remote sensing

MDPI

Article

SODE-Net: A Slender Rotating Object Detection Network Based on Spatial Orthogonality and Decoupled Encoding

Xiaozhi Yu [1], Wei Xiang [2,3,*], Lu Yu [2], Kang Han [2] and Yuan Yang [1]

[1] School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China; yuxiaozhi@stu.xaut.edu.cn (X.Y.)
[2] School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC 3086, Australia
[3] College of Science and Engineering, James Cook University, Cairns, QLD 4878, Australia
* Correspondence: w.xiang@latrobe.edu.au

Abstract

Remote sensing objects often exhibit significant scale variations, high aspect ratios, and diverse orientations. The anisotropic spatial distribution of such objects' features leads to the conflict between feature representation and boundary regression caused by the coupling of different attribute parameters: previous detection methods based on square-kernel convolution lack the overall perception of large-scale or slender objects due to the limited receptive field; if the receptive field is simply expanded, although more context information can be captured to help object perception, a large amount of background noise will be introduced, resulting in inaccurate feature extraction of remote sensing objects. Additionally, the extracted features face issues of feature conflict and discontinuous loss during parameter regression. Existing methods often neglect the holistic optimization of these aspects. To address these challenges, this paper proposes SODE-Net as a systematic solution. Specifically, we first design a multi-scale fusion and spatially orthogonal convolution (MSSO) module in the backbone network. Its multiple shapes of receptive fields can naturally capture the long-range dependence of the object without introducing too much background noise, thereby extracting more accurate target features. Secondly, we design a multi-level decoupled detection head, which decouples target classification, bounding-box position regression and bounding-box angle regression into three subtasks, effectively avoiding the coupling problem in parameter regression. At the same time, the phase-continuous encoding module is used in the angle regression branch, which converts the periodic angle value into a continuous cosine value, thus ensuring the stability of the loss value. Extensive experiments demonstrate that, compared to existing detection networks, our method achieves superior performance on four widely used remote sensing object datasets: DOTAv1.0, HRSC2016, UCAS-AOD, and DIOR-R.

Keywords: rotating object detection; multi-scale spatial orthogonal convolution; regression parameter decoupling

Academic Editor: Paolo Tripicchio

Received: 18 July 2025
Revised: 27 August 2025
Accepted: 28 August 2025
Published: 1 September 2025

Citation: Yu, X.; Xiang, W.; Yu, L.; Han, K.; Yang, Y. SODE-Net: A Slender Rotating Object Detection Network Based on Spatial Orthogonality and Decoupled Encoding. *Remote Sens.* **2025**, *17*, 3042. https://doi.org/10.3390/rs17173042

1. Introduction

Remote sensing object detection (RSOD) aims to classify and localize objects of interest (e.g., aircraft, ships, and vehicles) in remote sensing images, serving as a critical technology for scene interpretation in the remote sensing field. It is widely applied across various domains, including military reconnaissance, disaster relief, environmental monitoring, and urban planning. In recent years, with the rapid development of deep learning technology,

various deep learning-based methods have gradually become mainstream in remote sensing object detection due to their strong feature representation capabilities and have achieved breakthroughs in detection performance.

Deep learning-based RSOD methods can be broadly categorized into two-stage algorithms and one-stage algorithms. Two-stage algorithms first generate a series of regions of interest (Rols) that may contain objects, and then perform object/background classification and bounding box regression on these proposals [1,2]. However, remote sensing images contain numerous rotated objects. Traditional methods typically use horizontal anchor boxes to generate candidate regions, which leads to misalignment between targets and anchors, thereby introducing significant background interference. To address this, Ma and Ding et al. [3–6] incorporated target angle information into the Region Proposal Network (RPN) and designed specialized angle prediction networks, achieving precise localization using rotated bounding boxes. Enhancing the feature representation of targets is another primary approach to improving detection performance. For example, Fu et al. [7] enhanced feature representation through multi-scale information fusion; CAD-Net [8] strengthened the connection between object features and their corresponding scenes by learning the global and local context of regions of interest, thereby enhancing the network's feature representation; ReDet [9] uses group convolution to generate rotation-equivariant features, and then combines rotation-invariant Rol alignment to extract rotation-invariant features from rotation-equivariant features to achieve accurate detection of rotating objects. The single-stage algorithm does not need to generate candidate regions, but directly regresses the bounding box and category of the object from multiple locations of the image [10,11]. The optimization strategies for single-stage detectors can also be categorized into two parts: efficient feature extraction in the network and bounding box classification and regression. Xu et al. [12] employed a densely connected network to enhance feature extraction capability. AF-SSD [13] enhances key features by introducing spatial and channel dual-path attention. AFRE-Net [14] first creates a fine-grained feature pyramid network (FG-FPN) to provide richer spatial and semantic features and then generates stronger feature representations through a feature enhancement module. OASL [15] integrates orientation-aware spatial information into the classification and localization branches to enhance feature diversity, thereby establishing a solid foundation for improved detection performance.

Whether it is a two-stage or one-stage algorithm, high-quality feature extraction is a prerequisite for ensuring the final detection quality. As analyzed earlier, we can enhance the network's feature extraction through various methods such as multi-scale feature fusion, incorporating contextual information, or adding attention mechanisms. Additionally, deformable convolution can be employed to dynamically adjust the receptive field, thereby improving feature adaptability. However, these methods almost universally rely on small-sized convolutional kernels as the basic module for extracting image features. In remote sensing images, targets exhibit significant variations in pose, manifested as large-scale differences, high aspect ratios, and diverse directional arrangements. For backbone feature extraction, small-size square-kernel convolution is constrained by the images' limited receptive fields, capturing only localized information while maintaining fixed receptive fields across different objects. Although deformable convolutions can dynamically adjust receptive fields, they require meticulous parameter tuning to learn offset values and fail to consistently deliver large receptive fields. Consequently, both approaches lack the overall perception of large-scale or slender objects. Therefore, the use of large-kernel convolution to extract features has begun to attract people's attention. Recently, LSKNet [16] first introduced large-kernel convolution into remote sensing object detection. This method connects the feature maps generated by large-size convolution kernels along the channel direction to form features with rich scene context information. PKINet [17] further parallels

large-kernel convolutions of multiple sizes to extract dense texture features of different receptive fields, thereby further improving the network 's ability to detect multi-scale change objects. CPMFNet [18] adds a parallel dilated convolution layer structure to the backbone network, and dynamically adjusts the kernel size through dilation rates to ensure that a larger receptive field is provided under the premise of equal computation.

In summary, the aforementioned methods expand the receptive field by employing large-kernel convolutions or dilated convolutions, capturing more contextual information from the scene and modeling long-range dependencies for large-sized and high-aspect-ratio targets. Consequently, they significantly improve the detection performance of remote sensing objects without complex designs. However, in the task of detecting slender and rotated objects, existing networks face conflicts between feature representation and bounding box regression due to the coupling of different attribute parameters, which manifests as the following:

(1) Noise Introduction and Soaring Computational Load: Firstly, due to their large receptive fields, large-kernel convolutions inevitably capture more background information, especially for small and slender targets, thereby introducing significant background noise. Secondly, their computational load increases exponentially compared to small-kernel convolutions. While dilated or atrous convolutions can reduce computational costs, their sparse feature sampling struggles to accurately extract fine boundary features.

(2) Bounding Box Angle Discontinuity: When calculating the offset loss for rotated bounding boxes, the rotation angle is a critical regression parameter. However, angular values are periodic, leading to discontinuities at definition boundaries. Specifically, when an angle approaches a boundary, the predicted box and the ground truth box may be nearly equivalent in physical space, but the regression paths differ significantly depending on the rotation direction, resulting in large deviations in loss computation. (In taking the long-edge representation as an example, if the ground truth angle is 89° (clockwise) and the predicted angle is 91° (clockwise), the physical deviation is only 2°. However, since the predicted angle exceeds the defined range, it is recalculated as $-89°$ (counterclockwise), causing the actual angular deviation in loss computation to become 178°.)

(3) Coupling of Different Attribute Parameters: Traditional remote sensing detection networks only consider the feature differences required for the classification and regression tasks, and only decouple the classification and regression branches at the starting position of the detection head to avoid the interference caused by shared features. However, beyond the significant feature differences between different branches, the bounding box regression branch itself also exhibits variations in the target's attribute parameters. Specifically, the position and scale of the object need to be predicted based on rotation-invariant features, while the rotation angle prediction is based on rotation and other variable features, and there is still the problem of feature conflict.

To address the aforementioned issues, we must not only achieve more appropriate feature extraction but also maintain the consistency between features and tasks, as well as the continuity of loss in subsequent regression calculations—that is, considering the design and optimization of the network from a holistic perspective. Based on this, we propose SODE-Net, which primarily consists of a backbone network with an MSSO module, a decoupled detection head, and a phase-continuous encoding module for rotation angles. First, we designed the MSSO module in the backbone network to replace stages based on large square-kernel convolutions. This module naturally captures long-range dependencies of objects through multi-scale orthogonal strip-shaped receptive fields without introducing excessive background noise, thereby enabling the network to extract more precise features. Second, we construct a two-stage finely decoupled detection head. It first employs an oriented detection module to generate direction-sensitive features, and then

extracts rotation-invariant features from the direction-sensitive features to perform bounding box regression and classification separately. Furthermore, in the regression branch, we use two parallel sets of convolutional layers to predict the position and shape of the rotated bounding box, as well as its angle, achieving secondary decoupling to avoid feature conflicts between the two. Finally, we introduce a phase-continuous encoding module to independently encode the rotation angle, converting the angle value into its phase cosine value. This value remains continuous during periodic angle variations, thereby resolving the discontinuity issue near the boundary of the angle definition domain. The proposed SODE-Net is a general joint solution that combines rotated feature extraction and rotation angle regression. In summary, our contributions include the following:

(1) We designed a backbone network incorporating multi-scale fusion and spatially orthogonal convolution (MSSO) module, which combines the advantages of square-kernel convolutions and strip convolutions. It can efficiently extract features of objects with varying aspect ratios without introducing excessive background noise, particularly excelling in capturing elongated objects.

(2) We designed a detection head with a multi-level decoupled architecture. It first employs rotational filters to generate orientation-sensitive features, and then extracts rotation-invariant features from them to perform bounding box regression and classification separately. In the regression branch, two parallel convolutional groups are used to independently process the regression of box position/shape and angle, achieving secondary decoupling to further separate feature conflicts between them.

(3) We introduce a phase-continuous encoding module in the angle regression branch, which converts the rotation angle values of bounding boxes into cosine phase values. These values remain continuous throughout angular periodic variations, thereby eliminating discontinuity and instability in regression loss caused by the periodicity of angle changes.

(4) The experimental results on large remote sensing datasets DOTAv1.0, HRSC2016, UCAS-AOD, and DIOR-R show that our proposed SODE-Net outperforms other remote sensing image rotation target detection methods and achieves the results of SOTA.

2. Related Work

2.1. Remote Sensing Rotating Object Detection

Objects in aerial images exhibit diverse orientations. Using horizontal bounding boxes for representation can lead to issues such as imprecise localization and significant overlap among bounding boxes [19,20]. To address this, researchers have proposed rotated-object detection algorithms, which employ rotated bounding boxes with angles to detect targets. Current approaches primarily focus on rotated feature extraction and bounding box representation. RoI Transformer [4] transforms horizontal regions of interest (RoIs) into rotation-sensitive rotated RoIs and extracts rotation-invariant features through rotated RoI alignment (RRoI-Align). RRoI-Align can only align features in the spatial dimension but not in the direction dimension. Therefore, it is necessary to continuously adjust the feature direction through the region of interest. In this regard, ReDet [9] first employs group convolution to generate feature maps with N orientation channels, enabling the network to produce rotation-equivariant features. These features are then processed by the Rotation-Invariant Region of Interest Alignment (RiRoI Align) module, which achieves alignment in both spatial and orientation dimensions by rotating the RoIs and cyclically switching channels. However, the above methods are anchor-based, where anchor boxes cannot precisely adapt to object shapes and suffer from discontinuous rotation angle variations. Consequently, researchers have proposed point-set-based representation methods, which provide more fine-grained position representation and easy classification information through a set of point sets [21]. Point RCNN [22] first learns representative points of objects

during the Region Proposal Network (RPN) stage and generates pseudo-oriented bounding boxes (OBBs) for rotated RoIs (RRoIs). It then refines the corner points of each RRoI through PointReg for regression and further optimization. CFA [23] treats a set of sampled points on feature maps as a convex hull to represent object features. It then refines the convex hull via feature point resampling to better cover irregularly shaped objects, thereby mitigating feature aliasing of objects. However, point-set-based methods are more sensitive to image quality and annotation errors, especially for objects with high aspect ratios. Even minor deviations can cause a sharp drop in the Intersection over Union (IoU).

2.2. Large-Kernel Convolution

Objects in remote sensing images often exhibit significant scale variations and mostly possess high aspect ratios. Large-kernel convolutions with expanded receptive fields inherently excel in such scenarios. Compared to small-kernel convolutions, large-kernel convolutions broaden the receptive field, capture richer contextual information, and establish more effective long-range dependencies, thereby enhancing the model's feature representation. These techniques have been widely adopted in image processing and downstream tasks [24,25]. RepLKNet [26] proposed a CNN backbone based on large-kernel convolutions and systematically evaluated its performance in image classification, segmentation, and object detection tasks, demonstrating highly competitive results. Recently, LSKNet [16] introduced large kernel convolution into the field of remote sensing object detection for the first time. The network adds a large kernel convolution module to the backbone network in the form of residual connection, which effectively captures the context information in the remote sensing scene, thereby significantly improving the detection accuracy. Due to the large number of large kernel convolution parameters, the use of more flexible and smaller parameter expansion convolution is another way to expand the receptive field [27,28]. MDCT [29] employs multiple dilated convolutional layers in remote sensing detection networks to enlarge the receptive field, enhancing the backbone network's feature extraction capability for large-scale and slender objects. Furthermore, multi-head attention modules are embedded in the neck to further guide the network's focus on global information, thereby strengthening the overall representational capacity of the network. Nevertheless, while expanding receptive fields, both large-kernel and dilated convolutions inevitably introduce considerable noise, which significantly interferes with target recognition. To address this issue, this paper introduces the MSSO module to mitigate the incorporation of background noise, while maintaining the network's ability to model long-range dependencies and holistic perception through carefully designed combinations.

2.3. Angular Regression

The discontinuity of angle values is a unique problem in rotating object detection, primarily caused by the definition form of rotating bounding boxes. Compared to horizontal bounding boxes, rotating boxes include an angular parameter whose value can abruptly change at the domain boundaries, leading to sudden shifts in the regression path of predicted boxes and resulting in training instability [30,31]. Yang et al. [32] added the IoU constant factor to the traditional smooth L1 loss function to suppress the sudden increase in the loss function in the boundary case, thereby reducing the learning difficulty of the model. Yang et al. [30] classified the definition range of the angle from the perspective of angle classification. Due to the limited classification results, the situation in which the angle exceeds the definition range is avoided. However, optimizing the loss function does not fundamentally resolve the angle boundary issue, and the angle classification approach is inherently discrete, making fine-grained classification challenging. Recently, Yu et al. [33] proposed a Phase-Shifting encoder (PCS) that encodes the rotation angle of a rotated

bounding box into continuous sine or cosine phase variations. These phase values are then decoded back into discrete angle values through a decoder, establishing a one-to-one mapping between rotation angles and their corresponding trigonometric values. This method transforms angular variations into continuous changes in cosine values, thereby avoiding abrupt jumps in angle values caused by rotational periodicity at boundary points. By ingeniously mapping the periodically discontinuous angles to continuous function values, this approach simultaneously resolves both the boundary discontinuity issue and the square-like ambiguity problem. Given its effectiveness, we incorporate this encoder into the detection head to handle angle regression.

3. Methodology

In this paper, we systematically analyze the challenges faced by existing networks in slender rotating object detection, including difficulties in holistic object perception, the introduction of background noise, and the conflicts in feature and boundary representation caused by the coupling of different attribute parameters. Based on this, we designed a layer incorporating multiple shapes of receptive fields as the backbone stage for feature extraction and decoupled detection heads for parameter regression in the detector. Additionally, we introduced a rotary angle encoder to address the angle jump issue. The overall architecture of SODE-Net is illustrated in Figure 1. Specifically, we designed a multi-scale fusion and spatially orthogonal feature extraction (MSSO) module that incorporates multiple MSSO blocks in a residual connection manner and connects them in series to extract features through expanded, multi-scale combined receptive fields, as shown in the lower part of Figure 1. Based on these features, the Region Proposal Network (RPN) generates RoIs for the detection head. Subsequently, within the detection head, we employ a multi-level decoupling module to separate classification, bounding box regression, and rotation angle regression into three distinct branches, each processing features independently to avoid feature coupling issues caused by shared features, as shown in the right part of Figure 1. Finally, in the angle regression branch, we map the angle value to the phase cosine value through the rotary angle encoder, so that the predicted output becomes a continuous value. All modules will be described in detail below.

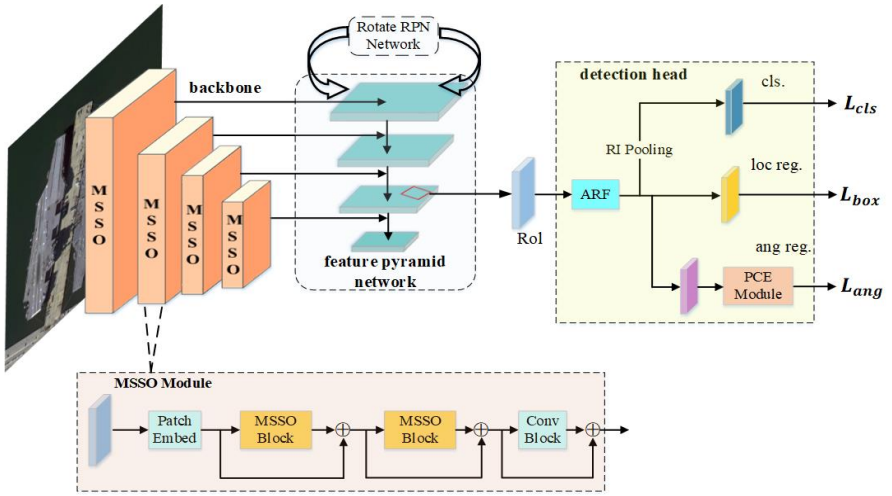

Figure 1. The overall architecture of SODE-Net consists of two parts: the MSSO backbone network and the fine decoupled detection head. Module MSSO consists of multiple cascaded MSSO-blocks with residual connections, which are used to extract features and generate RoIs through a rotating RPN. The fine-grained decoupling detection head decomposes the acquired features into three branches to decouple the classification, bounding box position, and angle representation parameters during regression.

3.1. Multi-Scale Fusion and Spatially Orthogonal (MSSO) Module

In RSOD tasks, objects with large-scale variations and high aspect ratios are widely present. Conventional convolutions, due to their limited receptive fields, can only focus on local information, thus lacking a holistic perception of large-scale or elongated objects. Based on this, some methods introduce large-kernel convolution to expand the receptive fields and capture more contextual scene information. However, this approach may introduce significant background noise, adversely affecting detection performance.

To address the aforementioned issues, this section proposes the MSSO Module. Within this module, we first employ the patch-embedding method to divide the feature map into multiple patches and then sequentially connect two MSSO blocks with residual connections. Specifically, each MSSO block first utilizes large-size square-kernel convolution to extract local contextual information, followed by multi-scale spatial orthogonal strip-shaped convolutions to capture long-range dependencies across different orientations while reducing background noise interference. The generated feature maps are subsequently applied as attention weights to the input features, thereby enhancing discriminative feature representation of key regions while suppressing background or redundant information. Compared to previous approaches, the strip convolutions can effectively extract fundamental features of objects with varying aspect ratios, and the multi-scale fusion design accommodates feature extraction for objects at different scales. The sequential architecture effectively combines the advantages of both square convolution and strip-shaped convolution without requiring additional information fusion modules, and the pipeline as illustrated in Figure 2. The detailed implementation is as follows.

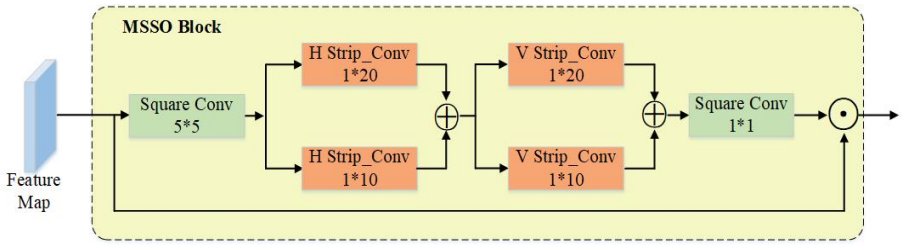

Figure 2. The MSSO block incorporates two sets of multi-scale strip-shaped convolutional layers that are spatially orthogonal in horizontal and vertical directions. These layers are connected in series with square-kernel convolutions to extract anisotropic features, while element-wise multiplication attention is employed to enhance the output features.

Given an input feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, with $\mathbf{C}$, we first apply a square-kernel convolution $\mathbf{K} \in \mathbb{R}^{C\times H\times W}$ to extract local contextual features, producing the feature map $\mathbf{F}$ defined as

$$\mathbf{F} = \mathrm{Conv}_{K_H\times K_W}(\mathbf{X}) \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}, \tag{1}$$

where $H \times W$ and $K_H \times K_W$ denote the feature map size and the convolution kernel size. The standard convolution layer primarily extracts local contextual information from the input feature.

Next, we use multi-scale horizontal and vertical orthogonal large-size strip-shaped convolutions to extract features along both spatial axes, which can be formally expressed as

$$\mathbf{F}^{\mathbf{W}} = \mathrm{Str}_{\mathrm{Conv}_{1\times K_{W1}}}(\mathbf{F}) + \mathrm{Str}_{\mathrm{Conv}_{1\times K_{W2}}}(\mathbf{F}) \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}, \tag{2}$$

$$\mathbf{F}^{\mathrm{H}} = \mathrm{Str}_{\mathrm{Conv}_{K_{H1}\times 1}}(\mathbf{F}^{\mathrm{W}}) + \mathrm{Str}_{\mathrm{Conv}_{K_{H2}\times 1}}(\mathbf{F}^{\mathrm{W}}) \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}, \tag{3}$$

where $1 \times K_{W1}$ and $1 \times K_{W2}$ denote the kernel sizes of the horizontal convolutions, while $K_{H_1\times 1}$ and $K_{H_2\times 1}$ represent the kernel sizes of the vertical convolutions. The combination of

the two sets of orthogonal strip-shaped convolutions can collect features across two spatial axes, and it is also good at capturing long-range dependence information in features.

To further enhance the interaction of features across the channel dimension, we apply a 1×1 pointwise convolution to the output of the orthogonal convolution layer, obtaining feature map $\mathbf{Y}$. Feature map $\mathbf{Y}$ is then used as attention weights applied to the original input $\mathbf{X}$, yielding the final output feature $\mathbf{A}$, expressed as follows:

$$\mathbf{Y} = \text{Sigmoid}(Conv_{1\times1}(F^H)), \tag{4}$$

$$\mathbf{A} = \mathbf{Y} \odot \mathbf{X}, \tag{5}$$

where $\odot$ denotes elementwise multiplication. Benefiting from the feature extraction of multi-scale orthogonal convolutions and the channelwise feature aggregation via pointwise convolution, each position in feature map $\mathbf{Y}$ encodes both horizontal and vertical characteristics across a broad spatial region. By applying $\mathbf{Y}$ as attention weights to the input $\mathbf{X}$, the network enhances its representation of elongated or narrow structures in the spatial dimension while reinforcing feature emphasis in object regions.

3.2. Deep Fine-Grained Decoupling Detection Head

Previous object detectors predominantly employ shared fully connected layers in their detection heads for both classification and localization tasks. However, the spatial correlation of fully connected layers is inherently limited, making them insensitive to positional variations and thus unsuitable for precise localization. To address this, we decouple the classification and localization tasks by introducing an oriented detection module (ODM) to mitigate the inherent conflicts between these two objectives. Furthermore, within the localization task itself, there is also feature coupling between position prediction and angle prediction. Consequently, we implement a secondary decoupling of these two subtasks. The overall workflow is illustrated in Figure 3.

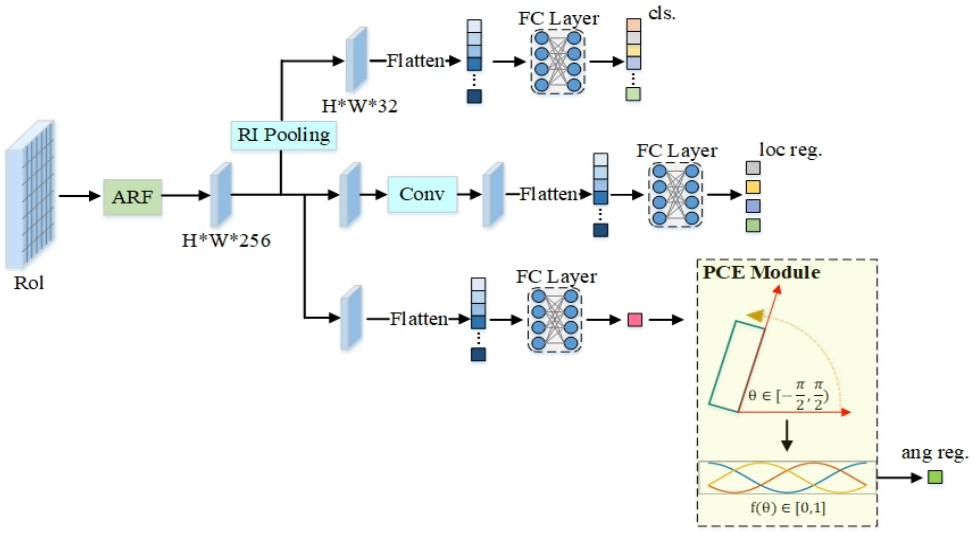

Figure 3. The deep fine-grained decoupling detection head consists of three branches, which utilize distinct features to predict the category, location and shape, and rotation angle, respectively. Among them, the RoI features are first modeled by the active rotating filters (ARFs) into orientation-aware features. Building upon this, the classification branch employs rotation-invariant (RI) pooling to extract rotation-invariant features for categorization. The angle branch utilizes the PCE module to convert angle values into continuous cosine values for angle regression.

ODM: This module first employs active rotating filters (ARFs) [34] to encode the orientation information. The ARF is a K×K×N filter that actively rotates N-1 times during the convolution process, generating feature maps with N orientation channels (N is 8 by default). For input feature map X and ARF feature F, the output of the i-th orientation in feature Y can be expressed as

$$Y^{(i)} = \sum_{n=0}^{N-1} F_{\theta_i}^{(n)} \cdot X^{(n)}, \quad \theta_i = i\frac{2\pi}{N}, \quad i = 0, \ldots, N-1, \tag{6}$$

where F_{θ_i} is the feature **F** rotated clockwise by θ_i, and $F_{\theta_i}^{(n)}$ and $X^{(n)}$ are the features of the n-th orientation channel in F_{θ_i} and **X**, respectively. By applying ARF to convolutional layers, we obtain orientation-sensitive features with explicitly encoded orientation information. While bounding box regression benefits from such orientation-sensitive features, object classification requires rotation-invariant features. To extract rotation-invariant features, we perform rotation-invariant pooling over the orientation-sensitive features, simply selecting the strongest response across all orientation channels as the output feature $\hat{X}$:

$$\hat{X} = \max\{X^{(n)} | n = 0, 1, \ldots, N-1\}. \tag{7}$$

In this way, we can align features of objects with varying orientations to achieve robust object classification. Compared to orientation-sensitive features, orientation-invariant feature is efficient with fewer parameters. For instance, a H × W × 256 feature map with 8 orientation channels is reduced to H × W × 32 after max-pooling. Finally, we feed the orientation-sensitive features and orientation-invariant features into two subnetworks dedicated to bounding box regression and classification, respectively.

Regression Parameter Decoupling: In the bounding box regression subtask, the prediction of box coordinates requires rotation-invariant features, while angle prediction demands rotation-equivariant features, which still presents an issue of coupled feature representation. Moreover, the subsequent phase-continuous encoding module only involves angle regression. Therefore, we further decompose the regression process of rotated bounding boxes into multiple branches. For different parameters within the bounding box, we group them according to their characteristics and assign separate branches in the detection head to predict each group. This approach enables independent and interference-free regression for distinct parameters, thereby achieving more accurate rotated object detection. The details are as follows.

The parameters of the rotated bounding box are divided into two groups: the box position and size (x, y, w, h), and the rotation angle (θ). In the bounding box regression subtask, we extract distinct feature vectors and from two sets of feature maps, which are then processed by fully connected layers to generate the final predictions. The formula is as follows:

$$d_x, d_y, d_w, d_h = W_1\_FC(G_1), \tag{8}$$

$$d_\theta = W_2\_FC(G_1), \tag{9}$$

where (d_x, d_y, d_w, d_h) and (d^θ) are the predicted offsets for positional parameters, shape parameters, and angular parameters, respectively. W_1_FC and W_1_FC are the learnable fully connected layer parameters. We construct separate loss functions for the two sets of parameters predicted from the grouped features G_1 and G_2, then use backpropagation to update the convolutional layer parameters. This process continuously enhances the convolutional layer's ability to extract features corresponding to these parameters, thereby achieving feature decoupling.

3.3. Phase-Continuous Encoding (PCE) Module

As mentioned in the previous section, the objects detected in remote sensing exhibit diverse orientations and are often characterized by high aspect ratios, making horizontal bounding boxes inadequate for accurate representation. Consequently, current mainstream approaches predominantly employ rotated object detectors, which obtain high-quality detection boxes that tightly enclose objects by incorporating rotation angles. However, rotated bounding boxes suffer from the drawback of discontinuous regression loss due to the periodicity inherent in the orientation angles of the bounding boxes.

Based on this, we introduce a phase-continuous encoding module to encode the angle information into a continuous cosine phase values, and then decode the phase information back to the discrete angle prediction through the decoder, thus solving the problem of angle jump at the boundary. In this way, the network further solves the problem that the angle value is discontinuous when calculating the loss on features extracted by the MSSO module, thereby enhancing the network's precision in capturing slender objects and stable regression under multi-angle variations in objects. The specific workflow is as follows:

Setting the rotation bounding box as the 'long edge 90' angle definition, the rotation angle is $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$ and $\varphi = 2\theta$; the angle-encoding formula is

$$x_n = \cos\left(\varphi + \frac{2n\pi}{N_{step}}\right), \tag{10}$$

where N_{step} denotes the number of phase-shifting steps, and $n = 1, 2, \ldots, N_{step}$. This encoder maps angles to cosine values, and since cosine is continuous over the range of angle variations, it resolves the discontinuity issue caused by direct angle regression. Correspondingly, the angle decoding formula can be expressed as

$$\varphi = -\arctan \frac{\Sigma_{n=1}^{N_{step}} x_n \sin\left(\frac{2n\pi}{N_{step}}\right)}{\Sigma_{n=1}^{N_{step}} x_n \cos\left(\frac{2n\pi}{N_{step}}\right)}, \tag{11}$$

where the output angle φ is within the range of $(-\pi, \pi]$ and is uniquely determined.

The rotating bounding box is represented by five parameters (x, y, w, h, θ), which are the center coordinates, width, height, and angle of the box. According to Formula (9), the coding output is in the range of $[-1, 1]$. In order to make the training more stable, we convert the output features:

$$F_{out} = 2 \times sigmoid(F_{in}) - 1, \tag{12}$$

where F_{in} represents the output features of the convolutional layer, and F_{out} denotes the predicted encoded data constrained to the range $[-1, 1]$. Subsequently, the L1-loss is applied to compute the loss for the angle regression branch:

$$L_{ang} = |X_{GT} - X_{pred}|, \tag{13}$$

where X_{GT} is obtained by applying phase encoding to the rotation angle of the ground truth (GT) bounding box.

Based on the above analysis, the total loss of the network comprises three components: the classification loss L_{cls}, the bounding box location loss L_{box}, and the rotation angle loss L_{ang}. The overall loss function can be expressed as follows:

$$L = L_{cls} + \lambda_1 L_{box} + \lambda_2 L_{ang}, \tag{14}$$

where λ_1 and λ_2 are the balancing weights for bounding box localization and rotation, respectively.

4. Experiments

4.1. Datasets

We conducted extensive experiments on three widely used remote sensing object detection datasets:

DOTAv1.0 [35] is a large-scale dataset for remote sensing detection, including 2806 images, 188,282 instances, and 15 categories. The number of images in the training set, validation set and test set is 1411, 458, and 937, respectively. The images of the dataset come from different sensor platforms. The image resolution is diverse, the target category is rich, the scale and angle are different, the number of small targets is large, and there is a wide range of dense arrangement. This provides strong data support for remote sensing object detection, but it also brings great challenges to accurate detection.

HRSC2016 [36] is a remote sensing dataset for ship detection, which contains 1061 aerial images with a size range of 300 × 300 and 1500 × 900. The numbers of images in the training set, validation set, and test set are 436, 181, and 444. The dataset covers different geographical environments and illumination conditions. The target background varies in interference, arbitrary direction, dense arrangement, and cloud occlusion, which provides data support for detection in complex environments.

UCAS-AOD [37] is a high-resolution aerial target detection dataset comprising 2420 images with 14,596 instances, categorized into two classes (aircraft and vehicles), and a certain number of background samples. The dataset covers diverse scenarios and perspectives, including varied geographical environments and lighting conditions, effectively testing algorithms' generalization capability in complex situations. Additionally, the uniform distribution of object orientations makes it particularly advantageous for studying the directional robustness of detection algorithms.

DIOR-R [38] is a large-scale dataset for remote sensing detection, including 23,463 images, 190,288 instances, and 20 categories. The targets exhibit significant scale variations, dense arrangements, and arbitrary orientation distributions. Due to its comprehensive diversity in both target types and directional distributions, it has become an essential benchmark dataset for rotated object detection research.

4.2. Implementation Details

For a fair comparison, we adopted the same data processing approach as previous mainstream methods [39,40]. Our model was built with oriented R-CNN as a baseline. All models were trained on the training set and validation set, and then tested on the test set, unless otherwise specified. For the DOTAv1.0 dataset, the single-scale and multi-scale training/testing strategies were adopted, respectively. In the single-scale strategy, we cropped the original image into 1024 × 1024 patches in the form of a sliding window, with a stride size of 200. And in multi-scale scenarios, the sliding window was set to three sizes at scales of (0.5, 1, 1.5). The model was trained with 12 epochs and tested with a test set, and the results on the DOTA dataset were submitted to the official website online server. For the DIOR-R dataset, we used the original image size of 800 × 800 and maintained the same training strategy as for the DOTAv1.0 dataset. For the detection accuracy on the HRSC2016 dataset, we adjusted the image size to 800 × 800 without changing the image scale, trained with 36 epochs, and used the mean average precision (mAP) as an evaluation metric. The FLOPs reported in this article were calculated with a 1024 × 1024 image input.

During the experiment, the AdamW optimizer was used for training, and weight decay was set to 0.05. The initial learning rate was set to 0.0001 and 0.0004 for DOTAv1.0 and HRSC2016, respectively.

4.3. Ablation Studies

4.3.1. Module Ablation Study

We conducted ablation experiments on Datasets HRSC2016 and DOTAv1.0 to validate the performance of the proposed MSSO module and PCE module, as well as to explore the optimal parameter settings that maximize their effectiveness. The ablation experiments on the DOTAv1.0 dataset were conducted using single-scale training and testing.

Table 1 presents the results of the baseline model on the DOTAv1.0 dataset with and without our proposed MSSO module and PCE module. On the DOTAv1.0 dataset, the baseline model achieves only 76.27% mAP due to the limited receptive fields of small-size convolutional kernels, which inherently lack the capability for holistic perception of large-scale or slender objects. When integrating the proposed MSSO module into the backbone network, the detector's performance improved by 3.18%, demonstrating that MSSO can more accurately extract features for objects with varying aspect ratios—particularly those with high aspect ratios—without introducing significant noise. Additionally, compared to the baseline network using ResNet50 as the backbone, incorporating the MSSO module in the backbone further reduces the model's parameter (40.6 M→30.6 M). After incorporating the PCE module, the rotation angle values are encoded as continuous cosine function values, which prevents a sharp increase at the boundary of the domain during angle loss calculation, thereby avoiding instability in the regression of the loss function at angular boundaries. As a result, the detector's accuracy improved by 0.37%. The combined use of the MSSO and PCE enhanced the performance of our network by 3.56% over the baseline model, achieving higher precision in rotated object detection.

Table 1. Influence of MSSO and PCE modules on DOTA-v1.0. (The best-performing outcomes are highlighted in bold).

MSSO Module	PCE Module	mAP (%)	#P (M)
✗	✗	76.27	40.6
✓	✗	79.45	**30.6**
✗	✓	76.64	40.6
✓	✓	**79.83**	**30.6**

Similar experimental results were obtained on the HRSC2016 dataset. As shown in Table 2, SODE-Net achieved superior performance through the combination of different modules. The MSSO module employs a multi-scale spatial orthogonal convolution block to more accurately extract features of objects with varying aspect ratios and rotation angles, while enhancing the feature representation of object regions through a spatial attention mechanism. Subsequently, for the regression task, the classification, localization, and angle prediction are decoupled, with the PCE module specifically handling the regression of rotation angles. Building upon the robust feature extraction of the backbone network, this approach further mitigates feature coupling during regression and avoids training instability caused by angular jumps to collectively improve overall performance. SODE-Net ultimately achieved a detection accuracy of 91.25% mAP on the HRSC2016 dataset while maintaining merely 30.6M model parameters.

Table 2. Influence of MSSO and PCE modules on HRSC2016. (The best-performing outcomes are highlighted in bold).

MSSO Module	PCE Module	mAP (%)	#P (M)
✗	✗	90.60	40.6
✓	✗	91.03	**30.6**
✗	✓	90.97	40.6
✓	✓	**91.25**	**30.6**

4.3.2. Detection Head Ablation Study

In the design of the detection head, we decoupled the regression parameters for position and angle into separate branches. Here, (x, y, w, h, θ) denote the coordinates of the geometric center point, width, height, and rotation angle of the target's rotated bounding box, respectively. The subscript f_c in the lower-right corner of the parentheses indicates a fully connected layer, *conv* denotes a convolutional layer, *cls* represents the classification branch, *reg* stands for the bounding box regression branch, *loc* represents the position regression branch, and *ang* stands for the angle regression branch. We first replaced the fully connected layers in the detection head with convolutional layers. Since convolutional operations incorporate spatial correlation information compared to fully connected layers, this modification led to a significant improvement in performance. Next, we decoupled classification from bounding box regression, resulting in a slight increase in detection accuracy. Finally, we decomposed the bounding box regression into a localization branch (x, y, w, h, θ) and an angle branch (θ), further alleviating intrinsic feature conflicts and achieving optimal detection performance. The results are shown in Table 3.

Table 3. Ablation study of the decoupling detection head. (The best-performing outcomes are highlighted in bold).

Different Head Structure	mAP (%)
$[(x, y, w, h, \theta)_{fc}]_{cls+reg}$	79.52
$[(x, y, w, h, \theta)_{conv}]_{cls+reg}$	79.67
$[(x, y, w, h, \theta)_{conv}]_{cls} + [(x, y, w, h, \theta)_{conv}]_{reg}$	79.76
$[(x, y, w, h, \theta)_{conv}]_{cls} + [(x, y, w, h)_{conv}]_{loc} + [(\theta)_{conv}]_{ang}$	**79.83**

4.3.3. Ablation Study on Phase-Shift Steps

In the rotation angle encoder and decoder, the angle value $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$. To ensure the uniqueness of angle values within this range, the phase-shift step size N_{step} must be an integer greater than or equal to 3. Therefore, we evaluated several common step values. We conducted experiments on the DOTAv1.0 dataset, and the results are shown in Table 4. It can be observed that the phase-shift step count has a limited impact on the results, and performance does not improve with increasing step numbers. Considering computational complexity, we selected $N_{step} = 3$ in this study.

Table 4. Results of the angular encoder with different phase-shift steps. (The best-performing outcomes are highlighted in bold).

Metrics	$N_{step} = 3$	$N_{step} = 4$	$N_{step} = 5$
DOTAv1.0 (mAP_{50})	**79.83**	79.75	79.68
DOTAv1.0 (mAP_{75})	57.28	**57.40**	57.36
DOTAv1.0 ($mAP_{50:95}$)	51.70	**51.74**	51.61

4.4. Main Results

4.4.1. Comparisons with Different Backbone Models

In this section, we compare MSSO with several other backbone networks for rotated object detection in remote sensing images on the DOTAv1.0 dataset. The compared models include the classic ResNet50 backbone, rotation convolution-based ReR50 backbone, and large-kernel convolution-based LSKNet-S. As evidenced by the comparison results, MSSO achieves the highest mAP (79.83%) while maintaining the second-smallest model size (13.5 M parameters). The results are shown in Table 5.

Table 5. Comparison of different backbone models on DOTAv1.0 dataset (with single-scale training and testing). Params (#P) and FLOPs are computed for backbone only. (The best-performing outcomes are highlighted in bold).

Model (Backbone)	#P	FLOPs	mAP (%)
ResNet50 [41]	23.3M	86.1	75.87
ReR50 [9]	**12.0M**	-	76.25
LSKNet-S [16]	14.4M	54.4	77.49
MSSO-Net	13.5M	**52.5**	**79.83**

4.4.2. Experimental Results and Analysis on DOTAv1.0

Table 6 presents the results of our proposed SODE-Net compared with other remote sensing image rotation detection methods on the DOTAv1.0 dataset. SODE-Net achieved state-of-the-art results with mAPs of 63.33%, 86.15%, 74.20%, and 79.46% for bridges (BR), large vehicles (LV), soccer ball fields (SBF), and helicopters (HC), respectively, and it also obtained the best average result across all categories at 81.97%. Additionally, SODE-Net has the smallest number of parameters (#P), and its computational cost (FLOPs) is also better than that of most models. As seen in Table 6, apart from our method, LSKNet-S achieves the best performance in the task of remote sensing rotated object detection, as it effectively captures long-range dependencies of large-sized objects and high-aspect-ratio targets through large-kernel convolutions. However, LSKNet-S introduces too much background noise by expanding the receptive field, and it also ignores the defect of the discontinuity of regression loss caused by sudden angular jumps. SODE-Net addresses these limitations, resulting in a 0.33% improvement in mAP.

Table 6. Performance comparisons on DOTAv1.0 (* denotes multi-scale training and testing, bold indicates the optimal value, and underline indicates the suboptimal value).

Methods	Model	#P	FLOPs	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
Rol Trans. [4]	R101	55.1 M	200 G	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
DAL [42]	R101	-	-	88.68	76.55	45.08	66.80	67.00	76.76	79.74	90.84	79.54	78.45	57.71	62.27	69.05	73.14	60.11	71.44
SLA [43]	R50	-	-	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
CFC-Net [44]	R101	-	-	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
O-RCNN [39]	R50	41.1 M	199 G	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87
S^2ANet [45]	R50	38.5 M	-	88.89	83.60	57.74	81.95	79.94	83.19	**89.11**	90.78	84.87	87.81	70.30	68.25	78.30	77.01	69.58	79.42
ReDet [9]	ReR50	34.0 M	-	88.81	82.48	60.83	80.82	78.34	86.06	88.31	90.87	**88.77**	87.03	68.65	66.90	79.26	79.71	74.67	80.10
TIOE-Det [46]	R50	41.1 M	198 G	89.76	85.23	56.32	76.17	80.17	85.58	88.41	90.81	85.93	87.27	68.32	70.32	68.93	78.33	68.87	78.69
RTMDet-R [47]	RTMDet-R	52.3 M	205 G	88.01	86.17	58.54	82.44	81.30	84.82	88.71	90.89	**88.77**	87.37	71.96	71.18	**81.23**	81.40	77.13	81.33
KFIoU [48]	Swin-T	58.8 M	206 G	89.44	84.41	62.22	82.51	80.10	86.07	88.68	90.90	87.32	88.38	72.80	71.95	78.96	74.95	75.27	80.93
AO2-DETR [49]	R50	74.3 M	304 G	**89.95**	84.52	56.90	74.83	80.86	83.47	88.47	90.87	86.12	**88.55**	63.21	65.09	79.09	**82.88**	73.46	79.22
LSKNet-S [16]	LSK-S	31.0 M	**161 G**	89.57	**86.34**	63.13	**83.67**	**82.20**	86.10	88.66	90.89	88.41	87.42	71.72	69.58	78.88	81.77	76.52	81.64
CPMFNet [18]	R50	42.3 M	218 G	89.53	82.92	54.63	77.58	79.37	84.87	88.23	**90.95**	86.95	84.93	68.20	70.39	76.35	72.01	66.68	78.23
PKINet-S [17]	PKINet-S	30.8 M	190 G	88.85	85.04	56.50	82.87	78.43	84.90	88.00	80.89	85.63	86.44	69.02	**73.58**	78.46	81.67	72.30	80.17
SODE-Net	MSSO-Net	**30.6 M**	178 G	88.77	85.65	55.62	83.61	77.89	84.73	88.06	90.87	86.37	86.53	69.28	73.25	78.75	75.85	72.27	79.83
SODE-Net *	MSSO-Net	**30.6 M**	178 G	89.26	84.72	**63.33**	82.38	82.15	**86.15**	88.93	90.89	88.23	87.86	**74.20**	72.18	78.90	80.86	**79.46**	**81.97**

We visualize some detection results of SODE-Net on DOTAv1.0, as shown in Figure 4. In the first row, first column image, the harbor objects exhibit extreme aspect ratios and diverse orientations. SODE-Net generates rotated bounding boxes that align precisely with the targets, fully enclosing the objects while minimizing background coverage. In the second row, first column image, the vehicles and ships are small in scale and densely docked at the harbor. Nevertheless, SODE-Net achieves accurate detection of these challenging objects. In the third row, first column image, the object is partially obscured by shadows. Meanwhile, in the images from the last two columns, the overall imaging conditions are poor with low contrast. Under these extreme circumstances, SODE-Net successfully detects vehicles, planes, and ships, demonstrating strong robustness and anti-interference capability in complex scenarios.

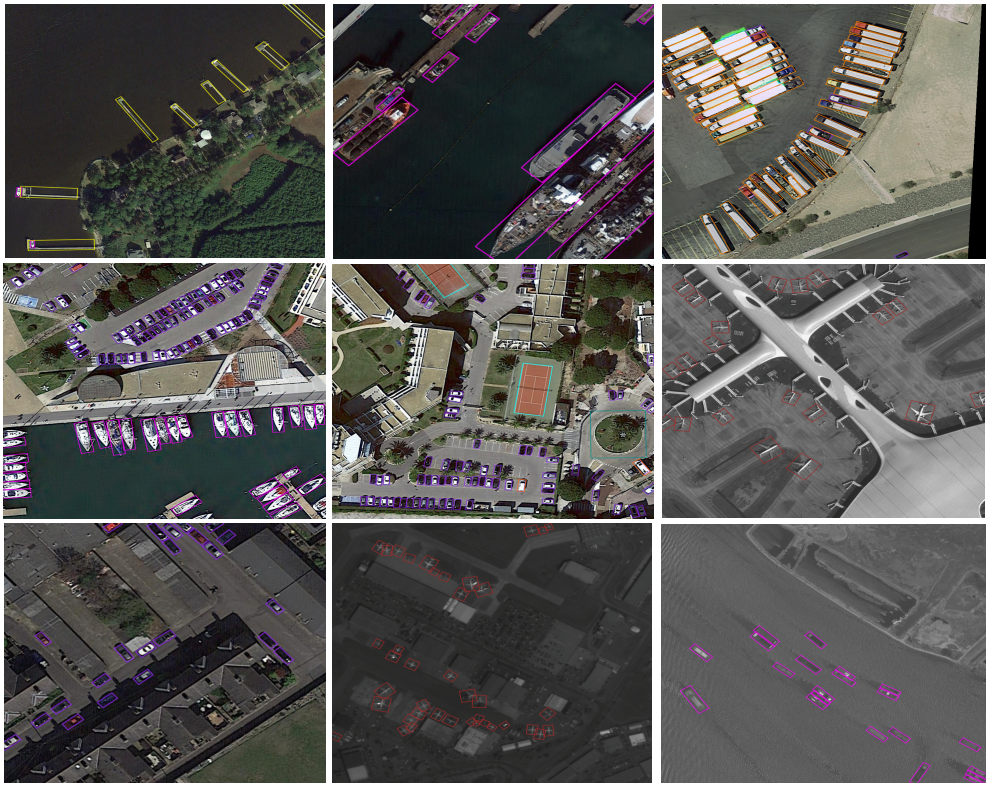

Figure 4. Partial detection results of SODE-Net on the DOTAv1.0 dataset.

These results demonstrate that incorporating MSSO into the backbone for feature extraction plays a critical role in detecting high-aspect-ratio objects that are prevalent in remote sensing images. SODE-Net decouples the classification, localization, and angle regression tasks, leading to more accurate predictions of object categories and rotated bounding box regression. Moreover, by phase-continuous encoding for angle regression, we ensure continuous angle prediction and further stabilize training. The synergistic effect of these two modules enables SODE-Net to achieve superior performance in rotated object detection for remote sensing image.

4.4.3. Experimental Results and Analysis on HRSC2016

Table 7 presents the detection results of our method on the HRSC2016 dataset. As can be observed, SODE-Net outperforms all other methods with a mAP of 91.25% under the PASCAL-VOC 2007 metrics. Figure 5 displays some detection results of our method on the HRSC2016 dataset. It can be seen that the HRSC2016 dataset primarily contains rotated multi-oriented ship targets with extreme aspect ratios. Under challenging conditions such as highly complex image backgrounds (first row, first column), similar appearances are

observed between the background and targets (first row, third column), and low color contrast (second row, third column), SODE-Net successfully distinguishes the targets from the background. This demonstrates that our proposed feature extraction method effectively separates background and target features, enabling the network to more accurately estimate target scales and boundaries.

Table 7. Comparisons of mAP with state-of-the-art methods on HRSC2016. (The best-performing outcomes are highlighted in bold).

Methods	Backbone	Input Size	mAP (%)
Rol Transformer [4]	ResNet101	512 × 800	86.20
DAL [42]	ResNet101	416 × 416	88.95
SLA [43]	ResNet101	768 × 768	89.51
CFC-Net [44]	ResNet101	800 × 800	89.70
O-RCNN [39]	ResNet101	1333 × 800	90.50
S^2ANet [45]	ResNet101	512 × 800	90.17
ReDet [9]	ReR50	512 × 800	90.46
G-Rep [50]	Swin-T	800 × 800	89.46
TIOE-Det [46]	ResNet101	800 × 800	90.16
LSKNet-S [16]	LSKNet-S	800 × 800	90.65
CPMFNet [18]	ResNet50	800 × 800	90.62
PKINet-S [17]	PKINet-S	800 × 800	90.70
SODE-Net	MSSO-Net	800 × 800	**91.25**

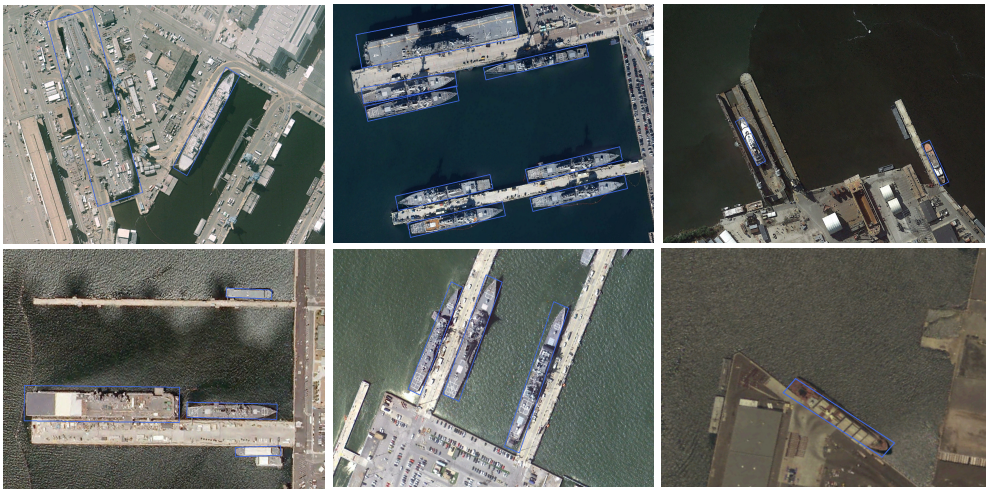

Figure 5. Partial detection results of SODE-Net on HRSC2016 dataset.

4.4.4. Experimental Results and Analysis on UCAS-AOD

Table 8 presents the detection results of the selected comparison methods and SODE-Net on the UCAS-AOD dataset. As shown, SODE-Net achieves a mAP of 91.06%, outperforming other methods. Figure 6 illustrates some detection results of SODE-Net on the UCAS-AOD dataset. It can be observed that the UCAS-AOD dataset contains airplanes and cars with varying orientations. In scenarios with densely arranged and arbitrarily oriented objects (first row, first column) and low color contrast (second row, first column), our method successfully distinguishes targets from the background and accurately localizes them with precise bounding boxes.

Table 8. Comparisons of mAP with state-of-the-art methods on UCAS-AOD. (The best-performing outcomes are highlighted in bold).

Methods	Backbone	Input Size	mAP (%)
Faster RCNN [1]	ResNet50	800 × 800	88.36
Rol Transformer [4]	ResNet50	800 × 800	89.02
DAL [42]	ResNet50	800 × 800	89.87
SLA [43]	ResNet50	800 × 800	89.44
CFC-Net [44]	ResNet50	800 × 800	89.49
TIOE-Det [46]	ResNet50	800 × 800	89.49
RIDet-O [51]	ResNet50	800 × 800	89.62
S^2ANet [45]	ResNet50	800 × 800	89.99
G-Rep [50]	Swin-T	800 × 800	90.16
SODE-Net	MSSO-Net	800 × 800	**91.06**

Figure 6. Partial detection results of SODE-Net on UCAS-AOD dataset.

4.4.5. Experimental Results and Analysis on DIOR-R

Table 9 presents the detection results of the selected comparison methods and SODE-Net on the DIOR-R dataset. As shown, SODE-Net achieves a mAP of 68.45%, outperforming other methods. Apart from our method, CPMFNet achieves the best performance in the task of remote sensing rotated object detection, as it utilize dilated convolutions to dynamically expand the receptive field, thereby enhancing the network's global perception capability for large-scale targets. However, simply increasing the receptive field introduces excessive background noise. SODE-Net effectively mitigates this issue through the integration of strip convolutions, resulting in a 0.60% improvement in mAP. Furthermore, SODE-Net maintains the smallest number of parameters (30.6 M).

Table 9. Comparisons with state-of-the-art methods on DIOR-R. (The best-performing outcomes are highlighted in bold).

Methods	Backbone	#P	mAP (%)
Faster RCNN-O [1]	ResNet50	41.1 M	59.54
Gliding Vertex [52]	ResNet50	-	60.60
TIOE-Det [46]	ResNet50	41.1 M	61.98
Rol Transformer [4]	ResNet101	55.1M	63.87
O-RCNN [39]	ResNet50	41.1 M	64.30
LSKNet-S [16]	LSKNet-S	31.0 M	65.90
CPMFNet [18]	ResNet50	42.3 M	67.85
PKINet-S [17]	PKINet-S	30.8 M	67.03
SODE-Net	MSSO-Net	**30.6 M**	**68.45**

5. Conclusions

This paper proposes SODE-Net, an end-to-end optimized network framework, to address the prevalent challenges in remote sensing detection—insufficient accurate feature extraction for multi-scale and slender targets, as well as feature coupling and discontinuous loss in parameter regression. Within this framework, we first designed a module incorporating spatially orthogonal convolutions for efficient feature extraction. Subsequently, the detection head decouples classification, bounding box position regression, and bounding box angle regression tasks to mitigate feature conflicts between tasks. Finally, a phase-continuous encoding module is introduced to convert the bounding box rotation angle into continuous cosine function values, ensuring loss continuity. Our experiments demonstrate that SODE-Net achieves state-of-the-art performance on three prominent remote sensing datasets.

Author Contributions: Conceptualization, W.X. and X.Y.; methodology, X.Y.; software, X.Y.; validation, L.Y.; formal analysis, W.X.; investigation, L.Y.; resources, W.X.; data curation, K.H.; writing—original draft preparation, X.Y.; visualization, K.H.; supervision, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: DOTA can be obtained from https://captain-whu.github.io/DOTA/index.html (accessed on 28 April 2025). HRSC2016 can be obtained from https://ieee-dataport.org/documents/hrsc2016-0 (accessed on 2 June 2025). UCAS-AOD can be obtained from https://aistudio.baidu.com/datasetdetail/53318 (accessed on 2 June 2025).

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

References

1. 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. [CrossRef]
2. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018 .
3. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
4. 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.
5. 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, Online, 2–9 February 2021; Volume 35, pp. 3163–3171.
6. Liu, J.; Tang, J.; Yang, F.; Zhao, Y. Fast arbitrary-oriented object detection for remote sensing images. *Eur. J. Remote Sens.* **2024**, *57*, 2431006. [CrossRef]
7. 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. [CrossRef]
8. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10015–10024. [CrossRef]
9. 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.
10. 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.
11. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21.
12. Xu, D.; Wu, Y. Improved YOLO-V3 with DenseNet for multi-scale remote sensing target detection. *Sensors* **2020**, *20*, 4276. [CrossRef]
13. Lu, X.; Ji, J.; Xing, Z.; Miao, Q. Attention and feature fusion SSD for remote sensing object detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5501309. [CrossRef]

14. Zhang, T.; Sun, X.; Zhuang, L.; Dong, X.; Sha, J.; Zhang, B.; Zheng, K. AFRE-net: Adaptive feature representation enhancement for arbitrary oriented object detection. *Remote Sens.* **2023**, *15*, 4965. [CrossRef]
15. Zhao, Z.; Li, S. OASL: Orientation-aware adaptive sampling learning for arbitrary oriented object detection. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *128*, 103740. [CrossRef]
16. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large selective kernel network for remote sensing object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 16794–16805.
17. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly kernel inception network for remote sensing detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27706–27716.
18. Bai, P.; Xia, Y.; Feng, J. Composite Perception and Multiscale Fusion Network for Arbitrary-Oriented Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5645916. [CrossRef]
19. Wen, L.; Cheng, Y.; Fang, Y.; Li, X. A comprehensive survey of oriented object detection in remote sensing images. *Expert Syst. Appl.* **2023**, *224*, 119960. [CrossRef]
20. Shi, P.; Zhao, Z.; Fan, X.; Yan, X.; Yan, W.; Xin, Y. Remote sensing image object detection based on angle classification. *IEEE Access* **2021**, *9*, 118696–118707. [CrossRef]
21. Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point set representation for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9657–9666.
22. Zhou, Q.; Yu, C. Point RCNN: An angle-free framework for rotated object detection. *Remote Sens.* **2022**, *14*, 2605. [CrossRef]
23. Guo, Z.; Zhang, X.; Liu, C.; Ji, X.; Jiao, J.; Ye, Q. Convex-hull feature adaptation for oriented and densely packed object detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5252–5265. [CrossRef]
24. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
25. Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; Zhou, J. Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 980–993.
26. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
27. Khalfaoui-Hassani, I.; Pellegrini, T.; Masquelier, T. Dilated convolution with learnable spacings. *arXiv* **2021**, arXiv:2112.03740.
28. Liu, S.; Chen, T.; Chen, X.; Chen, X.; Xiao, Q.; Wu, B.; Kärkkäinen, T.; Pechenizkiy, M.; Mocanu, D.; Wang, Z. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv* **2022**, arXiv:2207.03620.
29. Chen, J.; Hong, H.; Song, B.; Guo, J.; Chen, C.; Xu, J. MDCT: Multi-kernel dilated convolution and transformer for one-stage object detection of remote sensing images. *Remote Sens.* **2023**, *15*, 371. [CrossRef]
30. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 677–694.
31. Qian, W.; Yang, X.; Peng, S.; Yan, J.; Guo, Y. Learning modulated loss for rotated object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 19–21 May 2021; Volume 35, pp. 2458–2466.
32. 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, Virtually, 19–21 May 2019; pp. 8232–8241.
33. Yu, Y.; Da, F. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13354–13363.
34. 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.
35. 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.
36. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017; SciTePress: Setúbal, Portugal, 2017; Volume 2, pp. 324–331.
37. 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; IEEE: Piscataway, NJ, USA, 2015; pp. 3735–3739.
38. 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]
39. 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, BC, Canada, 11–17 October 2021; pp. 3520–3529.

40. Zeng, Y.; Chen, Y.; Yang, X.; Li, Q.; Yan, J. ARS-DETR: Aspect ratio-sensitive detection transformer for aerial oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5610315. [CrossRef]
41. 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.
42. 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, Virtually, 19–21 May 2021; Volume 35, pp. 2355–2363.
43. 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. [CrossRef]
44. 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. [CrossRef]
45. 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]
46. Ming, Q.; Miao, L.; Zhou, Z.; Song, J.; Dong, Y.; Yang, X. Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 241–255. [CrossRef]
47. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. Rtmdet: An empirical study of designing real-time object detectors. arXiv 2022. *arXiv* **2023**, arXiv:2212.07784.
48. Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; Tian, Q. The KFIoU loss for rotated object detection. *arXiv* **2022**, arXiv:2201.12558.
49. Dai, L.; Liu, H.; Tang, H.; Wu, Z.; Song, P. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2342–2356. [CrossRef]
50. Hou, L.; Lu, K.; Yang, X.; Li, Y.; Xue, J. G-rep: Gaussian representation for arbitrary-oriented object detection. *Remote Sens.* **2023**, *15*, 757. [CrossRef]
51. Ming, Q.; Zhou, Z.; Miao, L.; Yang, X.; Dong, Y. Optimization for oriented object detection via representation invariance loss. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19* , 8021505. [CrossRef]
52. 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.* **2020**, *43*, 1452–1459. [CrossRef]