 2017.

[12] M.-R. Hsieh, Y.-L. Lin, and W. H. Hu, “Drone-based object counting by spatially regularized proposal network,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, Oct. 2017, pp. 4165–4173.

[13] 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., Columbus, OH, USA, Jun. 2014, pp. 580–587.

[14] 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), Long Beach, CA, USA, Jun. 2019, pp. 2944–2953.

[15] 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.

[16] Z. Xiao, Y. Gong, Y. Long, D. Li, X. Wang, and H. Liu, “Airport detection based on a multiscale fusion feature for optical remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 9, pp. 1449–1473, Sep. 2017.

[17] X. Li and S. Wang, “Object detection using convolutional neural networks in a coarse-to-fine manner,” IEEE Geosci. Remote Sens. Lett., vol. 14, no. 11, pp. 2037–2041, Nov. 2017.

[18] X. Chen, L. Ma, and Q. Du, “Oriented object detection by searching corner points in remote sensing imagery,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

[19] 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.

[20] S. Liu, L. Zhang, H. Lu, and Y. He, “Center-boundary dual attention for oriented object detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–14, 2022.

[21] Z. He, Z. Ren, X. Yang, Y. Yang, and W. Zhang, “MEAD: A mask-guided anchor-free detector for oriented aerial object detection,” Int. J. Spech Technol., vol. 52, no. 4, pp. 4382–4397, Mar. 2022.

[22] Q. Q. Liu and J. B. Li, “Orientation robust object detection in aerial images based on R-NMS,” in Proc. 9th Int. Conf. Inf. Commun. Technol. (ICICT), vol. 154, Nanning, China, 2019, pp. 650–656.

[23] 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.

[24] J. Han, J. Ding, N. Xue, and G.-S. Xia, “ReDet: A rotation-equivariant detector for aerial object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2785–2794.

[25] C. Yingxue, D. Wenrui, L. Hongguang, W. Yufeng, L. Shuo, and Z. Xiao, “Arbitrary-oriented dense object detection in remote sensing imagery,” in Proc. IEEE/9th Int. Conf. Softw. Eng. Service Sci. (ICSESS), Beijing, China, Nov. 2018, pp. 436–440.

[26] 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.

[27] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G. S. Xia, and X. Bai, “Gilding 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.

[28] W. Li, W. Wei, and L. Zhang, “GSDet: Object detection in aerial images based on scale reasoning,” IEEE Trans. Image Process., vol. 30, pp. 4599–4609, 2021.

[29] R. Qin, Q. Liu, G. Gao, D. Huang, and Y. Yang, “MRDet: A multilevel network for accurate rotated object detection in aerial images,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–12, 2022.
[50] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, “Arbitrary-oriented scene text detection via rotation proposals,” IEEE Trans. Multimedia, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[51] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, “R2CNN: Rotational region CNN for orientation robust scene text detection,” 2017, arXiv:1706.09579.

[52] L. Zhou, H. Wei, H. Li, W. Zhao, Y. Zhang, and Y. Zhang, “Objects detection for remote sensing images based on polar coordinates,” 2020, arXiv:2001.02988.

[53] Z. Chen, K. Chen, W. Lin, J. See, H. Yu, Y. Ke, and C. Yang, “PloU Loss: Towards accurate oriented object detection in complex environments,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 195–211.

[54] X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, C. Ma, and C. Xu, “Dynamic refinement network for oriented and densely packed object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 11207–11216.

[55] X. Yang, L. Hou, Y. Zhou, W. Wang, and J. Yan, “Dense label encoding for boundary discontinuity free rotation detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 15814–15824.

[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, 2018, pp. 150–165.

[57] Y. Li, Q. Huang, X. Pei, L. Jiao, and R. Shang, “RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images,” Remote Sens., vol. 12, no. 3, p. 289, Jan. 2020.

[58] C. Lu, C. Xu, Z. Cui, D. Wang, T. Zhang, and J. Yang, “Feature-attentioned object detection in remote sensing imagery,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2019, pp. 3886–3890.

[59] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Seattle, WA, USA, Jun. 2016, pp. 2921–2929.

[60] J. Han, J. Ding, J. Li, and G.-S. Xia, “Align deep features for oriented object detection,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–11, 2022.

QUNYAN JIANG was born in 1998. She received the bachelor’s degree in automotive service engineering from the Tongji Zhejiang College, China, in 2020. She is currently pursuing the master’s degree with the College of Field Engineering, Army Engineering University of PLA. Her research interests include computer vision and model compression.

JUYING DAI was born in 1982. She received the bachelor’s and master’s degrees from the Nanjing University of Aeronautics and Astronautics, China, and the Ph.D. degree from the Army Engineering University of PLA, China. She is currently an Associate Professor at the Army Engineering University of PLA. Her research interests include signal processing and diagnosis.

TING RUI received the M.S. and Ph.D. degrees from the PLA University of Science and Technology, Nanjing, in 1998 and 2001, respectively. He is currently a Professor with the Army Engineering University of PLA. He mainly applies computer vision, machine learning, multimedia, and video surveillance. He has authored and coauthored more than 80 scientific articles.

JINKANG WANG received the bachelor’s degree in mechanical engineering from the Army Engineering University of PLA, China, in 2020, where he is currently pursuing the master’s degree in mechanical engineering. His current research interests include mechanics, machine learning, and computer vision.

Q. Jiang et al.: Context-Based Oriented Object Detector for Small Objects in Remote Sensing Imagery