mathematics

MDPI

Article

Improving Oriented Object Detection by Scene Classification and Task-Aligned Focal Loss

Xiaoliang Qian, Shaoguan Gao, Wei Deng and Wei Wang *

College of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; qxlzengli@zzuli.edu.cn (X.Q.); 332101050082@email.zzuli.edu.cn (S.G.); 2007046@zzuli.edu.cn (W.D.)

* Correspondence: wangwei-zzuli@zzuli.edu.cn

Abstract: Oriented object detection (OOD) can precisely detect objects with arbitrary direction in remote sensing images (RSIs). Up to now, the two-stage OOD methods have attracted more attention because of their high detection accuracy. However, the two-stage methods only rely on the features of each proposal for object recognition, which leads to the misclassification problem because of the intra-class diversity, inter-class similarity and clutter backgrounds in RSIs. To address the above problem, an OOD model combining scene classification is proposed. Considering the fact that each foreground object has a strong contextual relationship with the scene of the RSI, a scene classification branch is added to the baseline OOD model, and the scene classification result of input RSI is used to exclude the impossible categories. To focus on the hard instances and enhance the consistency between classification and regression, a task-aligned focal loss (TFL) which combines the classification difficulty with the regression loss is proposed, and TFL assigns lager weights to the hard instances and optimizes the classification and regression branches simultaneously. The ablation study proves the effectiveness of scene classification branch, TFL and their combination. The comparisons with 15 and 14 OOD methods on the DOTA and DIOR-R datasets validate the superiority of our method.

Keywords: oriented object detection; remote sensing image; scene classification branch; task-aligned focal loss

MSC: 68T45

Citation: Qian, X.; Gao, S.; Deng, W.; Wang, W. Improving Oriented Object Detection by Scene Classification and Task-Aligned Focal Loss. *Mathematics* **2024**, *12*, 1343. https://doi.org/10.3390/math12091343

Academic Editors: Xiangtao Zheng, Jinchang Ren, Ling Wang and Ivan Lorencin

Received: 25 March 2024
Revised: 17 April 2024
Accepted: 25 April 2024
Published: 28 April 2024

1. Introduction

Oriented object detection (OOD) has a significant role in remote sensing image (RSI) interpretation, which can localize objects with arbitrary direction more accurately by using an oriented rectangular box, comparing with horizontal object detection [1–3]. OOD can be applied in resource exploration, urban planning, modern agriculture, military reconnaissance and so on [4–10].

The existing OOD methods can be classified into three categories, including anchor-free methods, one-stage anchor-based methods and two-stage anchor-based methods. Anchor-free methods directly predict the bounding box and category score for each feature point through fully convolutional networks, such as the dynamic refinement network (DRN) [11], oriented representative points (Oriented RepPoints) [12], etc. One-stage methods predefine a number of anchors for each feature point and then use the fully convolutional network to predict the localization offset and category score of each predefined anchor, such as the single-shot alignment network (S^2A-Net) [13], dynamic anchor learning (DAL) [14], etc. [15–22]. Two-stage methods add a region proposal network (RPN), which predicts object proposals from a large number of predefined anchors, and then the localization offset and category score of each proposal are predicted according to its features obtained through a rotated region of interest alignment (RRoIAlign) operation, such as Oriented R-CNN [23], anchor-free oriented proposal generator (AOPG) [24], etc. [25–27].

On the whole, the detection accuracy of two-stage methods is better than the other two types of method; therefore, most works have focused on the two-stage methods. However, they suffer from a common problem, i.e., the misclassification caused by solely relying on the features of each proposal itself. Compared with natural scene images, the intra-class diversity, inter-class similarity and complexity of the background are more severe in remote sensing images. In this situation, solely using the features of proposal can easily lead to the misclassification problem. As shown in Figure 1a, a background region is incorrectly identified as a harbor.

(**a**) (**b**)

Figure 1. Illustration of misclassification problem. (**a**) Detection result of OOD model without scene classification. (**b**) Detection result of our method. The blue and yellow rectangles denote tennis courts and a harbor, respectively. The red circle denotes the misclassification object.

To address the above problem, in this paper an OOD model that fuses scene classification is proposed. As shown in Figure 2, a scene classification branch is added to the baseline OOD model, and the result of scene classification can help the object classification branch to exclude the impossible classes. As a matter of fact, each foreground object has a strong contextual relationship with the scene of RSI; consequently, making full use of scene classification to assist object recognition can effectively alleviate the aforementioned misclassification problem. As shown in Figure 1b, our method can effectively exclude the misclassification of the harbor with the help of scene classification.

Furthermore, to focus on the hard instances and enhance the consistency between classification and regression, in this paper a novel task-aligned focal loss (TFL) is proposed. Specifically, the classification difficulty of each instance is combined with the regression loss to obtain the TFL, which can increase the relative weights of hard instances and optimize the classification and regression of each instance simultaneously.

The major contributions of this paper can be summarized as follows:

1. A novel OOD model fusing scene classification is proposed to address the misclassification problem caused by solely relying on the features of each proposal itself. Owing to intra-class diversity and inter-class similarity, the misclassification of objects easily occurs if only the features of each proposal itself are used for object classification. Considering the contextual relationship between the foreground objects and the scene of an RSI, the scene classification branch is incorporated into the baseline OOD model help the object classification branch exclude impossible categories that do not exist in the RSI;

2. A novel TFL is proposed to focus on the hard instances and enhance the consistency between classification and regression. The TFL is obtained through combining the classification difficulty with regression loss, which can increase the relative weights of hard instances and achieve the simultaneous optimization of the classification and regression branches;
3. The superiority of the proposed method is demonstrated on two public RSI datasets, i.e., the DOTA and DIOR-R benchmarks.

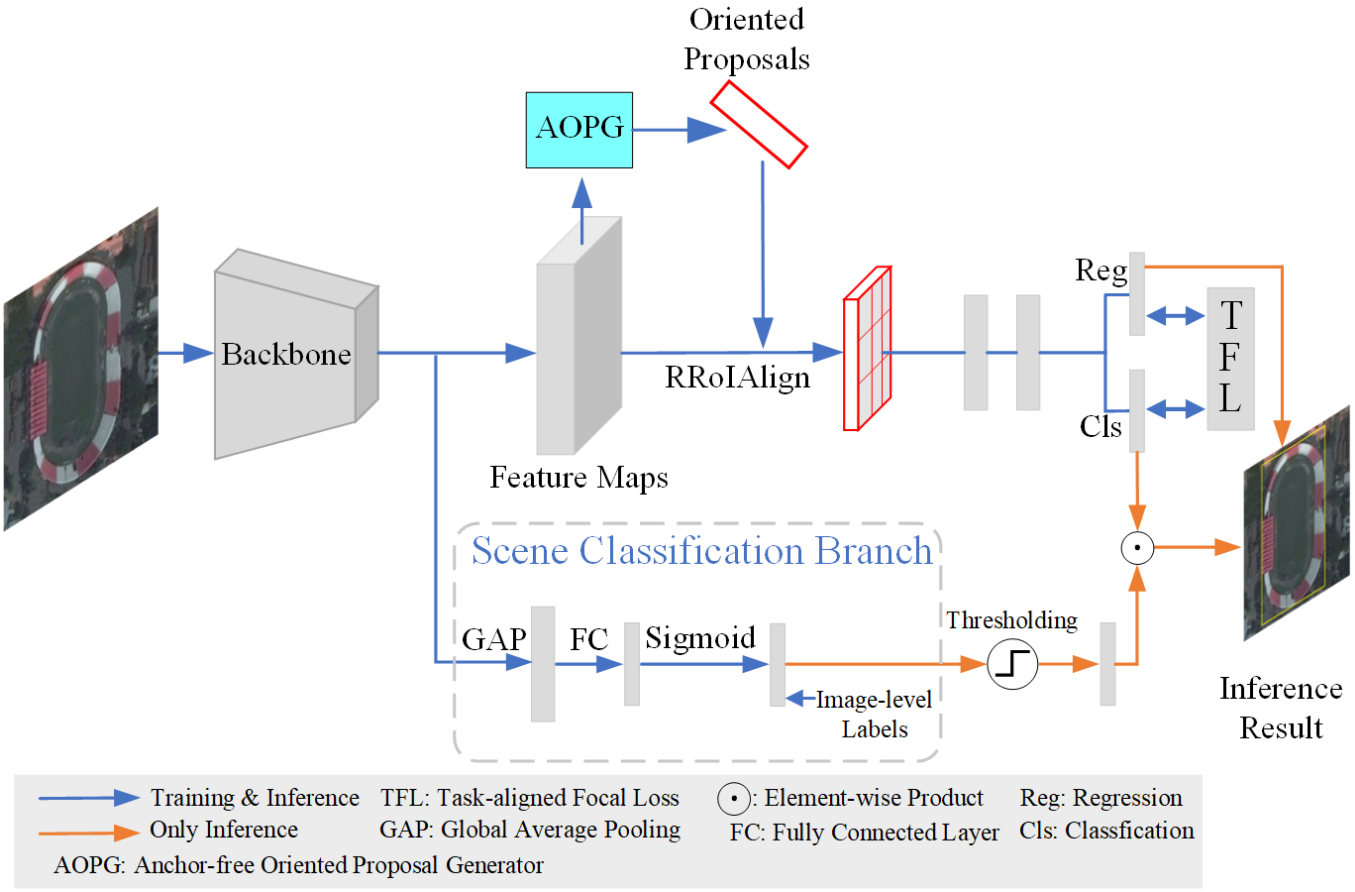

Figure 2. Framework of the proposed method.

2. Related Work

2.1. Oriented Object Detection Methods

Due to the development of deep learning [28–42], the OOD of RSI has made rapid progress in recent years [43–53]. The current OOD methods are generally divided into three categories, i.e., anchor-free methods, one-stage anchor-based methods and two-stage anchor-based methods, and their representative approaches are briefly introduced as follows.

Anchor-free methods: DRN; in [11], the authors proposed a feature selection module that allows the receptive fields of neurons to dynamically adapt to the shape and orientation of objects. Moreover, a dynamic refinement head is proposed to model the uniqueness and specificity of each sample by means of object-awareness to refine the features and thus make better inferences. Oriented RepPoints; in [12], the authors utilized adaptive point sets to represent the oriented boxes and designed an adaptive points assessment assignment strategy to select typical samples for better training.

One-stage anchor-based methods: S^2A-Net; in [13], the authors proposed an aligned convolution module to produce high-quality proposals that could relieve the inconsistency problem between classification and regression to some extent. DAL; in [14], the authors proposed a dynamic anchor learning strategy that combined spatial alignment with feature alignment capabilities for better label assignment. Refined single-stage detector (R3Det); in [54], the authors proposed a feature refinement module and a SkewIou loss to address the detection of objects with large aspect ratios and dense distributions in arbitrary directions.

Two-stage anchor-based methods: RoI Transformer; in [55], the authors converted horizontal proposals to oriented proposals through the proposed RoI learner for accurate OOD. Oriented R-CNN; in [23], the authors proposed an oriented RPN that simply generated high-quality proposals by designing an innovative oriented proposals representation. AOPG; in [24], the authors proposed a coarse location module to acquire oriented anchors and refine them to obtain high-quality oriented proposals. Dual-aligned oriented detector (DoDet); in [56], the authors designed a localization-guided detection head to mitigate the problem of inconsistency between classification and regression. On improving bounding box representations (OIBBR); in [57], the authors proposed a quadrant point representation to handle the boundary discontinuity problem.

2.2. RoI Pooling Methods

For the two-stage detection model, the region of interest (RoI) pooling operation is imposed on the feature maps of each proposal to obtain a feature representation with fixed size. For the original RoI pooling method [58], the feature maps of each proposal are approximately divided into 7×7 bins in the spatial dimension, and the max pooing operation is applied to obtain a 7×7 feature map, which is used as a channel of feature maps of each proposal. However, the original method suffers from quantization errors, i.e., the size of each proposal is not always evenly divisible by 7.

To address the above problem, an RoI Alignment (RoIAlign) operation [59] is proposed to replace original RoI pooling operation, and it is briefly introduced as follows. Firstly, the feature maps of each proposal are exactly divided into 7×7 bins in the spatial dimension, even if the size of the proposal is not evenly divisible by 7; in other words, the coordinates of each bin are decimals. Afterwards, each bin is exactly divided into four regions, and the pixel value of the center point of each region is calculated through the bilinear interpolation of four adjacent pixels of the center point. Finally, the value of each bin is obtained by taking the maximum value of the four center points, and the 7×7 feature representation of each proposal in the spatial dimension is obtained without the quantization error.

Note that the RoIAlign operation is designed for horizontal proposals,and cannot be directly applied to oriented proposals. Therefore, an RRoIAlign operation [55] is proposed to handle above restriction, which is a rotated version of RoIAlign operation, i.e., the only difference of them is rotation. So far, the RRoIAlign operation has been widely used in almost all OOD methods besides our method. To the best of our knowledge, no better RoI pooling method has been proposed to replace the RRoIAlign operation.

3. Proposed Method

3.1. Overview

The architecture of the proposed method is illustrated in Figure 2. First of all, the AOPG [24] is used to generate high-quality oriented proposals. Secondly, the classification and regression results of each proposal are derived from its features, obtained through the RROI Align operation, and the classification and regression branches are trained by the proposed TFL. Thirdly, a scene classification branch is added on the basis of shared backbone features, and it is trained by the image-level labels derived from the instance-level labels. In the inference stage, the predicted results of the scene classification branch are binarized by a predefined threshold, and they are then used to exclude the impossible categories for object classification. Finally, the refined category scores and predicted offsets

of all proposals are jointly used to determine the final detection results. The proposed scene classification branch and TFL are introduced in detail as follows.

3.2. Scene Classification Branch

As mentioned previously, the existing two-stage methods usually suffer from the misclassification problem caused by solely relying on the features of each proposal itself. Considering the fact that each foreground object has a strong contextual relationship with the scene of the RSI, a scene classification branch is added into the OOD model to assist the object classification.

As shown in Figure 2, the scene classification branch is constructed on the basis of a shared backbone. First of all, the feature maps output from the shared backbone are converted into a one-dimensional feature vector by using the global average pooling operation, which are irrelevant with the features obtained from the RRoI Align operation. Secondly, the feature vector is convolved by a fully connected layer to obtain the scene feature vector, denoted as $f \in \mathbb{R}^{C+1}$, where C indicates the number of categories. Finally, the scene classification result of input RSI, denoted as s, can be obtained through the following equation:

$$s_i = Sigmoid(f_i), s.t.\ i \in \mathbb{Z}\ , 1 \le i \le C+1, \tag{1}$$

where f_i indicates the ith element in f, s_i indicates the probability that input RSI belongs to the ith category, the $(C+1)$th category denotes background and $Sigmoid(\cdot)$ denotes the sigmoid activation function. It is worth noting that the softmax classifier is not suitable for the scene classification branch because a multi-classification task is involved here; consequently, as shown in Figure 2, the sigmoid function is adopted as a binary classifier for the classification of each category.

At this point, the training loss of the scene classification branch, denoted as L_{SCB}, is defined by using the binary cross entropy (BCE) loss:

$$L_{SCB} = \sum_{i=1}^{C+1} \left(l_i \log s_i + (1 - l_i) \log(1 - s_i) \right), \tag{2}$$

where l_i indicates the image-level label of the ith category of input RSI; $l_i = 1$ if input RSI contains the objects belonging to ith category, otherwise, $l_i = 0$. As mentioned previously, the classifier of the scene classification branch consists of multiple binary classifiers; therefore, the BCE loss of each category is accumulated to obtain the L_{SCB}.

3.3. Task-Aligned Focal Loss

The focal loss [60] has demonstrated that paying more attention to the hard instances can effectively enhance the detection capability; however, it cannot handle the inconsistency between the classification and regression. To focus on the hard instances and enhance the consistency between classification and regression simultaneously, a task-aligned focal loss, denoted as L_{TFL}, is proposed, and its formulation is given as follows:

$$L_{TFL} = L_{cls} + \alpha L_{reg} \tag{3}$$

$$L_{cls} = \begin{cases} -\log(p_i), & \text{if } y_i = 1 \text{ and } 1 \le i \le C \\ -(1 - p_i) \log(1 - p_i), & \text{if } y_i = 1 \text{ and } i = C+1 \end{cases} \tag{4}$$

$$L_{reg} = e^{(1-p_i)} L_{RIoU},\ s.t.\ y_i = 1 \text{ and } 1 \le i \le C, \tag{5}$$

where L_{cls} and L_{reg} denote the classification and regression loss in L_{TFL}, respectively, α denotes the relative weight of L_{reg} and is quantitatively analyzed in Section 4.2, $p_i \in p$ indicates the probability that each proposal belongs to the ith category, $p \in {}^{C+1}$ indicates the classification result of each proposal, y_i denotes the instance-level label of the ith category of each proposal, α is a predefined hyper-parameter and L_{RIoU} denotes the RIoU

loss [50], which proposed a new metric named rotated intersection of over union (RIoU) and optimized the RIoU for better bounding box regression (BBR).

The explanation of L_{TFL} is as follows. Each instance will be used for classification and regression if it belongs to the assembly of positive samples, i.e., $y_i = 1$ and $1 \leq i \leq C$. At this point, the classification difficulty of each instance, indicated by $e^{(1-p_i)}$, is adaptively adjusted with p_i and is used as the weight of the L_{RIOU}. On the one hand, the hard instances will be assigned larger weights compared with the easy instances in L_{reg}. On the other hand, the L_{RIOU} is optimized and the p_i is converged to 1 simultaneously, along with the minimization of L_{reg}; in other words, the consistency between classification and regression is enhanced. At this point, the traditional cross entropy, indicated by $-\log(p_i)$, is used for the training of classification of positive samples. The proposal will only be used for classification if it belongs to the assembly of negative samples, i.e., $y_i = 1$ and $i = C + 1$; consequently, the original focal loss, indicated by $-(1 - p_i)\log(1 - p_i)$, is used for the training of the classification of negative samples.

3.4. Overall Training Loss

The overall training loss of our OOD method, denoted as L, is formulated as follows:

$$L{=}L_{RPN} + L_{SCB} + L_{TFL} \tag{6}$$

$$L_{RPN} = L'_{cls} + L'_{reg}, \tag{7}$$

where L_{RPN} denotes the loss of the RPN, L'_{cls} denotes the training loss of the classification branch in RPN and is defined by using cross entropy, L'_{reg} denotes the training loss of the BBR branch in RPN and is defined by using the smooth L1 loss [58], and the details of L_{RPN} can be seen in the baseline OOD model [24].

3.5. Inference Stage

First of all, the scene classification result of the input RSI is predicted by the trained scene classification branch, and it is binarized through the following equation:

$$ts_i = \begin{cases} 1, \text{ if } s_i \geq T \\ 0, \text{ otherwise} \end{cases}, \ ts_i \in ts, \tag{8}$$

where $ts \in \mathbb{R}^{C+1}$ denotes the binarized scene classification result of input RSI, ts_i denotes the result of the ith category in ts, T is a predefined threshold and is quantitatively analyzed in Section 4.2. In other words, $ts_i = 0$ if the input RSI does not contain the objects belonging to the ith category, otherwise, $ts_i = 1$. Consequently, the ts can be used to exclude impossible categories for each proposal:

$$ps_i = p_i \times ts_i, \ ps_i \in ps, \tag{9}$$

where $ps \in \mathbb{R}^{C+1}$ denotes the final classification result of each proposal, ps_i denotes the result of the ith category in ps. The motivation of Equation (9) is as follows. The possibility that a proposal belongs to the ith category will be excluded if the predicted result of the scene classification branch indicates that the input RSI does not contain the scene of the ith category. Finally, the ps is used for non-maximumsuppression (NMS) [61].

4. Experiments

4.1. Experiment Setup

4.1.1. Datasets

The DOTA dataset [62] includes 2806 images and 188,282 instances with 15 categories, i.e., plane (PL), baseball (BD), bridge (BR), ground track and field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer fields (SBF), rings (RA), harbors (HA), swimming pools (SP) and helicopters (HC). For single-scale experiments, the raw RSI is cropped into several sub-images with

1024 × 1024 pixels by using a step with 824 pixels. For multi-scale experiments, the raw RSI is initially resized with three ratios, i.e., 0.5, 1 and 1.5, and then the resized RSIs are cut into multiple sub-images with 1024 × 1024 pixels by using a step with 524 pixels. All experimental results were obtained through the DOTA evaluation server.

The DIOR-R dataset [63] includes 23,463 images with 800 × 800 pixels and 192,518 instances with 20 categories, i.e., aircraft (APL), airport (APO), baseball field (BF), BC, BR, chimney (CH), highway service area (ESA), expressway toll station (ETS), dam (DAM), golf field (GF), GTF, HA, overpass (OP), SH, stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE) and windmill (WM).

Similar to other OOD methods [24,49,50,56,57], vertical and horizontal flipping are also used to augment the training samples for fair comparison.

4.1.2. Implementation Details

The stochastic gradient descent (SGD) algorithm was used to optimize our model with a momentum of 0.9 and a weight decay of 0.0001. The initial learning rate, batchsize and the number of epoch were set to 0.02, 8 and 12, respectively. The learning rate was modified to the 0.1-time previous stage at the eighth and eleventh epochs. The threshold of NMS is 0.1 (0.5) on the DOTA (DIOR-R) dataset [62,63].

The experiments were conducted on the mmdetection platform [64] within the PyTorch framework and run on four NVIDIA GeForce RTX 2080Ti (4×11-GB memory).

The mean average precision (mAP) of each category and all categories was used to assess the performance of the proposed method.

4.2. Parameter Analysis

The α (T) in Equation (3) (Equation (8)) is quantitatively analyzed on the DOTA dataset to assess its impact on the final result. As shown in Figures 3 and 4, the highest mAP are obtained at $\alpha = 0.5$ ($T = 0.1$); therefore, in this paper the α (T) is set to 0.5 (0.1). As shown in Equation (3), the α is used to control the relative weight of classification loss and BBR loss; therefore, a too-large or too-small value is not suitable for α.

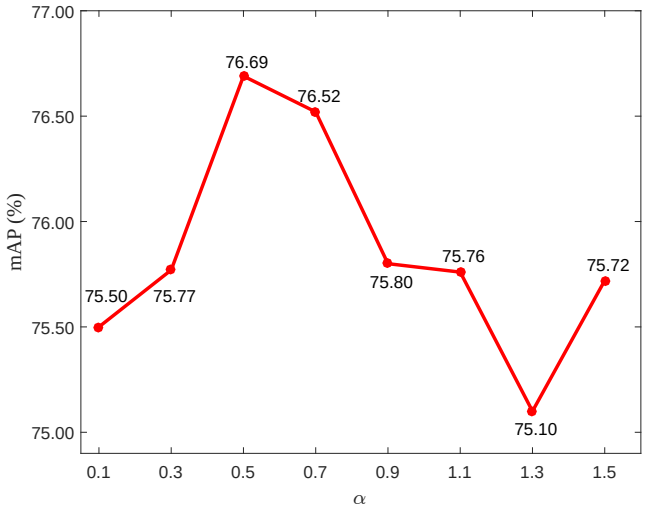

Figure 3. Parameter analysis of α on DOTA dataset.

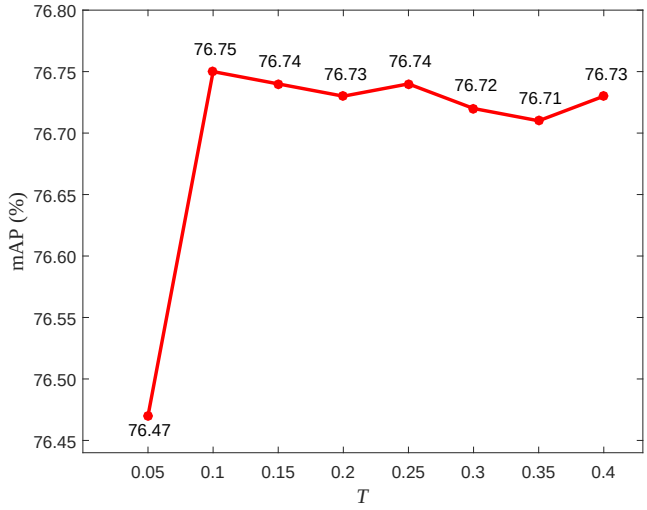

Figure 4. Parameter analysis of T on DOTA dataset.

4.3. Ablation Study

Ablation studies were conducted on the DOTA dataset to evaluate the contribution of the scene classification branch and task-aligned focal loss. The AOPG [24] was applied as the benchmark methodology. As shown in Table 1, the baseline + SCB (TFL) denotes the combination of the baseline method and scene classification branch (task-aligned focal loss), and the baseline +SCB + TFL denotes our method in which the scene classification branch and task-aligned focal loss are all used.

As shown in Table 1, compared with baseline, the mAP of baseline + SCB, baseline + TFL and baseline +SCB + TFL are increased by 1.51%, 1.45% and 1.77%, respectively, which verifies the effectiveness of SCB, TFL and their combination.

Table 1. Ablation study of SCB and TFL on the DOTA dataset.

Baseline	SCB	TFL	mAP
✓			75.24
✓	✓		76.75
✓		✓	76.69
✓	✓	✓	76.95

4.4. Comparisons with Other OOD Methods

To fully validate the capability of our method (denoted as SCTFL), it is compared with several other OOD models on two RSI benchmarks, and the details are as follows.

4.4.1. Results on the DOTA Dataset

As shown in Table 2, we compare our model with 15 other OOD methods on the DOTA dataset: RetinaNet-O [60]; Faster RCNN [31]; DRN [11], CenterMap-Net [65]; multi-category rotation detector for small, cluttered and rotated objects (SCRDet) [66]; S^2A-Net [13]; dynamic prior along with the coarse-to-fine assigner (DCFL) [67]; ROI Transformer [55]; task-collaborated detector (TCD) [68]; AOPG [24]; DODet [56]; Oriented R-CNN [23]; Oriented RepPoints [12]; OIBBR [57] and rotation proposal generation and optimization detector (RPGAOD) [69], and the mAP of our method is boosted by 8.52%, 7.90%, 6.25%, 5.21%, 4.34%, 2.83%, 2.69%, 2.34%, 1.77%, 1.71%, 1.46%, 1.08%, 0.98%, 0.70% and 0.48%, respectively, which shows our best performance on the DOTA dataset.

Table 2. Comparisons with 15 OOD methods on the DOTA dataset under the condition of single-scale training and testing. The AP of each category and the mAP of 15 categories are listed here. The best, sub-optimal and third best results are denoted in red, green and blue colors, respectively (the same below).

Methods	**Backbone**	**Epoch**	**PL**	**BD**	**BR**	**GTF**	**SV**	**LV**	**SH**	**TC**
Retina-O [60]	ResNet50	12	88.67	77.62	41.81	58.17	74.58	71.64	79.11	90.29
Faster-RCNN [31]	ResNet50	12	88.34	73.06	44.86	59.09	73.25	71.49	77.11	90.84
DRN [11]	H-104	120	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14
CenterMap-Net [65]	ResNet50	12	88.88	81.24	53.15	60.65	78.62	66.55	78.10	88.83
SCRDet [66]	ResNet50	12	89.34	80.65	52.09	68.36	60.32	72.41	90.85	87.94
S^2A-Net [13]	ResNet50	12	89.11	82.84	48.37	71.11	78.11	79.39	87.25	90.83
DCFL [67]	ResNet50	12	-	-	-	-	-	-	-	-
ROI Transformer [55]	ResNet50	12	88.65	82.60	52.53	70.87	77.93	76.67	86.87	90.71
TCD [68]	ResNet50	12	89.27	83.79	56.91	72.13	65.75	76.76	70.67	90.88
AOPG [24]	ResNet50	12	89.27	83.49	52.50	69.97	73.51	82.31	87.95	80.89
DODet [56]	ResNet50	12	83.94	84.31	51.39	71.04	79.04	82.86	88.15	90.90
Oriented R-CNN [23]	ResNet50	12	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90
Oriented RepPoints [12]	ResNet50	40	87.02	83.17	54.13	71.16	80.18	78.40	87.28	90.90
OIBBR [57]	ResNet50	12	89.55	83.66	54.06	73.93	78.93	83.08	88.29	80.89
RPGAOD [69]	ResNet50	12	89.34	83.53	54.39	76.59	78.09	81.44	87.64	90.83
SCTFL (Ours)	ResNet50	12	89.85	83.49	55.61	74.61	78.71	83.47	88.05	90.90
Methods	**Backbone**	**Epoch**	**BC**	**ST**	**SBF**	**RA**	**HA**	**SP**	**HC**	**mAP**
Retina-O [60]	ResNet50	12	82.18	74.32	54.75	60.60	62.57	69.57	60.64	68.43
Faster-RCNN [31]	ResNet50	12	78.94	83.90	48.59	62.95	62.18	64.91	56.18	69.05
DRN [11]	H-104	120	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70
CenterMap-Net [65]	ResNet50	12	77.80	83.61	49.36	66.19	72.10	72.36	58.70	71.74
SCRDet [66]	ResNet50	12	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
S^2A-Net [13]	ResNet50	12	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
DCFL [67]	ResNet50	12	-	-	-	-	-	-	-	74.26
ROI Transformer [55]	ResNet50	12	83.83	82.51	53.95	67.61	74.67	68.75	61.03	74.61
TCD [68]	ResNet50	12	86.95	83.81	60.16	71.21	76.54	71.95	70.89	75.18
AOPG [24]	ResNet50	12	87.64	84.71	60.01	66.12	74.19	68.30	57.80	75.24
DODet [56]	ResNet50	12	86.88	84.91	62.69	67.63	75.47	72.22	45.54	75.49
Oriented R-CNN [23]	ResNet50	12	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87
Oriented RepPoints [12]	ResNet50	40	85.97	86.25	59.90	70.49	73.53	72.27	58.97	75.97
OIBBR [57]	ResNet50	12	86.60	84.80	62.03	65.55	74.16	70.09	58.16	76.25
RPGAOD [69]	ResNet50	12	58.89	85.33	65.44	64.96	73.73	70.31	59.50	76.47
SCTFL (Ours)	ResNet50	12	86.87	84.35	66.31	67.10	74.55	68.21	62.20	76.95

4.4.2. Results on the DIOR-R Dataset

As shown in Table 3, we compare our method with 14 other OOD methods on the DIOR-R dataset: fully convolutional one-stage object detection (FCOS-O) [70], RetinaNet-O [60], Double-heads [71], Faster RCNN-O [31], Gliding Vertex [65], S^2A-Net [13], ROI Transformer [55], OIBBR [57], AOPG [24], Oriented R-CNN [23], TCD [68], DODet [56], Oriented RepPoints [12] and DCFL [67], and the mAP of our method is boosted by 15.89%, 9.73%, 7.83%, 7.74%, 7.22%, 4.38%, 3.41%, 3.08%, 2.87%, 2.65%, 2.24%, 2.18%, 0.57% and 0.48%, respectively, which shows our best performance on the DIOR-R dataset.

Table 3. Comparisons with 14 OOD methods on the DIOR-R dataset. The AP of each category and the mAP of 20 categories are listed here.

Methods	Backbone	Epoch	APL	APO	BF	BC	BR	CH	DAM	ETS	ESA	GF
FCOS-O [70]	ResNet50	12	48.70	24.88	63.57	80.97	18.41	68.99	23.26	42.37	60.25	64.83
RetinaNet-O [60]	ResNet50	12	61.49	28.52	73.57	81.17	23.98	72.54	19.94	72.39	58.20	69.25
Double-heads [71]	ResNet50	12	62.13	19.53	71.50	87.09	28.01	72.17	20.35	61.19	64.56	73.37
Faster RCNN-O [31]	ResNet50	12	62.79	26.80	71.72	80.91	34.20	72.57	18.95	66.45	65.75	66.63
Gliding Vertex [65]	ResNet50	12	65.35	28.87	74.96	81.33	33.88	74.31	19.58	70.72	64.70	72.30
S^2A-Net [13]	ResNet50	12	65.40	42.04	75.15	83.91	36.01	72.61	28.01	65.09	75.11	75.56
ROI Transformer [55]	ResNet50	12	63.34	37.88	71.78	87.53	40.68	72.60	26.86	78.71	68.09	68.96
OIBBR [57]	ResNet50	12	63.22	41.39	71.97	88.55	41.23	72.63	28.82	78.90	69.00	70.07
AOPG [24]	ResNet50	12	62.39	37.79	71.62	87.63	40.90	72.47	31.08	65.42	77.99	73.20
Oriented R-CNN [23]	ResNet50	12	62.00	44.92	71.78	87.93	43.84	72.64	35.46	66.39	81.35	74.10
TCD [68]	ResNet50	12	67.95	45.30	73.48	83.55	40.38	74.74	33.60	70.54	79.33	78.03
DODet [56]	ResNet50	12	63.40	43.35	72.11	81.32	43.12	72.59	33.32	78.77	70.84	74.15
Oriented RepPoints [12]	ResNet50	12	-	-	-	-	-	-	-	-	-	-
DCFL [69]	ResNet50	12	-	-	-	-	-	-	-	-	-	-
SCTFL (Ours)	ResNet50	12	69.45	43.78	77.42	88.26	46.67	72.54	35.17	79.43	63.95	72.79

Methods	Backbone	Epoch	GTF	HA	OP	SH	STA	STO	TC	TS	VE	VM	mAP
FCOS-O [70]	ResNet50	12	50.66	31.84	40.80	73.09	66.32	56.61	77.55	38.10	30.69	55.87	51.39
RetinaNet-O [60]	ResNet50	12	79.54	32.14	44.87	77.71	65.57	61.09	81.46	47.33	38.01	60.24	57.55
Double-heads [71]	ResNet50	12	81.97	40.68	42.40	80.36	73.12	62.37	87.09	54.94	41.32	64.86	59.45
Faster RCNN-O [31]	ResNet50	12	79.24	34.95	48.79	81.14	64.34	71.21	81.44	47.31	50.46	65.21	59.54
Gliding Vertex [65]	ResNet50	12	78.68	37.22	49.64	80.22	69.26	61.13	81.49	44.76	47.71	65.04	60.06
S^2A-Net [13]	ResNet50	12	80.47	35.91	52.10	82.33	65.89	66.08	84.61	54.13	48.00	69.67	62.90
ROI Transformer [55]	ResNet50	12	82.74	47.71	55.61	81.21	78.23	70.26	81.61	54.86	43.27	65.52	63.87
OIBBR [57]	ResNet50	12	83.01	47.83	55.54	81.23	72.15	62.66	89.05	58.09	43.38	65.36	64.20
AOPG [24]	ResNet50	12	81.94	42.32	54.45	81.17	72.69	71.31	81.49	60.04	52.38	69.99	64.41
Oriented R-CNN [23]	ResNet50	12	80.95	43.52	58.42	81.25	68.01	65.52	88.62	59.31	43.27	66.31	64.63
TCD [68]	ResNet50	12	78.22	48.46	56.00	85.89	65.07	68.99	81.61	57.22	48.72	63.79	65.04
DODet [56]	ResNet50	12	75.47	48.00	59.31	85.41	74.04	71.56	81.52	55.47	51.86	66.40	65.10
Oriented RepPoints [12]	ResNet50	12	-	-	-	-	-	-	-	-	-	-	66.71
DCFL [69]	ResNet50	12	-	-	-	-	-	-	-	-	-	-	66.80
SCTFL (Ours)	ResNet50	12	81.17	47.50	53.34	89.23	78.99	77.99	88.54	60.02	53.19	71.22	67.28

4.5. Evaluation of Generalizability

4.5.1. Evaluation under the Condition of Multi-Scale Training and Testing

To evaluate the generalizability of our method through varying spatial resolutions and object scales, seven OOD methods are compared with our method under the condition of multi-scale training and testing: DRN [11], S^2A-Net [13], TCD [68], AOPG [24], DODet [56], Oriented R-CNN [23] and OIBBR [57]. These provide multi-scale experiment results. For multi-scale training and testing, the raw RSI is initially resized with three ratios, i.e., 0.5, 1 and 1.5, and then the resized RSIs are cut into multiple sub-images with 1024 × 1024 pixels by using a step with 524 pixels. As shown in Table 4, our method achieves the best results among the eight OOD methods, which validates the generalizability of our method under the condition of multi-scale training and testing.

Table 4. Comparisons with 7 OOD methods on the DOTA dataset under the condition of multi-scale training and testing. The AP of each category and the mAP of 15 categories are listed here. The bold font denotes the best results (the same below).

Methods	Backbone	Epoch	PL	BD	BR	GTF	SV	LV	SH	TC
DRN [11]	H-104	120	89.45	83.16	48.98	62.24	70.63	74.25	83.99	90.73
S^2A-Net [13]	ResNet50	12	88.89	83.60	57.74	81.95	79.94	83.19	89.11	90.78
TCD [68]	ResNet50	12	71.77	80.56	58.18	89.78	88.31	77.84	83.88	68.04
AOPG [24]	ResNet50	12	89.88	85.57	60.90	81.51	78.70	85.29	88.85	90.89
DODet [56]	ResNet50	12	89.96	85.52	58.01	81.22	78.71	85.46	88.59	90.89
Oriented R-CNN [23]	ResNet50	12	89.84	85.43	61.09	79.82	79.71	85.35	88.82	90.88
OIBBR [57]	ResNet50	12	90.14	85.31	60.98	79.92	80.21	85.04	88.80	90.87
SCTFL (Ours)	ResNet50	12	89.79	84.67	60.97	79.39	79.31	85.46	88.36	90.88

Methods	Backbone	Epoch	BC	ST	SBF	RA	HA	SP	HC	mAP
DRN [11]	H-104	120	84.60	85.35	55.76	60.79	71.56	68.82	63.92	72.95
S^2A-Net [13]	ResNet50	12	84.87	87.81	70.30	68.25	78.30	77.01	69.58	79.41
TCD [68]	ResNet50	12	86.29	78.92	90.84	78.97	72.50	88.01	86.85	80.05
AOPG [24]	ResNet50	12	87.60	87.65	71.66	68.69	82.31	77.32	73.10	80.66
DODet [56]	ResNet50	12	87.12	87.80	70.50	71.54	82.06	77.43	45.54	74.47
Oriented R-CNN [23]	ResNet50	12	86.68	87.73	72.21	70.80	82.42	78.18	74.11	80.87
OIBBR [57]	ResNet50	12	86.45	88.04	70.88	71.72	82.99	80.55	73.19	81.00
SCTFL (Ours)	ResNet50	12	87.28	87.39	71.25	72.04	83.35	81.98	74.20	**81.09**

4.5.2. Evaluation in Densely Populated Scenes

To evaluate the generalizability of our method in densely populated scenes, it is compared with popular OOD methods in the four densely populated scenes of two RSI datasets, respectively. The four densely populated scenes of the DOTA dataset include SV, LV, SH and TC, and scenes of the DIOR-R dataset include APL, SH, TC and VE. As shown in Tables 5 and 6, our method has the highest mAP in the four densely populated scenes of the two RSI datasets, which indicates that our method has better generalizability in densely populated scenes.

Table 5. Comparisons with 14 OOD methods in the densely populated scenes of the DOTA dataset. The AP of each scene and the mAP of 4 scenes are listed here.

Methods	SV	LV	SH	TC	mAP
SCRDet [66]	68.36	60.32	72.41	90.85	72.98
TCD [68]	65.75	76.76	70.67	90.88	76.01
CenterMap-Net [65]	78.62	66.55	78.10	88.83	78.02
Faster RCNN [31]	73.25	71.49	77.11	90.84	78.17
Retina-O [60]	74.58	71.64	79.11	90.29	78.90
DRN [11]	73.48	70.69	84.94	90.14	79.81
ROI Transformer [55]	77.93	76.67	86.87	90.71	83.04
S^2A-Net [13]	78.11	78.39	87.25	90.83	83.64
AOPG [24]	73.51	82.31	87.95	90.89	83.66
Oriented RepPoints [12]	80.18	78.40	87.28	90.90	84.19
RPGAOD [69]	78.09	81.44	87.64	90.83	84.50
DODet [56]	79.04	82.86	88.15	90.90	85.23
Oriented R-CNN [23]	78.93	83.00	88.20	90.90	85.25
OIBBR [57]	78.93	83.08	88.29	90.89	85.29
SCTFL (Ours)	78.61	83.66	88.27	90.90	**85.36**

Table 6. Comparisons with 12 OOD methods in the densely populated scenes of the DIOR-R dataset. The AP of each scene and the mAP of 4 scenes are listed here.

Methods	APL	SH	TC	VE	mAP
FCOS-O [70]	48.70	73.09	77.55	30.69	57.50
RetinaNet-O [60]	61.49	77.71	81.46	38.01	64.66
ROI Transformer [55]	63.34	81.21	81.61	43.27	67.35
Double-heads [71]	62.13	80.36	87.09	41.32	67.72
Gliding Vertex [65]	65.35	80.22	81.49	47.71	68.69
Oriented R-CNN [23]	62.00	81.25	88.62	43.27	68.78
Faster RCNN-O [31]	62.79	81.14	81.44	50.46	68.95
OIBBR [57]	63.22	81.23	89.05	43.38	69.22
AOPG [24]	62.39	81.17	81.49	52.38	69.35
S^2A-Net [13]	65.40	82.33	84.61	48.00	70.08
DODet [56]	63.40	85.41	81.52	51.86	70.54
TCD [68]	67.95	85.89	81.61	48.72	71.04
SCTFL (Ours)	69.45	89.23	88.54	53.19	**75.10**

In addition, as shown in Tables 2–4, the overall performance of our method is the best in 29 object categories of the DOTA and DIOR-R datasets (a total of 35 categories eliminate 6 duplicated categories), which demonstrates that our method has excellent generalizability in different object categories, and it can validate certain matters to some extent, i.e., our method can provide good performance when novel object categories are introduced.

4.6. Subjective Evaluation

To intuitively evaluate the effectiveness of our method, some its detection results on the two benchmarks are visualized in Figures 5 and 6, respectively.

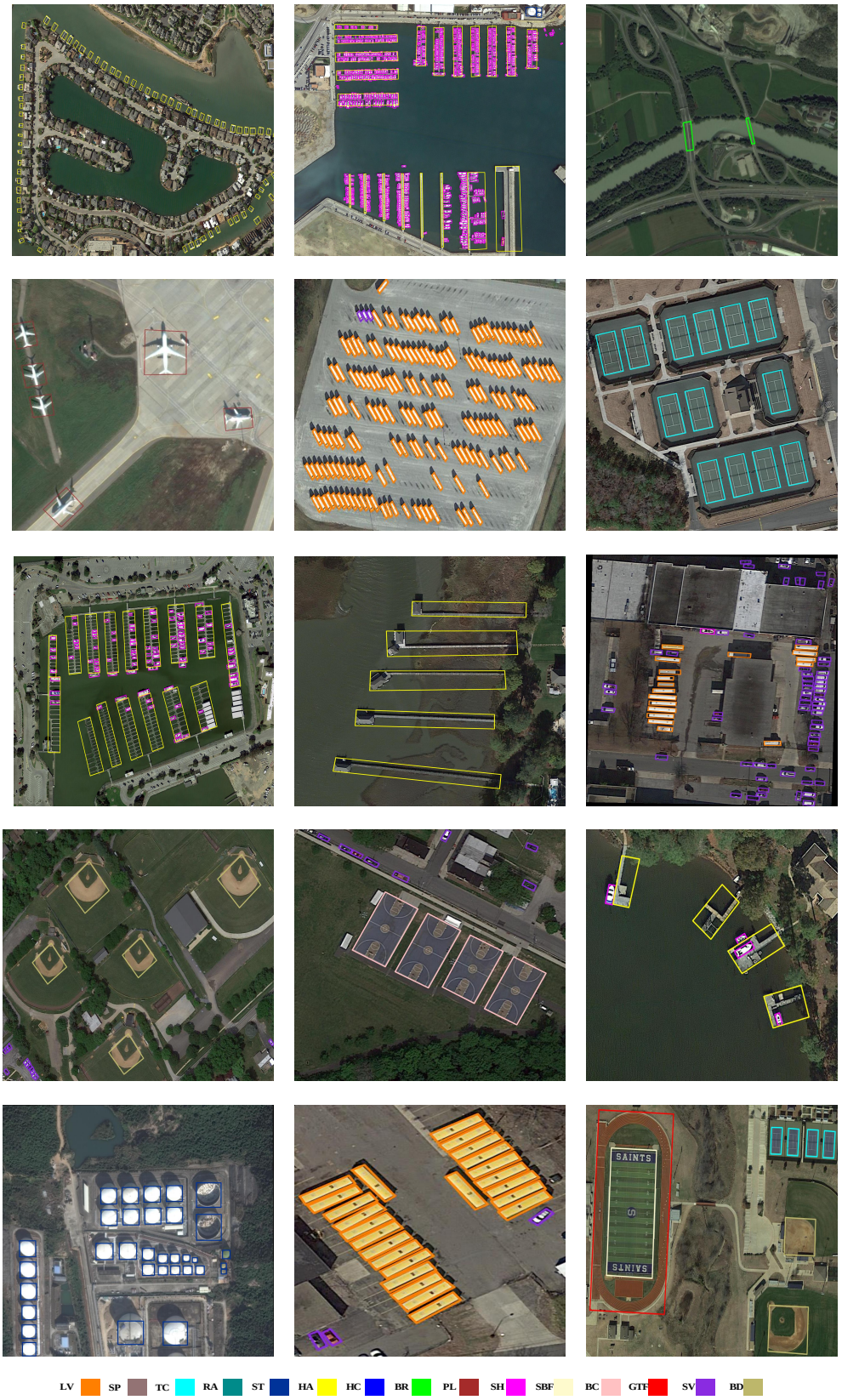

Figure 5. Visualizations of detection results on the DOTA dataset.

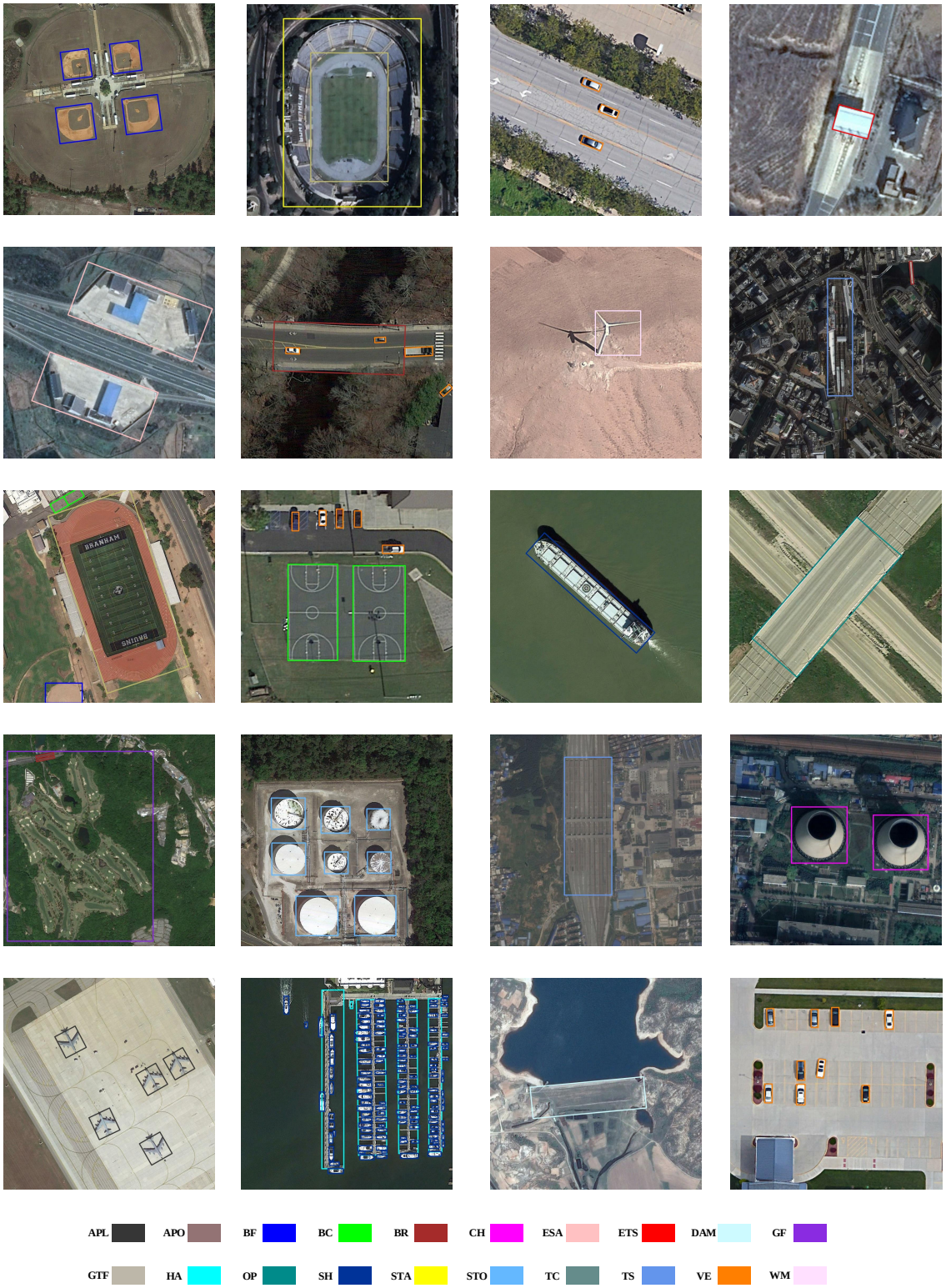

Figure 6. Visualization of detection results on the DIOR-R dataset.

5. Discussion

5.1. Analysis of Experiment Results

As shown in Table 2 (Table 3), our method achieved the best performances in 4 (9) categories, sub-optimal performances in 2 (3) categories and the third best performances in 1 (2) categories on the DOTA (DIOR-R) datasets, respectively. This is attributed to the

effectiveness of the proposed scene classification branch and task-aligned focal loss. For example, as shown in Table 3, our method gives the best performance in the ship category. A possible reason for this is that our method can exclude the misclassification of ships with the help of the scene classification branch since ships are closely related to a certain scene, e.g., sea. For another example, most of the OOD methods give poor performance in bridge and vehicle categories; however, our method shows the best performance in the above two categories. A possible reason for this is that the proposed task-aligned focal loss can assign larger weights to hard instances and enhance the consistency of classification and regression tasks. Furthermore, as shown in Tables 2 and 3, the mAP of all categories of our method is the highest among all OOD methods, which also demonstrates the effectiveness of the scene classification branch and task-aligned focal loss.

5.2. *Analysis of Failure Results*

Although the proposed method provides better performance compared with other OOD methods, it still obtains failure results on some occasions, and the reasons are summarized below.

1. Caused by inter-class similarity. It is difficult for OOD methods to distinguish objects with similar characteristics but are of different categories because of the inter-class similarity of RSI, and our method also cannot handle it if similar objects exist in one RSI simultaneously. As a matter of fact, the scene classification branch introduced by our method can only exclude the impossible categories that do not exist in the input RSI. As shown in Figure 7a, the large vehicles are mistakenly detected as small vehicles because of their similar characteristics.
2. Caused by shadows. The shadows of foreground objects have similar shapes to the foreground objects themselves; therefore, on some occasions they are mistakenly detected as a part of the foreground objects. As shown in Figure 7b, the shadows of a chimney and a windmill are mistakenly recognized as a part of the targets.

(**a**) (**b**)

Figure 7. Illustration of some failure results of our method. (**a**) failure results caused by inter-class similarity on the DOTA dataset. (**b**) failure results caused by shadows on the DIOR-R dataset. The meaning of colors of bounding boxes in (**a**) and (**b**) refers to Figure 5 and Figure 6, respectively.

6. Conclusions

This paper proposes an OOD model combining scene classification to handle the misclassification problem caused by only relying on the features of each proposal itself. The scene classification branch is added into the baseline OOD model, and the scene classification results are employed to help the object classification branch to exclude impossible categories in the RSI. In addition, a task-aligned focal loss is proposed to focus on hard instances and enhance the consistency between classification and regression, which combines the instance difficulty with regression loss to increase the relative weight of hard instance and optimize the classification and regression branches simultaneously. Ablation experiments show the effectiveness of the scene classification branch, task-aligned focal loss and their fusion. Comparisons with 15 and 14 OOD methods on the DOTA and DIOR-R datasets demonstrate the excellent performance of our method.

Author Contributions: Conceptualization, X.Q.; formal analysis, X.Q.; funding acquisition, X.Q.; methodology, X.Q. and S.G.; project administration, W.W.; resources, W.W.; software, S.G.; supervision, W.W.; validation, W.W. and W.D.; writing—original draft, S.G.; writing—review and editing, X.Q. and W.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (Grant No. 62076223), the Key Research Project of Henan Province Universities (Grant No. 24ZX005) and the Key Science and Technology Program of Henan Province (Grant No. 232102211018).

Data Availability Statement: The DOTA and DIOR-R datasets are available at the following URLs: https://captain-whu.github.io/DOTA/index.html (accessed on 15 March 2023) and https://gcheng-nwpu.github.io/ (accessed on 15 March 2023), respectively.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OOD	Oriented Object Detection
RSI	Remote Sensing Image
SCB	Scene Classification Branch
TFL	Task-aligned Focal Loss

References

1. Yao, X.; Shen, H.; Feng, X.; Cheng, G.; Han, J. R2IPoints: Pursuing Rotation-Insensitive Point Representation for Aerial Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5517010. [CrossRef]
2. Wu, X.; Hong, D.; Chanussot, J. Convolutional Neural Networks for Multimodal Remote Sensing Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5517010. [CrossRef]
3. Hong, D.; Yokoya, N.; Chanussot, J.; Zhu, X.X. An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing. *IEEE Trans. Image Process.* **2019**, *28*, 1923–1938. [CrossRef] [PubMed]
4. Zhang, L.; Zhang, Y.; Yan, H.; Gao, Y.; Wei, W. Salient object detection in hyperspectral imagery using multi-scale spectral-spatial gradient. *Neurocomputing* **2018**, *291*, 215–225. [CrossRef]
5. Liao, W.; Bellens, R.; Pizurica, A.; Philips, W.; Pi, Y. Classification of Hyperspectral Data Over Urban Areas Using Directional Morphological Profiles and Semi-Supervised Feature Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1177–1190. [CrossRef]
6. Gao, L.; Zhao, B.; Jia, X.; Liao, W.; Zhang, B. Optimized Kernel Minimum Noise Fraction Transformation for Hyperspectral Image Classification. *Remote Sens.* **2017**, *9*, 548. [CrossRef]
7. Du, L.; You, X.; Li, K.; Meng, L.; Cheng, G.; Xiong, L.; Wang, G. Multi-modal deep learning for landform recognition. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 63–75. [CrossRef]
8. Zhang, L.; Shi, Z.; Wu, J. A Hierarchical Oil Tank Detector With Deep Surrounding Features for High-Resolution Optical Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4895–4909. [CrossRef]
9. Stankov, K.; He, D.C. Detection of Buildings in Multispectral Very High Spatial Resolution Images Using the Percentage Occupancy Hit-or-Miss Transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4069–4080. [CrossRef]
10. Han, X.; Zhong, Y.; Zhang, L. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]

11. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic refinement network for oriented and densely packed object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11207–11216.
12. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1829–1838. [CrossRef]
13. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [CrossRef]
14. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic anchor learning for arbitrary-oriented object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volumr 35; pp. 2355–2363.
15. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271. [CrossRef]
16. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767/
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I 14*; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
18. Liu, W.; Ma, L.; Wang, J.; xsChen, H. Detection of Multiclass Objects in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 791–795. [CrossRef]
19. Chen, S.; Zhan, R.; Zhang, J. Geospatial Object Detection in Remote Sensing Imagery Based on Multiscale Single-Shot Detector with Activated Semantics. *Remote Sens.* **2018**, *10*, 820. [CrossRef]
20. Tang, T.; Zhou, S.; Deng, Z.; Lei, L.; Zou, H. Arbitrary-Oriented Vehicle Detection in Aerial Imagery with Single Convolutional Neural Networks. *Remote Sens.* **2017**, *9*, 1170. [CrossRef]
21. Tayara, H.; Chong, K.T. Object Detection in Very High-Resolution Aerial Images Using One-Stage Densely Connected Feature Pyramid Network. *Sensors* **2018**, *18*, 3341. [CrossRef] [PubMed]
22. Chen, Z.; Zhang, T.; Ouyang, C. End-to-End Airplane Detection Using Transfer Learning in Remote Sensing Images. *Remote Sens.* **2018**, *10*, 139. [CrossRef]
23. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3520–3529. [CrossRef]
24. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625411. [CrossRef]
25. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2016; Volume 29.
26. Zhong, Y.; Han, X.; Zhang, L. Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 281–294. [CrossRef]
27. Chen, C.; Gong, W.; Chen, Y.; Li, W. Object Detection in Remote Sensing Images Based on a Scene-Contextual Feature Pyramid Network. *Remote Sens.* **2019**, *11*, 548. [CrossRef]
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef]
29. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July, 2017; pp. 2117–2125.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, Las Vegas, NV, USA, 27–30 June 2016 ; pp. 770–778.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
32. Qian, X.; Li, J.; Cao, J.; Wu, Y.; Wang, W. Micro-cracks detection of solar cells surface via combining short-term and long-term deep features. *Neural Netw.* **2020**, *127*, 132–140. [CrossRef] [PubMed]
33. Qian, X.; Lin, S.; Cheng, G.; Yao, X.; Ren, H.; Wang, W. Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion. *Remote Sens.* **2020**, *12*, 143–163. [CrossRef]
34. Qian, X.; Huo, Y.; Cheng, G.; Yao, X.; Li, K.; Ren, H.; Wang, W. Incorporating the Completeness and Difficulty of Proposals Into Weakly Supervised Object Detection in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1902–1911. [CrossRef]
35. Li, L.; Yao, X.; Wang, X.; Hong, D.; Cheng, G.; Han, J. Robust few-shot aerial image object detection via unbiased proposals filtration. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5617011. [CrossRef]
36. Qian, X.; Wang, C.; Li, C.; Li, Z.; Zeng, L.; Wang, W.; Wu, Q. Multi-Scale Image Splitting Based Feature Enhancement and Instance Difficulty Aware Training for Weakly Supervised Object Detection in Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7497–7506. [CrossRef]
37. Qian, X.; Huo, Y.; Cheng, G.; Gao, C.; Yao, X.; Wang, W. Mining High-quality Pseudo Instance Soft Labels for Weakly Supervised Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5607615. [CrossRef]

38. Xie, X.; Cheng, G.; Feng, X.; Yao, X.; Qian, X.; Han, J. Attention Erasing and Instance Sampling for Weakly Supervised Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5600910. [CrossRef]
39. Qian, X.; Li, C.; Wang, W.; Yao, X.; Cheng, G. Semantic segmentation guided pseudo label mining and instance re-detection for weakly supervised object detection in remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *119*, 103301. [CrossRef]
40. Xie, X.; Lang, C.; Miao, S.; Cheng, G.; Li, K.; Han, J. Mutual-assistance learning for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 15171–15184. [CrossRef]
41. Zheng, X.; Cui, H.; Lu, X. Multiple Source Domain Adaptation for Multiple Object Tracking in Satellite Video. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5626911. [CrossRef]
42. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5606514. [CrossRef]
43. Ming, Q.; Miao, L.; Zhou, Z.; Dong, Y. CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5605814. [CrossRef]
44. Qian, X.; Zeng, Y.; Wang, W.; Zhang, Q. Co-saliency detection guided by group weakly supervised learning. *IEEE Trans. Multimed.* **2022**, *25*, 1810–1818. [CrossRef]
45. Wang, J.; Li, F.; Bi, H. Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4707013. [CrossRef]
46. Li, Z.; Hou, B.; Wu, Z.; Ren, B.; Ren, Z.; Jiao, L. Gaussian synthesis for high-precision location in oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5619612. [CrossRef]
47. Zhang, C.; Su, J.; Ju, Y.; Lam, K.M.; Wang, Q. Efficient inductive vision transformer for oriented object detection in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5616320. [CrossRef]
48. Cheng, G.; Li, Q.; Wang, G.; Xie, X.; Min, L.; Han, J. SFRNet: Fine-Grained Oriented Object Recognition via Separate Feature Refinement. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610510. [CrossRef]
49. Qian, X.; Zhang, N.; Wang, W. Smooth giou loss for oriented object detection in remote sensing images. *Remote Sens.* **2023**, *15*, 1259. [CrossRef]
50. Qian, X.; Wu, B.; Cheng, G.; Yao, X.; Wang, W.; Han, J. Building a bridge of bounding box regression between oriented and horizontal object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605209. [CrossRef]
51. Zheng, X.; Cui, H.; Xu, C.; Lu, X. Dual Teacher: A Semisupervised Cotraining Framework for Cross-Domain Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5613312. [CrossRef]
52. Xie, X.; Cheng, G.; Li, Q.; Miao, S.; Li, K.; Han, J. Fewer is more: Efficient object detection in large aerial images. *Science China Inf. Sci.* **2024**, *67*, 112106. [CrossRef]
53. Zheng, X.; Chen, X.; Lu, X. Visible-Infrared Person Re-Identification via Partially Interactive Collaboration. *IEEE Trans. Image Process.* **2022**, *31*, 6951–6963. [CrossRef] [PubMed]
54. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35; pp. 3163–3171.
55. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.
56. Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; Han, J. Dual-aligned oriented detector. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618111. [CrossRef]
57. Yao, Y.; Cheng, G.; Wang, G.; Li, S.; Zhou, P.; Xie, X.; Han, J. On improving bounding box representations for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5600111. [CrossRef]
58. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
59. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
60. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
61. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th international conference on pattern recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; IEEE: New York, NY, USA, 2006; Volume 3, pp. 850–855. [CrossRef]
62. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hong Kong, China, 20–24 August 2018; pp. 3974–3983.
63. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
64. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
65. Wang, J.; Yang, W.; Li, H.C.; Zhang, H.; Xia, G.S. Learning center probability map for detecting objects in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4307–4323. [CrossRef]

66. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241. [CrossRef]
67. Xu, C.; Ding, J.; Wang, J.; Yang, W.; Yu, H.; Yu, L.; Xia, G.S. Dynamic Coarse-to-Fine Learning for Oriented Tiny Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7318–7328.
68. Zhang, C.; Xiong, B.; Li, X.; Kuang, G. TCD: Task-collaborated detector for oriented objects in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4700714. [CrossRef]
69. Qiao, Y.; Miao, L.; Zhou, Z.; Ming, Q. A Novel Object Detector Based on High-quality Rotation Proposal Generation and Adaptive Angle Optimization. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5617715. [CrossRef]
70. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
71. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June, 2020; pp. 10186–10195. [CrossRef]