RoI Fusion Strategy With Self-Attention Mechanism for Object Detection in Remote Sensing Images

Yuxi Zhang, Yongcheng Wang, Ning Zhang, Zheng Li, Zhikang Zhao, Yunxiao Gao, Chi Chen, and Hao Feng

***Abstract*—In remote sensing image (RSI) object detection, the oriented bounding box (OBB) can accurately locate objects with arbitrary orientation and obtain orientation information. The detection based on OBB is still a challenging task. In RSI, the distribution of objects is extremely uneven, which causes aggregation to occur. Some researchers believe that the characteristic of dense distribution is a reason for the difficulty of object detection. However, there are no in-depth experimental studies on this. This paper proposes an OBB-based dense object determination method, which determines the dense objects in datasets by two conditions consisting of interclass distance, intraclass distance, minimum distance between objects, and minimum edge length of objects. The experimental results of dense and non-dense object detection concludes that the characteristics of dense distribution in RSI do not easily cause the objects to be more difficult to detect. To make full use of the object features, we propose a second-stage detection head named RoIF-Net, in which we extract region of interest (RoI) from the input image and fuse it with the RoI extracted from feature maps to add detail features, and construct a feature induction module based on self-attention mechanism to achieve position regression and category classification. This structure can be used in any two-stage network to enhance detection capabilities. Using our method on three credible and challenging datasets, DOTA, DIOR-R, and UCAS-AOD, we obtained 81.80%, 68.49%, and 90.25% mAP, respectively, reaching SOTA based on OBB detection, proving the effectiveness and advancement of our method.**

***Index Terms*—Dense object, object detection, remote sensing images (RSI), region of interest (RoI) fusion, self-attention.**

I. Introduction

In RECENT years, the acquisition cost of remote sensing images (RSIs) has gradually decreased and the resolution has become higher and higher. Its application potential in various fields is gradually gaining attention and needs to be explored. Object detection in RSI can play an important role in many fields, such as military reconnaissance, postdisaster reconstruction, environmental protection, urban and rural planning, economic evaluation, among others. Since RSI are characterized by complex surface environments and large differences in object scales, it is extremely challenging to perform object detection on them.

In the RSI object detection datasets, two forms are generally used to label the objects. One is horizontal bounding box (HBB), which is the smallest external horizontal rectangular box that can contain the object [1], [2], [3], [4], and the other is oriented bounding box (OBB), which is a rectangular box with corresponding angle according to the rotation direction of the object [3], [5], [6], [7]. In generic scenes on natural images, HBBs are more commonly used. However, in RSI, the special bird's-eye view causes the objects in them to have arbitrary rotation directions. In this case, using the HBB introduces unnecessary background information and makes it difficult to obtain accurate object pose, so the OBBs are usually used to annotate and detect the objects in RSI. The visual annotations of HBBs and OBBs are shown in Fig. 1. At present, there have been numerous related research works on the oriented object detection in RSI based on OBB [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. However, the complex background of RSI can cause more difficult feature recognition and bring a considerable challenge to the object detection on it, which is still a key problem that needs to be solved for RSI-oriented object detection.

In addition, researchers have pointed out in some works that objects are more difficult to detect when they are densely arranged together [19], [20], [21], [22], [23], [24], [25], [26], [27]. In natural image scenes, dense can lead to overlap between objects, which results in the absence of object features, which greatly affects the detection results. In contrast, overlap almost rarely occurs due to the overhead view in RSI. No research work has made careful experiments on whether closely spaced objects in RSI are more difficult to detect. First, there is no clear and reasonable determination of dense objects, and second, there is no comparative analysis of the detection results of dense and nondense objects. Therefore, we believe that whether dense objects are more difficult to detect in remote sensing object detection is an inconclusive issue.

In order to determine whether dense objects in RSI are more difficult to detect for the network, we design the determination conditions for dense objects based on understanding of dense objects and implement the classification of objects in the datasets into dense and nondense objects. Then we determine whether

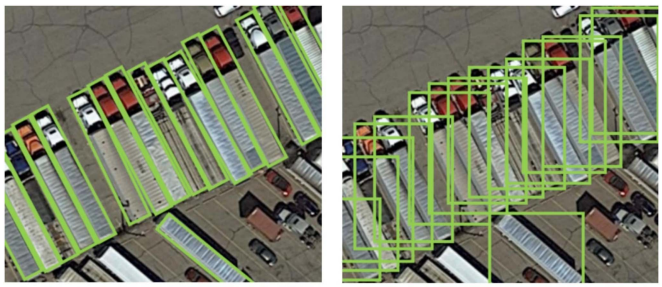
Manuscript received 21 March 2023; revised 10 June 2023; accepted 18 June 2023. Date of publication 26 June 2023; date of current version 11 July 2023. This work was supported by the Chinese Academy of Sciences. *(Corresponding author: Yongcheng Wang.)*

Yuxi Zhang, Zheng Li, Zhikang Zhao, Yunxiao Gao, Chi Chen, and Hao Feng are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhangyuxi18@mails.ucas.ac.cn; lizheng20@mails.ucas.ac.cn; zhaozhikang20@mails.ucas.ac.cn; gaoyunxiao19@mails.ucas.ac.cn; chenchi21@mails.ucas.ac.cn; fenghao21@mails.ucas.ac.cn).

Yongcheng Wang is with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China (e-mail: wangyc@ciomp.ac.cn).

Ning Zhang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: cdd_ningzhang@tsinghua.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3289585



(a) (b)

Fig. 1. (a) Objects annotated by OBB. (b) Objects annotated by HBB.

dense objects in RSI are more difficult to detect based on the detection results for both dense and nondense objects. We design a dense object determination method with two determination conditions, which is based on OBB labeling. When any object in the image is judged, if this object and any other object in the image satisfy both conditions, both two objects are judged as dense objects. One of the determination conditions is that the ratio of the interclass distance to the intraclass distance of the two objects is less than a specific threshold, and the other determination condition is that the minimum distance between the two objects is less than an expression related to the minimum edge length of that object. After determining whether objects in the datasets are dense objects using the determination method, we detect them. By analyzing the detection results, we found that dense objects have a higher recall compared to nondense objects. It is concluded that dense is not a fundamental factor that makes the object difficult to detect in RSI object detection.

When using OBB for object detection, a two-stage network is often used to obtain better detection results, in which the extracted region of interest (RoI) needs to be fed into the second stage detection head. In order to extract RoI containing rich detail features and make full use of them, we propose RoIF-Net. The classical RoI extractions, such as RoI pooling [28] and RoI align [29], are performed on feature maps output from the backbone network, which is rich in high-level semantic features but lacks such detail information as low-level spatial features after being extracted by layer after layer of convolutional networks. Other improved RoI extraction methods [49], [50] expand the RoI range without taking the issue into account. We propose RoI fusion module to enrich the low-level spatial features by adding RoI extraction on the original images. Noting that the original image has higher resolution compared to the feature map, we perform RoI extraction on the original image by resampling it to a 28×28 patch, then generating a 7×7 patch by channel rearrangement, then expand it to the same dimension as the feature map, and finally add it to the RoI extracted from the feature map to obtain the final RoI. After obtaining the RoI with more detail information, it is detected in the second stage. The simple fully connection layer structure in most methods [10], [34], [42], [51] cannot fully utilize the more complex feature information. In order to fully exploit and utilize the feature information of the fused RoI, we construct a feature induction module to discriminate and generalize the features of RoI, using the self-attention mechanism of Transformer [30] and combining convolutional and fully connection layers to enhance the discriminative ability of the network for complex features. The aforementioned is our design of RoIF-Net, a new second-stage detection structure, which is used to fully utilize the advantages of the two-stage detection network and improve the classification and positioning accuracy of the second stage detection head.

The main contributions of this article are as follows.

1) The dense object determination method is proposed, which defines dense objects by interclass distance, intraclass distance, minimum distance, and minimum edge length. By this method, the objects are classified into dense and nondense objects. Further, our experimental results show that these defined dense objects are not relatively more difficult to be detected in RSI.
2) The second-stage detection head RoIF-Net consisting of RoI fusion module and feature induction module is proposed, which increases the detail information of RoI by extracting RoI from the original image and improves the feature discrimination and generalization ability through the self-attention mechanism, and this structure can be applied to any two-stage detection network to improve the detection accuracy.
3) The proposed method achieves SOTA for rotated object detection on three strongly credible and highly challenging RSI object detection datasets: DOTA, DIOR-R, and UCAS-AOD with 81.80%, 68.49%, and 90.25% mAP, respectively.

The rest of this article is structured as follows. Section II covers the recent work on RSI object detection. Section III presents a detailed introduction of the proposed dense object definition method and RoIF-Net. Section IV is about the demonstration and analysis of the experimental results. Section V presents the conclusion and outlook of this article.

II. Related Works

A. Dense Object Detection

Some objects in the image are densely packed together and many works mention that these objects are difficult to detect [19], [20], [21], [22], [23], [24], [25], [26], [27]. In natural image, commodity detection on supermarket shelves [21], [22], face and pedestrian detection in crowded scenes [23], [24], etc., may be performed on a large number of dense objects. The distribution of objects in RSI is extremely uneven, and many scenes such as airports, parking lots, and ship ports have a high number of dense objects clustered together, as shown in Fig. 2. Yingxue et al. [20] select only aircraft, vehicles, and ships in the DOTA [3] dataset for detection, which have a higher probability of dense alignment. Shu et al. [19] obtain object center point before generating accurate object bounding box in order to detect dense buildings. Ming et al. [27] proposed a coordinate attention module to deal with the problem of severe performance degradation caused by minor position deviations in dense small object detection. Li et al. [26] enhance the shallow feature information of small and dense objects by jump connecting

(a) (b)

Fig. 2. (a) Densely distributed airplanes in RSI. (b) Densely distributed cars and tennis courts in RSI.

the manually extracted shallow features to the deep network after processing. Deng and Yang [25] proposed a multistep sampling strategy to improve the probability of dense objects being sampled during the training process. However, there is no clear determination of dense objects in these works, and no separate analysis of detection results for dense and nondense objects, but only qualitative statements that dense objects are relatively difficult to detect. Whether dense is a factor causing the difficulty of object detection remains to be confirmed. The study of dense objects contributes to the further development of RSI object detection techniques. In this article, a method to determine dense objects is proposed. The objects in the dataset are divided into dense objects and nondense objects. According to their detection results in the experiments, whether they are difficult to detect is analyzed.

B. RSI Object Detection Based on Deep Learning

The annotation method commonly used in object detection is the minimum external HBB. In natural image object detection datasets, horizontal boxes are almost always used to annotate objects. The remote sensing object detection datasets with high credibility, such as NWPU VHR-10 [1], RSOD [2], DOTA [3], and DIOR [4], have horizontal box annotation. The network detects the object by generating a HBB, the position of which is used to locate the object. The position of the rectangular box can be represented by four parameters, which are typically the coordinates of the center point, the length and the width of the rectangular box. It is also necessary to classify the objects in the detection box. Therefore, the final result of the detection network generally consists of a regression part and a classification part. In natural image object detection, classical networks, such as SSD [31], RCNN series [28], [29], [32], [33], [34], YOLO series [35], [36], [37], [38], and RetinaNet [39], are all HBB-based object detection networks. These networks can be directly applied to RSI object detection, but the effect needs to be improved.

RSI are captured by aerial imaging devices from a top-down view, where the object is on the earth's surface, and has an arbitrary rotation angle in this view. If an HBB is used to locate an object, the box will contain a large amount of background when the object direction deviates from the horizontal or vertical angle and will contain other objects when the objects are densely distributed, and the above-mentioned phenomenon is more obvious when the object aspect ratio is large. The OBB with angle can solve the above problem, and the object can be included to the maximum extent when the rectangular box direction is the same as the object direction. In RSI object detection, more and more datasets use OBB to label objects, such as DOTA [3], UCAS-AOD [5], HRSC2016 [6], and DIOR-R [7]. Determining an OBB is simply a matter of adding the angle to the HBB, so adding the angle information prediction branch to a network structure based on HBB detection can achieve OBB object detection. In order to achieve more accurate orientation detection, some detectors improve the network structure specifically for angle prediction [40], [41], [42], [43], [44], [45], [46], [47], [48]. DCL [44] and CSL [45] convert the angle detection from a regression problem to a classification problem. Oriented RCNN [42] determines OBB by the minimum external HBB and the distance from the vertex of the OBB to the midpoint of the edge of that HBB. CenterMap [41] determines the OBB by generating a foreground region heat map that has high heat value in the center region of the object and low heat value in the edge region. GWD [46] and KLD [47] learn OBB with angular information by regression loss based on Gaussian model.

C. Two-Stage Detection Network

The two-stage network is an important object detection strategy based on deep learning. In the two-stage detection network, the first stage performs preliminary localization and classification of the object, then the RoI at the corresponding location is extracted on the feature map based on the initially obtained detection box, and finally, the second stage performs more accurate position regression and classification of the RoI. Two-stage detection network is proposed for the first time in RCNN [32], in which selective search is used to extract RoI. In fast RCNN [28], RoI pooling is proposed to provide uniform size input for the second stage. Then in faster RCNN [33], region proposal network is used instead of selective search for the first-stage detection, and the most commonly used two-stage detection network is built. In RSI object detection, two-stage detection networks are also widely adopted and continuously improved for their high accuracy [10], [42], [49], [50], [51]. Li et al. [49] and Gong et al. [50] obtain object context information by extracting and fusing a wider range of RoI. RoI Transformer [51] adds angle prediction for RoI based on HBB in the second stage, and then extracts RoI based on OBB for final detection. Oriented RCNN [42] directly extracts RoI based on the OBB in the second stage according to the angle prediction in the first stage.

Based on the two-stage detection network, a number of other variant forms have been further developed, which can be categorized as two-stage detection networks in a broad sense. Cai and Vasconcelos [34] proposed Cascade RCNN to discriminate positive samples by incremental intersection-over-union (IoU) thresholds in multiple detection stages, and multistage detection networks were thus generated and developed. The key for the two-stage detector to be able to achieve the second detection is to align the features according to the detection box obtained from

the first-stage detection before performing the second detection, which is the function achieved by RoI extraction. In deformable convolution [52], [53], the shape of the convolution kernel is not fixed and can vary. Inspired by this, the refinement detector was proposed. In the refinement detector, the shape of the deformable convolution kernel is set according to the shape of the detection box obtained in the first stage, and feature alignment is achieved by convolution using such a kernel. S^2A-NET [54] and R^3DET [55] use this method, avoiding the RoI extraction step, and the detector is implemented by a fully convolutional network.

In the second-stage detection network proposed in this article, RoI extraction on the original image is added in RoI fusion module to enrich the feature information, especially the low-level features with detail information. A feature induction module is designed to discriminate and generalize the features in RoI using a self-attention mechanism, which complements the detail features added in RoI and enhances the network's discrimination of confusable features. The increased capability of the second-stage detection network allows for a higher level of detection accuracy across the detector.

III. Proposed Method

In this section, we describe the proposed dense object determination method and the second-stage detection head in detail. First, dense object determination method is detailed in Section III-A. Next, the overall network structure of two-stage detector is introduced in Section III-B. Finally, the designed RoIF-Net is introduced in Section III-C.

A. Determination Method for Dense Object

In previous research work, there is no specific definition of dense objects to determine dense objects, let alone to analyze the detection results of dense objects. In order to be able to analyze the detection effect of the network on dense objects, we designed the method for determining dense objects. According to this method, the objects in the dataset can be divided into dense and nondense objects so that the detection effect can be analyzed in the network test using evaluation metrics for both dense and nondense objects. Because the object detection dataset uses rectangular boxes to label the objects, we need to use the object location information from the rectangular box annotations to identify dense objects in the dataset. If the HBB annotations are used to determine dense objects, it will result in a situation where the rectangular boxes are dense or even overlapping while the objects in the boxes are still far away from each other. This is because the HBB does not contain any object pose information, and the box may contain a large amount of background in addition to the object, and the area of the rectangular box cannot be approximated as the object area.

Therefore, we use OBB annotation to determine the dense object. The OBB has the angular information of the object, which can closely contain the object compared to the HBB, which contains less background, and the area of the OBB can be approximated as the area of the object. In the determination method we designed, the OBB represents the object and is used for judgment. When considering how to perform dense object determination, if only the minimum distance between objects is used to determine, it cannot represent the complex position relationship between objects and the judgment result is not satisfactory. If the number of objects in a region is counted to determine the dense area, it is not possible to quantify the relationship between an object and its surrounding objects. In order to quantify whether an object is dense or not and to take into account the position relationship of all pixels between objects as much as possible, we designed two conditions for the determination. If an object and any other object in the same image satisfy these two conditions, the object is considered as dense.

The first determination condition is based on the ratio of the interobject distance to the intraobject distance. In pattern discriminant analysis, for a pattern class $\{a_i\}_{i\,=\,1,2,\ldots,Ka}$, the intraclass distance is

$$\overline{D^2(\{a_i\},\{a_j\})} = \frac{1}{K_a}\sum_{j=1}^{K_a}\left[\frac{1}{K_a-1}\sum_{i=1,i\neq j}^{K_a}(a_j-a_i)^2\right] \tag{1}$$

the smaller the intraclass distance is, the higher the degree of aggregation of this pattern class. If there is another pattern class $\{b_i\}_{i\,=\,1,2,\ldots,Kb}$, its interclass distance with the previous pattern class is

$$\overline{D^2(\{a_i\},\{b_j\})} = \frac{1}{K_b}\sum_{j=1}^{K_b}\left[\frac{1}{K_a}\sum_{i=1}^{K_a}(b_j-a_i)^2\right] \tag{2}$$

which can be used as a measure of the separability of these two pattern classes, and the larger interclass distance indicates that their separability is better. We consider each object on the image as a pattern class, and each pixel within the object as a sample in this pattern class. We calculate the intraclass distance of each object and the interclass distance of every two objects, and use the ratio of the interclass distance to the intraclass distance to determine the denseness, and the smaller the interclass distance relative to the intraclass distance, the more intensive the two objects are. Since we can only use the rectangular box annotation in the dataset to identify the object, we recognize the OBB as the object to calculate the intraclass and interclass distance. The distance of sample points within one object class and the distance of sample points belonging to different object classes are shown in Fig. 3. In the determination process, for any object on the image, its intraclass distance is calculated as D_w. If there is another object on the image, their interclass distance is D_b, which is less than the threshold value we set, then both two objects meet the first determination condition for dense objects. The first determination condition of our design can be expressed by

$$D_b/D_w < T. \tag{3}$$

In this formula, the larger the threshold T is, the larger the ratio of interclass distance to intraclass distance that satisfies the condition is, and the less intensive it is. Conversely, the smaller the threshold T, the higher the denseness. The determination of the dense object is relatively subjective, and the threshold T can be set autonomously according to the different demands on the denseness.

Fig. 3. References in dense object determination method. Green box is the object range. Green dots in the box represent the sample points in the object. $D(a_i, a_j)$ denotes the distance between two sample points in a single object. $D(a_i, b_j)$ denotes the distance between two sample points belonging to different objects. l denotes the minimum side length of the object. And d denotes the minimum distance between two objects.

The second criterion is related to the minimum distance between objects and the minimum side length of objects. The minimum distance is the value of the closest distance between two objects. We think that the minimum distance needs to be limited when determining dense objects. Considering that the larger the object size is, the larger its feature scale is, the minimum distance restriction in the determination condition should be relaxed for that object. We use the minimum side length of the object as the factor limiting the minimum distance. In addition, we consider that the minimum distance limit should not increase in equal proportion to the object size because the number of pixels increases with the object size, and the more pixels between objects, the less dense they are. So, we square the minimum side length of the object to reduce the rate of increase of the minimum distance limit. We still consider the OBB as the object to calculate the minimum distance and the minimum edge length. The minimum side length of an object and the minimum distance between two objects are shown in Fig. 3. In the determination process, for an object on the image whose minimum edge length is l, if there exists another object on the image and the minimum distance between them is d, and these two objects satisfy each other with

$$d < a\sqrt{l} \tag{4}$$

then both two objects satisfy the second determination condition. In this formula, similar to T in (3), a is used as a moderator to adjust the severity of this condition to meet different subjective needs. The smaller a is, the stricter the minimum distance restriction and the more dense the object is.

An object in an image is considered as a dense object when it satisfies both of the determination conditions we designed. The first condition, (3), determines whether an object is dense or not by comparing the dispersion of the object with another object and its own dispersion, and this form of determination considers the denseness from the totality of the two object areas.

TABLE I
NUMBER OF DENSE AND NONDENSE OBJECTS IN THE DOTA, DIOR-R, AND UCAS-AOD DATASETS UNDER THE DETERMINATION METHOD WITH DIFFERENT T AND a VALUES

Dataset	Value of T	Value of a	Dense	Nondense
DOTA	5.25	5	20 863	7989
	7.75	5	22 740	6112
	10.25	5	23 217	5635
	7.75	3	18 968	9884
		5	22 740	6112
		7	23 659	5193
DIOR-R	5.25	5	74 729	49 662
	7.75	5	85 338	39 053
	10.25	5	87 962	36 429
	7.75	3	80 254	44 137
		5	85 338	39 053
		7	88 243	36 148
UCAS-AOD	5.25	5	1497	3239
	7.75	5	2248	2488
	10.25	5	2471	2265
	7.75	3	1406	3330
		5	2248	2488
		7	2434	2302

The second determination condition, (4), ensures that the closest points between two objects can be within a threshold value. The combination of these two conditions considers the denseness both from the whole object area and from a single pixel point, which can determine dense objects in a more reasonable way. The severity of the determination conditions can be adjusted according to different subjective requirements. We divide the objects in the DOTA [3], DIOR-R [7], and UCAS-AOD [5] datasets into dense and nondense objects according to the designed determination method with different values of T and different values of a. The results are given in Table I. As can be seen from the table, the number of dense objects in the datasets increases with increasing threshold T and moderator a. This is due to the fact that the larger the threshold T or the moderator a is, the more lenient the determination conditions are, as described earlier. Furthermore, in Tables II and III, we give the division between the DOTA and the DIOR datasets for each category of objects when the threshold T is 7.75 and the moderator a is 5. The results of dividing dense and nondense objects in some images with three different thresholds T and three different moderators a are shown in Figs. 4 and 5. The comparison of the results visually demonstrates that the dense object determination condition is more relaxed when T or a is larger. And it can be seen that the determination method we designed can ideally distinguish dense objects from nondense objects. The effectiveness of this method is proved.

B. Two-Stage Detection Network as Baseline

In this article, we propose the second-stage detection structure RoIF-Net, which is part of a two-stage detection network. The two-stage detection network detects the object in the image

TABLE II
NUMBER OF DENSE AND NONDENSE OBJECTS IN EACH CATEGORY IN THE DOTA DATASET UNDER THE DETERMINATION METHOD WITH T VALUE OF 7.75 AND a VALUE OF 5

Category	Dense	Nondense
plane	1940	591
baseball diamond	45	169
bridge	78	385
ground track field	98	46
small vehicle	4252	1186
large vehicle	3682	705
ship	8504	456
tennis court	533	227
basketball court	117	15
storage tank	2473	415
soccer-ball field	112	41
roundabout	5	174
harbor	772	1318
swimming pool	95	345
helicopter	34	39
sum total	22 740	6112

TABLE III
NUMBER OF DENSE AND NONDENSE OBJECTS IN EACH CATEGORY IN THE DIOR-R DATASET UNDER THE DETERMINATION METHOD WITH T VALUE OF 7.75 AND a VALUE OF 5

Category	Dense	Nondense
airplane	5539	2673
airport	70	596
baseball field	803	2631
basketball court	1894	252
bridge	483	2101
chimney	484	547
dam	30	508
expressway toll station	53	635
expressway service area	822	263
golf field	195	380
ground track field	450	1435
harbor	2680	422
overpass	553	1225
ship	31 879	3304
stadium	215	457
storage tank	20 474	2887
tennis court	6248	1095
train station	44	465
vehicle	12 417	14184
windmill	5	2993
sum total	85 338	39 053

twice, and the first detection generates relatively coarse localization and classification results, followed by further localization correction and accurate classification in the second-stage detection network based on the first generated coarse results. The overall structure of the classical two-stage detection network is shown in Fig. 6. The image to be detected is fed into the network as an input, and first, the backbone network is used for feature extraction, then the extracted feature map is fused in the neck network to generate a multiscale feature map [56], next the first-stage detection head is used to detect on the multiscale feature map to obtain preliminary results. The detection results of the first stage generally include the regression and classification results of the object boxes. During the training process, the generated detection boxes need to be matched with the annotated real objects, and the position regression losses are calculated based on the mutually matched detection boxes and real boxes. These detection boxes that can match to the real boxes are classified as foreground and those that are not matched are classified as background, and the classification loss is obtained according to the foreground and background, and the network parameters are updated by these losses so that the network gradually learns how to detect the objects. RoI extraction is performed based on the object box obtained from the first-stage detection, and the extracted RoI is fed into the second-stage detection head for adjustment, which includes position regression adjustment and accurate category classification. During the training process, the regression loss is calculated again in the same way, and the classification loss is calculated according to the object class, from which the network parameters are then updated. The detection result after the second-stage detector head adjustment is the final result of the whole two-stage detector.

In a two-stage detection network, the loss function generally consists of two major parts, one for the losses generated by the first-stage detection head and the other for the losses generated by the second-stage detection head. As mentioned above, the loss generated at each stage contains regression loss and classification loss, and the network learns the object location and size through regression loss and the object class through classification loss. The overall loss function is given by

$$\begin{aligned}\text{Loss} = &\frac{\lambda_1}{N_F}\left(\sum_i L_c\left(c_i^F, l_i^*\right) + \sum_i [l_i^* \geq 1] L_r\left(x_i^F, g_i^*\right)\right) \\ &+ \frac{\lambda_2}{N_S}\left(\sum_i L_c\left(c_i^S, l_i^*\right) + \sum_i [l_i^* \geq 1] L_r\left(x_i^S, g_i^*\right)\right).\end{aligned} \tag{5}$$

The equation consists of two parts, which are the first-stage loss and the second-stage loss. In the losses of these two stages, λ_1 and λ_2 are the loss balance coefficients, which are generally 1, N_F and N_S are the number of positive samples in the two stages, i represents each sample, L_c and L_r are the classification and the regression loss functions, c_i^F and c_i^S are the classification predictions in the two stages, l_i^* is the classification label, $[l_i^* > 1]$ is the Iverson bracket indication equation, which means the value is 1 when i is a positive sample, x_i^F and x_i^S are the location predictions in the two stages, and g_i^* is the location label.

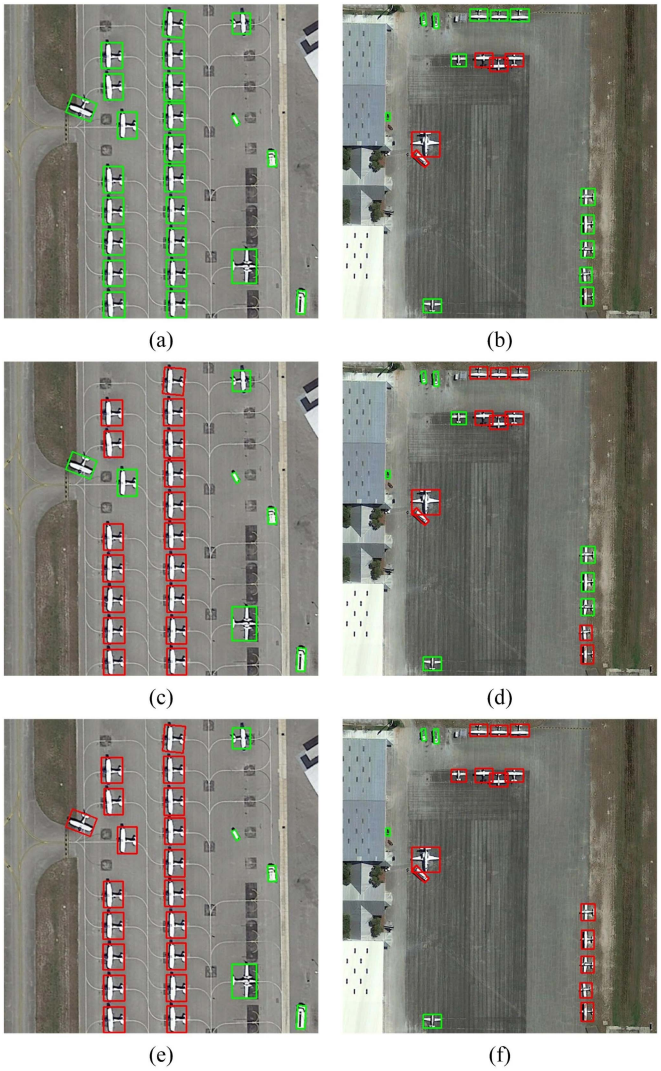


Fig. 4. Dense objects and nondense objects judged by the determination method with different values of T. Objects boxed in red are dense objects and those boxed in green are nondense objects. (a) and (b) Results of the determination of dense and nondense objects at a T value of 5.25. (c) and (d) Determination results at a T value of 7.75. (e) and (f) Determination results at a T value of 10.25.

Fig. 5. Dense objects and nondense objects judged by the determination method with different values of moderator a. Objects boxed in red are dense objects and those boxed in green are nondense objects. (a) and (b) Results of the determination of dense and nondense objects at a moderator a of 3. (c) and (d) Determination results at a moderator a of 5. (e) and (f) Determination results at a moderator a of 7.

C. Structure of RoIF-Net

The second stage outputs the final detection results, which plays an important role in the two-stage detection network and is a key factor for the two-stage detection network to obtain high accuracy. Noting the importance of the second stage, in order to give full play to its role in the overall detection network, we designed the second-stage network structure RoIF-Net, as shown in Fig. 7. RoIF-Net is divided into two parts. One is the RoI fusion module, which simultaneously performs RoI extraction on the feature map and the original image and fuses them together. The other is the feature induction module based on the self-attention mechanism, which is able to discriminate and generalize the features and generates the final adjustment results.

In the RoI fusion module, we extract RoI not only on the feature map but also on the original image, which is to be able to obtain more detail features. The feature map obtained after backbone extraction has sufficient high-level features; however, the most original detail information will disappear after the complex network, and the supplement of detail information is beneficial to the second-stage detection network for more accurate object localization and classification. In this structure, according to the detection boxes generated by the first stage, RoI extraction is performed on the feature map obtained from backbone and the original image as the input to the whole network, respectively. When extracting RoI on the feature map, as in the classical two-stage detection network Faster RCNN [33], we resample its corresponding range into a 7×7 patch, regardless of the detection box size. When extracting RoI on the original image, we resample the range corresponding to the detection box into a 28×28 patch in order to avoid losing too much information since the original image size is much

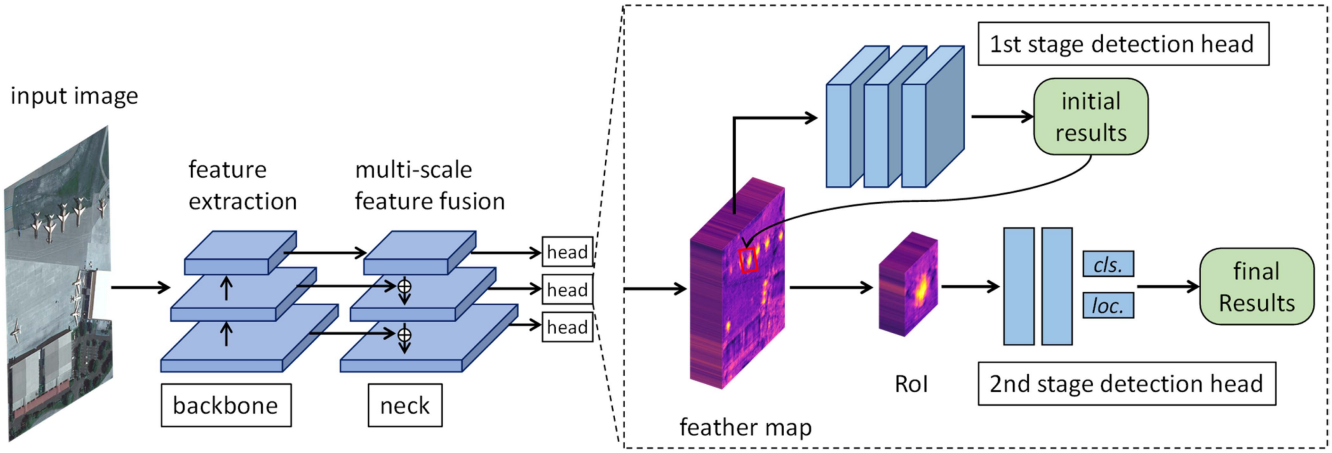


Fig. 6. Overall framework of a two-stage detection network. Classical two-stage detection network consists of backbone, neck, the first-stage detection head, and the second-stage detection head. Image to be detected is input into the backbone, and the second-stage detection head generates the final results.

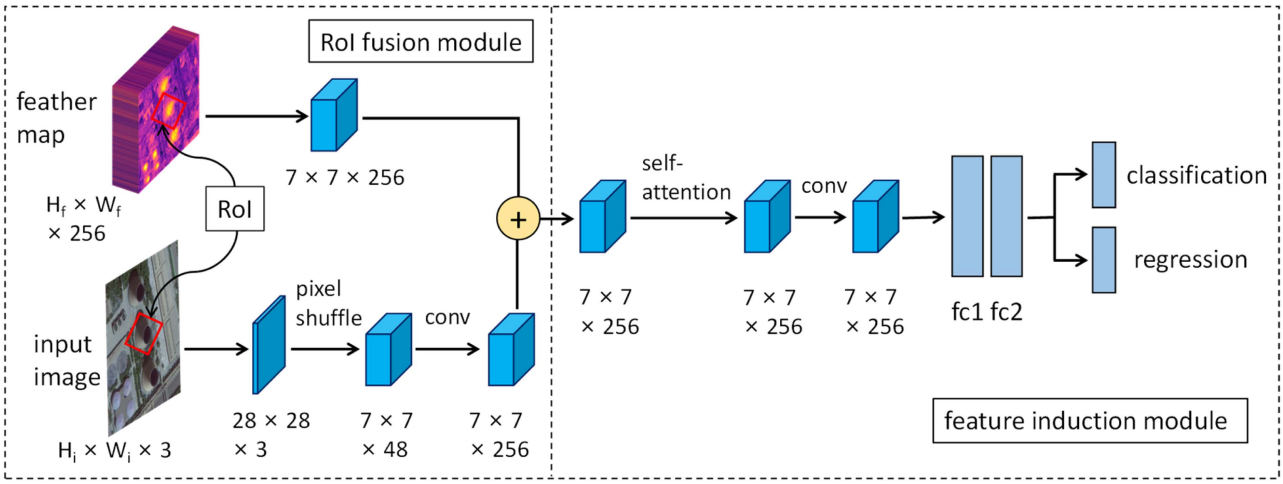


Fig. 7. Overall structure of RoIF-Net. Structure is a second-stage detection head consisting of an RoI fusion module and a feature induction module that outputs detection results.

larger relative to the feature map (at least four times in general). Next, in order to fuse the two RoIs, we inverse the subpixel convolution strategy [57], reducing the size of the RoI on the original image by expanding the number of channels without losing information, then expanding the dimensionality to that of the RoI on the feature map with a 1×1 convolution kernel, and finally fusing the extracted two RoIs by summing operations.

In the feature induction module, we constructed it using the transformer structure based on the self-attentive mechanism [30], as shown in Fig. 8. This part makes full use of RoI with fused detail features to enhance the discrimination of confusable features and perform feature generalization, which improves the classification and localization accuracy of the second-stage detection network. In this structure, first we expand the RoI obtained in the previous structure into 49 256-dimensional feature vectors and use them as 49 tokens of the transformer. Three feature matrices Q, K, V are generated from the input token, in which each 256-dimensional feature vector is decomposed into four 64-dimensional feature vectors. Q and K^T are multiplied to get the self-attention weight matrix. Since there is a positional relationship between feature points in an image, it is important to add positional information to the feature points used as token inputs. We use the positional encoding matrix in Swin Transformer [58], which contains the relative positional information between every two feature points and helps the network to judge the spatial location of features. The self-attention matrix is obtained by adding the weight matrix with the position matrix and then multiplying it with the V matrix after performing softmax. The four 64-dimensional vectors are synthesized into a 256-dimensional feature vector, and then the expanded spatial dimension is restored to obtain the RoI after the self-attentive mechanism. Then using the residual mechanism [59], it is summed with the original RoI. Finally, the final regression and classification results are obtained after passing through two convolution layers and two fully connected layers.

RoIF-Net adds detail features by fusing the RoI extracted from the original image to provide more information for the final regression and classification, uses a self-attention mechanism to enhance the discrimination of confusing features, and finally performs feature generalization to achieve high-quality

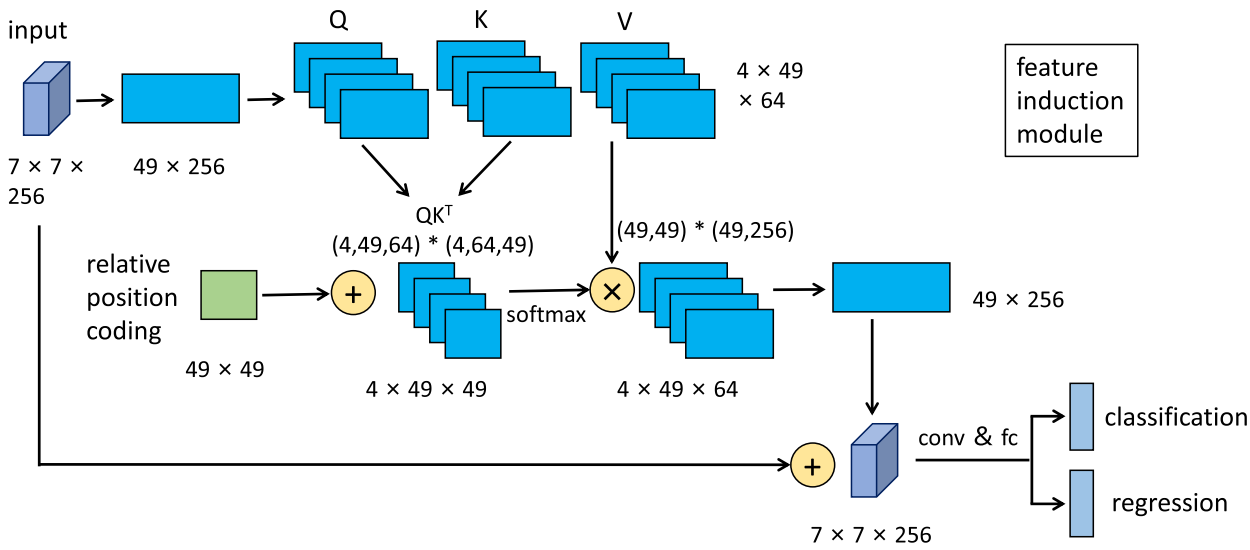


Fig. 8. Overall structure of the feature induction module.

adjustment of the detection box. Since the second stage is relatively independent in the detector, RoIF-Net can theoretically be applied to any two-stage detection network. Simply replacing the second-stage detection network with RoIF-Net can make the original two-stage detector a step up in detection effectiveness.

IV. Experiments and Analysis

A. Datasets

1) DOTA-v1.0: DOTA-v1.0 [3] is a large aerial remote sensing dataset, which contains 2806 aerial images collected from Google Earth, satellite JL-1, etc. It has 188 282 ground objects annotated on it for object detection tasks, some of which are arranged very densely. These objects cover 15 common categories, namely 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). There are two types of annotation in this dataset: HBB and OBB. In this article, we use OBB annotation for dense object determination and all experiments. The whole dataset has been divided into three parts by the publisher: the training set, the validation set, and the test set, and their ratio is 3:1:2.

The images in the DOTA dataset vary in size, with a large gap between the minimum size of 800 × 800 pixels and the maximum size of 4000 × 4000 pixels. To avoid the loss of image information caused by resizing, we cropped the original images into a series of 1024 × 1024 patches as input. The experiments performed on this dataset in this article all use the multiscale data augmentation method, using three scale factors (0.5, 1.0, 1.5) to resize the original image and the crop step is 512. If the instances are segmented at the time of cropping, we decide whether to use them or not according to the method in [3]. In the test, we map the detection results to the original size image before evaluation.

2) DIOR-R: DIOR-R [7] is a large-scale publicly available remote sensing dataset for object detection, which is an extended version of the DIOR [4] dataset that uses OBBs to annotate objects in images. It contains 23 463 images covering a wide range of scenes on which 192 518 object instances belonging to 20 common object classes are annotated, namely airplane (APL), airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DAM), expressway toll station (ETS), expressway service area (ESA), golf field (GF), ground track field (GTF), harbor (HA), overpass (OP), ship (SH), stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM). The image size in the dataset is uniformly 800 × 800 pixels, and the spatial resolution ranges from 0.5 to 30 m, so the objects on the images have large-scale differences. The training and validation set of DIOR-R contains 11 725 images and 68 073 instances, and the test set includes 11 738 images and 124 445 instances.

3) UCAS-AOD: UCAS-AOD [5] is a publicly available high-definition aerial photography dataset for object detection, in which images are captured in selected regions of the world using Google Earth. The dataset contains 1510 images of approximately 1300 × 700 in size, which are labeled with two types of objects: car and plane. The labels are in the form of HBB and OBB, and in this article, we use OBB annotations for our experiments. The images are randomly divided into a training set, a validation set, and a test set in the ratio of 5:2:3.

B. Implementation Details

The backbone network used in our following experiments is first pretrained on ImageNet [60], and the network is initialized using the parameters obtained from the pretraining. In the training phase, two Nvidia RTX3090 GPUs are used to perform the experiments, and the batch size of a single GPU is set to 2, for a total of 4. When ResNet50 [59] is used as the backbone, the SGD optimizer is used to perform gradient updates of the model parameters, where the initial learning rate is set to 0.005, the

learning rate is reduced to 1/10 of the original at each decay, and the momentum and weight decays are set to 0.9 and 0.0001, respectively. When ConvNeXT [61] or Swin [58] is used as the backbone network, the AdamW [62], [63] optimizer is used to perform gradient updates, where the initial learning rate is set to 0.0001, the learning rate is reduced to 1/10 of the original at each decay, and the weight decay is set to 0.05. When DOTA dataset is used for experiments, the number of training times is set to 12 epochs, and the learning rate is decayed after the 8th and 11th epochs, respectively. In the training process, we use data enhancement strategies, such as random flipping, random rotation, and multiscale scaling, to increase the complexity of the dataset. When using DIOR-R, the number of training times is still 12 epochs, and the learning rate decay is still performed after the 8th and 11th epochs, and data enhancement strategies such as random flipping and random rotation are used. When using UCAS-AOD, the number of training times is 36 epochs, and the learning rate decay is performed after the 24th and 33th epochs. In the testing phase, we use a single Nvidia RTX3090 GPU for inference. We keep the bounding boxes with confidence scores greater than 0.05 and set the IOU threshold of NMS to 0.1. At the same time, considering that an image contains a limited number of objects, we set the maximum number of objects in each image to 2000.

C. Evaluation Metrics

When evaluating the effectiveness of detection networks, a uniform set of criteria is needed. Average precision (AP) is the most authoritative evaluation metric in object detection, and the calculation of this value is related to two basic and credible evaluation metrics: precision and recall. To judge that the network correctly detects the object, two conditions need to be satisfied. The first condition is that the IoU between the detection box and the ground truth box is greater than 0.5, and the second is that the network classifies the object in the detection box correctly. On this basis, the precision and recall are determined by the formulas

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

where TP is the number of objects determined by the network to be positive samples and detected correctly, FP is the number of objects determined by the network to be positive samples but detected incorrectly, and FN is the number of ground truth objects determined by the network to be negative samples. When evaluating the detection results, as the confidence score decreases, more and more objects are detected, and the recall is higher. Generally speaking, the precision decreases with the increase of recall. AP is the precision integral of the recall from 0 to 1, which can be expressed as

$$\text{AP} = \int_0^1 (\text{precision}) d(\text{recall}). \tag{8}$$

TABLE IV
RECALL OF DENSE AND NONDENSE OBJECTS IN THE DIOR-R DATASET UNDER THE DETERMINATION METHOD WITH DIFFERENT T VALUES

Value of T	Dense or not	Object number	Recall
5.25	✔	74 729	77.95
	✖	49 662	69.75
7.75	✔	85 338	76.74
	✖	39 053	70.16
10.25	✔	87 962	76.07
	✖	36 429	71.31

The values of 11 recall rates (0, 0.1, …, 1) are generally used for calculation. AP can make a valid evaluation of the detection results of one class of objects. And when evaluating the detection results of multiple classes of objects, mAP, which is the average of AP of multiple classes of objects, can be used.

In this article, we are going to evaluate the detection results of dense and nondense objects separately. The determination of density involves objects other than the one to be determined. When determining dense objects on the detection results, other objects are not necessarily ground truth objects, so the determination results are not credible. When calculating the precision of dense objects and nondense objects separately, it is necessary to know the number of dense and nondense objects in the detection results. Since the dense objects determined on the detection results are not credible, the precision in this case does not provide a valid assessment of the detection results. Instead, the number of dense and nondense objects in the ground truth needs to be known when calculating the recall separately. The determination of dense or nondense objects on the ground truth is credible, so the recall can still effectively evaluate the detection results. As aforementioned, we evaluate the detection results of dense and nondense objects using only the recall without using the precision and the AP that includes the precision.

D. Analysis of Dense Object Detection Results

We classify the objects in the DIOR-R [7] dataset into dense and nondense objects according to the determination method proposed in Section III-A with three thresholds T representing different densities: 5.25, 7.75, and 10.25. RoI Transformers [51] are used to detect and calculate their recall, respectively. The results of the experiment are given in Table IV. The recall rate of dense objects is 8.20% higher than that of nondense objects when T is 5.25, 6.58% higher when T is 7.75, and 4.76% higher when T is 10.25. Similarly, we performed the same experiments using three different modulators a, 3, 5, and 7, and the results are given in Table V. The recall rate of dense objects is 5.60% higher than that of nondense objects when a is 3, 6.58% higher when a is 5, and 8.93% higher when a is 7. It can be seen that the recall rate of dense objects is higher under both relatively strict and lenient determination conditions. In addition, for the dense and nondense objects discriminated in the DOTA [3], DIOR-R [7], and UCAS-AOD [5] datasets under the relatively moderate T value of 7.75 and a value of 5, we use two different detection

TABLE V
RECALL OF DENSE AND NONDENSE OBJECTS IN THE DIOR-R DATASET UNDER THE DETERMINATION METHOD WITH DIFFERENT a VALUES

Value of a	Dense or not	Object number	Recall
3	✔	80 254	76.66
	✖	44 137	71.06
5	✔	85 338	76.74
	✖	39 053	70.16
7	✔	88 243	77.27
	✖	36 148	68.34

TABLE VI
RECALL OF DENSE AND NONDENSE OBJECTS IN DOTA, DIOR-R, AND UCAS-AOD DATASETS

Dataset	Dense or not	Object number	Recall	
			FR-O	RoI Trans.
DOTA	✔	22 740	91.46	94.98
	✖	6112	87.55	89.77
DIOR-R	✔	85 338	71.56	76.74
	✖	39 053	65.42	70.16
UCAS-AOD	✔	2248	93.37	96.17
	✖	2488	89.59	93.49

networks, Faster RCNN [33] and RoI Transformer, to detect and calculate their recall separately. In this experiment, we use the training set for training and validation set for testing on the DOTA dataset, and the trainval set for training and test set for testing on the DIOR-R and UCAS-AOD datasets. In order to avoid the imbalance of training samples between dense and nondense objects in the datasets, which will lead to the imbalance of network learning and affect the judgment of results, we balance the training samples to make the number of dense and nondense objects the same before training. The results of the experiment are given in Table VI. On the DOTA dataset, the recall of dense objects obtained by Faster RCNN is 3.91% higher than the nondense objects, and the recall of dense objects obtained by RoI Transformer is 5.21% higher than the nondense objects. On the DIOR-R dataset, the recall of dense objects obtained by Faster RCNN is 6.14% higher than the nondense objects, and the recall of dense objects obtained by RoI Transformer is 6.58% higher than the nondense objects. On the UCAS-AOD dataset, the recall of dense objects obtained by Faster RCNN is 3.78% higher than the nondense objects, and the recall of dense objects obtained by RoI Transformer is 2.68% higher than the nondense objects. It is known from this experiment that the overall recall of dense objects is somewhat higher compared to nondense objects when tested on different datasets using different networks. This result is different from our intuitive understanding and from other works that describe dense objects as harder to detect. In other works, it is only qualitatively stated that dense objects are more difficult to detect without relevant experimental proof, while our results rely on experiments and are relatively more credible.

As shown in Fig. 9, we display the feature maps extracted by backbone and sent to the detection head in the hot map. In order to show the characteristics of the feature maps more

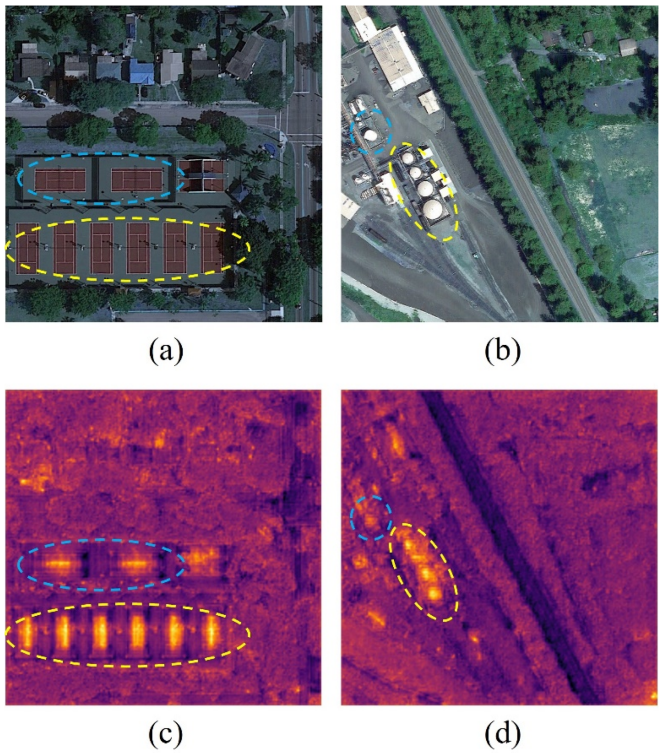


Fig. 9. Dense and nondense objects on the input image and feature heat map. (a), (b) Input images. (c), (d) Feature heat maps. Those in the yellow circle are dense objects, while those in the blue circle are nondense objects.

comprehensively, we average the values on all feature channels and convert them into a hot map. From the circles marked in the figure, we can see that the area with dense objects has higher values, while the area with sparse objects has lower values, indicating that the network has a higher response to the area with dense objects. The network achieves the object detection task based on the recognition of various different features. The features in the region with a large number of objects are richer and denser, so the network has a high response to this region. In addition, in a natural image with an imaging perspective of front or side view, multiple objects on it may be at different depth positions, resulting in mutual occlusion phenomena that cause the loss of object features. This phenomenon is more likely to occur in the area with dense targets, which has a great adverse impact on the detection of dense objects. Due to its special overhead view in RSI, the imaging targets are objects on the ground surface and rarely exist to obscure each other, and the object features are basically complete with few missing cases. From the above-mentioned analysis, it can be concluded that the densely distributed objects in the object detection of RSI are less likely to be difficult to detect.

E. Ablation Experiments of RoIF-Net

In this section, we perform ablation experiments to verify the effectiveness of the proposed second-stage detection head RoIF-Net. The experiments are all trained on the training set of the DOTA [3] dataset and tested on the test set. We use mAP as a criterion to evaluate the performance of the method.

TABLE VII
RESULTS OF ABLATION EXPERIMENTS FOR ROI FUSION MODULE AND FEATURE INDUCTION MODULE ON DOTA DATASET

baseline	RFM	FIM	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
✔			89.39	85.59	60.98	81.33	80.80	85.60	88.40	90.90	87.54	87.72	71.52	73.38	79.10	82.22	72.98	81.16
✔	✔		89.34	84.59	60.93	82.54	80.83	85.89	88.48	90.84	85.80	87.81	72.31	72.41	79.26	81.89	73.73	81.11
✔		✔	89.36	85.62	61.02	82.36	80.95	85.68	88.51	90.90	87.11	87.68	73.13	74.29	79.29	81.57	72.86	81.35
✔	✔	✔	89.23	86.22	62.05	82.02	80.69	86.37	88.26	90.87	87.07	87.52	74.44	74.02	79.55	83.06	75.63	81.80

RFM stands for RoI fusion module and FIM stands for feature induction module.

1) Ablation Experiments of RoI Fusion Module and Feature Induction Module: The overall structure of RoIF-Net is divided into two parts: the RoI fusion module and the feature induction module. In the RoI fusion module, the RoI are extracted from the original image and the feature maps, and they are summed and fused together to increase the detail features. The feature induction module uses a self-attentive mechanism to discriminate and generalize the features to produce the final regression and classification results. In this section, we present and analyze the results of the ablation experiments of these two modules. In the experiment, RoI Transformer [51] is used as the baseline and ConvNeXT [61] as the backbone. We design four groups of experiments, in which the first group uses the traditional second-stage detection structure in the original method, the second group uses only the RoI fusion module, the third group uses only the feature induction module, and the fourth group uses both two modules, and each group is trained and tested the network separately. The test results are given in Table VII. It can be seen from the table that the RoI fusion module does not improve mAP when used alone and even has a 0.05% drop, the feature induction module has a slight 0.19% improvement in mAP when used alone, and only when these two modules are used together does mAP improve significantly by 0.64%. When the RoI fusion module is used alone, the added detail features on the original image are not further extracted and generalized, which is hardly helpful for the final regression and classification of the second stage network. The feature induction module mainly consists of a self-attention mechanism, which works poorly on the rich high-level semantic features extracted from the feature map, while it works better on the detail features extracted from the original image. The two modules complement each other and are used together to fully utilize the capabilities of the RoIF-Net.

2) Ablation Experiments of RoIF-Net in Different Two-Stage Detector: As described in Section III-C, the RoIF-Net we designed can be placed in an arbitrary two-stage detector. In this section, we use different two-stage detection networks as baseline, and change the second stage to RoIF-Net for ablation experiments. In this experiment, we use three two-stage detection networks, i.e., Faster RCNN [33], Oriented RCNN [42], and RoI Transformer [51], with different backbone. The experimental results are given in Table VIII, from which it can be seen that the detection results of the network improved by 0.65%, 0.41%, 0.56%, and 0.64% of mAP after using the RoIF-Net we designed, respectively. This result indicates that the RoIF-Net stimulates the potential of the second-stage detection network, which can be effective in improving the accuracy in different two-stage detection networks with strong universality.

TABLE VIII
RESULTS OF ABLATION EXPERIMENTS FOR ROIF-NET BASED ON DIFFERENT TWO-STAGE DETECTION NETWORKS ON THE DOTA DATASET

Detector	Backbone	mAP		
		Baseline	RoIF-Net	Improvement
FR-O	R50	74.63	75.28	+0.65
Oriented RCNN	ConvNeXT	81.09	81.50	+0.41
RoI Trans.	Swin	80.64	81.20	+0.56
RoI Trans.	ConvNeXT	81.16	81.80	+0.64

TABLE IX
EFFECTIVENESS OF OUR PROPOSED ROIF-NET WITH FASTER RCNN AS BASELINE ON THE DOTA DATASET

baseline	+Ours	Params(M)	GFLOPs	FPS	mAP
✔		41.22	216.50	18.35	74.63
✔	✔	42.08	243.54	15.89	75.28

3) Analysis of the Computational Complexity of RoIF-Net: We use Faster RCNN as the baseline for experiments and analyze the computational complexity of the proposed method using the number of model parameters, floating point operations (FLOPs), and frames per second (FPS). The experimental results are given in Table IX, from which it can be seen that using our proposed second-stage detection structure RoIF-Net on the basis of Faster RCNN, the number of model parameters and FLOPs increase by 0.86M and 27.04G, respectively, and the FPS has a reduction of 2.46, while the mAP improves by 0.65%. This is due to the addition of RoI extraction and convolution operations in the RoI fusion module and self-attention and convolution operations in the feature induction module. These two modules improve the detection effect, but reduce the detection efficiency.

F. Comparison With Advanced Methods

In this section, we compare the method proposed in this article with other classical and advanced methods on the internationally credible and challenging public datasets DOTA [3], DIOR-R [7], and UCAS-AOD [5]. In the experiments of this section, our method uses the two-stage detection network RoI Transformer [51] as baseline and replaces the second-stage detection head with RoIF-Net. The datasets are described in Section IV-A, and the experimental parameters are set in Section IV-B.

1) Comparison Results on DOTA: On the DOTA dataset, we compared with a variety of advanced methods as well as classical methods and the results are given in Table X. As can be seen from the table, our proposed RoIF-Net is able to achieve 81.80% mAP when using ConvNeXT [61] as the backbone network, which outperforms all the results in the table to the current SOTA level.

TABLE X
COMPARISON WITH STATE-OF-THE-ART METHODS ON DOTA DATASET

Method	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
FR-O [3]	R-101	79.42	77.13	17.70	64.05	35.30	38.02	37.16	89.41	69.64	59.28	50.30	52.91	47.89	47.40	46.30	54.13
RetinaNet-O [39]	R-101	88.82	81.74	44.44	65.72	67.11	55.82	72.77	90.55	82.83	76.30	54.19	63.64	63.71	69.73	53.37	68.72
DRN [8]	H-104	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
FFA [9]	R-101	90.10	82.70	54.20	75.20	71.00	79.90	83.50	90.70	83.90	84.60	61.20	68.00	70.70	76.00	63.70	75.70
CoF-Net [66]	R-50	89.60	83.10	48.30	73.60	78.20	83.00	86.70	90.20	82.30	86.60	67.60	64.60	74.70	71.30	78.40	77.20
Oriented Rep. [14]	Swin	89.11	82.32	56.71	74.95	80.70	83.73	87.67	90.81	87.11	85.85	63.60	68.60	75.95	73.54	63.76	77.63
FoRDet [18]	VGG-16	89.62	85.88	47.55	81.45	80.63	81.84	88.08	90.87	88.27	86.41	72.42	67.69	73.91	72.67	64.63	78.13
TransConvNet [15]	Swin	89.25	84.67	55.72	75.23	80.23	82.43	89.58	90.64	86.14	88.70	69.34	69.95	71.75	74.27	68.37	78.41
TIOE-Det [67]	–	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
S^2A-Net [54]	R-50	88.89	83.60	57.74	81.95	79.94	83.19	89.11	90.78	84.87	87.81	70.30	68.25	78.30	77.01	69.58	79.42
ReDet [10]	ReR-50	88.81	82.48	60.83	80.82	78.34	86.06	88.31	90.87	88.77	87.03	68.65	66.90	79.26	79.71	74.67	80.10
G-Rep [13]	Swin	88.15	81.64	61.30	79.50	80.94	85.68	88.37	90.90	85.47	87.77	71.01	67.42	77.19	81.23	75.83	80.16
DEA [11]	ReR-50	89.92	83.84	59.65	79.88	80.11	87.96	88.17	90.31	88.93	88.46	68.93	65.94	78.04	79.69	75.78	80.37
CGCDet [68]	R-50	89.42	84.49	59.83	80.78	79.53	84.75	88.55	90.79	87.81	87.06	69.72	71.09	79.38	80.96	75.32	80.70
AFF-Det [69]	R-50	88.96	85.57	61.64	79.90	76.41	85.20	88.59	90.82	87.24	86.73	69.69	69.93	79.15	83.48	77.58	80.73
Oriented RCNN [42]	R-50	89.84	85.43	61.09	79.82	79.71	85.35	88.82	90.88	86.68	87.73	72.21	70.80	82.42	78.18	74.11	80.87
OSKDet [16]	R-101	90.03	86.94	61.24	81.48	79.63	85.72	88.52	90.84	89.26	87.55	68.38	71.24	78.89	79.95	73.97	80.91
KFIoU [17]	Swin	89.44	84.41	62.22	82.51	80.10	86.07	88.68	90.90	87.32	88.38	72.80	71.95	78.96	74.95	75.27	80.93
RoIF-Net (ours)	Swin	89.29	84.85	60.97	82.39	80.45	85.97	88.55	90.85	86.39	87.98	72.45	71.54	79.07	82.21	75.05	81.20
RoIF-Net (ours)	ConvNeXT	89.23	86.22	62.05	82.02	80.69	86.37	88.26	90.87	87.07	87.52	74.44	74.02	79.55	83.06	75.63	81.80

The results marked in red and blue are the best and second best in each column, respectively.

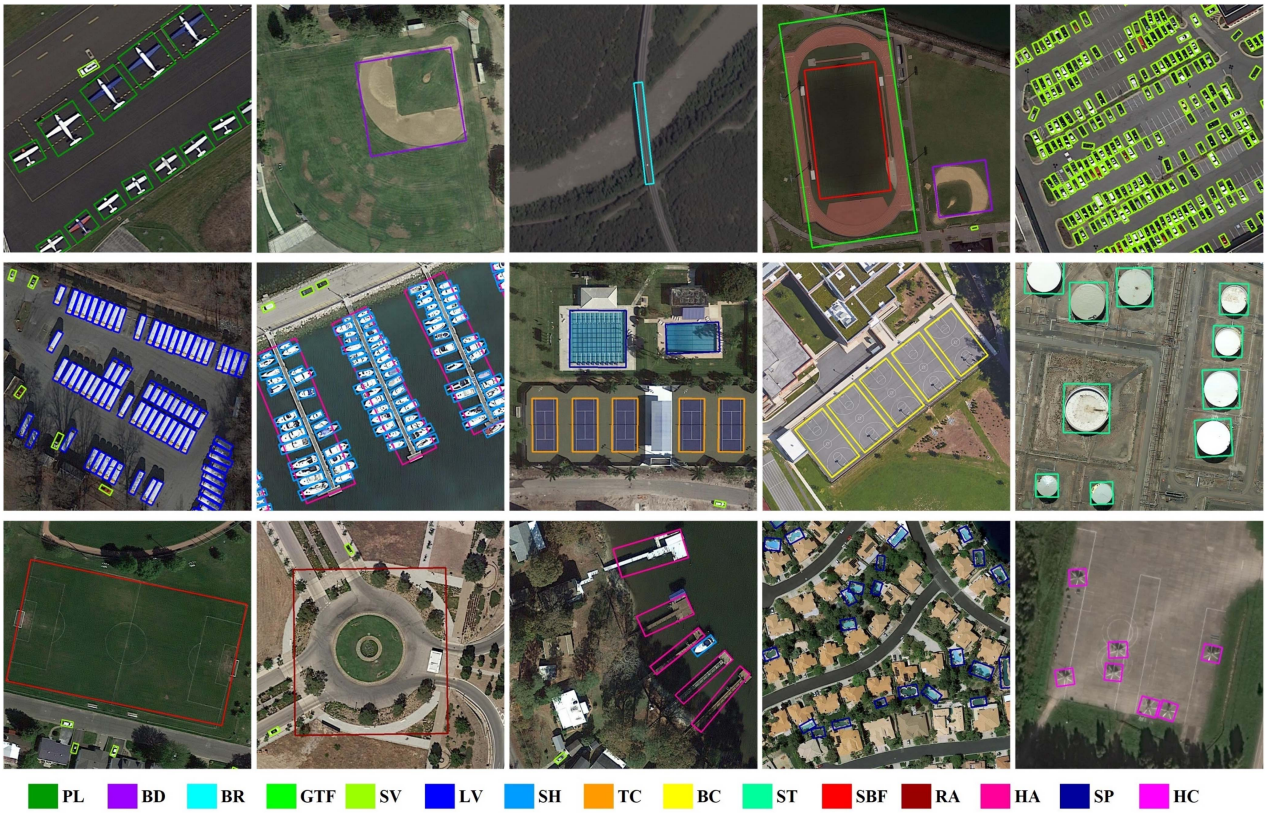


Fig. 10. Some visualization results from our method on DOTA. Confidence threshold is set to 0.3. Boxes in different colors represent different categories of objects.

Out of all 15 detection categories, we have the best or the second best results in 7 categories, which proves the advantage of our method in OBB object detection. In addition, when using Swin [58] as the backbone, it was also able to achieve 81.20% mAP, which is still better than other methods and in the next best level. The above-mentioned results show the progressiveness of our method. Some visual detection results on the DOTA dataset are shown in Fig. 10. It can be seen from the figure that in the DOTA dataset, although the background in the image is complex, the size difference of the object is large, and the object has arbitrary direction, each type of object can still be detected well, and the visualization has achieved satisfactory results.

2) Comparison Results on DIOR-R: We also compare with several classical and advanced methods on the DIOR-R dataset, and the results are given in Table XI. The DIOR-R dataset has 20 categories and is relatively more challenging. As can be seen from the table, our proposed method RoIF-Net is able to achieve 65.12% mAP results when using ResNet50 [59] as the backbone, which is better than all other methods and reaches the SOTA. In addition, it was able to achieve an impressive 68.49%

TABLE XI
COMPARISON WITH STATE-OF-THE-ART METHODS ON DIOR-R DATASET

Method	Backbone	APL	APO	BF	BC	BR	CH	DAM	ETS	ESA	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP
FR-O [33]	R-50	62.79	26.80	71.72	80.91	34.20	72.57	18.95	66.45	66.75	66.63	79.24	34.95	48.79	81.14	64.34	71.21	81.44	47.31	50.46	65.21	59.54
RetinaNet-O [39]	R-50	61.49	28.52	73.57	81.17	23.98	72.54	19.94	72.39	58.20	69.25	79.54	32.14	44.87	77.71	67.57	61.09	81.46	47.33	38.01	60.24	57.55
Gliding Ver. [43]	R-50	65.35	28.87	74.96	81.33	33.88	74.31	19.58	70.72	64.70	72.30	78.68	37.22	49.64	80.22	69.26	61.13	81.49	44.76	47.71	65.04	60.06
RoI Trans. [51]	R-50	63.34	37.88	71.78	87.53	40.68	72.60	26.83	78.71	68.09	68.96	82.74	47.71	55.61	81.21	78.23	70.26	81.61	54.86	43.27	65.52	63.87
AOPG [7]	R-50	62.39	37.79	71.62	87.63	40.90	72.47	31.08	65.42	77.99	73.20	81.94	42.32	54.45	81.17	72.69	71.31	81.49	60.04	52.38	69.99	64.41
DODet [12]	R-50	63.40	43.35	72.11	81.32	43.12	72.59	33.32	78.77	70.84	74.15	75.47	48.00	59.31	85.41	74.04	71.56	81.52	55.47	51.86	66.40	65.10
RoIF-Net (ours)	R-50	72.99	39.03	72.88	82.58	40.83	73.75	29.19	69.46	78.71	74.65	83.97	47.39	55.45	82.35	80.74	63.85	82.67	55.03	49.44	67.40	65.12
RoIF-Net (ours)	ConvNeXT	72.17	43.95	72.23	89.65	43.94	72.66	34.07	74.93	88.67	78.59	84.71	51.13	57.51	81.27	83.03	71.19	89.83	63.22	50.57	66.48	68.49

The results marked in red and blue are the best and second best in each column, respectively.

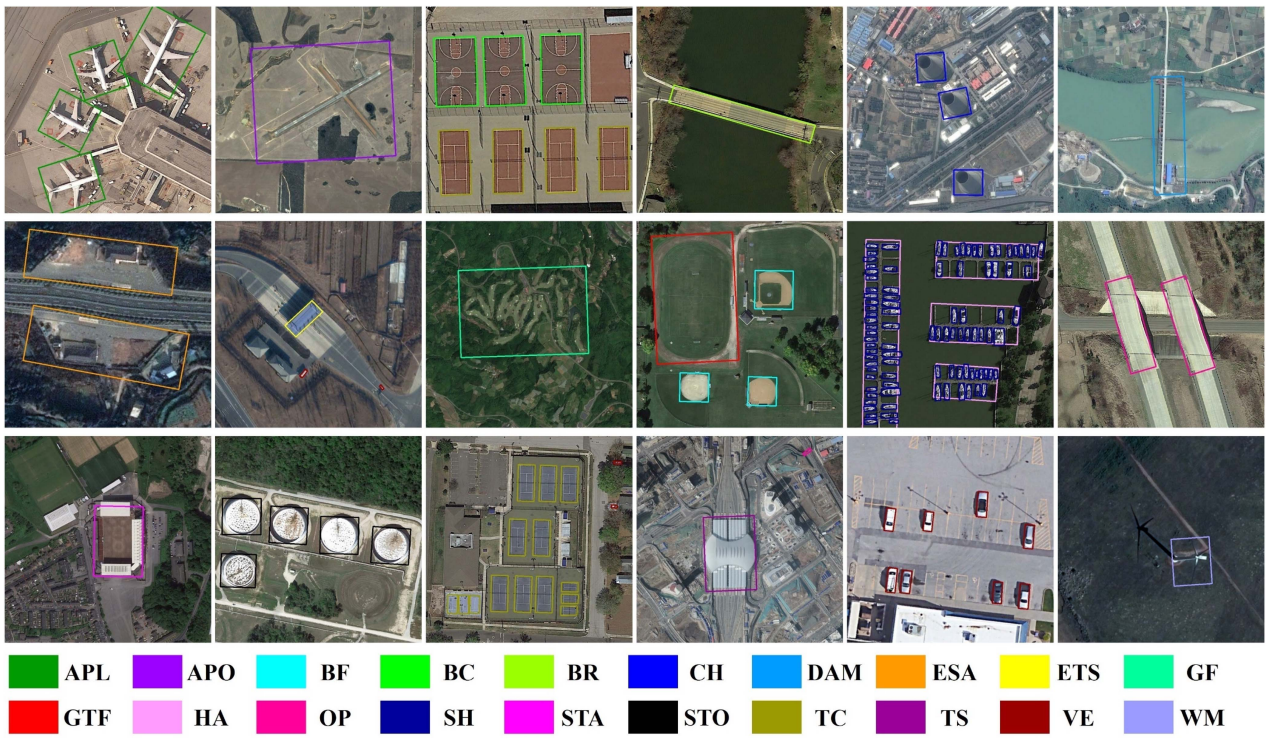


Fig. 11. Some visualization results from our method on DIOR-R. Confidence threshold is set to 0.3. Boxes in different colors represent different categories of objects.

mAP when using ConvNeXT [61] as the backbone, a result that is significantly ahead of all other results in the table. Among all the 20 detection categories, we have the best results in 11 categories. In some categories such as APL, ESA, GTF, STA, and TC, we have a large mAP lead compared to other methods. These aforementioned results illustrate the advancement of our method. Some visual detection results of our method on this dataset are shown in Fig. 11. It can be seen from the figure that although the categories are diverse and the detection is difficult, our method produces few errors in the detection of objects with arbitrary directions, and the visualization achieves desired results.

3) Comparison Results on UCAS-AOD: Similarly, on the test set of the UCAS-AOD, we compared with other methods and the results are given in Table XII. This dataset has only two types of objects, so the detection difficulty is relatively small. As it is shown in the table, our proposed RoIF-Net can achieve an excellent mAP result of 90.25%, which is better than the other methods. In both detection categories, our method is the best in the detection results for plane and the second best for car. Some of the visualized detection results on this dataset are shown in Fig. 12. As can be seen from this, good results are obtained for the detection of cars and planes with arbitrary orientations in different scenarios.

TABLE XII
COMPARISON WITH STATE-OF-THE-ART METHODS ON UCAS-AOD DATASET

Method	car	plane	mAP
YOLOv3-O [37]	74.63	89.52	82.08
RetinaNet-O [39]	84.64	90.51	87.57
Faster RCNN-O [33]	86.67	90.61	88.64
GF-CSL [64]	88.76	90.56	89.66
Point RCNN [65]	89.60	90.48	90.04
Oriented Rep. [14]	89.51	90.70	90.11
G-Rep [13]	89.64	90.67	90.16
RoIF-Net (ours)	89.83	90.67	90.25

The results marked in red and blue are the best and second best in each column, respectively.

Fig. 12. Some visualization results from our method on UCAS-AOD. Confidence threshold is set to 0.3. Yellow boxes represent detected cars and the blue boxes represent detected planes.

V. Conclusion and Discussion

In this article, we design a dense object determination method based on OBB annotation, according to which the objects in the datasets are classified as dense and nondense objects. Their detection results show that dense objects in RSI are not more difficult to detect compared to nondense objects. Our work still has certain limitations: the determination method of dense objects can further be optimized and improved, and the effect of dense distribution on object feature recognition under different environmental conditions can be studied in depth. The important contribution of this work is to provide an idea to quantify the denseness of objects, which hopefully will help to enrich and deepen the study of dense objects in future work. We propose the RoIF-Net to improve the detection effectiveness of two-stage network based on OBB, which adds detail information by fusing the RoI extracted from the original image and the feature maps, and constructs a feature induction module to realize the final position regression and category classification. We demonstrate the effectiveness of our proposed method through extensive experiments on the DOTA, DIOR-R, and UCAS-AOD datasets, and the OBB detection experimental results achieve SOTA on these datasets. However, this method is only applicable in the two-stage detection method and increases the computational complexity, which causes a decrease in detection efficiency. In the future work, how to efficiently utilize detail features to make the network avoid the background influence in identifying the object features is of great research value.

References

[1] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014, doi: 10.1016/j.isprsjprs.2014.10.002.

[2] 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, doi: 10.1109/TGRS.2016.2645610.

[3] 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.*, 2018, pp. 3974–3983, doi: 10.1109/CVPR.2018.00418.

[4] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020, doi: 10.1016/j.isprsjprs.2019.11.023.

[5] 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.*, 2015, pp. 3735–3739, doi: 10.1109/ICIP.2015.7351502.

[6] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2017, vol. 2, pp. 324–331, doi: 10.5220/0006120603240331.

[7] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411, doi: 10.1109/TGRS.2022.3183022.

[8] X. Pan et al., "Dynamic refinement network for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11204–11213, doi: 10.1109/CVPR42600.2020.01122.

[9] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 294–308, 2020, doi: 10.1016/j.isprsjprs.2020.01.025.

[10] 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.*, 2021, pp. 2785–2794, doi: 10.1109/CVPR46437.2021.00281.

[11] D. Liang et al., "Anchor retouching via model interaction for robust object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5619213, doi: 10.1109/TGRS.2021.3136350.

[12] G. Cheng et al., "Dual-aligned oriented detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618111, doi: 10.1109/TGRS.2022.3149780.

[13] L. Hou, K. Lu, X. Yang, Y. Li, and J. Xue, "G-Rep: Gaussian representation for arbitrary-oriented object detection," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 757, doi: 10.3390/rs15030757.

[14] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1829–1838.

[15] X. Liu, S. Ma, L. He, C. Wang, and Z. Chen, "Hybrid network model: TransConvNet for oriented object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2090, doi: 10.3390/rs14092090.

[16] D. Lu, D. Li, Y. Li, and S. Wang, "OSKDet: Orientation-sensitive keypoint localization for rotated object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1182–1192.

[17] X. Yang et al., "The KFIoU loss for rotated object detection," 2022, *arXiv:2201.12558*, doi: 10.48550/arXiv.2201.12558.

[18] T. Zhang et al., "Foreground refinement network for rotated object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610013, doi: 10.1109/TGRS.2021.3109145.

[19] Z. Shu, X. Hu, and J. Sun, "Center-point-guided proposal generation for detection of small and dense buildings in aerial imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 7, pp. 1100–1104, Jul. 2018, doi: 10.1109/LGRS.2018.2822760.

[20] 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.*, 2018, pp. 436–440, doi: 10.1109/ICSESS.2018.8663939.

[21] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5227–5236.

[22] Z. Ji, Q. Kong, H. Wang, and Y. Pang, "Small and dense commodity object detection with multi-scale receptive field attention," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1349–1357.

[23] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12214–12223.

[24] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and R. V. Babu, "Locate, size, and count: Accurately resolving people in dense crowds via detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2739–2751, Aug. 2021.

[25] Z. Deng and C. Yang, "Multiple-step sampling for dense object detection and counting," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 1036–1042.

[26] J. Li, H. Zhang, R. Song, W. Xie, Y. Li, and Q. Du, "Structure-guided feature transform hybrid residual network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5610713.

[27] 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, 2021, Art. no. 2664.
[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.
[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.
[30] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–11.
[31] W. Liu et al., "SSD: Single shot MultiBox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.
[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.*, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.
[33] 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, doi: 10.1109/TPAMI.2016.2577031.
[34] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[35] 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.*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
[36] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.
[37] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.
[38] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[39] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.
[40] 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, doi: 10.1109/JSTARS.2020.3036685.
[41] J. Wang, W. Yang, H. C. Li, H. Zhang, and G. S. Xia, "Learning center probability map for detecting objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4307–4323, May 2021, doi: 10.1109/TGRS.2020.3010051.
[42] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3500–3509, doi: 10.1109/ICCV48922.2021.00350.
[43] 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, doi: 10.1109/TPAMI.2020.2974745.
[44] 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.*, 2021, pp. 15814–15824, doi: 10.1109/CVPR46437.2021.01556.
[45] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 677–694.
[46] 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.
[47] X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 18381–18394.
[48] X. Zheng, W. Zhang, L. Huan, J. Gong, and H. Zhang, "AProNet: Detecting objects with precise orientation from aerial images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 181, pp. 99–112, 2021, doi: 10.1016/j.isprsjprs.2021.08.023.
[49] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.
[50] Y. Gong et al., "Context-aware convolutional neural network for object detection in VHR remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 34–44, Jan. 2020.
[51] 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.*, 2019, pp. 2844–2853.
[52] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773, doi: 10.1109/ICCV.2017.89.
[53] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9300–9308, doi: 10.1109/CVPR.2019.00953.
[54] J. Han, J. Ding, J. Li, and G. S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511, doi: 10.1109/TGRS.2021.3062048.
[55] 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, vol. 35, no. 4, pp. 3163–3171.
[56] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.
[57] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 1874–1883.
[58] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[60] 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.*, 2009, pp. 248–255.
[61] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.
[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[63] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–10.
[64] J. Wang, F. Li, and H Bi, "Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707013, doi: 10.1109/TGRS.2022.3175520.
[65] Q. Zhou and C. Yu, "Point RCNN: An angle-free framework for rotated object detection," *Remote Sens.*, vol. 14, no. 11, 2022, Art. no. 2605, doi: 10.3390/rs14112605.
[66] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617, doi: 10.1109/TGRS.2022.3233881.
[67] 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. Photogrammetry Remote Sens.*, vol. 196, pp. 241–255, 2023, doi: 10.1016/j.isprsjprs.2023.01.001.
[68] Y. Wang et al., "Learning oriented object detection via naive geometric computing," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3242323.
[69] P. Zhen et al., "Towards accurate oriented object detection in aerial images with adaptive multi-level feature fusion," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 19, no. 1, pp. 1–22, 2023, doi: 10.1145/3513133.

Yuxi Zhang received the bachelor's degree in optoelectronic information engineering from the Harbin Institute of Technology, Weihai, China, in 2018. He is currently working toward the Ph.D. degree in optical engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research interests include image processing, deep learning, and object detection in remote sensing images.

Yongcheng Wang received the bachelor's degree in measurement and control technology and instruments from Jilin University, Changchun, China, in 2003, and the Ph.D. degree in optical engineering from the Chinese Academy of Sciences, Changchun, China, in 2010.

He is currently a Researcher with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include image engineering and space payload embedded systems.

Ning Zhang received the bachelor's degree in communication engineering from Northeastern University, Qinhuangdao, China, in 2017, and the Ph.D. degree in optical engineering from the Chinese Academy of Sciences, Changchun, China, in 2022.

She is currently a Research Associate with Tsinghua University Beijing. Her research interests include remote sensing image super-resolution and change detection.

Zheng Li received the bachelor's degree in optoelectronic information engineering from the Changchun University of Science and Technology, Changchun, China, in 2020. He is currently working toward the Ph.D. degree in optical engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research interests cover image processing, deep learning, and object detection in remote sensing images.

Zhikang Zhao received the bachelor's degree in optoelectronic information engineering from the Ocean University of China, Qingdao, China, in 2019. He is currently working toward the Ph.D. degree in optical engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research interests include image processing, deep learning, and remote sensing image super-resolution.

Yunxiao Gao received the bachelor's degree in measurement and control technology and instruments from Qufu Normal University, Jinan, China, in 2018. He is currently working toward the Ph.D. degree in optical engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, Changchun, China.

His research interests cover image processing, deep learning, and object detection in remote sensing images.

Chi Chen received his bachelor's degree in optoelectronic information engineering from Anhui University, Hefei, China, in 2021. He is currently working toward the Ph.D. degree in optical engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Beijing, China.

His research interests include image processing and hyperspectral image super-resolution.

Hao Feng received the bachelor's degree in measurement and control technology and instruments from the East China University of Technology, Nanchang, China, in 2020. He is currently working toward the Ph.D. degree in optical engineering with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His research interests include hyperspectral remote sensing image processing and deep learning.