remote sensing

MDPI

Article

Semantic-Edge-Supervised Single-Stage Detector for Oriented Object Detection in Remote Sensing Imagery

Dujuan Cao , Changming Zhu *, Xinxin Hu and Rigui Zhou

College of Information and Engineering, Shanghai Maritime University, Shanghai 201306, China;
202030310126@stu.shmtu.edu.cn (D.C.); 202130310147@stu.shmtu.edu.cn (X.H.); rgzhou@shmtu.edu.cn (R.Z.)
* Correspondence: cmzhu@shmtu.edu.cn

Abstract: In recent years, significant progress has been made in arbitrary-oriented object detection. Different from natural images, object detection in aerial images remains its problems and challenges. Current feature enhancement strategies in this field mainly focus on enhancing the local critical response of the target while ignoring the target's contextual information, which is indispensable for detecting remote sensing targets in complex backgrounds. In this paper, we innovatively combine semantic edge detection with arbitrary-oriented object detection and propose a feature enhancement network base on a semantic edge supervision module (SES) that realizes an attention-like mechanism in three dimensions of space, channel, and pyramid level. It helps the network pay attention to the edge features of targets at multiple scales to obtain more regression clues. Furthermore, to solve the problem of dense objects with different directions in remote sensing images, we propose a rotation-invariant spatial pooling pyramid (RISPP) to extract the features of objects from multiple orientations. Based on the two feature enhancement modules, we named the network SE^2-Det; extensive experiments on large public datasets of aerial images (DOTA and UCAS-AOD) validate our approach's effectiveness and demonstrate our detector's superior performance.

Keywords: aerial images; oriented object detection; feature enhancement; convolutional neural network

check for updates

Citation: Cao, D.; Zhu, C.; Hu, X; Zhou, R. Semantic-Edge-Supervised Single-Stage Detector for Oriented Object Detection in Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 3637. https://doi.org/10.3390/rs14153637

Academic Editor: Pedro Melo-Pinto

Received: 10 June 2022
Accepted: 27 July 2022
Published: 29 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid development of deep neural networks, more and more excellent object detection algorithms based on convolutional neural networks have been proposed. The generic object detection mainly utilizes the horizontal detection box algorithm, which is divided into two-stage detectors [1–5] and one-stage detectors [6–9] according to the algorithm detection process.

However, current horizontal detectors still have many limitations for some practical scenarios, such as remote sensing image analysis. The oriented object detection task in remote sensing image analysis aims to detect and classify objects in remote sensing images, such as planes, ships and bridges. Because the remote sensing image is taken from the top view, the target to be detected has different directions. Applying a horizontal detector to detect these objects will cause the area covered by boxes to contain a large amount of redundant background, further leading to problems such as overlapping multiple detection boxes. In order to locate the target more accurately, this task needs to predict a set of oriented bounding boxes (obbs). The oriented object detection of remote sensing image analysis is applied in many military and civilian fields, such as topographic survey and emergency rescue. Still, in addition to the characteristics of arbitrary direction mentioned, the task has challenges such as small and dense, and large aspect ratios.

The current state-of-the-art methods for oriented object detection solve the domain-specific problems proposed above from multiple perspectives. They can be summarized in four directions: feature enhancement and alignment methods for dense and high aspect ratio remote sensing targets [10,11], denoising and attention modules for complex backgrounds [12,13], specific sampling strategies for overhead view target [14,15], and novel angle representations and loss functions [16,17] for field-unique regression issues. Our work focuses on the first and second directions mentioned above to improve single-stage oriented object detectors by introducing a novel attention-like mechanism network and rotation-invariant feature enhancement module to tackle different challenges in remote sensing image analysis.

Existing representative methods of feature enhancement and attention mechanism in remote sensing image analysis can be divided into the following three aspects: using effective attention mechanisms lets the network focus on targets' salient features, which solves the noise and boundary blur problem [14,18]; using the task-aware feature activation method enhances the decoupled detection head, which solves the feature incompatibility problem [18]; using more robust convolution extracts rotation-sensitive features, which solve the problem of the arbitrary orientation of targets [11,19].

Nevertheless, it is found that there are three significant shortcomings in these methods. Firstly, few detectors manage to pay attention to the contextual features of remote sensing targets, which are pretty crucial according to prior knowledge [20] (see the first row of Figure 1). Secondly, most attention modules are inadequately supervised, so they cannot directly guide the network (see the second row of Figure 1). Besides, the rotation-invariant features, which are essential to solve the problem of targets with various orientations [21], are indirectly extracted by pooling rotation-sensitive features.

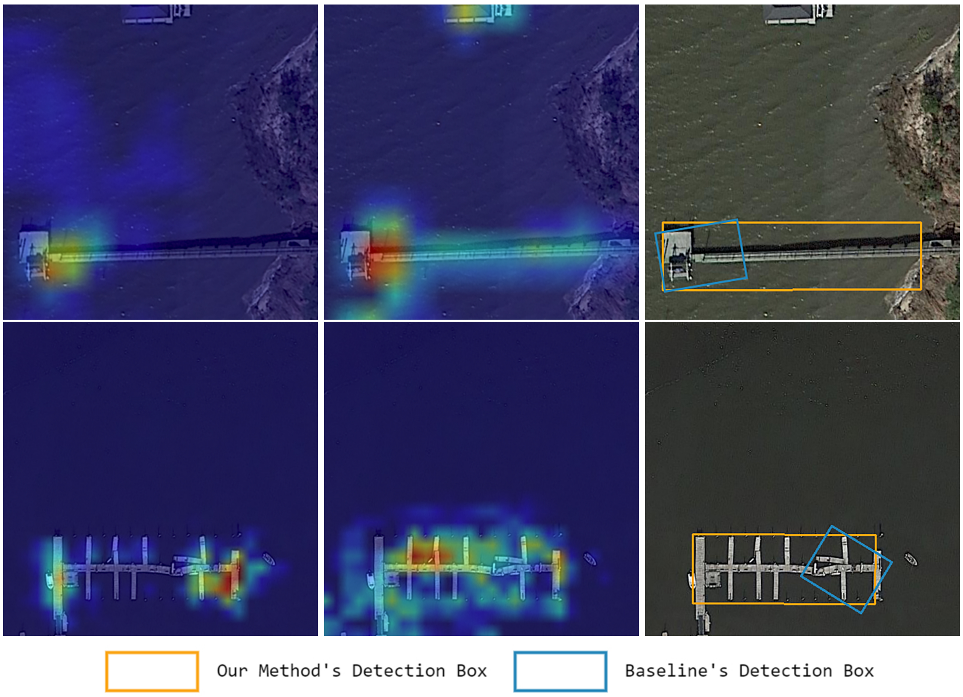

Figure 1. Baseline feature map (**left**), feature map of our method (**middle**), and detection result (**right**). The first line: baseline only focuses on the local key features of the target with a large aspect ratio, so the final regression box cannot completely contain the entire target. Our method focuses on the target global edge features and provides rich contextual information to obtain accurate detection boxes. The second line: our method implements an attention-like mechanism in the form of strong supervision to guide the network to learn the global features of objects with different shapes in a targeted manner and obtain accurate regression boxes.

To better pay attention to the context information of remote sensing objects, we introduce semantic edge detection into arbitrary-oriented object detection for the first time. In this way, we can instruct the network to learn the targets' edge features purposefully.

After careful analysis of the difference between remote sensing object detection and conventional object detection, it is found that the remote sensing scene has a very high degree of fit with edge detection. Objects in natural images are easily occluded from a heads-up view, resulting in overlapping edges and bounding boxes between different objects. CASENet [22] proposed a multi-label loss function for this issue, allowing a pixel to belong to multiple categories. Combining semantic edge detection with general object detection is still very challenging. While compared to the remote sensing targets in the top-down view, the edges of both inter-class and intra-class targets are separated and do not overlap. Therefore, semantic edge detection can better help arbitrary-oriented object detection from a bird's-eye view.

In this paper, we propose a novel feature enhancement network base on a semantic edge supervision module (SES), which aims to guide the network to pay attention to the contextual features of the target from complex backgrounds, thereby giving the network more regression clues. Besides benefiting from its unique semantic characteristics compared with binary edge detection, the proposed semantic edge supervision can achieve the effect of an attention-like mechanism, which guides the network more pertinently in the channel, space, and pyramid levels. Different from general attention mechanisms [23–25], its advantage is to help the network learn the context information of different instances among the same category in the spatial dimension. The edge maps of different categories can be decoupled in the channel dimension. Meanwhile, it can enrich the communication between the features among different scales in hierarchical dimensions and guide the network to learn the importance of different categories at each pyramid level. In general, our network suppresses redundant backgrounds and highlights the edge response of the target in the form of strong supervision. In this way, the network can accurately and completely pay attention to the remote sensing target to be detected. More detailed experimental results can be found in Section 3.2.2.

Based on the above analysis, we innovatively guide the network to focus on the edge features of the target to further extract the features required for regression. However, the key features of the target are also indispensable, which play a decisive role in the localization and classification of the target [18,26]. In remote sensing, the target to be detected has the characteristics of changing directions. The key features should also have rotation invariance, so our goal is to extract the rotation-invariant features. We propose a rotation-invariant spatial pooling pyramid (RISPP) based on the spatial pooling pyramid proposed in [27] to enhance the rotation-invariant feature of the target. This module focuses on the fusion of features under multiple rotation angles to obtain more comprehensive rotation-invariant features. It is worth noting that the rotation-invariant features of the target are often located at the center of the target, so the rotation-invariant feature enhancement method can also focus on the center response of the target to further assist in locating the center of the target. Finally, our work enhances the comprehensive features suitable for remote sensing objects from two aspects, the SES module and the RISPP module. Our proposed architecture and method achieve competitive performance with state-of-the-art approaches on two public remote sensing datasets, DOTA [28] and UCAS-AOD [29]. The contributions of this paper are as follows:

- We creatively introduce semantic edge detection technology into oriented object detection and propose a semantic-edge-supervision feature enhancement network (SE^2-Det) based on the core ideas in its field. This module guides the network in a strong supervision form in the channel, space, and pyramid level dimensions. It also effectively solves the problems of complex background and lack of context regression clues in remote sensing object detection.
- In order to more effectively and comprehensively extract the rotation-invariant features of remote sensing targets in any direction, we propose a rotation-invariant spatial pooling pyramid (RISPP) and achieve remarkable results.

The rest of this paper is organized as below. Section 2 introduces the work related to this paper from three aspects: general object detection, arbitrary orientation object detection in remote sensing, and semantic edge detection. It also introduces our method in detail. Section 3 describes the datasets used in the experiment and demonstrates all the ablation experiments related to our innovations, as well as the excellent performance of our method on different datasets. Section 4 objectively discusses the limitations and future direction of the proposed network. Conclusions are given in Section 5.

2. Material and Methods

2.1. Related Work

This section reviews some excellent object detection frameworks in two directions. First, the generic target detection method based on a horizontal bounding box (hbb) has made remarkable achievements in natural image object detection. Secondly, we introduce several object detection methods based on an oriented bounding box (obb), which are directly related to our paper. Finally, since we introduce edge detection to remote sensing image analysis for the first time, we also introduce the related content of semantic edge detection. In particular, some efforts have attempted to have edge detection assist their tasks, such as semantically segmented domains.

2.1.1. Generic Object Detection

More and more high-performance object detection algorithms have been proposed in recent years. Horizontal detection algorithms are primarily used in conventional object detection tasks and are divided into two-stage and one-stage detectors. RCNN series of algorithms can be called the classic two-stage detectors. RCNN [1] first introduces a convolutional neural network into two-stage detection. On its basis, many improved algorithms have also been proposed, such as Fast RCNN [2], Faster RCNN [3], Mask RCNN [4], and Cascade RCNN [5]. Two-stage detectors are generally accurate but slow due to the presence of the two-stage.

One-stage detectors do not generate proposals before classification and regression and are widely used because of their fast speed. YOLO [6] series and SSD [9] series are representatives of one-stage, but the accuracy of the one-stage detector is slightly lower than the two-stage detector. RetinaNet [8] introduces the focal loss function for the class imbalance problem, which combines the one-stage detector's speed advantage and the two-stage detector's accuracy advantage.

2.1.2. Arbitrary-Oriented Object Detection in Aerial Images

Horizontal detectors have limitations in practical applications, such as detecting small and dense remote sensing targets with arbitrary directions and large scales. In order to adapt to these tasks, more and more oriented remote sensing detectors for different aspects are proposed. Because of targets' small and dense characteristics in complex backgrounds, some detectors achieve the effect of denoising by improving the attention mechanism. Unlike the commonly used spatial and channel attention [23–25], SCRDet [12] proposes a supervised multi-dimensional attention mechanism. Nevertheless, it does not consider the interaction between intra-class and inter-class objects, which is especially important for complex scenes. Based on this, SCRDet++ [13] uses mask supervision to guide features at multiple scales and act on the classification score, realizing inter-class decoupling and intra-class denoising. However, the above methods do not directly provide regression feature clues for regression tasks. CFC-Net [18] adopts the idea of decoupling to solve the problem of incompatible classification and regression features. Although the polarization function obtains more regression clues in the regression head, the regression clues lack pertinence and stability, and it is not easy to obtain the global context information of the target. S2A-Net [11] uses ARF [30] to extract rotation-sensitive features, which solves the problem of arbitrary orientation to a certain extent. Nevertheless, it will generate additional orientation channels, significantly impacting the number of model parameters and inference speed.

2.1.3. Semantic Edge Detection

With the continuous advancement of deep neural networks, edge detection has also evolved from simple low-level filtered edge detection, such as Sobel [31] or Canny [32], to deep binary or semantic edge detection. The most relevant work of this paper is semantic edge detection. CASENet [22] solves the problem of edge classification through a multi-label learning framework and uses a new nested architecture to fully let the low-level features assist the high-level features in semantic classification. DFF [33] improves the limitations of the fixed weight fusion model through a new dynamic feature fusion strategy. In semantic segmentation, some people adopt multi-task learning methods to improve the segmentation performance by using the auxiliary edge detection information. The semantic segmentation network Gated-SCNN [34] adopts a two-stream structure, and the gating mechanism of the shape branch uses edge information to improve the segmentation effect of small structure objects. Different from Gated-SCNN assisting semantic segmentation through binary edge detection, RPCNet [35] first combines semantic edge detection and semantic segmentation tasks and improves the boundary pixel accuracy of semantic segmentation by constraining the loss through edge consistency loss. In order to help salient object detection, EGNet [36] models the salient edge and target information and uses the complementarity between the two tasks to optimize the two tasks jointly. We introduce semantic edge detection into remote sensing detection for the first time, which is fundamentally different from the above methods.

2.2. Methods

2.2.1. Overall Pipeline

The proposed SE^2-Det is built based on the basic single-stage detector. The overall framework is shown in Figure 2. It consists of four parts: feature extraction backbone, contextual feature enhancement module based on the SES module, RISPP module and detection head for classification and regression tasks.

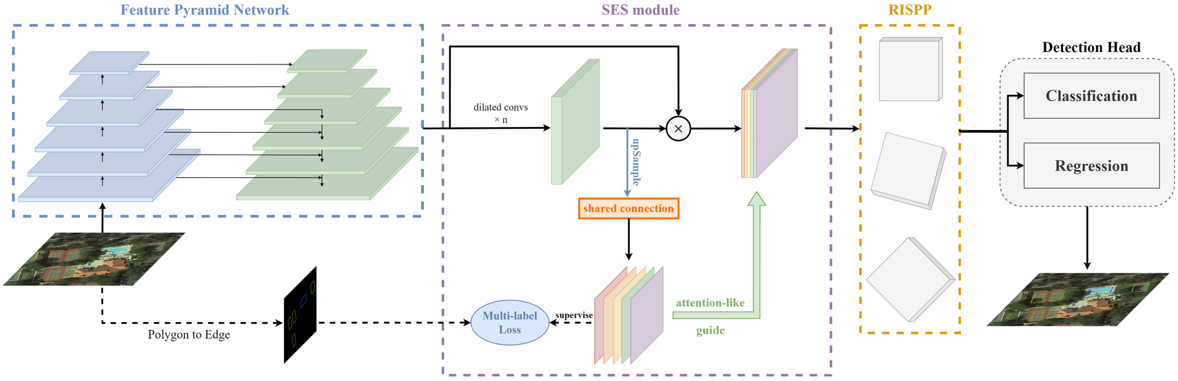

Figure 2. The overall architecture of the SE^2-Det. Take an RGB image as input and output, where the output image has OBB prediction boxes. The backbone (ResNet [37] + FPN [38]) extracts common features; then, contextual edge features and rotation-invariant features are extracted through SES module and RISPP module, respectively. Finally, the features of the two branches are fused and sent to the classification head and the regression head to obtain the OBB prediction result.

The specific process of the framework is as follows: firstly, we use the currently common ResNet-based feature pyramid network (FPN) as the feature extraction network to extract multi-scale features from RGB aerial images. On the one hand, the ResNet network uses skip connection, which combines shallow and deep semantic features well. On the other hand, the feature pyramid network takes full advantage of different scale information to overcome the challenges of target prediction with different sizes. Therefore, the ResNet-based feature pyramid network combines shallow and deep feature fusion and multiresolution prediction to better deal with large-scale changes and small and dense problems of object detection for remote sensing. For the extracted multi-scale features, we first feed them into the proposed SES module to guide the network to learn the edge and context features of the target and decouple the different inter-class and intra-class instances. Then, to further supplement the critical features of remote sensing targets, we feed the features into the proposed RISPP module to extract the rotation-invariant features of the target. So far, we have strengthened the target feature's center and edge response through the above two modules and obtained a rich and accurate feature description. Finally, we send it to the detection head for classification and regression, and the output is an RGB image with OBB prediction boxes. The following is a detailed introduction to the proposed modules.

2.2.2. Semantic Edge Supervision for Contextual Feature Enhancement

In remote sensing, aerial imagery often has more complex backgrounds than natural images, and the target has the characteristics of large-scale change, high aspect ratio, small and dense. After our observations, basic one-stage detectors tend to have the following three problems:

(i) It is challenging to capture the target's global features and only focus on the critical local features of the target while ignoring the significant target edge and context features, which is very unfavorable for regression, especially those remote sensing targets with high aspect ratios, as shown in the second row of the first column of Figure 3;

(ii) The network invariably pays attention to the apparent boundaries in the image, but it is worth noting that these boundaries are often the background, thus causing a lot of redundant background interference, as shown in the second row of the second column of Figure 3;

(iii) It is challenging to pay attention to small targets in complex backgrounds. Due to the lack of targeted attention, a large amount of background noise drowns out the small targets, as shown in the third column of the second row of Figure 3.

Figure 3 shows the above three problems. The comparison in the first column shows that the baseline only pays attention to some harbor features. Our method not only pays attention to the global boundary features of the harbor but also provides rich contextual information. The second column represents the background boundary interference problem. The bridge to be detected is only located above the lake in the center of the picture. Because the network is sensitive to all boundaries, the baseline pays more attention to the road boundary where the bridge deck extends to the land, which seriously interferes with the detection of the original bridge deck. From the comparison in the third column, in the baseline feature map, many high-response areas are distributed on the white line of the parking area, focusing only on the local features of the vehicle head. After introducing the proposed attention-like mechanism, our method not only has the effect of eliminating noise but also pays attention to the global features of the large vehicle.

In order to deal with the above challenges, various attention mechanisms in the spatial and channel dimensions have been widely proposed [18,23–25,39,40], and achieve significant results in model accuracy performance and feature response. However, we find that there are still some shortcomings. Firstly, these attention mechanisms make the high response of the feature map expand outward from the center area of the target without paying extra attention to the edge features of the inter-class and intra-class targets. Thus, the ability to capture the contextual edge information of the target is relatively weak. Some research [41,42] has demonstrated that the discriminant features required to locate the target are often not wholly distributed on the ground truth; this also confirms the importance of the context edge information of the target. Secondly, they all work through back-propagating from the head of the network during the training phase. However, as the depth of the network deepens, the propagation chain also gradually increases. This long-distance supervision may cause the effect of the attention mechanism to be weakened. Therefore, we call these attention mechanisms general "inadequately supervised" attentions in the domain, and it is challenging to achieve domain-specific effects.

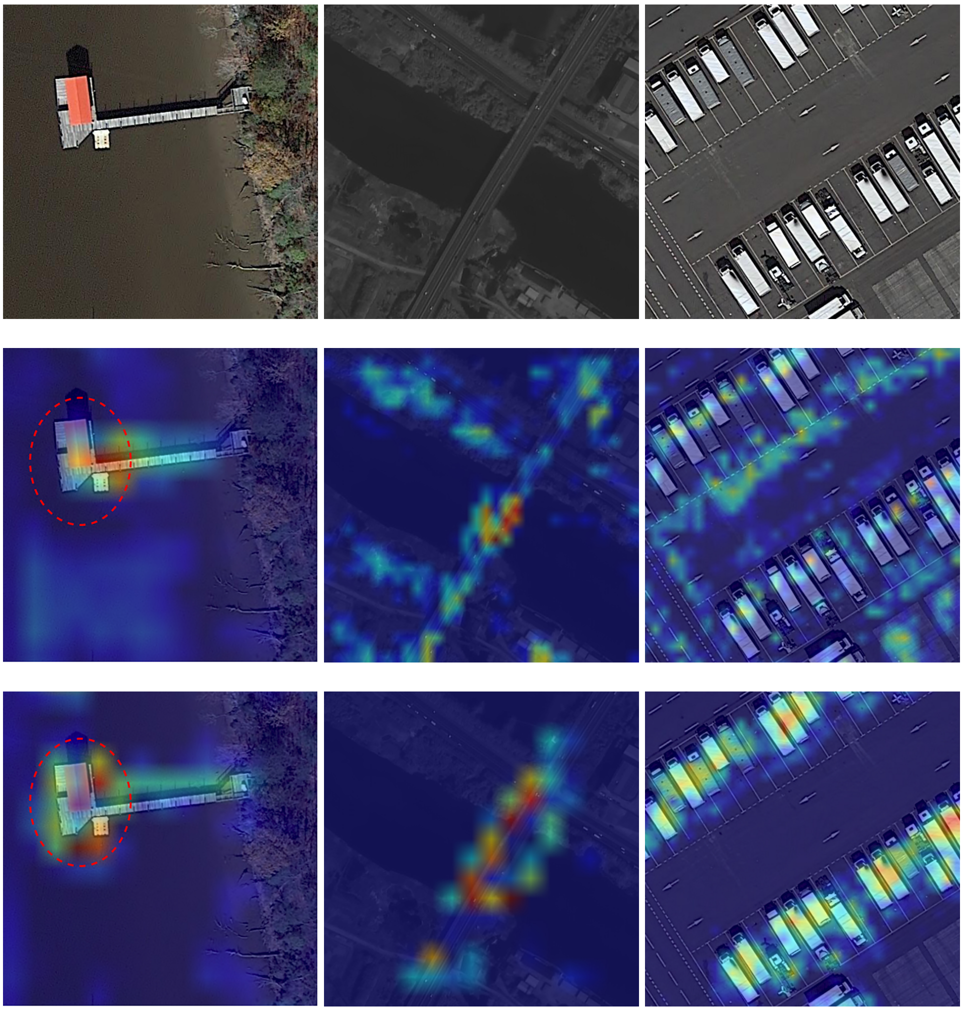

Figure 3. The original image (the first row), the baseline feature map (the second row), and the feature map of our method (the third row). The first column represents the problem of only focusing on local features and ignoring global context information. The second column represents the background boundary interference problem. The third column represents the difficult problem of small object detection with a dense arrangement and different orientations.

Based on the above analysis of domain-specific problems and previous methods, we innovatively combine semantic edge detection with oriented object detection and design a semantic edge supervision network with a high domain fit. This module is fundamentally different from the attention mechanism mentioned above. We solve all the problems with excellent adaptability between semantic edge detection and remote sensing object detection. The following will introduce the module's implementation details and specific processes.

Figure 4a shows the detailed flow of the semantic edge supervision module, in which we use pyramid levels P_2 to P_4 for semantic edge detection, where P_i has $1/2^i$ resolution of the input image. On the one hand, because the receptive field of the shallow feature map is small and the semantic information is weak, the conventional object detection tasks are predicted based on the pyramid level P_3 to P_7. In our work, we also perform edge detection on the P_2 feature map, but considering the computational overhead of the feature map at this scale, we do not send it into the detection head. This advantage is that more edge details can be extracted without increasing the computational overhead of the head. Meanwhile, it can indirectly guide the upper-layer features to suppress background information and provide more accurate edge localization and structural information to retain more contextual details for the final edge activation map. On the other hand, the maximum downsampling stride of the features in edge detection [22,43–45] is eight. It can be inferred that feature maps with too small scales are challenging to restore the detailed features of the edge, so out of the above analysis, we decide to use pyramid levels P_2 to P_4 for semantic edge detection.

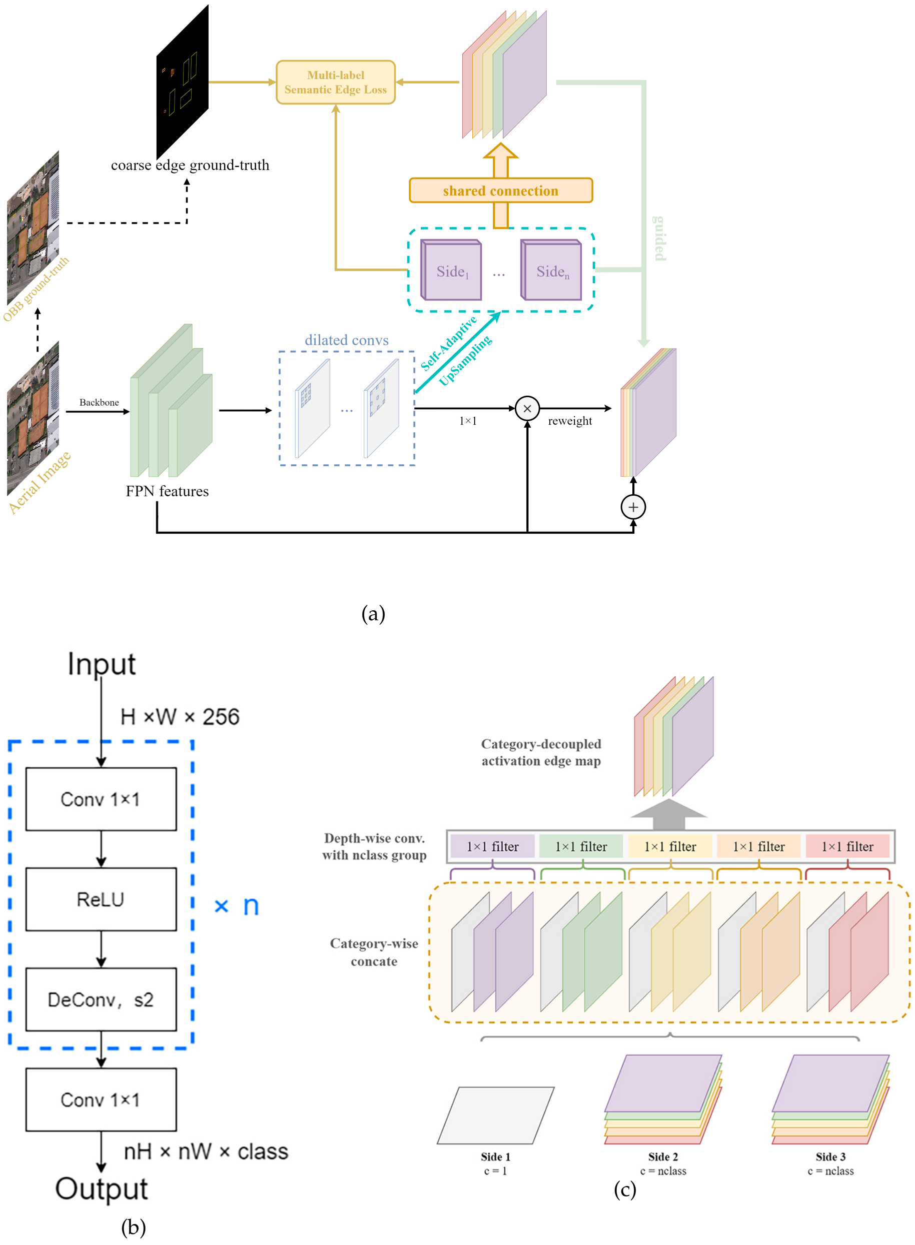

Figure 4. (**a**) Structure of semantic edge supervision module. The dilated convolution module expands the receptive field of feature maps; the self-adaptive upsampling module obtains the semantic edge activation map, and the bypass activation map obtains the fusion activation map via the shared connection operation; the pixel-wise loss is utilised to compute coarse edge ground truth with bypass activation maps and the fused activation map in different strategies. (**b**) Self-adaptive decoder block. (**c**) Shared connection.

The specific implementation details are as follows: we first consider expanding the receptive field for the above stack of pyramid blocks output. Due to the uniqueness of edge detection, a larger receptive field is generally required to cover the target's edge information. Therefore, we first use multiple dilated convolutions to adapt the subsequent edge detection to obtain edge information of different scales. The dilation rate follows the strategy in [43], which also well copes with the challenge of large-scale changes in remote sensing targets. Then, we use two convolution layers to split the feature into two branches.

The first branch obtains the semantic edge activation map through the output of a self-adaptive upsample block (SAU), a new method proposed by us, as shown in Figure 4b. This module aims to optimize the checkerboard effect caused by upsampling using only a single transposed convolution for traditional semantic edge detection [46]. Specifically, for the input feature $\mathrm{F} \in \mathbb{R}^{C \times H \times W}$, we first reduce the channel dimension by using a 1×1 convolution and ReLU activation function. We use the transposed convolution with a stride of 2 for upsampling. So the factor of each upsampling is 2. We repeat the above operation n times until the size of the edge map is consistent with the original image. Finally, we use a 1×1 convolution with the number of channels equal to the number of categories to extract category-aware features. So far, we have obtained edge activation maps of multiple scales.

$$\begin{aligned}
\mathbb{A}_{P_2} &= \mathrm{A}_{\mathrm{P}_2}^{(1)}, \\
\mathbb{A}_{P_3} &= \{\mathrm{A}_{\mathrm{P}_3}^{(1)}, \mathrm{A}_{\mathrm{P}_3}^{(2)}, \ldots, \mathrm{A}_{\mathrm{P}_3}^{(\mathrm{n})}\}, \\
\mathbb{A}_{P_4} &= \{\mathrm{A}_{\mathrm{P}_4}^{(1)}, \mathrm{A}_{\mathrm{P}_4}^{(2)}, \ldots, \mathrm{A}_{\mathrm{P}_4}^{(\mathrm{n})}\},
\end{aligned} \tag{1}$$

where $\mathbb{A}_{P_k}$ represents the edge activation map of the k-th pyramid level obtained by the decoder output, $\mathrm{A}_{\mathrm{P}_\mathrm{k}}^{(\mathrm{n})}$ represents the n-th class channel of the current bypass activation map. For $\mathrm{A}_{\mathrm{P}_2}$, due to its limited receptive field, it is difficult to obtain sufficient semantic information, so this bypass only outputs a single-channel activation map to provide edge details. The number of channels in $\mathrm{A}_{\mathrm{P}_3}$ and $\mathrm{A}_{\mathrm{P}_4}$ is equal to the number of categories in the dataset, so in theory, different instances of each category have been decoupled to different channels. Next, we perform a shared connection operation on $\{\mathrm{A}_{\mathrm{P}_2}, \mathrm{A}_{\mathrm{P}_3}, \mathrm{A}_{\mathrm{P}_4}\}$ (Figure 4c); the module fuses the edge activation maps obtained above at different feature scales and category channels. Specifically, we first copy the single-channel edge map of $\mathrm{A}_{\mathrm{P}_2}$ by n, denoted as F, and then concatenate it with each class of the other two bypasses, which is expressed as follows:

$$\mathrm{A}_{\mathrm{fuesd}} = \mathit{Concat}\left(\mathrm{F}, \mathrm{A}_{\mathrm{P}_3}^{(1)}, \mathrm{A}_{\mathrm{P}_4}^{(1)}, \mathrm{F}, \mathrm{A}_{\mathrm{P}_3}^{(2)}, \mathrm{A}_{\mathrm{P}_4}^{(2)}, \ldots, \mathrm{F}, \mathrm{A}_{\mathrm{P}_3}^{(\mathrm{n})}, \mathrm{A}_{\mathrm{P}_4}^{(\mathrm{n})}\right). \tag{2}$$

Finally, the activation map $\mathrm{A}_{\mathrm{fuesd}}$ with $3n$ channels is obtained by concatenating. Then, through a depth-wise convolution with n groups and the convolution kernel size of 1, we can obtain the fused category-aware edge activation map. The above operations can further help the decoupling between classes and reweighting the feature maps between different pyramid levels, thus allowing the network to learn the importance of each category between different pyramid levels. Finally, the poly-based coarse edge ground truth and all the above activation maps are used to calculate the pixel-wise loss in a deeply supervised manner to guide the network to focus on different instances in each class in the spatial dimension. The poly-based coarse edge ground truth is generated from the annotations of the original dataset. In order to further explain the similarities and differences between the role of the proposed attention-like mechanism and the conventional attention method, we also list the attention mechanism in [24], and its expression is shown in Equation (3):

$$\begin{aligned}
\mathrm{Y} &= \mathrm{M}_\mathrm{s}(\mathrm{M}_\mathrm{c}(X)) \\
&= w_s \odot \bigcup_{i=1}^{C} x^{(i)} \cdot w_c^{(i)},
\end{aligned} \tag{3}$$

where $\mathrm{X}, \mathrm{Y} \in \mathbb{R}^{C \times H \times W}$ represents the feature maps before and after attention. Compared to the attention mechanism in [24], the channel attention M_c and the spatial attention M_s act on the input feature map in an insufficient supervised form. $w_c^{(i)} \in \mathbb{R}^{C \times 1 \times 1}$ represents the attention weight of each channel dimension, $\bigcup$ represents channel concatenation, and

C is the number of channels. In CBAM [24], the importance of each channel is obtained by using global max pooling and global average pooling, followed by a fully connected layer. Then, the channel attention is obtained by multiplying the channel weights one by one with each channel's feature $x^{(i)}$ and concatenating them. Subsequently, perform pooling, dimensionality reduction and sigmoid operations based on the channel again to obtain the response weight of the spatial dimension $w_s \in \mathbb{R}^{1\times H\times W}$, where $\odot$ represents the element-wise product.

$$\begin{aligned} \mathrm{Y} &= \mathbb{SES}(sc(\|X\|)) \\ &= \bigcup_{n=1}^{N}\bigcup_{k=1}^{K}\bigcup_{c=1}^{C} x_{k,c}^{(n)} \cdot w_{k,c}^{(n)}. \end{aligned} \tag{4}$$

Our method is significantly different from the attention methods described above. The expression is shown in Equation (4), where $\|\cdot\|$ represents our proposed self-adaptive upsampling, the factor of each up-sampling is 2, sc represents the shared connection operation, and $\mathbb{SES}$ stands for semantic edge supervision.

The main differences are as follows: first, in the channel dimension, the self-adaptive upsample block first reduces the channel dimensions through a 1×1 convolution and ReLU activation function, then uses the transposed convolution with a stride of 2 for upsampling, repeating the above until the edge map size is the same as the original. Quantitatively, we use a 1×1 convolution with the same number of channels as the number of categories to obtain edge activation maps with the same number of channels as the number of categories. Therefore, the above operations can achieve the effect of decoupling each category to its corresponding category channel. Finally, the channel supervise is performed by calculating the multi-label pixel-wise loss on the poly-based coarse edge ground truth and the multi-scale edge activation map obtained for different categories of channels, and then reweight is applied to each category of channels. Specifically, assuming that the number of categories in the current picture is $\mathrm{n} \in [0, \mathrm{N}]$, these n channels will be strengthened, and the remaining channels will be weakened as the background.

Next, we use the shared connection operation to achieve the effect of hierarchical attention under the action of supervision. To expand, we have obtained the activation maps of multiple scales after reweighting the channels in the previous step. Assuming that the activation maps of one of the categories in multiple bypasses are $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_\mathrm{k} \in \mathbb{R}^{N\times H\times W}$, we concatenate them to obtain the activation map of dimension kN. Then, we reduce its dimension via depth-wise convolution to obtain the fused activation map $\mathrm{A}_{\mathrm{fused}} \in \mathbb{R}^{N\times H\times W}$. At this time, the activation maps of multiple scales are reweighted and communicated to obtain the importance of different pyramid levels.

On the spatial dimension: we propose a semantic edge label, a coarse edge label derived from the oriented annotation bounding boxes of the original dataset. The target edges to be detected are supervised, not all edges. The background area, especially the non-object with an object-like shape, does not introduce edge guidance, weakens the background, highlights the target edge, and achieves some noise reduction effect. Specifically, we calculate the multi-label pixel-wise loss between the poly-based coarse edge ground truth of the target to be detected and multiple bypass activation maps for supervision. The spatial responses $\mathbf{W}_s \in \mathbb{R}^{N\times H\times W}$ within N categories are obtained. After attention to the channel and hierarchy dimensions, it is multiplied by the element with the previously obtained features. The effect of suppressing the background, highlighting the foreground contour in a strongly supervised form, and capturing the target global context information, can be achieved.

$$\mathrm{Y} = \sigma(\mathrm{F}_{\mathrm{P_k}}^{'}) \odot \mathrm{F}_{\mathrm{P_k}} + \mathrm{F}_{\mathrm{P_k}}. \tag{5}$$

After that, to take advantage of this guiding role, we express another branch in the form of Equation (5). σ represents the sigmoid activation function, $\mathrm{F}_{\mathrm{P_k}}$ represents FPN feature, $\mathrm{F}_{\mathrm{P_k}}^{'}$ represents FPN feature after expanding the receptive field and $\odot$ represents the element-wise product. We first use the sigmoid activation function to convert the FPN feature after expanding the receptive field F'_{P_k} into the attention-like weights, multiply it element-wise by itself F_{P_k} and finally perform the residual operation. The effect of this kind of attention can directly guide the adjacent network to pay attention to the edge features of the target and capture the complete object-wise context information through the above operations, so it is more influential and targeted. The final effect is shown in the third row of Figure 3. Meanwhile, we have conducted detailed ablation experiments for this innovation, and the results can be found in Section 3.2.2.

2.2.3. Rotation-Invariant Spatial Pooling Pyramid for Feature Enhancement of Arbitrary-Orientated Target

We tackle the challenging problem above and capture global contextual informative features in remote sensing images with complex backgrounds. Nevertheless, the key features of the target are still lacking, which is very unfavorable for the classification task. In order to solve this problem, we consider the particularity of remote sensing images that the target has arbitrary orientation characteristics in the top view. Therefore, we believe that the rotation-invariant feature of the target can be extracted as a critical feature in this field. However, it should be noted that the convolution filter cannot extract rotation-invariant features, so extracting general deep features for the same type of targets in different directions is difficult. Current works address these problems from two aspects, in which [30,47] improve the convolution operation itself so that they can extract rotation-sensitive features; [11,19] extract rotation-invariant features by pooling feature maps with rotation channels.

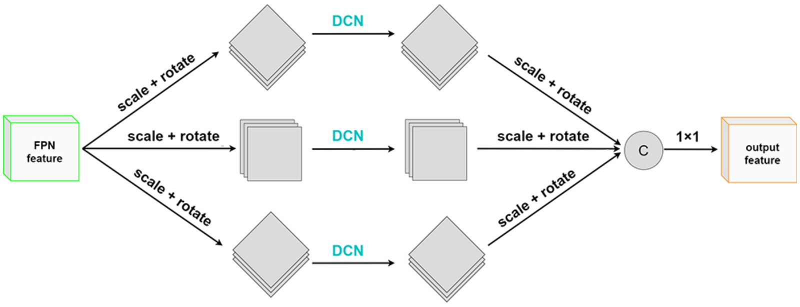

Figure 5. The structure of rotation-invariant spatial pooling pyramid.

In order to solve the above problems, we propose a rotation-invariant spatial pooling pyramid based on [27], which extracts rotation-invariant features by fusing feature maps of multiple rotation angles without adding additional direction channels. As shown in Figure 5, the overall process is as follows: for the single-scale feature $F \in \mathbb{R}^{C \times H \times W}$ extracted from the SES module, we first use n convolution kernels as 1, and the convolution with $\frac{C}{2}$ channels divides the original features into $F_1, F_2, \ldots, F_n$. Next, one of the branches will be rotated θ, 90°, 180°, or 270°. Then, we calculate the scaling ratio according to the angle and scale the feature map to avoid the loss of partial offset of the feature map after rotation. Subsequently, we use the deformable convolution [48] to extract the features of the target at this angle. This is owing to the following considerations:

(i) The offset learning in deformable convolution can better extract the representative features of targets in any direction and has better feature extraction ability for targets with large aspect ratios and different shapes;

(ii) Because of padding, the oriented feature map has obvious boundaries, bringing additional background noise to the ordinary convolution operation. The use of deformable convolution can alleviate this problem.

Through the above operations, we use multiple branches to extract features from the same feature map at different rotation angles. We rotate and scale the feature map back to its original state and concatenate the n branches together in the channel dimension. Finally, the rotation-invariant features can be obtained by restoring and compressing the number of channels through a 1 × 1 convolution.

2.2.4. Multi-Task Loss Function

The multi-task loss function consists of three parts. In addition to the loss used for classification and regression tasks in the original single-stage detector, we introduce an additional edge detection task loss to correspond to the proposed semantic edge detection module. Therefore, the overall definition of the multi-task loss function is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(p_i, p_i^*) + \mathcal{L}_{reg}(v_i, v_i^*) + \mathcal{L}_{edge}. \tag{6}$$

The classification loss $\mathcal{L}_{cls}$ uses focal loss [8], p_i represents the probability of whether the i-th anchor predicted by the network is the target, and p_i^* represents its ground truth labels. We use the smooth L1 loss [2] as the regression loss $\mathcal{L}_{reg}$, where v_i and v_i^* represent the predicted box and ground-truth box, each box is in vector form, and the box is encoded according to the format in [11]. For the edge detection task in the SES module, since multiple pyramid levels are involved, the overall edge loss of multiple bypasses is as follows:

$$\mathcal{L}_{edge} = \sum_{i=1}^{S} \alpha^{(i)} \mathcal{L}_{side}^{(i)} \tag{7}$$

in which S represents the number of bypasses used for supervision, $\alpha^{(i)}$ provides a trade-off coefficient for each bypass, which is set to 1 by default, and the total edge loss is the sum of the losses of the S bypasses. Next, for the loss calculation of each bypass, considering the extreme imbalance between edge and non-edge labels and the semantics of the edges to be predicted, we use weighted cross-entropy loss for multiple category labels, which is expressed explicitly as follows:

$$\begin{aligned}\mathcal{L}_{side}^{(i)} &= \sum_k \mathcal{L}_k(\mathbf{W}^{(\mathbf{i})}) \\ &= \sum_k \sum_{\mathbf{p}} \Big\{ -\beta \overline{\mathbf{Y}}_k(\mathbf{p}) \log \mathbf{Y}_k(\mathbf{p} \mid \mathbf{I}; \mathbf{W}^{(\mathbf{i})}) \\ &\quad -(1-\beta)\big(1-\overline{\mathbf{Y}}_k(\mathbf{p})\big) \log\Big(1-\mathbf{Y}_k(\mathbf{p} \mid \mathbf{I}; \mathbf{W}^{(\mathbf{i})})\Big) \Big\}.\end{aligned} \tag{8}$$

In the above formula, $\mathbf{W}^{(\mathbf{i})}$ represents the parameter of the network in the i-th bypass, and k represents the number of categories. We generate the coarse edge ground truth from the polygon annotations in the dataset, where $\overline{\mathbf{Y}}_k(\mathbf{p})$ represents the value at the k-th class pixel p on the ground truth. $\mathbf{Y}_k(\mathbf{p} \mid \mathbf{I}; \mathbf{W}^{(\mathbf{i})})$ represents the pixel point $\mathbf{p}$ in the input image $\mathbf{I}$ after the network inference about the k-th semantic category forecast result. The introduction of the non-edge pixel percentage β is used to solve the problem of label imbalance. In summary, the predicted value of each bypass and the ground truth calculate a binary cross-entropy loss for all categories k to determine whether it is an edge. Finally, we accumulate the multiple bypass output losses obtained above and obtain the overall multi-label loss for multi-category.

3. Results

In this section, we describe the two datasets used in the experiment and the main results of the experiment. Details of the ablation experiment design are described and analyzed.

3.1. Datasets

The DOTA[28] is a large-scale public dataset for object detection in the aerial image. There are 2806 images in total. The pixels of each image is between 800 × 800 and 4000 × 4000 pixels. The instance objects in the image are divided into 15 categories with a large variation in object scale, aspect ratio, and arbitrary orientation. These fifteen categories include Bridges (BR), Harbor (HA), Ship (SH), Plane (PL), Helicopter (HC), Small vehicle (SV), Large vehicle (LV), Baseball diamond (BD), Ground track field (GTF), Tennis court (TC), Basketball court (BC), Soccer-ball field (SBF), Roundabout (RA), Swimming pool (SP), and Storage tank (ST). The experts in remote sensing image analysis used horizontal and oriented bounding boxes to label each instance respectively, annotating 188,282 instances. The annotation boxes we use are all oriented bounding boxes, and the proposed semantic edge annotations are coarse edge maps obtained from the oriented annotation boxes. The entire dataset is divided into 1/2 training set, 1/6 validation set and 1/3 testing set, and the final test results are obtained through the official evaluation server. During training, since the original DOTA image is too large, we split the image into 1024 × 1024 sub-images with an overlap of 200 pixels. The network is trained with the SGD optimizer in training. The learning rate is 0.0025, weight decay and momentum are 0.0001 and 0.9; we train 12 epochs with batch size 2 on GeForce RTX A4000, and the 9th and 11th epochs are an epoch decreasing learning rate. To avoid overfitting, we apply random horizontal flipping.

UCAS-AOD [29] is a dataset containing two types of oriented aerial images: 510 car images and 1000 aircraft images. The pixels of each image is 659 × 1280. The instances in each image are annotated with horizontal and oriented bounding boxes containing 14,596 instances. There is no division standard for the original dataset. To make a fair comparison with the comparison method, we adopted the same division ratio of 5:2:3. In training and inferencing, all images are resized to 800 × 800. We use the SGD optimizer, and the initial learning rate is set to 0.01, and the weight decay and momentum are 0.0001 and 0.9.

3.2. Main Results

In this section, we conduct a series of ablation experiments on DOTA-v1.0 to verify the effectiveness of our proposed method. Unless otherwise stated, we all train on the training set of the DOTA dataset and evaluate on the validation set, and the image size is 1024 × 1024. We use RetinaNet based on ResNet50 as the baseline. To evaluate the performance of object detectors, we adopt the typical measure of mean Average Precision (mAP) and Frame Per Second (FPS).

3.2.1. Evaluation of Different Components

We evaluate all proposed modules to verify the contribution and effectiveness of each module. The experimental results are shown in Table 1, and the abbreviations are explained as follows: Bridges (BR), Ground track field (GTF), Large vehicle (LV), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HA), and Swimming pool (SP).

Table 1. Component-wise experiment.

Methods	SES	RISPP	BR	GTF	LV	BC	ST	SBF	RA	HA	SP	mAP	FPS
Baseline	-	-	37.9	64.8	67.5	60.2	60.3	49.0	62.9	58.4	50.8	63.49	18.2
Ours	×	✓	39.4	63.4	68.2	61.8	63.8	50.7	62.9	60.0	52.5	63.68	7.9
	✓	×	41.1	69.0	69.0	63.2	63.7	52.9	67.0	60.2	54.4	65.42	14.1
	✓	✓	41.7	68.4	69.8	63.3	64.5	53.2	66.4	61.1	54.5	65.95	6.0

First, to verify the effectiveness of the SES and RISPP modules themselves, we build model variants containing only themselves based on the baseline, respectively. Introducing the SES module can improve detection performance by 1.93%, indicating that the SES module can guide the network to focus on the target itself and provide better regression and contextual guidance information. For small and dense target swimming pools, it increases by 3.6%; while for high aspect ratio ground track field, it increases by 4.2%. After only introducing the RISPP module to extract the rotation-invariant features, the detection performance is improved by 0.19%, indicating that the extracted rotation-invariant features can help the network obtain key features beneficial to classification and localization. There is a 3.5% increase for the small and dense storage tanks. Finally, under the joint action of the two modules, the overall improvement is 2.46%, indicating that the two feature enhancement components have achieved a complementary effect and can achieve 65.95% mAP. In terms of time performance, Table 1 shows that the proposed two modules perform well in time and that the fused model greatly shortens the detection time. So far, we have verified the effectiveness of each proposed module. Next, we will perform a more detailed ablation analysis of each module.

3.2.2. Evaluation of Semantic Edge Supervision

In this section, we design detailed ablation experiments from three perspectives to verify the effectiveness of the proposed feature enhancement network based on semantic edge supervision. As shown in Table 2, we first evaluate the attention-like mechanism of this module from three dimensions: space, channel, and pyramid level. When the sides' channel is set to 1, it means that only space supervision is carried out to verify the effect of space supervision. Number 15 in the table corresponds to the number of categories of DOTA in the dataset for ablation experiments. Therefore, when the sides' channel is set to 15, it means that category channel supervision is introduced. Finally, the introduction of shared connection operation means the introduction of pyramid hierarchical attention.

Table 2. Ablation of attention-like mechanism.

Sides' Channel	Shared Connection	mAP
–	–	63.49
1, 1	×	64.30(+0.81)
15, 15	×	64.43(+0.13)
15, 15	✓	64.94(+0.51)

To avoid the interference of other variables, we first only output the single-channel activation map on the $P3$ and $P4$ bypasses of the pyramid level for spatial supervision. Compared with the baseline, it can be observed that the detection performance is improved by 0.81%, indicating that this attention-like mechanism can help the network pay attention to the contextual information between different instances in the spatial dimension. Next, we expand the number of channels of the bypass activation map to the number of categories and try to decouple the edge maps of different categories in the channel direction to verify the effect of channel dimension attention. Experiments show an increase of 0.13% based on spatial attention. Finally, we fuse the above-mentioned multi-scale and multi-channel edge activation maps through shared connections, which further improves the 0.51% mAP. It further verifies the importance of multi-scale feature interaction in the hierarchical dimension. It verifies that the network can learn the importance of different categories at each pyramid level under strong supervision.

Table 3. SAU and different supervised strategies ablation.

Supervision Strategy	Sides' Channel	SAU	mAP	FPS
Deep	1, 15, 15	×	64.60	14.1
Side3, Fuse	1, 15, 15	×	64.82	14.4
Side3, Fuse	15, 15	×	64.61	16.8
Side3, Fuse	1, 15, 15	✓	65.21	13.6

The rationality of different supervision strategies and the effectiveness of the self-adaptive upsample module (SAU) are verified in Table 3. We add a pyramid level P_2 to assist edge supervision in this experiment. First, we try to supervise the edge activation maps at all scales without using the SAU module, and the overall detection performance only reaches 64.60%. At the same time, according to the research in [49], we try to use SAU to build a buffer on each depth-supervised bypass to obtain better performance, but the result is reduced by 0.12%. Subsequently, we try to change the supervision strategy. It is found that the activation map after only supervising side3 and the shared connection fusion is better, with an improvement of 0.22%. Next, we try to verify the effect of the P_2 single-channel activation map under this supervised strategy and find that the detection performance is decreased by 0.21% without P_2 providing detailed information. This verifies that there are more edge details in the low-level feature maps, and more contextual details are preserved for the final edge activation map. It also verifies the rationality and necessity of selecting pyramid levels P_2 to P_4 for semantic edge supervision. Finally, to verify the superiority of the proposed SAU module over traditional semantic edge detection using only a single transposed convolution for upsampling, we add this module based on the above and achieve a performance of 65.21% mAP and 13.6 FPS. Therefore, the final setting achieves the best performance in terms of accuracy and time performance.

Table 4. Ablation of each module in SES module.

Dilated Conv	SES	Decoder	mAP	FPS
–	–	–	63.49	18.2
×	✓	raw	64.82 (+1.33)	14.4
ASPP{6, 8, 16}	✓	raw	64.55	12.4
2,2,2,4	✓	raw	64.99	13.2
1,2,5	✓	raw	65.42 (+0.6)	14.1
1,2,5	✓	SAU	65.52 (+0.1)	12.8

We have completed the ablation experiments within each component independently in the above experiments, while we also verify the interaction and effectiveness between all components in Table 4. First, we add the optimal SES module using the raw decoder to the baseline, improving 1.33% mAP and reducing 3.8 FPS. Then, we add different dilated convolution assist edge tasks before SES and find that the dilation rate of 1, 2, and 5 has the best effect, increasing to 0.6% mAP, verifying that expanding the receptive field can improve the edge detection performance. The dilated convolution block further strengthens the attention-like mechanism's role and deals with the large-scale change of remote sensing targets. Finally, we add the SAU module, which increases by 0.1% and achieves good performance in time, verifying that this module is still effective after a dilated convolution block. It can be observed that SES suppresses redundant backgrounds in a strong supervised form and highlights the edge responses of the objects to be detected. It enables the network to focus on the target, improving the detection performance. Therefore, SES has the most noticeable improvement to the whole with attention-like mechanism and supervision strategies and is the core component of the entire module.

3.2.3. Evaluation of RISPP

As shown in Table 5, we also design corresponding ablation experiments to compare the effect of RISPP applying different convolution types. All comparative experimental angles are set to 90°, 180° , and 270° to avoid feature loss after feature rotation. The abbreviations in Table 5 are explained as follows: Bridges (BR), Ground track field (GTF), Basketball court (BC), Harbor (HA), and Swimming pool (SP). When using 1 × 1 convolution, the overall mAP is only 63.07%. Subsequently, after changing to 3 × 3 convolution, the overall increase is 0.57%, where the large aspect ratio of the basketball court is improved by 4.8%. Finally, after replacing the convolution with a deformable convolution that is more suitable for this module, it is increased by 0.04% compared to the 3 × 3 convolution. Compared to using the 1 × 1 convolution, there is a 3.4% increase for the large aspect ratio harbor. In terms of time performance, this setting also achieves the best performance. The above ablation experiments fully verify that offset learning in deformable convolution can better extract the representative features of objects in any direction and has better feature extraction ability for objects with large aspect ratios and different shapes. Moreover, by observing the feature map and combining this module with the effect of deformable convolution on small targets with complex backgrounds, it is concluded that padding makes the feature map after rotation have apparent boundaries. These newly generated boundaries bring additional background noise to ordinary convolution operations, which we successfully mitigate using deformable convolution.

Table 5. Ablation of RISPP.

Conv. Type	BR	GTF	BC	HA	SP	mAP	FPS
1 × 1	37.6	63.3	56.8	56.6	50.6	63.07	8.5
3 × 3	38.6	61.9	61.6	59.3	52.1	63.64	8.3
3 × 3, deformable	39.4	63.4	61.8	60.0	52.5	63.68	7.9

3.2.4. Results on DOTA

We compare the proposed method with other state-of-the-art methods on the DOTA dataset. As shown in Table 6, our method achieves an mAP of 76.42 % and 6.0 FPS, surpassing all the methods compared. We use the single-stage RetinaNet as the baseline for improvement, which surpasses some single-stage detectors and far surpasses other two-stage detectors. The abbreviations in this table are explained as follows: Plane (PL), Baseball diamond (BD), Bridges (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).

Table 6. Comparisons with other state-of-the-art methods on DOTA dataset.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Two-Stage:																	
ICN [50]	R-101	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
RoI-Trans. [51]	R-101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet [12]	R-101	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
FADet [52]	R-101	90.21	79.58	45.49	76.41	73.18	68.27	79.56	90.83	83.40	84.68	53.40	65.42	74.17	69.69	64.86	73.28
Gliding Vertex [53]	R-101	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
Mask OBB [54]	RX-101	89.56	85.95	54.21	72.90	76.52	74.16	85.63	89.85	83.81	86.48	54.89	69.64	73.94	69.06	63.32	75.33
FFA [55]	R-101	90.1	82.7	54.2	75.2	71.0	79.9	83.5	90.7	83.9	84.6	61.2	68.0	70.7	76.0	63.7	75.7
CenterMap OBB [56]	R-101	89.83	84.41	54.60	70.25	77.66	78.32	87.19	90.66	84.89	85.27	56.46	69.23	74.13	71.56	66.06	76.03
Single-stage:																	
PIoU [57]	DLA-34	80.9	69.7	24.1	60.2	38.3	64.4	64.8	90.9	77.2	70.4	46.5	37.1	57.1	61.9	64.0	60.5
P-RSDet [58]	R-101	89.02	73.65	47.33	72.03	70.58	73.71	72.76	90.82	80.12	81.32	59.45	57.87	60.79	65.21	52.59	69.82
A2S-Det [59]	R-101	89.59	77.89	46.37	56.47	75.86	74.83	86.07	90.58	81.09	83.71	50.21	60.94	65.29	69.77	50.93	70.64
O2-DNet [60]	H-104	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
DAL [15]	R-101	88.61	79.69	46.27	70.37	65.89	76.10	78.53	90.84	79.98	78.41	58.71	62.02	69.23	71.32	60.65	71.78
DRN [61]	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
BBAVector [62]	R-101	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32
CFC-Net [18]	R-50	89.08	80.41	52.41	70.02	76.28	78.11	87.21	90.89	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50
SLA [14]	R-50	88.33	84.67	48.78	73.34	77.47	77.82	86.53	90.72	86.98	86.43	58.86	68.27	74.10	73.09	69.30	76.36
SE^2-Det(Ours)	R-101	89.31	85.67	50.53	72.82	79.99	73.96	85.85	90.69	84.73	83.23	64.84	67.83	72.56	76.59	67.85	76.42

Some detection results of DOTA are shown in Figure 6. It can be observed that the presented method is friendly to the detection of densely arranged small objects in different directions, especially the small vehicles, ships, and large vehicles in the first row. All objects in the image are successfully identified and have accurate regression boxes. It benefits from the background denoising effect of the attention-like mechanism and the sensitive capture of the target orientation by the rotation-invariant features. Represented by detecting large vehicles, bridges, and harbors, our method can well detect and regress large aspect ratio targets. It is mainly due to the introduction of edge supervision, which enables the network to pay attention to the target's global characteristics and context characteristics from a complex background and has more regression clues. Overall, our method also copes well with remote sensing targets' intra-class and inter-class scale changes. Figure 6 shows the intra-class scale variation for the harbors, the roundabouts and the storage tanks. Inter-class scale variation is also easy to find, such as small vehicles and basketball courts. The excellent performance of our method on the scale variation is mainly due to the feature enhancement network based on semantic edge supervision that realizes an attention-like mechanism in the three dimensions of space, channel, and pyramid level.

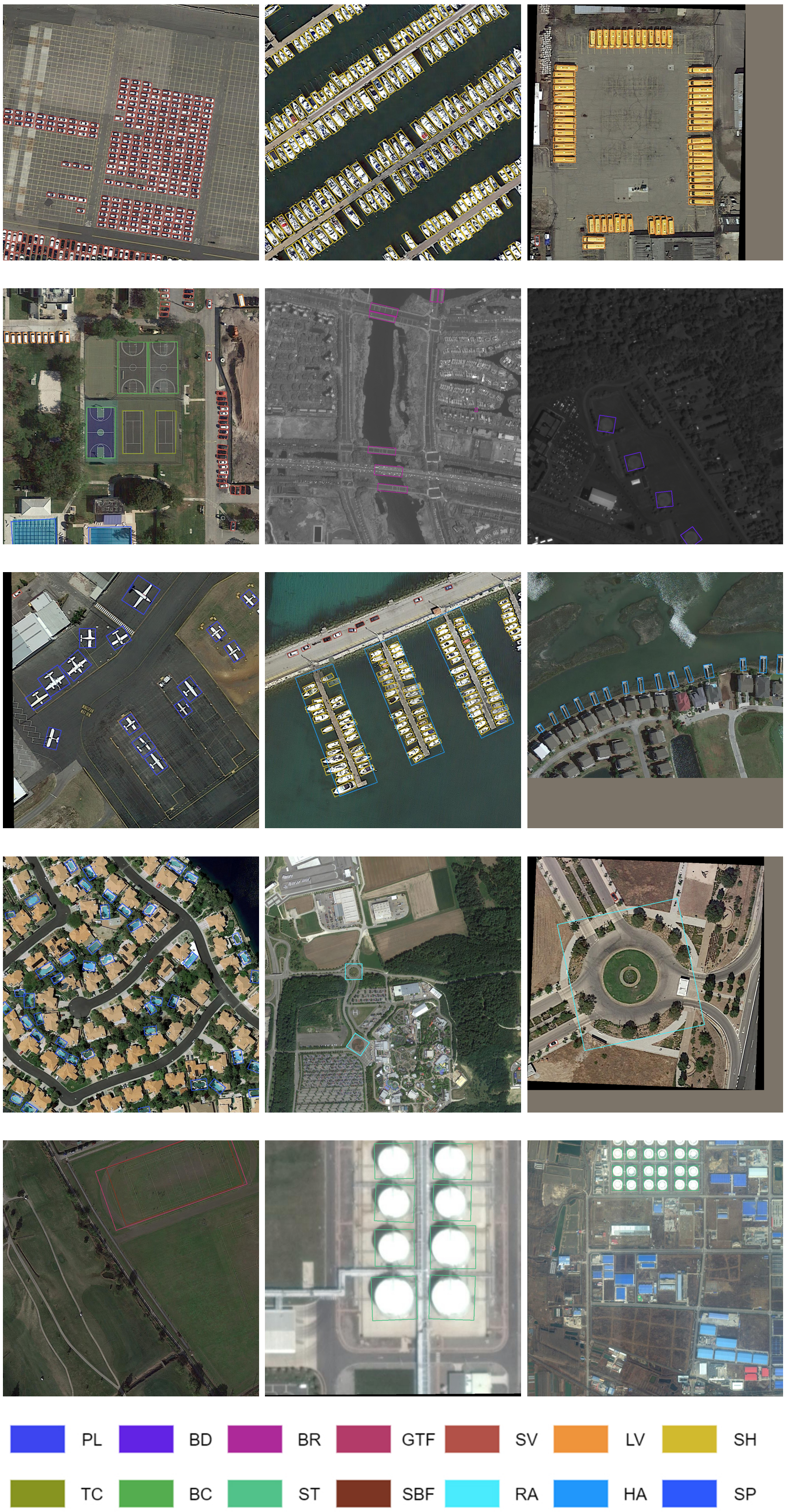

Figure 6. Visualization of predictions on the DOTA dataset using our method, SE^2-Det.

3.2.5. Results on UCAS-AOD

We also conduct experiments on the UCAS-AOD dataset to verify the generality of our method. Table 7 shows that our method exceeds many advanced methods, achieving a 2.43% improvement over the baseline RetinaNet-R to 90.00% mAP. Our method is not only 1.64% higher than the classical two-stage Faster-RCNN but also 0.51% higher than the oriented object detection method, CFC-Net. Among all the methods compared, our method achieves the highest level of plane class detection, indicating that the proposed SES and RISPP modules make our method more robust to targets in different directions. Some detection results on UCAS-AOD are shown in Figure 7.

Table 7. Comparisons with high-quality detection performance on UCAS-AOD dataset.

Methods	Car	Plane	mAP
RetinaNet-R [8]	84.64	90.51	87.57
Faster-RCNN [3]	86.87	89.86	88.36
RoI-Trans. [51]	87.99	89.90	88.95
SLA [14]	88.57	90.30	89.44
CFC-Net [18]	89.29	88.69	89.49
DAL [15]	89.25	90.49	89.87
SE^2-Det(Ours)	89.24	90.71	90.00

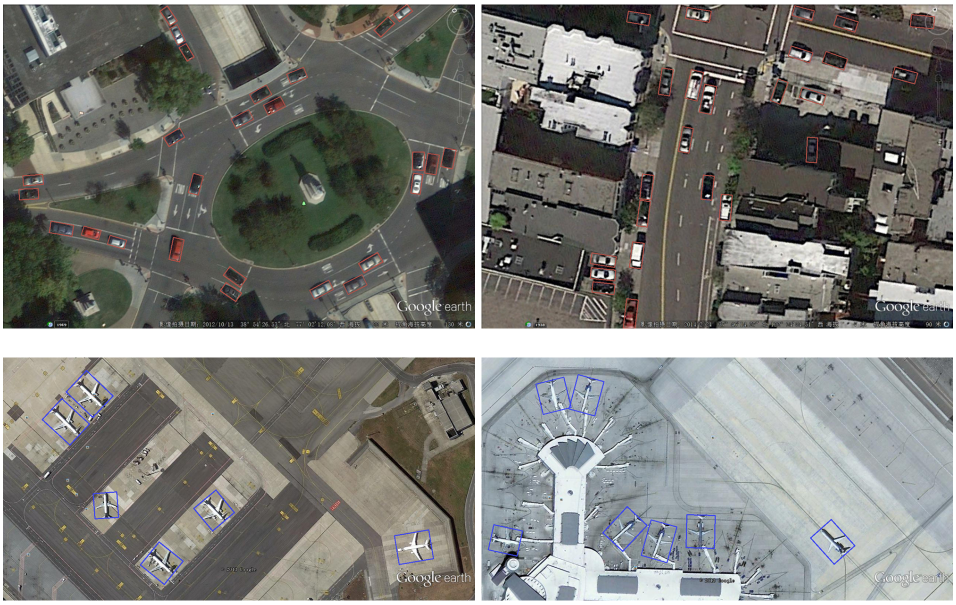

Figure 7. Visualization of predictions on the UCAS-AOD dataset using our method SE^2-Det.

4. Discussion

Since we introduce edge detection to remote sensing for the first time, we also conduct statistical analysis for further discussion to verify the effectiveness of edge supervision. From Figure 8, we can see that most categories' average deviation between regression boxes and ground truth has improved after edge supervision. The improvement of Basketball court (BC), Large vehicle (LV), and Harbor (HA) are more pronounced, showing that the regression boxes' quality for high aspect ratio targets is generally improved after edge supervision [63].

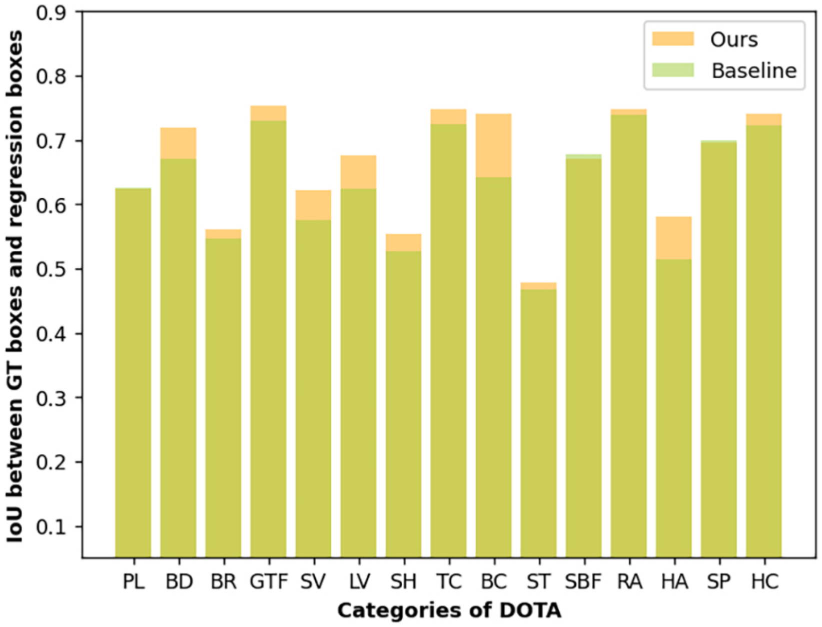

Figure 8. Average IoU of each category of ground truth boxes and regression boxes before and after edge supervision.

We verify that our method surpasses the comparison detector and the effectiveness of the proposed module, but the proposed network can have a better development direction. On the one hand, Figure 8 shows that after the introduction of edge detection, the mean deviation between regression box and ground truth has been significantly improved in many categories. Still, it does not perform well in some categories, such as the Soccer ball field (SBF), which needs further research. On the other hand, the semantic edge labels introduced in this paper are rough edge graphs derived from oriented bounding boxes. This means that the edge ground truth used for supervision is the object's bounding rectangle, not the object's actual edge, which may limit the effect of edge supervision. Changing the coarse edge to the actual edge of the target to be detected may solve this problem, which will be our future research direction.

5. Conclusions

We pioneeringly introduce semantic edge detection into oriented remote sensing object detection and propose a simple and effective semantic-edge-supervision feature enhancement network (SE^2-Det). The SES module's core idea is to guide the network to pay attention to the multi-scale contextual features of each category target through the attention-like mechanism of three dimensions: space, channel, and level. While suppressing the background, it strengthens the edge response between different instances in each category, which solves the lack of global response for complex backgrounds and single targets (especially targets with high aspect ratios and different shapes) in remote sensing scenes. Meanwhile, we further enhance the rotation-invariant features of remote sensing targets through the RISPP module to capture the essential features specific to remote sensing targets. We conduct detailed ablation and comparative experiments on two public remote sensing datasets, DOTA and UCAS-AOD, and achieve competitive performance.

Author Contributions: Conceptualization, D.C. and X.H.; methodology, D.C. and X.H.; software, X.H.; validation, D.C. and X.H.; formal analysis, X.H.; investigation, D.C. and X.H.; resources, C.Z. and R.Z.; data curation, D.C., C.Z. and X.H.; writing—original draft preparation, D.C.; writing—review and editing, D.C.,C.Z. and X.H.; visualization, X.H.; supervision, C.Z. and R.Z.; project administration, D.C.; and funding acquisition, C.Z. and R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the ’Science and technology innovation action plan’ Natural Science Foundation of Shanghai under grant number 22ZR1427000, National Natural Science Foundation of China (CN) under grant number 61602296. Furthermore, this work is also sponsored by the ‘Chenguang Program’ supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission under grant number 18CG54. The authors would like to thank their supports.

Data Availability Statement: The source code of the paper is available at https://github.com/Virusxxxxxxx/SE2-Det, accessed on 5 July 2022.

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

References

1. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
2. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
3. 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.* **2017**, *39*, 1137–1149.
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
5. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
7. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9626–9635.
8. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327.
9. 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, Amsterdam, The Netherlands, 11–14 October 2016; Lecture Notes in Computer Science; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
10. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 3163–3171.
11. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602511.
12. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8231–8240.
13. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022, *early access*.
14. Ming, Q.; Miao, L.; Zhou, Z.; Song, J.; Yang, X. Sparse Label Assignment for Oriented Object Detection in Aerial Images. *Remote Sens.* **2021**, *13*, 2664.
15. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic Anchor Learning for Arbitrary-Oriented Object Detection. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2355–2363.
16. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 11830–11841.
17. Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; Yan, J. Learning High-Precision Bounding Box for Rotated Object Detection via Kullback-Leibler Divergence. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Vol. 34, pp. 18381–18394.
18. 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.* **2022**, *60*, 5605814.
19. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A Rotation-Equivariant Detector for Aerial Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2786–2795.
20. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-Layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2148–2161.
21. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection. *IEEE Trans. Image Process.* **2019**, *28*, 265–278.

22. Yu, Z.; Feng, C.; Liu, M.Y.; Ramalingam, S. CASENet: Deep Category-Aware Semantic Edge Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1761–1770. ISSN: 1063-6919.
23. Hao, X.; Shan, C.; Xu, Y.; Sun, S.; Xie, L. An Attention-based Neural Network Approach for Single Channel Speech Enhancement. In Proceedings of the ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6895–6899. ISSN: 2379-190X.
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
25. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
26. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of Localization Confidence for Accurate Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 784–799.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916.
28. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
29. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
30. Zhou, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Oriented Response Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 519–528.
31. Kittler, J. On the accuracy of the Sobel edge detector. *Image Vis. Comput.* **1983**, *1*, 37–42.
32. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698.
33. Hu, Y.; Chen, Y.; Li, X.; Feng, J. Dynamic feature fusion for semantic edge detection. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI'19, Macao, China, 10–16 August 2019; AAAI Press: Macao, China, 2019; pp. 782–788.
34. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5229–5238.
35. Zhen, M.; Wang, J.; Zhou, L.; Li, S.; Shen, T.; Shang, J.; Fang, T.; Quan, L. Joint Semantic Segmentation and Boundary Detection Using Iterative Pyramid Contexts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13666–13675.
36. Zhao, J.X.; Liu, J.J.; Fan, D.P.; Cao, Y.; Yang, J.; Cheng, M.M. EGNet: Edge Guidance Network for Salient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8779–8788.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
38. Lin, T.Y.; DollÃąr, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. ISSN: 1063-6919.
39. Sun, Y.; Ye, J. FEDet: Feature Enhancement Single Shot Detector. In Proceedings of the 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, UK, 19–23 August 2019; pp. 843–850.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6000–6010.
41. Song, G.; Liu, Y.; Wang, X. Revisiting the Sibling Head in Object Detector. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11560–11569. ISSN: 2575-7075.
42. Zhang, X.; Wan, F.; Liu, C.; Ji, X.; Ye, Q. Learning to Match Anchors for Visual Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3096–3109.
43. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1395–1403. ISSN: 2380-7504.
44. Acuna, D.; Kar, A.; Fidler, S. Devil Is in the Edges: Learning Semantic Boundaries From Noisy Annotations. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11067–11075. ISSN: 2575-7075.
45. Yu, Z.; Liu, W.; Zou, Y.; Feng, C.; Ramalingam, S.; Kumar, B.V.K.V.; Kautz, J. Simultaneous Edge Alignment and Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 388–404.
46. Odena, A.; Dumoulin, V.; Olah, C. Deconvolution and Checkerboard Artifacts. *Distill* **2016**, *1*, e3.

47. Weiler, M.; Cesa, G. General E(2)-equivariant steerable CNNs. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Number 1286; Curran Associates Inc.: Red Hook, NY, USA, 2019; pp. 14357–14368.
48. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773. ISSN: 2380-7504.
49. Liu, Y.; Cheng, M.M.; Fan, D.P.; Zhang, L.; Bian, J.W.; Tao, D. Semantic Edge Detection with Diverse Deep Supervision. *Int. J. Comput. Vis.* **2022**, *130*, 179–198.
50. Azimi, S.M.; Vig, E.; Bahmanyar, R.; KÃűrner, M.; Reinartz, P. Towards Multi-class Object Detection in Unconstrained Remote Sensing Imagery. In Proceedings of the Computer Vision—ACCV, Perth, Australia, 2–6 December 2018; Springer: Cham, Switzerland, 2019; pp. 150–165.
51. 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.
52. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; Yang, J. Feature-Attentioned Object Detection in Remote Sensing Imagery. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3886–3890. ISSN: 2381-8549.
53. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding Vertex on the Horizontal Bounding Box for Multi-Oriented Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1452–1459.
54. Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; Yang, W. Mask OBB: A Semantic Attention-Based Mask Oriented Bounding Box Representation for Multi-Category Object Detection in Aerial Images. *Remote Sens.* **2019**, *11*, 2930.
55. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308.
56. 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.* **2021**, *59*, 4307–4323.
57. Chen, Z.; Chen, K.; Lin, W.; See, J.; Yu, H.; Ke, Y.; Yang, C. PIoU Loss: Towards Accurate Oriented Object Detection in Complex Environments. In Proceedings of the Computer Vision—ECCV, Glasgow, UK, 23–28 August 2020; Springer: Cham,Switzerland, 2020; pp. 195–211.
58. Zhou, L.; Wei, H.; Li, H.; Zhao, W.; Zhang, Y.; Zhang, Y. Arbitrary-Oriented Object Detection in Remote Sensing Images Based on Polar Coordinates. *IEEE Access* **2020**, *8*, 223373–223384.
59. Xiao, Z.; Wang, K.; Wan, Q.; Tan, X.; Xu, C.; Xia, F. A2S-Det: Efficiency Anchor Matching in Aerial Image Oriented Object Detection. *Remote Sens.* **2021**, *13*, 73.
60. Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; Sun, X. Oriented objects as pairs of middle lines. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 268–279.
61. 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.
62. Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; Metaxas, D. Oriented Object Detection in Aerial Images with Box Boundary-Aware Vectors. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 2149–2158. ISSN: 2642-9381.
63. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.