remote sensing

MDPI

Article

SREDet: Semantic-Driven Rotational Feature Enhancement for Oriented Object Detection in Remote Sensing Images

Zehao Zhang [1], Chenhan Wang [1], Huayu Zhang [1], Dacheng Qi [1], Qingyi Liu [2], Yufeng Wang [2] and Wenrui Ding [2,*]

[1] School of Electronic Information Engineering, Beihang University, Beijing 100191, China; zhangzehao@buaa.edu.cn (Z.Z.); wangchenhan1023@bupt.edu.cn (C.W.); huayuzhang@buaa.edu.cn (H.Z.); dc_qi@buaa.edu.cn (D.Q.)

[2] Institute of Unmanned System, Beihang University, Beijing 100191, China; lqy671@buaa.edu.cn (Q.L.); wyfeng@buaa.edu.cn (Y.W.)

* Correspondence: ding@buaa.edu.cn

Abstract: Significant progress has been achieved in the field of oriented object detection (OOD) in recent years. Compared to natural images, objects in remote sensing images exhibit characteristics of dense arrangement and arbitrary orientation while also containing a large amount of background information. Feature extraction in OOD becomes more challenging due to the diversity of object orientations. In this paper, we propose a semantic-driven rotational feature enhancement method, termed SREDet, to fully leverage the joint semantic and spatial information of oriented objects in the remote sensing images. We first construct a multi-rotation feature pyramid network (MRFPN), which leverages a fusion of multi-angle and multiscale feature maps to enhance the capability to extract features from different orientations. Then, considering feature confusion and contamination caused by the dense arrangement of objects and background interference, we present a semantic-driven feature enhancement module (SFEM), which decouples features in the spatial domain to separately enhance the features of objects and weaken those of backgrounds. Furthermore, we introduce an error source evaluation metric for rotated object detection to further analyze detection errors and indicate the effectiveness of our method. Extensive experiments demonstrate that our SREDet method achieves superior performance on two commonly used remote sensing object detection datasets (i.e., DOTA and HRSC2016).

Keywords: oriented object detection; remote sensing images; feature enhancement; error diagnosis

Citation: Zhang, Z.; Wang, C.; Zhang, H.; Qi, D.; Liu, Q.; Wang, Y.; Ding, W. SREDet: Semantic-Driven Rotational Feature Enhancement for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 2317. https://doi.org/10.3390/rs16132317

Academic Editor: Shaohui Mei

Received: 24 May 2024
Revised: 19 June 2024
Accepted: 21 June 2024
Published: 25 June 2024

1. Introduction

Oriented object detection in remote sensing images aims to utilize rotated bounding boxes to accurately determine the position and category of the object of interest [1,2]. It has gradually evolved into a significant domain within computer vision [3] and serves as a foundation task for various applications, such as smart cities, maritime rescue, and battlefield surveillance [4–9]. Due to the characteristics of overhead perspective and remote photography [10], remote sensing images typically have several characteristics: (1) objects are distributed with arbitrary orientations and variant appearances; (2) dense small-scale objects, such as vehicles and ships, often tend to cluster together closely; and (3) remote sensing images contain a significant amount of background information.

Analyzing the first characteristic of remote sensing imagery, regular convolutional networks cannot guarantee the precision of features when the object is rotated [11], as shown in Figure 1A. We input an image and its 90-degree-rotated version into the network, and the resulting feature map visualizations are depicted in Figure 1(A2). We observe that the feature maps exhibit accurate and well-represented responses under normal input conditions. However, when the object is rotated, the extracted features show missing components and weakened responses. The last two characteristics of remote sensing images introduce noise

in object detection, including both interferences between objects and background noise, as shown in Figure 1B. Dense arrangements of objects may encounter interclass feature coupling and intraclass feature boundary blurring, leading to less prominent feature responses for some objects, as seen in the yellow circle of Figure 1(B2). Using the effective objects in DOTAv-1.0 [12] as reference, we leverage the corresponding segmentation information from iSAID [13] to perform pixel-wise analysis, where the background pixels account for 96. 95% of the total. The abundance of background information may cause the similarity of background areas to the erroneously activated objects, as observed in the red circle of Figure 1(B2).

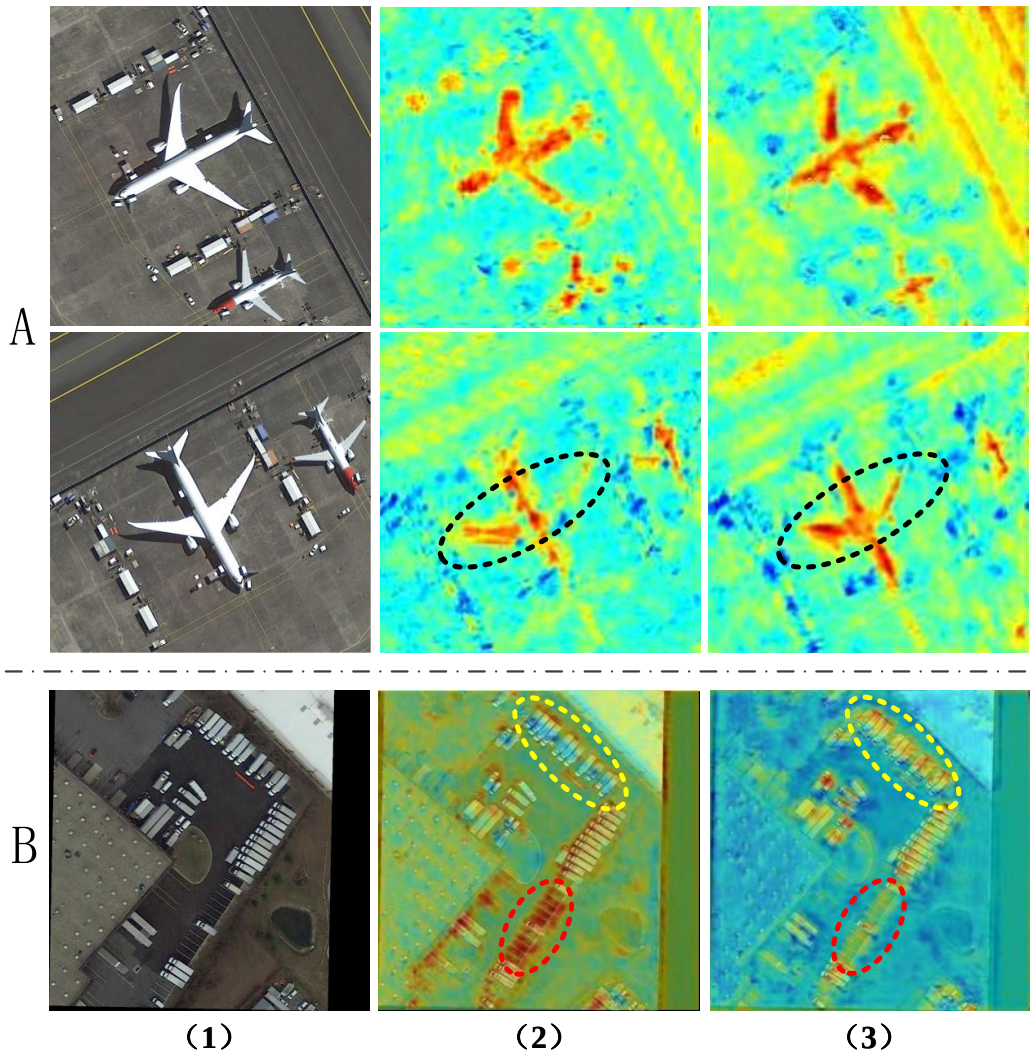

Figure 1. Challenges in Feature Extraction: Poor Rotation Handling, Feature Overlapping, Enhancement Errors, and Weak Responses. Columns indicate the images (**left**) and their feature maps produced by RetinaNet (**middle**) and our model (**right**). Specifically, (**A1**,**B1**) represent the images selected from the DOTA dataset, (**A2**,**B2**) represent the feature maps generated by ResNet50+FPN and (**A3**,**B3**) represent the feature maps extracted by the ResNet50 + MRFPN and ResNet50 + FPN + SFEM variants of our method.

Currently, rotation-invariant feature extraction methods focus on two main approaches. One involves improving the network extraction structure, such as designing rotation-invariant network architectures using group convolution networks [14]. However, these methods require complex network design and are challenging to train. The other approach involves integrating feature maps from different angles [15], but the feature maps lack semantic information communication between different scales, leading to low information

utilization. For rotation object detection, in addition to extracting rotation-invariant features, optimizing the feature maps is also crucial. Representative feature enhancement and attention mechanism methods can be categorized into three aspects. Firstly, effective attention mechanisms, such as channel attention and spatial attention, are introduced to focus on the salient features of the object, addressing noise and boundary blur problems [16,17]. However, utilizing information from feature pooling operations to generate weights and reconstruct feature maps does not guarantee the reliability of the weighting. This approach may still result in the activation of incorrect channels or spatial locations. Additionally, supplementary supervision, such as object box boundaries, center points, or masks, can be helpful to strengthen features [18–20]. However, these supervision signals may not be comprehensive enough, leading to feature overlap issues. Finally, considering the distinctions between the regression and classification tasks within the detection heads, it can be effective to design different feature activation methods to decouple features and address feature incompatibility problems [21]. However, this approach may fail to suppress interference from background noise, resulting in many false positives during the detection process.

In this paper, to further address the aforementioned issues, we design a multi-rotation feature pyramid network (MRFPN) architecture for oriented object detection in remote sensing imagery. This architecture enhances the rotation-invariant characteristics of objects while strengthening the contextual information and semantic consistency between feature maps by acquiring a more comprehensive set of rotation-invariant feature fusion across various rotation angles and scales. The feature response from our proposed architecture is depicted in Figure 1(A3). Furthermore, we introduce a novel component, named the Semantic-driven Feature Enhancement Module (SFEM), to obtain more precise and reliable semantic information to enhance feature maps. It approximates the decoupling of features from different object categories and enforces constraints, achieving feature denoising in the spatial domain. This component reduces inter-class feature coupling and intra-class interference and alleviates background interference to achieve robust rotation detection, as illustrated in Figure 1(B3). Finally, we propose a new evaluation metric for oriented object detection to assess the model response to different types of errors, further demonstrating the effectiveness of the proposed modules. The main contributions of this work are summarized as follows:

- We propose a semantic-driven rotational feature enhancement method for oriented object detection, effectively addressing the significant rotational feature variations and complex backgrounds in remote sensing object detection.
- We introduce a multi-rotation feature pyramid network to extract rotation-invariant features and maintain the consistency of multiscale semantic information. This module utilizes multi-angle and multiscale feature maps combined with deformable convolutions to represent remote sensing objects.
- We innovatively integrate the semantics information into oriented object detection by designing the semantic-driven feature enhancement module in an implicit supervision paradigm. It enhances features along the channel and spatial dimensions, effectively addressing inter-class coupling and background interference in feature maps.
- We introduce a novel evaluation metric for oriented object detection that refines different error types, which can reflect the sensitivity of the model to various types of errors. Extensive experiments demonstrate the superiority of the proposed method.

2. Related Work

2.1. Arbitrary Oriented Object Detection

In recent years, object detection in remote sensing images has become increasingly popular. Unlike general object detection, objects in remote sensing images can be oriented in arbitrary directions. With the continuous advancement of deep learning technology, many excellent methods have emerged to detect rotated objects [22]. Oriented R-CNN [23] employs a novel box encoding system called midpoint offset to constrain directed candidate areas effectively, while R3Det [24] addresses feature misalignment during refinement with

a feature refinement module. Yolov8 [25] designed an entirely new backbone extraction network, significantly enhancing the ability to extract target features and improving performance in remote sensing rotation object detection tasks. Zhang et al. [26] and Yang et al. [27] integrate image super-resolution methods for detecting small objects within vast backgrounds, even on low-resolution inputs. Additionally, an adaptive detection system based on early exit neural networks [28] reduces training costs by allowing high-confidence samples to exit the model early, thus improving the efficiency of detecting complex remote sensing images. Besides convolutional networks, transformer architectures have also made significant contributions. Ma et al. [29] are the first to attempt and implement an end-to-end transformer-based framework for oriented object detection. Building on the DERT network, Carion et al. [30] and Dai et al. [31] propose an adaptive oriented proposal refinement module, which effectively enhanced the capability of rotation detection in remote sensing targets. Additionally, Yu et al. [32] introduce a method called spatial transformation decoupling, providing a simple yet effective solution for oriented object detection using the ViT framework.

2.2. Rotation Invariant Feature Extraction

The varied orientations of objects in remote sensing imagery highlight the crucial need for extracting rotation-invariant features. To tackle this challenge, researchers have proposed two main approaches. The first involves modifying the convolution operation itself to enable the extraction of rotation-sensitive features. Cohen et al. [33] first propose the concept of group convolution, integrating four-fold rotational equivariance into CNNs. Hoogeboom et al. [34] expand group convolution to hexagonal lattices, incorporating sixfold rotational equivariance. This adaptation enables more efficient handling of rotations, improving feature recognition across various orientations. Following this, ReDet [14] constructed a backbone to extract rotation-invariant features. Pu et al. [35] design adaptive rotated convolution, which can adaptively rotate to effectively extract target features. Mei et al. [36] propose using polar coordinate transformation to convert rotational changes into translational changes, thereby mitigating the rotation sensitivity issue in CNN networks. The second approach extracts rotation-invariant features through feature mappings with rotational channels. Han et al. [37] and Deng et al. [11] utilize convolutional kernels at different angles to generate feature maps in various directions, thereby enriching the orientation information represented in the feature maps. Zheng et al. [38] propose an object-oriented rotation-invariant semantic representation framework to guide the network in learning rotation-invariant features. Finally, Cao et al. [15] construct a rotation-invariant spatial pooling pyramid by rotating feature maps to extract rotation-invariant features.

2.3. Semantic Information Feature Enhancement

Remote sensing images often include complex background details that can introduce noise into feature maps, potentially impacting object detection performance. Traditional channel or spatial attention mechanisms may not always accurately enhance regions corresponding to actual objects. To overcome this challenge, several approaches have been developed that leverage semantic information to enhance feature maps. Yang et al. [19] and Li et al. [39] utilize binary masks as supervisory information to spatially weight feature maps according to predicted probability maps, aiming to focus the model's attention on relevant areas. Yu et al. [7] use a deep segmentation network to enhance the relationship between roads and vehicles, incorporating this into a visual attention mechanism with spatiotemporal constraints to detect small vehicles. Correspondingly, Yang et al. [40] introduce multi-mask supervision to implicitly generate weight information, decoupling the features of different objects. Song et al. [20] use regions enclosed by the midpoints of the edges of the object's bounding box as masks, acquiring weight information for object and non-object areas through supervised learning of spatial feature encoding. Cao et al. [15] develope semantic edge supervision features using object box boundary information, effectively addressing the challenges of complex backgrounds and the lack of contextual cues in remote

sensing object detection. Liu et al. [18] transform object boxes into two-dimensional Gaussian expressions to obtain center and boundary masks, enhancing the network to improve object localization accuracy while suppressing interference from complex backgrounds. Finally, Zhang et al. [41] introduce a multistage enhancement network that enhances tiny objects at both the instance level and the feature level across different stages of the detector.

3. Method

The proposed SREDet is based on a fundamental single-stage detector. The complete framework, as depicted in Figure 2, consists of four main components: a feature extraction backbone, the MRFPN for extracting rotation-invariant features across multiple angles and scales, the SFEM for feature enhancement driven by semantic segmentation information, and the oriented detection head for classification and regression tasks.

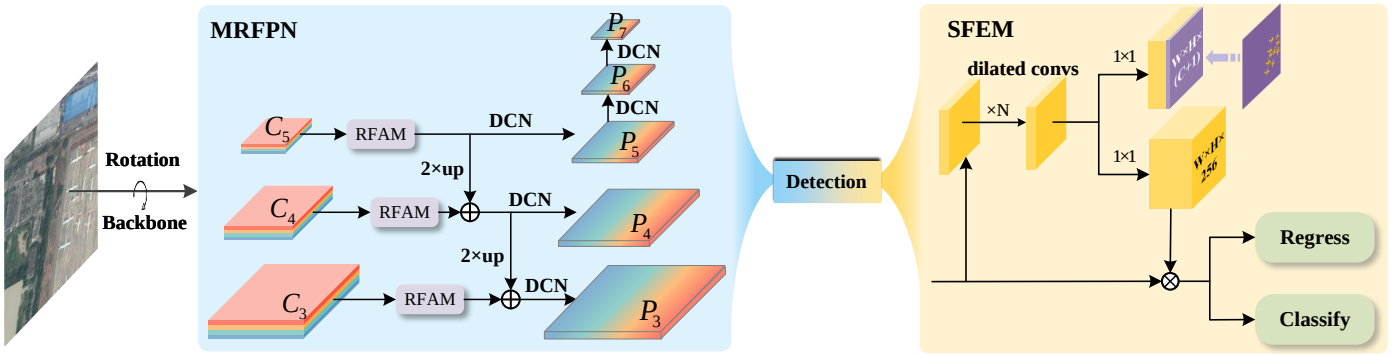

Figure 2. Overall architecture of the proposed SREDet model. SREDet primarily consists of four parts: First, the backbone feature extraction network is utilized for initial feature extraction. Subsequently, multi-angle (different colors represent different angle feature maps) and multiscale feature maps are fused through the MRFPN to extract rotation-invariant features. Features are then fed into the SFEM module to suppress background noise and enhance foreground objects. Finally, the processed features are passed to both classification and regression heads to obtain oriented bounding box prediction results.

3.1. Multi-Rotation Feature Pyramid Network

We consider the distinct characteristics of remote sensing imagery, specifically the fact that objects exhibit arbitrary directional features in the overhead view. Therefore, we believe the extraction of rotation-invariant features of objects essential for oriented object detection in remote sensing applications. Traditional convolutional neural networks cannot directly extract features that remain consistent across object rotations, leading to discrepancies in the features extracted from the same object with different angles. To overcome this limitation, two primary strategies have been developed: Deng et al. [11] and Weiler et al. [42] modify the convolution operation architecture to support the extraction of rotation-invariant features, and Han et al. [37] integrate multi-angle features to augment directional information and extract rotation-invariant characteristics.

In response to the previously mentioned issue, we propose a feature pyramid network that integrates multiscale and multi-angle features, as shown in Figure 2. This network aims to reduce discrepancies in feature extraction for the same object from different angles and to approximate rotation-invariant features as closely as possible. Let I be the input image. The output after passing through the backbone extraction network is as follows:

$$F_{\theta_i} = B(T_{\theta_i}(I)), \tag{1}$$

where represents the backbone network, θ_i represents different rotation angles, and T represents the rotation operation.

After obtaining multi-angle feature maps through the backbone network, to ensure feature map consistency, we designed a rotation feature alignment module (RFAM) as

seen in Figure 3. This module maps rotated features back to their original states and concatenates n branches together in the channel dimension. Finally, the features are fused using convolution with a kernel size of 1:

$$C_i = Conv(Concat[T_{-\theta_1}(F_{\theta_1})_i, T_{-\theta_2}(F_{\theta_2})_i, \cdots, T_{-\theta_n}(F_{\theta_n})_i])(i = 3, 4, 5). \quad (2)$$

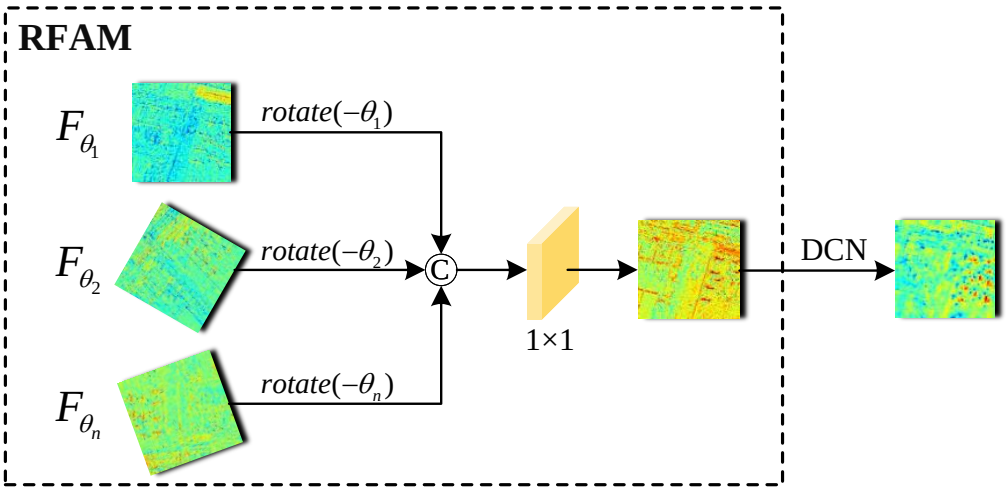

Figure 3. Structure of Rotation Feature Alignment Module. This module maps the features from different orientations back to the original direction, and extracts features more closely aligned with the object through deformable convolution.

To merge semantic information from different levels and extract features of objects with varying aspect ratios and shapes, we enhanced the feature pyramid network (FPN) [43] architecture by substituting the conventional convolutional layers with deformable convolutions. The outputs of different levels can be represented by the following formulas:

$$\begin{aligned} P_i &= DCN(C_i + Interpolation(C_{i+1}))(i = 3, 4), \\ P_5 &= DCN(C_5), \\ P_i &= DCN(P_{i-1})(i = 6, 7), \end{aligned} \quad (3)$$

where DCN represents deformable convolution and p_i represents the outputs of MRFPN.

3.2. Semantic-Driven Feature Enhancement Module

In remote sensing scenarios, background information is abundant, which can lead to inadvertent amplification of features similar to certain categories of objects. This phenomenon, in turn, generates a significant number of false positive samples during detection. Furthermore, the characteristic presentation of objects as densely packed small objects in remote sensing imagery leads to mutual interference among object features. This interference culminates in the blurring of feature maps and a diminished activation level.

To enhance feature maps, attention mechanisms such as channel attention, spatial attention, and hybrid attention are commonly employed to reweight the feature maps, thus highlighting significant areas while suppressing irrelevant ones. However, this approach is predicated on computing responses based on the spatial and channel characteristics of the feature maps and does not guarantee the effectiveness and reliability of the areas being enhanced or suppressed. To further enhance the reliability of the regions being augmented, Li et al. [39] and Cao et al. [44] utilize mask information obtained from bounding boxes to assist in the enhancement of feature maps. Yang et al. [19] employe an explicit feature map enhancement approach, whereby the probability predicted by the mask is directly multiplied on the original feature maps. In contrast, yang et al. [40] adopt an implicit feature map enhancement method, using convolution to generate weights from the feature maps of the layer preceding the mask prediction, which are the same dimensions as the

original feature maps, and then apply these weights to the original feature maps, as seen in Figure 4. In this paper, we define the method of directly multiplying the predicted semantic information probabilities with the feature maps in the spatial domain as explicit feature enhancement. Conversely, the approach of generating a set of weights from the semantic information feature maps and then weighting the feature maps accordingly is defined as implicit enhancement.

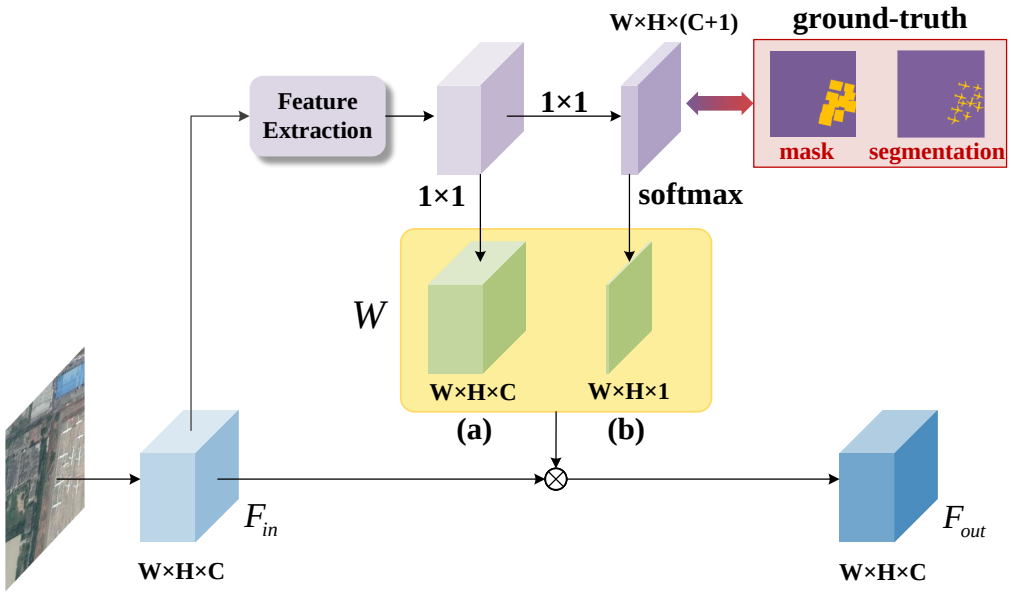

Figure 4. Different Semantic Formats and Enhancement Strategies. This figure shows two types of semantic annotation and two distinct enhancement methods, where (**a**,**b**) demonstrate the implicit and explicit enhancement, respectively. F_{in} and F_{out} represent the feature map before and after enhancement, and W indicates the weights generated by different strategies.

However, we believe that the use of bounding boxes to generate mask information still presents certain inadequacies:

- The overlap of bounding boxes for objects can still lead to mixing features within and between classes.
- The shape of some objects cannot be closely aligned with the bounding boxes, resulting in masks that incorporate excessive background information. This not only complicates the task of mask prediction but may also inadvertently enhance certain background regions.

We use semantic segmentation information to resolve the aforementioned issues and employ an implicit feature map enhancement approach. The features of objects belonging to different categories are decoupled into their respective channels and the features of both objects and backgrounds are separately enhanced and weakened within the spatial domain. The architecture of SFEM is illustrated in the provided Figure 2. To enhance the network's accuracy in predicting semantic segmentation without incurring additional computational costs, we employ dilated convolutions to expand the receptive field of the feature maps, thereby furnishing more abundant semantic information. The process of feature extraction can be represented as follows:

$$F' = conv_{d_n}(\cdots conv_{d_1}(F, W_1, b_1)\cdots, W_n, b_n), \tag{4}$$

we use convolution with a kernel size of 1 to adjust the number of channels and employ a sigmoid function as the activation function to generate feature weights:

$$F_{out} = \underbrace{sigmoid(conv_{1\times1}(F'))}_{W_{SFEM}} \odot F. \tag{5}$$

From another perspective, it can be considered that the network decouples the features of different categories into their respective channels. Without loss of generality, assuming that the dataset includes a total of L categories and that the test image contains the first L_0 categories, the output can be represented as follows:

$$F_{out} = \bigcup_{i=1}^{L_0} \bigcup_{n=1}^{C_i} w_n^i \odot x_n^i \cup \bigcup_{j=L_0+1}^{L} \bigcup_{m=1}^{C_j} w_m^j \odot x_m^j \cup \bigcup_{t=1}^{C_{bg}} w_t^{bg} \odot x_t^{bg}, \tag{6}$$

where $F_{out} \in \mathbb{R}^{C\times H\times W}$ is the element-wise product. C_i represents the number of channels belonging to the i-th category, and w_n^i and x_n^i denote the weight and feature of the i-th category along the n-th channel. The meanings of the remaining symbols can be deduced following the same logic.

3.3. Identifying Oriented Object Detection Errors

The primary evaluation metric for oriented object detection in remote sensing images is the mean average precision (mAP). Although mAP succinctly summarizes model performance, it is challenging to discern what errors constrain the model's performance. For example, a false positive may result from misclassification, incorrect orientation, inaccurate localization, or background confusion. Inspired by Bolya D [45], we introduce oriented object detection errors in rotation detection.

3.3.1. Defining Main Error Types

To comprehensively assess the error distribution within the component mAP, false positive and false negative samples are classified into five distinct types, as shown in Figure 5. We use the rotational intersection-over-unit ($RIoU$) metric to quantify the overlap between two rotated bounding boxes, where $RIoU_{\max}$ denotes the maximum $RIoU$ between a false positive sample and its corresponding ground truth(GT). Additionally, t_b represents the threshold for background objects, conventionally set to 0.1, while t_f signifies the threshold for foreground objects. Since we primarily focus on the mAP_{50} metric of the model, t_f is generally set to 0.5 unless otherwise noted.

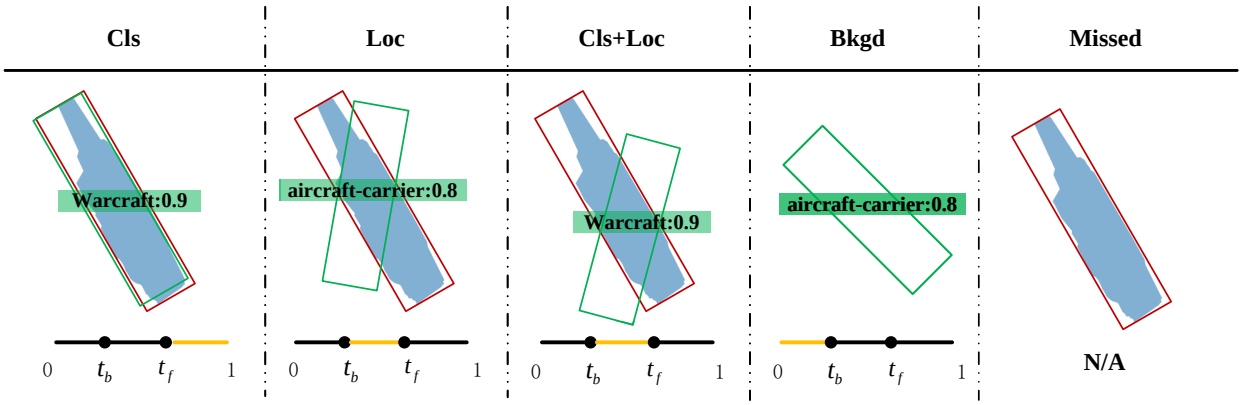

Figure 5. Definition of error types. Red boxes denote the GT of the object, green boxes represent false positive samples, and the actual situation of $RIoU$ for each error type is indicated by yellow highlighted line segments.

- **Classification Error:** $RIoU_{\max} \geq t_f$, but the predicted category is incorrect.
- **Localization Error:** The predicted category is correct, but $t_b \leq RIoU_{\max} \leq t_f$.

- **Cls and Loc Error:** $t_b \leq RIoU_{\max} \leq t_f$, and the predicted category is incorrect.
- **Background Error:** The background was falsely detected as the object, $RIoU_{\max} \leq t_b$.
- **Missed Error:** All undetected GT instances.

3.3.2. Setting Evaluation Metrics

Simply counting the numbers of each type of error does not scientifically demonstrate its impact on the performance of the model. To evaluate the impact of each type of error on the model, we will modify the errors by type and recalculate the mean average precision to obtain ΔmAP. The error between mAP_{mod} and the original mAP will serve as the evaluation metric:

$$\Delta mAP = mAP_{\text{mod}} - mAP. \tag{7}$$

We will modify each error type according to the following procedure:

- **Modify Classification Error:** Modify the predicted incorrect categories to the correct categories. If duplicate detections occur, remove object boxes with low confidence.
- **Modify Localization Error:** Replace the predicted object boxes with the corresponding GT object boxes. If duplicate detections occur, remove object boxes with low confidence.
- **Modify Cls and Loc Error:** Due to the inability to determine which GT object box matches the predicted object box, remove it from false positives.
- **Modify Background Error:** Remove all prediction boxes that misclassify background as objects.
- **Modify Missed Error:** When calculating mAP, subtract the number of ground truths missed from the total GT. From another perspective, it can be said that the model has performed precise detection on all missed objects.

3.4. Loss Function

Our loss function mainly consists of three components; besides the classification and regression losses from the original single-stage detector, we incorporate a semantic segmentation task loss to supervise the SFEM module. Therefore, the total loss definition for SREDet is as follows:

$$\mathcal{L} = \mathcal{L}_{cls}(l_i, l_i^*) + \mathcal{L}_{reg}(t_i, t_i^*) + \mathcal{L}_{seg}(p_i, p_i^*), \tag{8}$$

where $\mathcal{L}_{cls}$ denotes the classification loss, l_i signifies the probability as predicted by the network that an anchor is an object, and l_i^* represents the corresponding ground truth label. Our network employs focal loss [46] as the classification loss. The L1 loss is used as the regression loss $\mathcal{L}_{reg}$, and t_i and t_i^* denote the predicted bounding box and the ground truth bounding box, respectively. Each box is represented in vector form, and the boxes are encoded following the format specified in (9) and (10):

$$\begin{gathered} t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a \\ t_w = \log(w/w_a), t_h = \log(h/h_a), t_\theta = \theta - \theta_a, \end{gathered} \tag{9}$$

$$\begin{gathered} t_x^* = (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a \\ t_w^* = \log(w^*/w_a), t_h^* = \log(h^*/h_a), t_\theta^* = \theta^* - \theta_a. \end{gathered} \tag{10}$$

In the SFEM for the task of semantic segmentation detection, we utilize Dice loss [47]. Given its application across various pyramid layers, the overall semantic segmentation loss is formulated as follows:

$$\mathcal{L}_{seg} = \sum_{i=1}^{S} \varepsilon^i \left(1 - \frac{2 \times \left|p_i^* \cap p_i\right|}{\left|p_i^*\right| + \left|p_i\right|}\right), \tag{11}$$

where S represents the number of feature maps used for supervision, ε^i represents the weight coefficients associated with each feature map. p_i^* represents the set of pixels in

the semantic segmentation mask of ground truth, and p_i represents the set of pixels in the predicted semantic segmentation mask. The intersection $\left|p_i^* \cap p_i\right|$ counts the pixels common to both the prediction and the ground truth, while $\left|p_i^*\right|$ and $|p_i|$ count the pixels in the ground truth and the predicted semantic segmentation masks, respectively.

4. Results

In this section, we provide a comprehensive description of the two datasets employed in our experiments along with a detailed discussion of the principal results obtained. Furthermore, we meticulously outline and analyze the design of our ablation studies, shedding light on their significance and impact.

4.1. Datasets

4.1.1. DOTA and iSAID

DOTA [12] is one of the commonly used datasets for remote sensing image detection, currently available in three versions. Since only the DOTA-v1.0 dataset has been annotated with segmentation information by some scholars to form the iSAID dataset [13], we select DOTA-v1.0 as the experimental data. The dataset comprises 2806 high-resolution aerial images, covering various complex scenes and shooting angles. It contains a rich variety of targets, including planes (PL), baseball diamonds (BD), bridges (BR), ground track fields (GTF), small vehicles (SV), large vehicles (LV), ships (SH), tennis courts (TC), basketball courts (BC), storage tanks (ST), soccer-ball fields (SBF), roundabouts (RA), harbors (HA), swimming pools (SP) and helicopters (HC), totaling 188,282 annotated objects. We used DOTA's standard rotated bounding boxes as a reference and filtered the corresponding segmentation labels from the iSAID dataset. We divided the dataset according to the original DOTA data partitioning method, with the dataset split into 1/2 training set, 1/6 validation set, and 1/3 test set. The results in Table 1 were obtained by training on the training and validation sets and then predicting the test set, with the test results acquired through the official evaluation server. Results in the remaining tables are, by default, based on training on the training set and testing on the validation set.

4.1.2. HRSC2016

HRSC2016 is a publicly available remote sensing dataset specifically designed for ship detection. It includes images from six prominent harbors, featuring two primary scenarios: ships at sea and ships near the shore. The dataset comprises a total of 1061 images and 2976 object instances. The training set contains 436 images, the validation set includes 181 images, and the test set comprises 444 images. Image sizes range from 300 × 300 pixels to 1500 × 900 pixels, with the majority exceeding 1000 × 600 pixels. The original dataset provides ship targets labeled with oriented bounding boxes and we annotated all targets with semantic segmentation to facilitate model training.

4.2. Implementation Details

We conducted experiments on multiple baselines. For one approach, we selected networks such as RetinaNet [46] and Faster R-CNN [48] as baseline networks, using ResNet101 as the default backbone. To maintain consistency, the experiments were trained and tested on the MMrotate platform [49]. We used the SGD optimizer, setting momentum and weight decay to 0.9 and 0.0001, respectively. A MultiStepLR strategy was adopted, starting with a learning rate of 0.0025. The training spanned 24 epochs, with the learning rate automatically reduced to 1/10 of its original value at epochs 16 and 22. Rotated non-maximum suppression was applied to the predicted rotated bounding boxes to minimize redundancy, with a confidence score threshold of 0.1 and an IoU threshold of 0.1. For another approach, we conducted experiments using YOLOv8 as the baseline, employing the default YOLOv8 framework configuration [25]. The initial learning rate was set to 0.01, and the final learning rate was 0.001. The momentum was configured at 0.937, and the weight decay was set to 0.0005. We implemented a warmup period of 3.0 epochs, during

which the initial momentum was set to 0.8 and the initial bias learning rate was 0.1. All training and testing experiments were conducted on an RTX A6000 with a batch size of 2.

DOTA dataset comprises large-scale images, so during the training and testing phases, the images were divided into 1024 × 1024 patches with a 200-pixel overlap. Various data augmentation techniques were employed, specifically including random flipping (with a probability of 0.25), random rotation (with a probability of 0.25), and random color transformation (with a probability of 0.25). When YOLOv8 was being trained, mosaic augmentation was also introduced. Additionally, to further enhance network performance, multiscale training and testing were applied. When training on the HRSC2016 dataset, the number of training epochs was set to 72. The initial learning rate was set to 0.0025, and it was reduced to 0.1 of its original value at epochs 48 and 66. The data processing method used was the same as that applied to the DOTA dataset.

4.3. Main Results

4.3.1. Results on DOTA

We evaluate the proposed method against other state-of-the-art approaches on the DOTA dataset. The results are presented in Table 1. Our method achieved a mAP_{50} of 76.34%, and under the approach of multiscale training and testing, our method achieved a mAP_{50} of 79.31%. When comparing metrics in different categories, our detectors performed the best for the categories of small vehicles, ships, and tennis courts, and achieved the second-best results for the ground field tracks, basketball courts, roundabouts, and helicopters.

Table 1. Comparison to state-of-the-art methods on the DOTA-v1.0 dataset. R-101 denotes ResNet-101 (likewise for R-50 and R-152), RX-101 denotes ResNeXt-101 and H-104 denotes Hourglass-104. The best result is highlighted in bold, and the second-best result is underlined. * denotes multiscale training and multiscale testing.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{50}
FR-O [12]	R-101	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.40	52.52	46.69	44.80	46.30	52.93
RRPN [50]	R-101	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
RetinaNet-R [46]	R-101	88.92	67.67	33.55	56.83	66.11	73.28	75.24	90.87	73.95	75.07	43.77	56.72	51.05	55.86	21.46	62.02
CADNet [51]	R-101	87.80	82.40	49.40	73.50	71.10	63.50	76.60	90.90	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
O2-DNet [52]	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
CenterMap-Net [53]	R-50	88.88	81.24	53.15	60.65	78.62	66.55	78.10	88.83	77.80	83.61	49.36	66.19	72.10	72.36	58.70	71.74
BBAVector [54]	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
SCRDet [19]	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
DRN [55]	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
Gliding Vertex [56]	R-101	89.89	85.99	46.09	**78.48**	70.32	69.44	76.93	90.71	79.36	83.80	57.79	68.35	72.90	71.03	59.78	73.39
SRDF [20]	R-101	87.55	84.12	52.33	63.46	78.21	77.02	**88.13**	90.88	86.68	85.58	47.55	64.88	65.17	71.42	59.51	73.50
R3Det [24]	R-152	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74
FCOSR-S [57]	R-50	89.09	80.58	44.04	73.33	79.07	76.54	87.28	90.88	84.89	85.37	55.95	64.56	66.92	76.96	55.32	74.05
S2A-Net [37]	R-50	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
SCRDet++ [40]	R-101	89.20	83.36	50.92	68.17	71.61	80.23	78.53	90.83	86.09	84.04	65.93	60.80	68.83	71.31	**66.24**	74.41
Oriented R-CNN [23]	R-50	88.79	82.18	52.64	72.14	78.75	**82.35**	87.68	90.76	85.35	84.68	61.44	64.99	67.40	69.19	57.01	75.00
MaskOBB [58]	RX-101	89.56	**89.95**	**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
CBDA-Net [18]	R-101	89.17	85.92	50.28	65.02	77.72	82.32	87.89	90.48	86.47	85.90	**66.85**	66.48	67.41	71.33	62.89	75.74
DODet [59]	R-101	89.61	83.10	51.43	72.02	79.16	81.99	87.71	90.89	86.53	84.56	62.21	65.38	71.98	70.79	61.93	75.89
SREDet(ours)	R-101	89.36	85.51	50.87	74.52	**80.50**	74.78	86.43	**90.91**	87.40	83.97	64.36	69.10	67.72	**73.65**	65.93	**76.34**
SREDet(ours) *	R-101	90.23	86.75	54.34	80.81	80.41	79.37	87.02	90.90	88.28	86.84	70.16	70.68	74.43	76.11	73.42	79.32

This performance can be attributed to the rotation-invariant features' sensitivity to capturing the orientation of objects and the feature enhancement effects realized through semantic information. Represented by its detection capabilities for swimming pools, helicopters, and planes, our method effectively identifies and regresses objects with irregular shapes. This is primarily due to incorporating semantic segmentation information as supervision, allowing the network to focus precisely on the object and contextual features against complex backgrounds, providing more regression clues. Additionally, our method performs well with densely arranged objects, such as cars, benefiting from SFEM which reduces the coupling of intra-class features, thereby highlighting crucial features. We also observed that for ground field tracks, roundabouts, and baseball diamonds, utilizing semantic segmentation information is more efficient than object bounding box masks. The primary reason for this is that masks might include background information or other objects, causing feature confusion or erroneous enhancement. Our method also adeptly handles the challenges posed by arbitrary orientations, irregular shapes, dense arrangements, and varying scales of remote sensing objects, achieving precise rotation object detection.

From the visualized detection images, as seen in Figure 6, it can be observed that our network achieves excellent detection results for various types of objects. As seen in the first row of images, the network can accurately detect harbors of different shapes and sizes. This is primarily due to the MRFEN module's ability to extract features of varying scales and shapes. From the fourth column of images, it is evident that the network exhibits effective detection performance on dense objects, largely attributed to the SFEM, which alleviates the feature overlap among similar objects and enhances the feature maps of small objects.

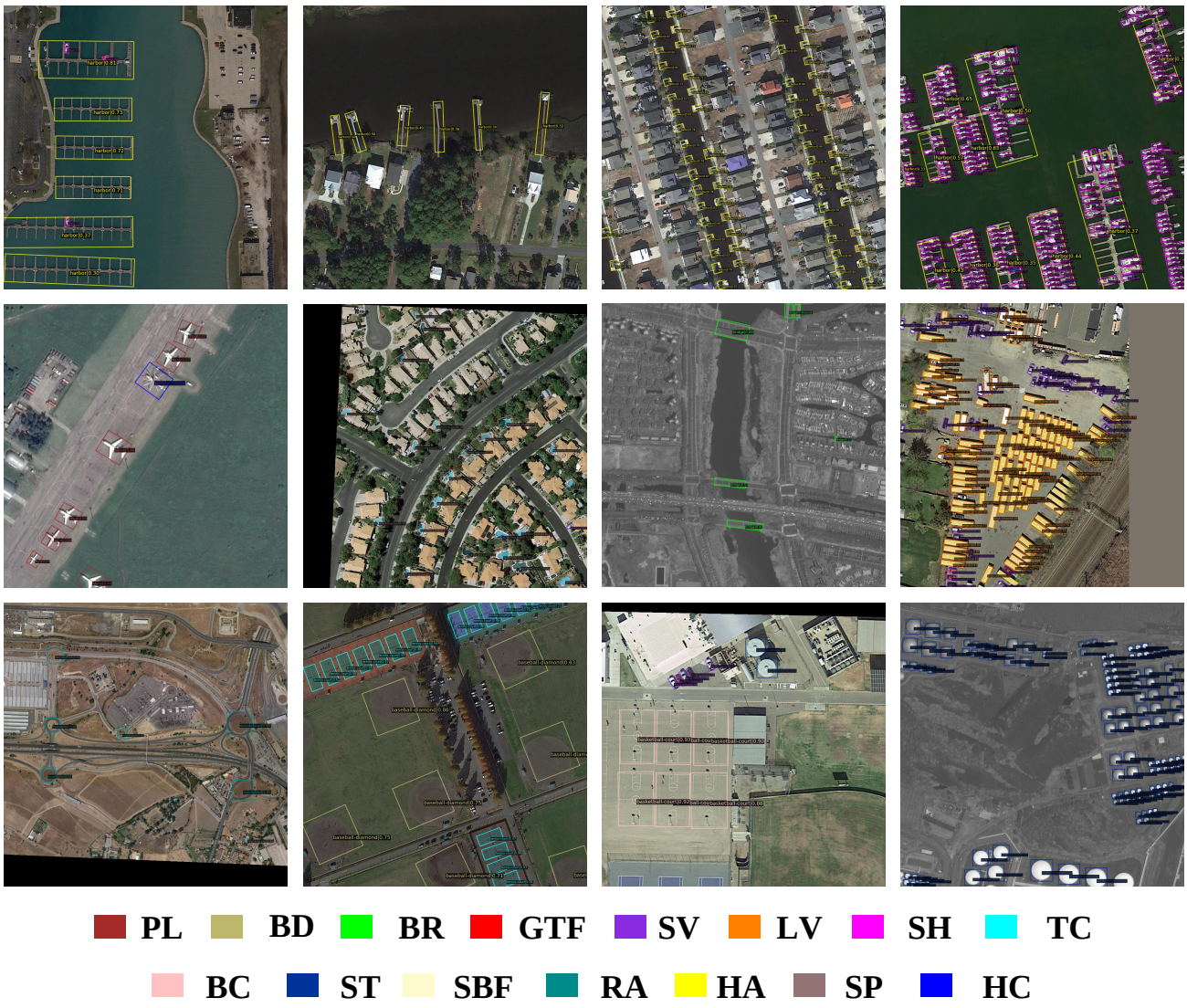

Figure 6. Visualization of Detection Results. Visualization of predictions on the DOTA dataset using our method, SREDet.

4.3.2. Ablation Study

We conducted ablation studies on the proposed modules to determine their respective contributions and effectiveness. All experiments employed simple random flipping as an augmentation technique to avoid overfitting. The results of these experiments are depicted in Table 2, while the error-type metrics proposed in this paper are presented in Table 3.

Table 2. Results of ablation experiments on DOTA dataset.

	MRFPN	SFEM	mAP_{50}	BR	GTF	SV	LV	SH	BC	ST	SBF	HA	SP	HC
baseline	–	–	63.4	41.3	60.3	65.6	69.8	78.1	55.2	59.7	50.5	58.8	52.9	40.3
Ours	✓		68.1	43.7	66.7	67.5	75.6	85.8	**66.1**	**66.0**	48.1	**67.6**	58.0	55.8
		✓	68.8	47.5	68.7	68.4	77.5	86.2	60.4	64.4	57.9	62.6	**59.3**	57.1
	✓	✓	**69.7**	**47.8**	**70.2**	**68.6**	**78.1**	**86.6**	65.7	65.8	**58.1**	67.2	58.9	**57.4**

Table 3. Error type metrics of ablation experiments on DOTA dataset.

	MRFPN	SFEM	$E_{cls}\downarrow$	$E_{loc}\downarrow$	$E_{cls\&loc}\downarrow$	$E_{bkg}\downarrow$	$E_{miss}\downarrow$
baseline	–	–	2.27	8.87	0.10	6.14	7.52
Ours	✓		**1.72**	7.59	0.11	5.84	**6.76**
		✓	1.81	**7.33**	**0.08**	5.56	7.13
	✓	✓	1.75	7.43	0.09	**5.51**	6.77

Firstly, to ascertain the effectiveness of the MRFPN and SFEM modules individually, model variants that solely incorporated each module were developed on the basis of the baseline. The integration of the MRFPN module led to a 4.7 increase in detection performance, particularly for objects such as basketball courts, storage tanks, and harbors. This improvement suggests that the multiscale rotation-invariant features extracted by the module facilitate the network's effective detection of objects varying in scale and orientation. Incorporation of the SFEM module resulted in a 5.4 improvement in detection capabilities for objects like large vehicles, swimming pools, helicopters, and ships, indicating that the SFEM module effectively intensifies object features and mitigates feature overlap among closely spaced objects. Finally, the combined application of both modules yielded an overall improvement of 6.3, demonstrating that the two feature enhancement components produce a synergistic effect. The multiscale rotation-invariant features extracted by MRFPN benefit the semantic segmentation tasks within the SFEM module, whereas the SFEM module can suppress noise in the features extracted by MRFPN.

We compared the responses of different types of errors to various improvement strategies, as seen in Table 3. In general, all improvement strategies were observed to contribute to a decrease in classification errors, regression errors, false positives in background detection, and missed objects detections, thus validating the effectiveness of our proposed modules. When only MRFPN was introduced, E_{bkg} was 5.84 and E_{miss} was 6.76. Similarly, with the introduction of SFEM alone, E_{bkg} was 5.56 and E_{miss} was 7.13. The comparison reveals that MRFPN can provide richer features (most notably by reducing classification errors) and reduce missed detections, but it may misidentify some backgrounds as objects. On the other hand, SFEM can suppress background noise and enhance object features (most notably reducing regression errors), but this approach can lead to increased missed objects. However, when both modules are applied together, they simultaneously reduce false positives from the background and missed detections, suggesting that the two components work together synergistically.

4.3.3. Detailed Evaluation and Performance Testing of Components

In this section, we mainly explore the impact of different styles of semantic labels (SemSty) and feature enhancement methods(Enh-Mtds) on network performance. Expl and Impl represent the explicit and implicit enhancement methods mentioned in Methods 3.2 of this article, respectively. Mask refers to the semantic mask obtained from object bounding boxes, and Seg indicates semantic segmentation information. Based on the experimental results in Table 4, we observe that under the same semantic annotation of Mask, the implicit method outperforms the explicit method by 1.5. Similarly, under the semantic annotation of Seg, the implicit method surpasses the explicit method by 1.4. Thus, when choosing the same type of semantic label, the implicit enhancement method is superior to the explicit enhancement method. This advantage primarily stems from the implicit method's ability

to decouple features of different objects into separate channels, facilitating the classification and regression of various categories of objects. In contrast, the explicit enhancement method is highly dependent on the accuracy of semantic segmentation, where any misclassification or omission in segmentation directly impacts network performance.

Table 4. Results of Semantic Supervision with Different Strategies.

	Enh-Mtds		**SemSty**											
	Expl	**Impl**	**Mask**	**Seg**	mAP_{50}	**PL**	**BD**	**GTF**	**BC**	**SBF**	**RA**	**HA**	**SP**	**HC**
baseline	–	–	–	–	63.4	88.5	74.9	60.3	55.2	50.5	63.9	58.8	52.9	40.3
Ours	✓		✓		66.5	88.7	77.1	62.9	58.6	54.2	61.0	60.9	57.2	52.2
	✓			✓	67.4	88.9	76.8	63.2	60.8	53.6	63.9	61.3	57.6	54.5
		✓	✓		68.0	88.8	77.1	**70.0**	**62.1**	53.3	61.5	62.3	58.7	56.1
		✓		✓	**68.8**	**89.2**	**77.4**	68.7	60.4	**57.9**	**64.1**	**62.6**	**59.3**	**57.1**

Furthermore, we also observe that under the explicit enhancement approach, Seg annotation improves performance by 0.9 compared to Mask annotation, and under the implicit enhancement method, Seg annotation leads to a 0.8 improvement over Mask annotation. Therefore, using Seg for supervision is superior to Mask under the same feature enhancement method, mainly due to the precise semantic information reducing background contamination and inter-class feature overlap. Specifically, using Seg annotation significantly outperforms mask annotation for objects like roundabouts. This is primarily because in the original DOTA dataset annotations the labeling for RA is not uniform, including objects like small or large vehicles, leading to inter-class feature overlap when using the Mask directly as semantic supervision. The mask may include part of the background information for objects with irregular shapes, such as swimming pools and helicopters, affecting the network's regression performance.

We compared the responses of different feature enhancement strategies to various types of errors, as seen in Table 5, all improvement strategies led to reductions in classification errors, regression errors, false positives from the background, and missed detections of objects, which demonstrates the effectiveness and versatility of the methods. Notably, using Masks as supervisory information with explicit feature enhancement best improved the issue of missed detections. However, among the four strategies, this approach showed the least improvement in false positives from the background, primarily because using Mask as semantic supervisory information reduces the difficulty of semantic segmentation but also increases the risk of incorrect segmentation.

Table 5. Error type metrics of Semantic Supervision with Different Strategies.

	Enh-Mtds		**SemSty**						
	Expl	**Impl**	**Mask**	**Seg**	$E_{cls} \downarrow$	$E_{loc} \downarrow$	$E_{cls\&loc} \downarrow$	$E_{bkg} \downarrow$	$E_{miss} \downarrow$
baseline	–	–	–	–	2.27	8.87	0.10	6.14	7.52
Ours	✓		✓		**1.74**	7.76	**0.06**	6.51	**6.91**
		✓	✓		1.75	7.73	0.07	5.70	7.25
	✓			✓	1.76	7.64	0.07	6.05	7.15
		✓		✓	1.81	**7.33**	0.08	**5.56**	7.13

Regarding false positives from the background, under the same style of semantic annotation, models using implicit enhancement methods outperform those with explicit frameworks. This advantage is mainly because semantic segmentation information does not directly affect network features. Instead, it indirectly generates weights for spatial feature enhancement and decoupling between different types of features, mitigating the direct impact of semantic segmentation errors. Concerning regression errors, using the same feature enhancement method is superior to using a Mask. The main reason is that

using Seg for semantic supervisory information can provide more accurate enhancement areas, which aids the network's regression tasks.

We provide a detailed visualization of different strategies for feature enhancement, as seen in Figure 7. From the visualization results of object boxes, it is evident that when using Masks as semantic guidance, false detections occur (as indicated by the red circles in the figure). Additionally, for some object detection cases, the results are suboptimal, failing to completely enclose the objects (as indicated by the green circles in the figure). This is primarily attributed to the utilization of Mask as a semantic guide, which introduces erroneous semantic information. In (e), for example, areas of the sea without ships are segmented as harbors, directly impacting the generation of feature weights and resulting in poor detection outcomes.

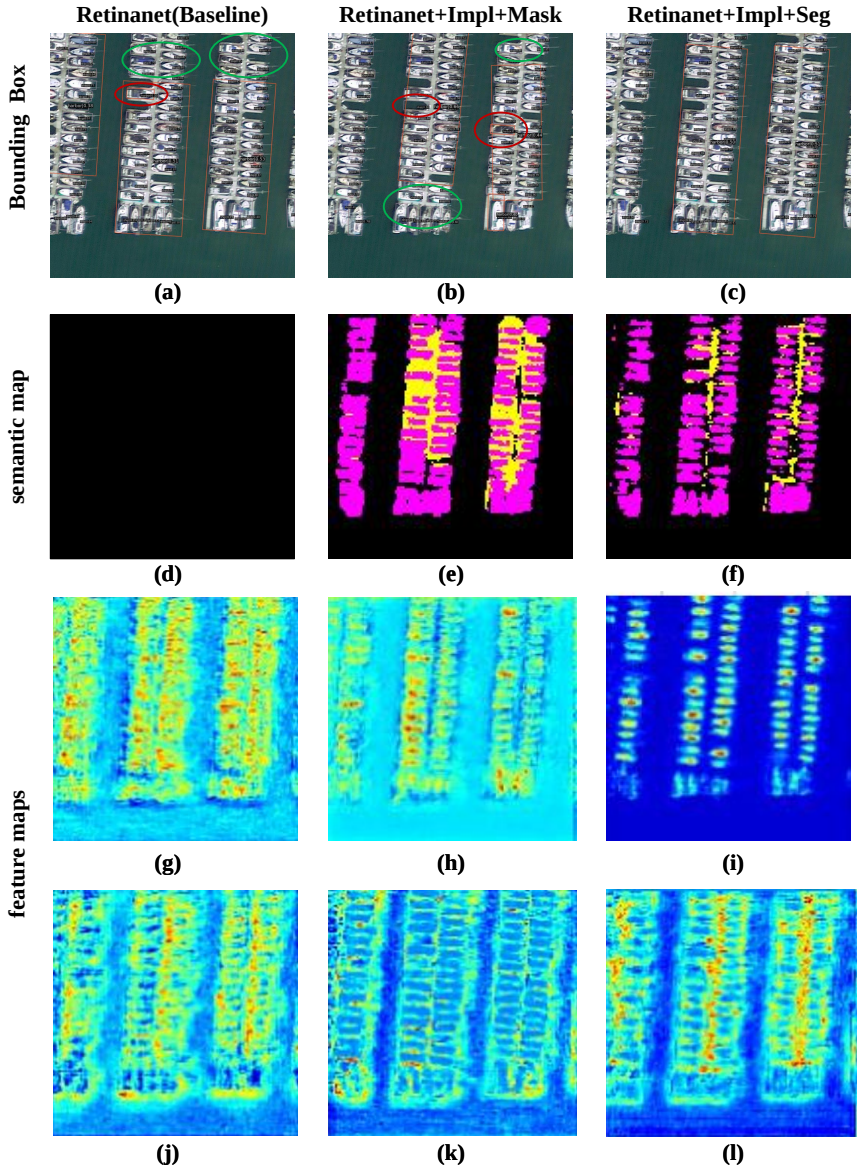

Figure 7. Visualization of Different Strategies. (**a**–**l**) The first row (**a**–**c**) and second rows (**d**–**f**) present the visualization results of detecting object boxes and outputting semantic maps. The last two rows (**g**–**l**) indicate the visualization results of different channel feature maps. The first column represents the experimental results of the baseline, while the second and third columns illustrate the results obtained by employing Mask and Segmentation as semantic guidance information for implicit feature map enhancement, respectively.

Regarding the feature maps, employing implicit enhancement effectively decouples features of different categories into different channels, as demonstrated by (h) and (k) as well as (i) and (l). It is apparent that (h) and (i) enhance features belonging to the category of ships, while (k) and (l) enhance features characteristic of harbors. Furthermore, a comparison of feature maps reveals that for images containing dense objects, using Segmentation as semantic supervision is more effective, yielding clearer and more responsive feature maps.

In the MRFPN, we tested different numbers of feature layers and compared the use of standard convolutions with deformable convolutions, as seen in Table 6. The experiments revealed that when using standard convolutions, there was no significant difference in performance between using four and five feature layers. However, after employing DCN for feature extraction, additional feature layers improved the network's performance. This improvement is primarily attributed to the DCN's enhanced capability to extract features from irregular targets.

Table 6. Ablative study of MRFPN with different strategies.

MRFPN Layers	Use DCN	mAP_{50}
$\{p_3, p_4, p_5\}$	–	67.98
$\{p_3, p_4, p_5, p_6\}$	–	68.08
$\{p_3, p_4, p_5, p_6\}$	✓	68.09
$\{p_3, p_4, p_5, p_6, p_7\}$	–	68.08
$\{p_3, p_4, p_5, p_6, p_7\}$	✓	**68.11**

In our experimental analysis of different strategies within the SFEM module, as seen in Table 7. When an equal number of dilated convolutions are stacked at each layer of the feature map, enhancing features across all feature maps yields better outcomes than enhancing only a subset of feature maps. When enhancing the same set of feature maps, appropriately stacking a certain number of dilated convolutions can enhance the model's detection performance. The primary reason is that multiple layers of dilated convolutions introduce a larger receptive field to the SFEM module, enabling the acquisition of more comprehensive contextual information.

Table 7. Ablative study of SFEM with different strategies.

Enhanced Layers	Stacked Dilated Convolution	mAP_{50}
$\{p_3, p_4, p_5\}$	$\{1, 1, 1\}$	68.1
$\{p_3, p_4, p_5\}$	$\{4, 3, 2\}$	68.5
$\{p_3, p_4, p_5, p_6, p_7\}$	$\{1, 1, 1, 1, 1\}$	68.3
$\{p_3, p_4, p_5, p_6, p_7\}$	$\{4, 4, 3, 2, 2\}$	**68.8**

We proposed a method for implicitly generating weights using semantic segmentation information to enhance feature maps. Therefore, the accuracy of semantic segmentation directly affects the network's performance. In the SFEM module, we tested three different losses, as seen in Table 8. By comparison, it can be seen that without adjusting the loss weights, focal loss performs best on the DOTA dataset for the class imbalance in remote sensing images. However, considering that Dice loss has a stronger ability to distinguish target regions, and based on our statistics, background pixels account for 96.95% of the dataset. We introduced weights to Dice loss by setting the classification weight of background pixels to 1 and foreground pixels to 20. The experimental results showed that this approach achieved the best performance.

Table 8. Ablation study of the SFEM with different loss functions. BG represents the background class weight and FG represents the foreground class weight.

Loss	Weights	mAP_{50}	mAP
Focal loss [46]	–	68.80	50.11
CE loss [60]	BG{1},FG{1}	67.89	49.87
Dice loss [47]	BG{1},FG{1}	67.99	50.07
Dice loss [47]	BG{1},FG{20}	**68.83**	**50.28**

We conducted comparative experiments to test the SFEM module with different base models, including the two-stage detection algorithm Faster R-CNN and the single-stage object detection model YOLOv8, as seen in Table 9.

Table 9. Performance of SFEM with Various Base Models.

Base Model	Backbone	with SFEM	mAP_{50}	mAP
Faster R-CNN [48]	ResNet101	–	70.24	53.12
Faster R-CNN [48]	ResNet101	✓	**71.12**	**53.28**
yolov8-m [25]	CSPDarknet	–	74.75	57.32
yolov8-m [25]	CSPDarknet	✓	75.36	58.06
yolov8-l [25]	CSPDarknet	–	75.08	57.81
yolov8-l [25]	CSPDarknet	✓	**75.84**	**58.47**

All models were trained on the training set and tested on the validation set. Our module achieved an improvement of 0.88 on mAP_{50} over Faster R-CNN, which is less pronounced compared to the single-stage detector. The main reason is that the RPN operation in the two-stage algorithm helps the network focus on the key feature regions of the target, rather than detecting over the entire feature map. Our module achieved improvements of 0.61 and 0.76 on mAP_{50} over YOLOv8-m and YOLOv8-l, respectively. It is worth noting that, for a fair comparison, no pre-trained models were used during training, and the default data augmentation method of YOLOv8 was applied.

4.3.4. Results on HRSC2016

The experimental results of HRSC2016 are presented in Table 10 as follows.

Table 10. Comparison with state-of-the-art methods on the HRSC2016 dataset.

Methods	Backbone	Size	mAP_{50}
R2CNN [61]	ResNet101	800 × 800	73.1
R2PN [50]	VGG16	/	79.6
OLPD [62]	ResNet101	800 × 800	88.4
RoI-Trans [63]	ResNet101	512 × 800	86.2
R3Det [24]	ResNet101	800 × 800	89.3
RetinaNet(baseline) [46]	ResNet101	800 × 800	84.6
RRD [64]	VGG16	384 × 384	84.3
BBAVectors [54]	ResNet101	800 × 800	89.7
SDet [65]	ResNet101	800 × 800	89.2
SREDet (ours)	ResNet101	800 × 800	89.8

With the modules proposed, our SERDet achieved an exemplary performance of 89.9%. Compared to specific ship detectors, SERDet shows an improvement of 5.2 over the baseline model. Simultaneously, we present the visualization results of ship detection, as seen in Figure 8, wherein it is evident that our proposed network is proficient in effectively detecting ships. SREDet exceeds the performance of other leading two-stage and single-stage detectors in the comparison.

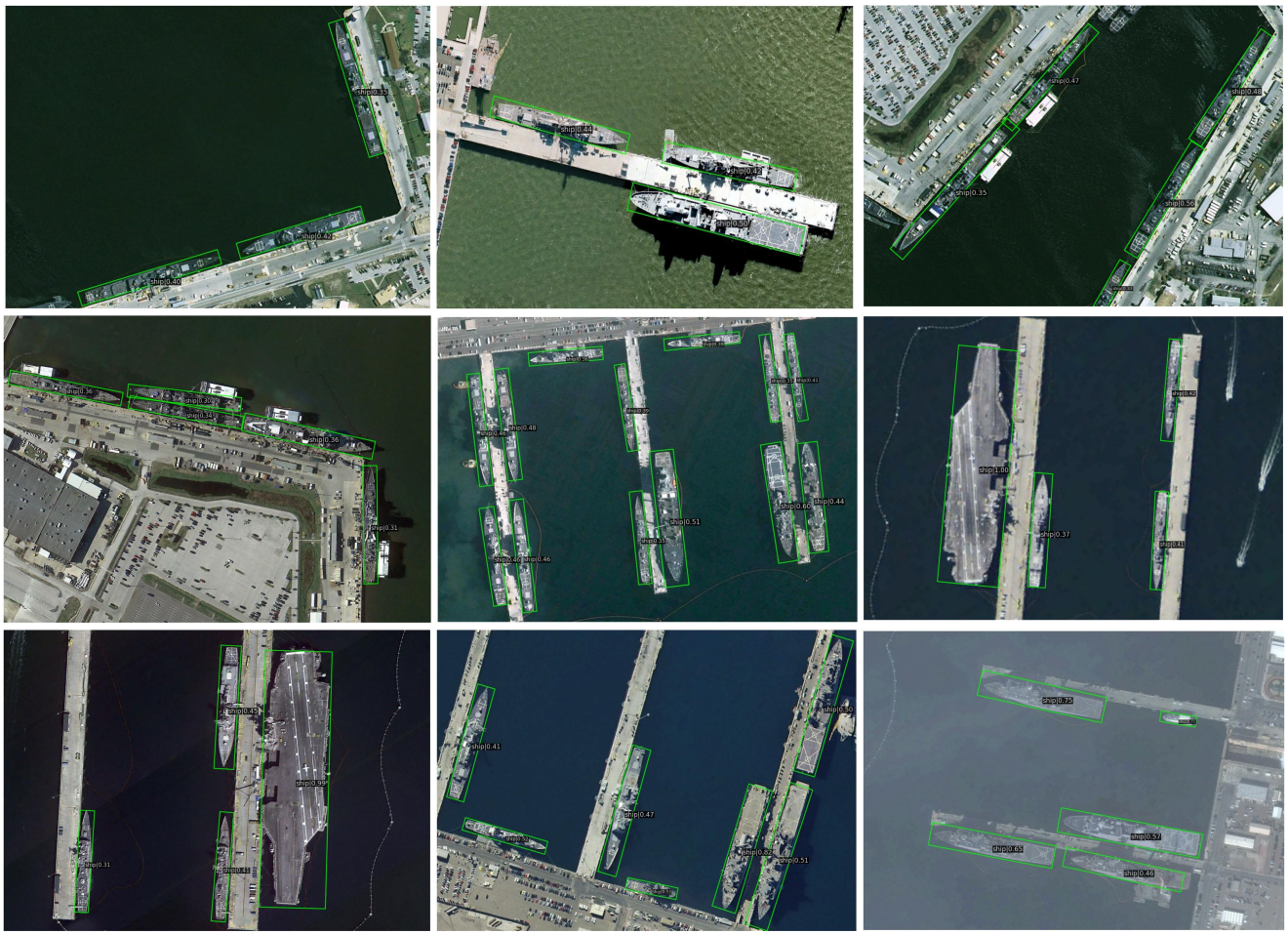

Figure 8. Visualization of Detection Results. Visualization of predictions on the HRSC2016 dataset using our SREDet method.

5. Conclusions

This study proposes a semantic-enhanced rotation object detection network, SREDet, targeting remote sensing image data. On the DOTA and HRSC2016 datasets, our network achieved mAP_{50} scores of 79.32% and 89.84%, respectively, surpassing other advanced methods compared in this study. First, the MRFPN module is designed to extract rotation-invariant features by fusing multi-angle feature maps. Second, the SFEM module, which utilizes semantic segmentation information for feature enhancement, is introduced. This module decouples the features of different object categories into separate channels. We compared our approach on several baselines, including Faster R-CNN and YOLOv8, and SFEM consistently improved detection accuracy, demonstrating the effectiveness of our proposed method. Finally, we introduced error-type analysis methods from general object detection, providing more refined evaluation metrics for rotated object detection. These metrics can demonstrate the network's ability to handle different types of errors, guiding further network improvements. However, the application of semantic segmentation information in this study is not comprehensive, as it does not consider the dependencies between different semantics. These relationships could further optimize the object representation in the network. In the future, we will investigate ways to integrate information from both semantic segmentation and object detection streams, designing better network structures to enhance rotation object detection capabilities. Furthermore, we will refine the proposed error-type evaluation metrics, focusing on angle error analysis, to provide a more comprehensive evaluation system.

Author Contributions: This research was a collaborative effort among seven authors, each contributing significantly to various aspects of the project. Y.W. was responsible for the conceptualization of experimental ideas and the design of the overall methodology. H.Z. played a crucial role in verifying the experimental design and conducting the initial analysis of the data. D.Q. took charge of the detailed analysis and interpretation of the experimental results. Q.L. contributed by exploring and investigating the experimental outcomes in depth. C.W. managed the organization and preprocessing of the data, ensuring its readiness for analysis. Z.Z. was instrumental in drafting the initial manuscript and in the primary visualization of the experimental results, including the creation of figures. Finally, W.D. was responsible for reviewing and revising the initial manuscript, providing critical feedback and making necessary modifications to enhance the quality of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China under Grant U20B2042 and 62076019.

Data Availability Statement: Publicly available datasets were used in this paper.

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

References

1. 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]
2. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
3. Han, W.; Chen, J.; Wang, L.; Feng, R.; Li, F.; Wu, L.; Tian, T.; Yan, J. Methods for small, weak object detection in optical high-resolution remote sensing images: A survey of advances and challenges. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 8–34. [CrossRef]
4. Yang, L.; Jiang, H.; Cai, R.; Wang, Y.; Song, S.; Huang, G.; Tian, Q. Condensenet v2: Sparse feature reactivation for deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3569–3578.
5. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Expert Syst. Appl.* **2022**, *197*, 116793. [CrossRef]
6. Gao, T.; Niu, Q.; Zhang, J.; Chen, T.; Mei, S.; Jubair, A. Global to local: A scale-aware network for remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615614. [CrossRef]
7. Yu, R.; Li, H.; Jiang, Y.; Zhang, B.; Wang, Y. Tiny vehicle detection for mid-to-high altitude UAV images based on visual attention and spatial-temporal information. *Sensors* **2022**, *22*, 2354. [CrossRef]
8. Pu, Y.; Liang, W.; Hao, Y.; Yuan, Y.; Yang, Y.; Zhang, C.; Hu, H.; Huang, G. Rank-DETR for high quality object detection. *arXiv* **2024**, arXiv:2310.08854.
9. Wang, Y.; Ding, W.; Zhang, B.; Li, H.; Liu, S. Superpixel labeling priors and MRF for aerial video segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2590–2603. [CrossRef]
10. Yang, L.; Chen, Y.; Song, S.; Li, F.; Huang, G. Deep Siamese networks based change detection with remote sensing images. *Remote Sens.* **2021**, *13*, 3394. [CrossRef]
11. Deng, C.; Jing, D.; Han, Y.; Deng, Z.; Zhang, H. Towards feature decoupling for lightweight oriented object detection in remote sensing images. *Remote Sens.* **2023**, *15*, 3801. [CrossRef]
12. 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.
13. Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.S.; Bai, X. isaid: A large-scale dataset for instance segmentation in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 28–37.
14. 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.
15. 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. [CrossRef]
16. 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]
17. 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.
18. Liu, S.; Zhang, L.; Lu, H.; He, Y. Center-boundary dual attention for oriented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5603914. [CrossRef]

19. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
20. Song, B.; Li, J.; Wu, J.; Chang, J.; Wan, J.; Liu, T. SRDF: Single-Stage Rotate Object Detector via Dense Prediction and False Positive Suppression. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5616616. [CrossRef]
21. Ming, Q.; Miao, L.; Zhou, Z.; Dong, Y. CFC-Net: A critical feature capturing network for arbitrary-oriented object detection in remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5605814. [CrossRef]
22. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep learning-based object detection techniques for remote sensing images: A survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
23. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3520–3529.
24. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 3163–3171.
25. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics, 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 1 December 2023).
26. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605415. [CrossRef]
27. Yang, L.; Han, Y.; Chen, X.; Song, S.; Dai, J.; Huang, G. Resolution adaptive networks for efficient inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2369–2378.
28. Yang, L.; Zheng, Z.; Wang, J.; Song, S.; Huang, G.; Li, F. An Adaptive Object Detection System based on Early-exit Neural Networks. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *16*, 332–345. [CrossRef]
29. Ma, T.; Mao, M.; Zheng, H.; Gao, P.; Wang, X.; Han, S.; Ding, E.; Zhang, B.; Doermann, D. Oriented object detection with transformer. *arXiv* **2021**, arXiv:2106.03146.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
31. 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]
32. Yu, H.; Tian, Y.; Ye, Q.; Liu, Y. Spatial transform decoupling for oriented object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6782–6790.
33. Cohen, T.; Welling, M. Group equivariant convolutional networks. In Proceedings of the International Conference on Machine Learning. PMLR, New York, NY, USA, 20–23 June 2016; pp. 2990–2999.
34. Hoogeboom, E.; Peters, J.W.; Cohen, T.S.; Welling, M. Hexaconv. *arXiv* **2018**,arXiv:1803.02108
35. Pu, Y.; Wang, Y.; Xia, Z.; Han, Y.; Wang, Y.; Gan, W.; Wang, Z.; Song, S.; Huang, G. Adaptive rotated convolution for rotated object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 6589–6600.
36. Mei, S.; Jiang, R.; Ma, M.; Song, C. Rotation-invariant feature learning via convolutional neural network with cyclic polar coordinates convolutional layer. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5600713. [CrossRef]
37. 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]
38. Zheng, S.; Wu, Z.; Du, Q.; Xu, Y.; Wei, Z. Oriented Object Detection For Remote Sensing Images via Object-Wise Rotation-Invariant Semantic Representation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5625515. [CrossRef]
39. Li, Y.; Huang, Q.; Pei, X.; Jiao, L.; Shang, R. RADet: Refine feature pyramid network and multi-layer attention network for arbitrary-oriented object detection of remote sensing images. *Remote Sens.* **2020**, *12*, 389. [CrossRef]
40. 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**, *45*, 2384–2399. [CrossRef]
41. Zhang, T.; Zhang, X.; Zhu, X.; Wang, G.; Han, X.; Tang, X.; Jiao, L. Multistage Enhancement Network for Tiny Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611512. [CrossRef]
42. Weiler, M.; Cesa, G. General e (2)-equivariant steerable cnns. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8792–8802.
43. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
44. Cao, L.; Zhang, X.; Wang, Z.; Ding, G. Multi angle rotation object detection for remote sensing image based on modified feature pyramid networks. *Int. J. Remote Sens.* **2021**, *42*, 5253–5276. [CrossRef]
45. Bolya, D.; Foley, S.; Hays, J.; Hoffman, J. Tide: A general toolbox for identifying object detection errors. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part III 16; Springer: Berlin/Heidelberg, Germany, 2020, pp. 558–573.
46. 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.

47. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
48. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
49. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; et al. Mmrotate: A rotated object detection benchmark using pytorch. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10 October 2022; pp. 7331–7334.
50. 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]
51. 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]
52. 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. [CrossRef]
53. Wang, J.; Yang, W.; Li, H.C.; Zhang, H.; Xia, G.S. Learning center probability map for detecting objects in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4307–4323. [CrossRef]
54. 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 IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 2150–2159.
55. 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.
56. 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]
57. Li, Z.; Hou, B.; Wu, Z.; Ren, B.; Yang, C. FCOSR: A simple anchor-free rotated detector for aerial object detection. *Remote Sens.* **2023**, *15*, 5499. [CrossRef]
58. 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. [CrossRef]
59. Cheng, G.; Yao, Y.; Li, S.; Li, K.; Xie, X.; Wang, J.; Yao, X.; Han, J. Dual-aligned oriented detector. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
60. Zhang, Z.; Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 14334–14345.
61. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational region CNN for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
62. Shu, Z.; Hu, X.; Sun, J. Center-point-guided proposal generation for detection of small and dense buildings in aerial imagery. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1100–1104. [CrossRef]
63. 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.
64. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.s.; Bai, X. Rotation-sensitive regression for oriented scene text detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918.
65. Ren, Z.; Tang, Y.; He, Z.; Tian, L.; Yang, Y.; Zhang, W. Ship detection in high-resolution optical remote sensing images aided by saliency information. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5623616. [CrossRef]