remote sensing

MDPI

Article

Rotation-Equivariant Feature Learning and Dual-Head Decoupling for Oriented Remote Sensing Object Detection †

Xue Zhang, Yanxia Wu, Dan Lin *, Guoyin Zhang, Changting Shi and Xueheng Sun

School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China; crystal_zhang@hrbeu.edu.cn (X.Z.); wuyanxia@hrbeu.edu.cn (Y.W.); zhangguoyin@hrbeu.edu.cn (G.Z.); shichangting@hrbeu.edu.cn (C.S.); sunxueheng@hrbeu.edu.cn (X.S.)

* Correspondence: danlin@hrbeu.edu.cn

† This paper is an extended version of our paper published in the 9th Chinese Conference on Pattern Recognition and Computer Vision (PRCV2026), Harbin, China, 22–25 August 2026.

Highlights

What are the main findings?

- RedhDet improves oriented remote sensing object detection by combining implicit rotation-equivariant learning with task-specific head decoupling, yielding consistent gains in detection and localization accuracy.
- LSIEN stabilizes equivariant feature propagation while preserving spatial details, whereas the fully connected head and convolutional head better serve classification and oriented bounding box regression, respectively.

What are the implications of the main findings?

- Rotation robustness can be effectively enhanced at the backbone level through implicit equivariant learning without restrictive equivariant kernels or specialized pre-trained weights.
- Separating classification and localization according to their distinct feature requirements provides an effective design for detecting dense, elongated, and arbitrarily oriented objects.

Abstract

Oriented remote sensing object detection (ORSOD) remains challenging because objects usually appear at arbitrary rotations, with large aspect-ratio variations and dense spatial layouts. Although recent detectors have achieved notable progress, most still rely on standard convolutional backbones, whose limited inherent rotation equivariance weakens robustness to rotated targets and reduces localization reliability. To address these issues, this paper proposes RedhDet, an ORSOD framework based on rotation-equivariant feature learning and dual-head decoupling. Specifically, a Ladder-Structured Implicitly Equivariant Network (LSIEN) is developed as the backbone to extract rotation-equivariant features in a soft manner, without relying on strictly constrained kernels or equivariant pre-trained weights. In addition, a decoupled dual-head architecture is introduced, in which a fully connected branch is used for classification and a convolutional branch is used for bounding box regression, thereby better matching the distinct requirements of semantic recognition and precise localization. Experimental results on the DOTA-1.0 and HRSC2016 datasets demonstrate the effectiveness of the proposed method. On the DOTA-1.0 test set, RedhDet improves mAP_{50} by 2.09% and mAP_{75} by 14.84% over the baseline. On HRSC2016, gains of 14.12% and 51.85% are achieved for mAP_{50} and mAP_{75}, respectively. With multi-scale training and testing, the proposed method achieves 80.40% mAP on DOTA-1.0, indicating its effectiveness and competitiveness for ORSOD.

Check for updates

Academic Editor: Haopeng Zhang

Received: 2 September 2026
Revised: 23 October 2026
Accepted: 23 October 2026
Published: 26 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: rotation-equivariant; dual-head; arbitrary orientation; oriented object detection; remote sensing images

1. Introduction

Object detection in remote sensing imagery has emerged as a critical research direction in computer vision, owing to its significant value in both military and civilian applications [1]. Compared with natural-scene images, remote sensing images typically exhibit larger scale variation, denser object layouts, and more arbitrary object orientations. In such scenarios, horizontal bounding boxes (HBBs) often introduce substantial redundant background and fail to describe the true geometry of elongated or rotated instances. By contrast, oriented bounding boxes (OBBs) provide a more faithful representation of object shape, direction, and spatial extent [2–4]. The release of large-scale benchmarks (e.g., DOTA [5], DIOR-R [6]) has further accelerated the transition from HBB-based detection to OBB-based detection, making oriented remote sensing object detection (ORSOD) one of the central topics in modern remote sensing interpretation.

Driven by the rapid development of deep learning, ORSOD has achieved remarkable progress in recent years. A dominant line of research extends horizontal object detectors by redesigning bounding-box parameterization [7], sample assignment [8], feature alignment [3,9,10], or loss functions [11]. More recently, transformer-based frameworks such as AO2-DETR [4] have further simplified the detection pipeline by introducing set prediction for arbitrary-oriented objects. Despite these advances, most existing methods are still constructed by adapting horizontal detection frameworks to the oriented setting, rather than explicitly endowing the feature extractor with a principled rotation equivariance.

An ideal ORSOD model should preserve geometric consistency under image rotation, such that the predicted boxes rotate accordingly while maintaining stable semantic recognition. However, this property is difficult to achieve with standard convolutional neural networks (CNNs), which are translation-equivariant but not naturally rotation-equivariant. Therefore, most existing methods rely on rotation augmentation, angle-sensitive heads, or feature alignment strategies to improve orientation robustness [3,10,12]. Although these methods improve localization accuracy and efficiency, they mainly enhance alignment or refinement, rather than explicitly enforcing rotation equivariance in backbone features.

From a geometric deep learning perspective, a more principled solution is to explicitly encode rotation equivariance into the feature extraction process. ReDet [13] is a representative attempt in this direction. It incorporates an E(2)-Equivariant CNN (E2CNN) [14] into the detector backbone and introduces Rotation-invariant RoI Align (RiRoI Align) to derive rotation-invariant region features from rotation-equivariant feature maps. This design demonstrates that equivariant representations are highly beneficial for aerial object detection, especially when objects appear in arbitrary directions. However, the underlying E2CNN impose representation-dependent constraints on convolution kernels. Although such constraints guarantee equivariance, they also restrict the design space of convolution kernels, to some extent limiting the model's representational capacity [15].

To overcome the rigidity of explicit steerable networks, Implicitly Equivariant Networks (IEN) [16] induce equivariance through auxiliary loss constraints while preserving the standard CNN architecture. This idea avoids handcrafted steerable kernel parameterization and retains the implementation flexibility of conventional backbones. Such a paradigm is attractive for ORSOD, where preserving the expressive power and compatibility of mainstream detectors is highly desirable. However, when adapted to oriented detection, implicit-equivariance schemes may still encounter several practical limitations.

First, layer-wise equivariance supervision based on parallel processing branches can accumulate approximation errors across stages. Second, group pooling tends to emphasize invariance at the cost of orientation-sensitive fine details, which are crucial for densely packed small objects in remote sensing images. Third, designs that favor square inputs are not well matched to remote sensing imagery, whose aspect ratios are often highly variable. Figure 1 compares the proposed method with IEN across these three aspects. As a result, there remains a clear need for a rotation-equivariant ORSOD framework that preserves CNN flexibility, avoids unnecessary information loss, and better fits the geometric characteristics of remote sensing data.

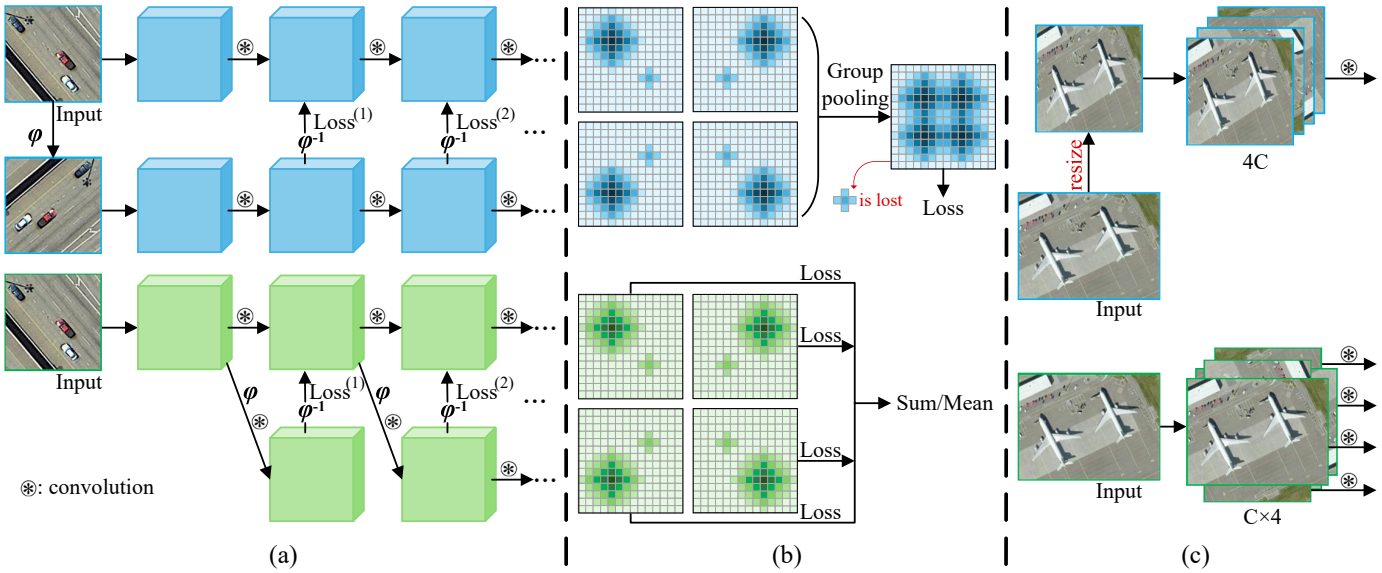


Figure 1. The limitations of IEN (**top**, blue) compared with our method (**bottom**, green), where φ denotes rotation and $\circledast$ represents convolution. (**a**) Error accumulation, (**b**) loss of fine-grained features, (**c**) aspect ratio limitation.

To address the above limitations, we propose the Ladder-Structured Implicitly Equivariant Network (LSIEN), a backbone tailored for ORSOD. LSIEN is designed to learn rotation-equivariant representations in an implicit manner while preserving the architectural flexibility of standard CNNs. Unlike previous implicit-equivariant designs, the proposed ladder-structured formulation follows the principle of rotation equivariance more faithfully by deriving the equivariance constraint of each layer from the corresponding feature map of the preceding layer. In this way, LSIEN mitigates the error accumulation caused by parallel layer-wise supervision and avoids the restrictive kernel design and pretrained-weight dependence commonly associated with explicit equivariant networks.

Although LSIEN enables the backbone to produce rotation-equivariant features and RiRoI Align further normalizes region features, classification and localization in ORSOD still exhibit fundamentally different representation requirements. Classification prefers spatially aggregated and rotation-invariant semantic cues, whereas box regression and angle prediction rely more heavily on fine-grained spatial configurations and local geometric structures. Treating these tasks with a unified fully connected head (fc-head) tends to over-abstract geometric information, which is particularly unfavorable for accurate localization. To alleviate this inconsistency, we further introduce a dual-branch decoupled-head, in which an fc-head is used for classification and a convolutional head (conv-head) is adopted for bounding box regression. This decoupled design allows each task to exploit the most suitable feature transformation paradigm, thereby improving both semantic discrimination and geometric sensitivity. Based on the proposed LSIEN backbone and the dual-head architecture, we propose a rotation-

equivariant dual-head detector, termed RedhDet. In summary, our main contributions in this paper are summarized as follows.

(1) We propose RedhDet, a rotation-equivariant dual-head detector for ORSOD. It achieves robust oriented object detection through the synergy of implicit equivariant feature learning, rotation-aligned region representation, and task-specific head decoupling.
(2) We propose LSIEN, a backbone specifically tailored for ORSOD to achieve implicit rotation equivariance without relying on equivariant pretrained weights. Through the proposed ladder-structured design, LSIEN more strictly conforms to the rotation-equivariance paradigm and effectively overcomes the main limitations of previous implicit-equivariant methods, including error accumulation, loss of fine-grained details caused by group pooling, and restrictions on input aspect ratios.
(3) We introduce a dual-head decoupling architecture, in which classification and OBB regression are performed by an fc-head and a conv-head, respectively. This design better aligns with the distinct feature requirements of semantic recognition and precise localization, thereby alleviating the inconsistency between classification and regression in ORSOD.
(4) Extensive experiments on DOTA-1.0, HRSC2016, and DIOR-R demonstrate the effectiveness of the proposed method. The results demonstrate that RedhDet achieves competitive performance on the ORSOD task.

This manuscript is an extended version of our conference paper [17]. Compared with the conference version, the present work includes an expanded technical presentation, additional experiments on the DIOR-R dataset, and a detailed hyperparameter sensitivity analysis. The rest of this paper is organized as follows. Section 2 presents the related work, while Section 3 describes the proposed RedhDet method. Section 4 reports extensive experiments and analyses. Finally, Section 5 concludes this paper.

2. Related Work

2.1. Oriented Object Detection in Remote Sensing Images

Benefiting from the release of large-scale remote sensing benchmarks and the rapid advances in computer vision, research on ORSOD has grown substantially in recent years. The arbitrary orientation of objects poses significant challenges to models in terms of feature extraction and alignment, label assignment, and parameter regression. From the perspective of feature extraction and alignment, SCRDet [2] and SCRDet++ [18] enhance multi-scale and contextual feature modeling for small, cluttered, and rotated objects. RoI Transformer [3] addresses the geometric mismatch at the proposal level by transforming horizontal RoIs into rotated ones. R^3Det [9] and S^2A-Net [10] further improve one-stage detection through progressive refinement and feature alignment. Oriented R-CNN simplifies the two-stage pipeline by directly generating high-quality oriented proposals. Another important line of work focuses on label assignment. DAL [19] introduces a matching degree to evaluate the localization potential of anchors more comprehensively. SASM [20] and SDDLA [21] further incorporate object shape information into dynamic label assignment, enabling more reliable positive-sample selection for oriented objects with diverse geometries. GGHL [22] introduces Gaussian-guided heatmap supervision in anchor-free settings to provide more geometry-aware sample assignment. A third line of research focuses on bounding-box regression and object representation. To address the boundary discontinuity of angle regression, existing methods adopt angle-classification schemes such as CSL [8] and DCL [23], as well as metric-aligned losses such as GWD [24], KLD [25], and KFIoU [11]. Gliding Vertex [7] derives rotated anchor boxes by sliding the four vertices along HBBs. BBAVectors [26] redesigns OBB parameterization by representing oriented objects with

four quadrant-aware vectors in the Cartesian coordinate system. In addition, anchor-free and DETR-based methods, such as Oriented RepPoints [27], AOPG [6], AO2-DETR [4], and ARS-DETR [28], further improve oriented detection by reducing handcrafted priors and introducing more flexible object representations.

2.2. Rotation-Equivariant and Rotation-Invariant Networks

Rotation equivariance and rotation invariance provide a principled way to model geometric transformations in visual recognition. Early studies extended conventional CNNs from pure translation symmetry to larger transformation groups through group convolutions [29], harmonic filters [30], and steerable kernels [14], thereby enabling feature responses to transform predictably under image rotations. More recently, IEN [16] provides an alternative route by enforcing equivariant behavior through learning objectives while preserving the architectural freedom of standard CNNs. In oriented object detection, rotation-related modeling can be roughly categorized into instance-level rotation invariance and backbone-level rotation equivariance. Early detectors mainly obtained rotation-invariant region features through rotated RoI warping or feature alignment, but their robustness remained limited when the backbone itself was not rotation-equivariant [3]. ReDet [13] addresses this issue by incorporating E2CNN into the backbone and proposing RiRoI Align to extract rotation-invariant RoI features from rotation-equivariant feature maps. Beyond explicit steerable backbones, recent methods also explore other routes for rotation-aware feature learning. ARC [31] adaptively rotates convolution kernels to better capture arbitrarily oriented objects. CPCCL [32] introduces cyclic polar coordinates convolution to convert rotational variation into a more tractable form for invariant feature learning. ORSR [33] guides the network to learn object-wise rotation-invariant semantic representations for remote sensing object detection. In addition, FPNFormer [34] models rotation-invariant semantics and rotation-equivariant orientation cues through transformer-based cross-scale interaction. SREDet [35] improves orientation robustness by multi-rotation feature fusion and semantic-driven feature enhancement. FRED [36] further decouples rotation-invariant classification from rotation-equivariant localization and pursues full end-to-end rotation equivariance. Although existing rotation-aware methods have improved orientation robustness through adaptive convolution, rotation-invariant representation learning, feature interaction, or explicit equivariant modeling, they generally rely on specific transformation strategies or specialized architectures, which may limit their compatibility with conventional CNN backbones. In contrast, we propose LSIEN to learn implicit rotation-equivariant representations without restrictive equivariant kernels or dedicated equivariant structures, thereby preserving the flexibility of standard CNNs while enhancing robustness to arbitrary rotations in ORSOD.

2.3. Classification-Localization Inconsistency and Decoupled Detection Heads

Recent studies showed that object classification and localization are intrinsically inconsistent in detection frameworks, because the former mainly depends on category-discriminative semantic cues, whereas the latter requires stronger sensitivity to spatial layout and boundary details [37]. In general object detection, localization-classification sequencing was introduced to reduce classification ambiguity caused by inaccurate localization [38]. RA-YOLOX [39] improves real-time detection through an aligned decoupled head. TGADHead [40] further incorporates task-guided attention and task-correlation modeling to enhance task-specific feature learning while preserving beneficial interaction between branches. Similarly, Han et al. [41] argued that complete task isolation is not always optimal and showed that moderate decoupling combined with adaptive feature extraction can yield better detection performance. Decoupling has also been extended

to multiscale feature aggregation, as exemplified by WPDFPN [42], which explicitly constructs task-specific pyramid features to alleviate classification-localization conflict during feature fusion. In ORSOD, this topic has also received growing attention. Chen et al. [43] introduced feature decoupling and post-classification regression for oriented object detection. Huang et al. [44] incorporated a task-decoupled head into an oriented detector with Gaussian-distribution-based label assignment. A recent multiscale task-decoupled network [45] further confirms the effectiveness of task-specific branches for oriented SAR ship detection. Overall, these studies suggest that improving detection accuracy requires not only stronger feature representations, but also head designs that better match the distinct representation and optimization requirements of classification and localization.

3. Proposed Method

This section elaborates our proposed method for RedhDet, and Figure 2 illustrates its overall architecture for ORSOD. The framework consists of three key components: rotation-equivariant feature learning, rotation-invariant transformation and alignment, and a dual-branch decoupled head. Specifically, the proposed LSIEN backbone learns rotation-equivariant features by imposing ladder-structured implicit equivariance constraints during training, thereby improving the stability of feature responses under object rotation. Based on the multiscale features generated by the FPN, the RPN and RoI Transformer [3] produce rotated proposals, which are further processed by RiRoI Align [13] to obtain spatially aligned and orientation-normalized RoI features. Finally, a dual-branch decoupled head is employed to alleviate the conflict between classification and localization, where the fc-head focuses on semantic discrimination and the conv-head emphasizes precise geometric regression. The detailed computation is described below.

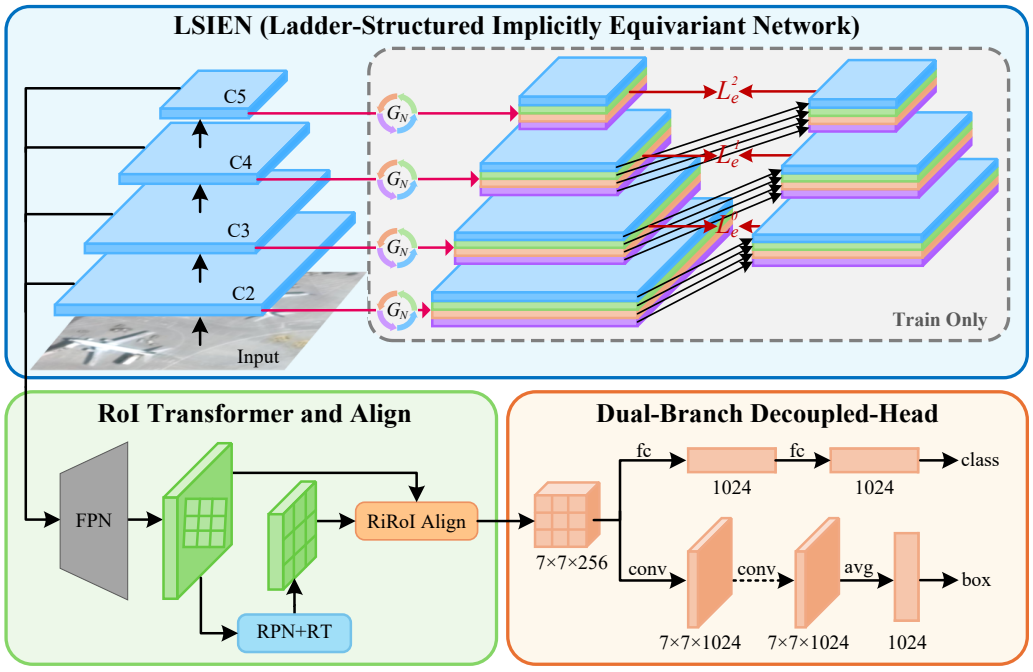


Figure 2. The overall architecture of the proposed RedhDet detector, which incorporates the LSIEN as the backbone, RoI transformer and align module, and the dual-branch decoupled-head.

3.1. Ladder-Structured Implicitly Equivariant Network

Rotation equivariance refers to the property that when the input undergoes a rotation, the function's output rotates in the same manner. Formally, a function f is rotation-equivariant if for any input x and rotation group G, it satisfies

$$f(\varphi_g(x)) = \varphi_g(f(x)) \quad \forall g \in G, \tag{1}$$

where $\varphi_g(\cdot)$ denotes the group action.

Implicit equivariance refers to enhancing the model's equivariance by introducing an auxiliary loss term during training. This is achieved by incorporating a feature-space auxiliary equivariance constraint into the objective function, which encourages the feature responses of transformed inputs to remain consistent with the correspondingly transformed feature responses. Our proposed LSIEN exhibits implicit equivariance, which is illustrated in the top panel of Figure 2. As illustrated in the diagram, the input image is subjected to a transformation at each layer of the ResNet backbone. The rotation transformation group G_N is defined as:

$$G_N = \left\{ R\left(\frac{2\pi k}{N}\right) \middle| k = 0, 1, \ldots, N-1 \right\}, \tag{2}$$

where N denotes the number of discrete rotation elements in G_N, and thus $2\pi/N$ represents the angular interval between two adjacent rotations. Here, $R(\theta)$ denotes the 2D rotation matrix for angle θ:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{3}$$

For convenience, G_4 and G_8 denote 4-fold and 8-fold rotation transformations, respectively, and Figure 2 visualizes the G_4 case for clarity.

Subsequently, in the ResNet, the feature X_i at the C_i-th layer ($2 \leq i \leq m$) is transformed by the group operation G_N into an expanded set of features Xp_i. It contains N times as many features as the original and is expressed as:

$$Xp_i = \left\{ \varphi_g^k(X_i) \middle| k = 0, 1, \ldots, N-1 \right\} \quad g \in G_N, \tag{4}$$

where $\varphi_g^k(\cdot)$ represents the k-th group action in G_N. Each feature in Xp_i is independent, and they are allowed to have different spatial sizes. Specifically, rotations by integer multiples of $\pi/2$ are implemented through direct tensor permutations without interpolation, whereas rotations by non-integer multiples of $\pi/2$ are performed using bilinear interpolation with zero padding and an expanded output canvas. Consequently, the rotated feature maps may have different spatial dimensions depending on the rotation angle.

The equivariance loss in IEN [16] does not strictly adhere to Equation (1), since the inputs on both sides are not identical. Since convolution is not rotation-equivariant, it introduces equivariance error. As a result, the equivariance error in IEN propagates and accumulates through successive convolution layers, leading to impure equivariance loss. In contrast, the proposed LSIEN mitigates this issue by formulating the auxiliary equivariance constraint in a manner that more closely follows the definition of rotation equivariance. The equivariance loss $\mathcal{L}_e$ is defined as:

$$\mathcal{L}_e = \sum_{i=2}^{m} \sum_{x_{p_i} \in Xp_i} \frac{\| F_i(x_{p_i}) - x_{p_{i+1}} \|^2}{N \times W_{i+1} \times H_{i+1} \times C_{i+1}}, \tag{5}$$

where m is the total number of feature layers, and the downsampling function $F_i(\cdot)$ maps the feature map from the C_i-th layer to the $C_{(i+1)}$-th layer. W_{i+1}, H_{i+1}, and C_{i+1} denote the

width, height, and number of channels of the feature map $x_{p_{i+1}}$, respectively. Substituting Equation (4) into Equation (5) yields the following:

$$\mathcal{L}_e = \sum_{i=2}^{m} \sum_{k=0}^{N-1} \frac{\| \mathrm{F}_i(\varphi_g^k(\mathrm{X}_i)) - \varphi_g^k(\mathrm{F}_i(\mathrm{X}_i)) \|^2}{N \times W_{i+1} \times H_{i+1} \times C_{i+1}}. \tag{6}$$

It can be observed that the innermost difference term in the above expression aligns with the definition of rotation equivariance given in Equation (1).

The proposed LSIEN backbone, inspired by the IEN [16] framework but significantly reconfigured as shown in Figure 1, incorporates several key innovations. First and most importantly, a ladder-shaped network architecture is employed to ensure that the rotation equivariance loss at each layer is derived from the same feature map of the previous layer. This structure more closely follows the definition of rotation equivariance and helps reduce error propagation across successive stages. Second, group pooling as used in IEN is intentionally omitted. This operation introduces partial invariance, which tends to compromise fine-grained details and is particularly undesirable for detecting small objects in remote sensing imagery. Finally, to prevent domain shift when the model learns rotation invariance from geometrically transformed images yet is applied to original images, we adopt a multi-channel design that preserves native image dimensions. The LSIEN network successfully extracts and outputs rotation-equivariant features, thereby providing essential support for subsequent spatial and orientation alignment.

3.2. RoI Transformer and Align

Although LSIEN improves backbone rotation equivariance, an intermediate mechanism is still required to convert dense feature maps into instance-level representations for subsequent classification and localization. Directly applying horizontal RoI extraction to remote sensing objects with arbitrary orientations usually introduces severe background redundancy and geometric mismatch. Therefore, following the standard two-stage oriented detection pipeline, we adopt the RoI Transformer [3] to convert horizontal proposals into rotated proposals, and then employ RiRoI Align [13] to obtain normalized RoI features with improved geometric consistency. Given a horizontal proposal $r = (x, y, w, h)$, the RoI Transformer learns a geometric transformation to predict an oriented proposal $\tilde{r} = (x_r, y_r, w_r, h_r, \theta_r)$, where (x_r, y_r), (w_r, h_r), and θ_r denote the center coordinates, side lengths, and orientation angle of the rotated proposal, respectively. In practice, this transformation can be implemented through a regression branch that predicts the offsets from the horizontal RoI to its oriented counterpart. By replacing coarse horizontal proposals with rotated ones, the extracted regions become better aligned with the actual object geometry, thereby reducing the interference caused by irrelevant background and providing a more appropriate basis for instance-level feature extraction.

However, rotated proposal extraction alone is still insufficient to guarantee stable rotation-invariant RoI features. Conventional rotated RoI warping mainly performs alignment in the spatial domain, while the orientation channels of equivariant feature maps remain inconsistent across rotated instances. To address this issue, we further adopt RiRoI Align [13], which performs both spatial alignment and orientation alignment. Orientation channels are constructed using group-equivariant convolutions based on the cyclic rotation group C_N, where each channel group corresponds to a discrete orientation. The resulting feature maps have dimensions $K \times N \times H \times W$, where K denotes the number of channels per orientation. These orientation channels are preserved during inference because the equivariant feature extraction layers remain active, allowing RiRoI Align to perform orientation-channel switching and interpolation. Let F denote the rotation-equivariant

feature map produced by the backbone and FPN, and let F_r denote the warped RoI feature associated with $\tilde{r}$. The output of RiRoI Align is formulated as

$$\hat{F}_r = \mathcal{I}(\mathcal{S}(F_r, s), \theta_r), \quad s = \left\lfloor \frac{\theta_r N}{2\pi} \right\rfloor, \tag{7}$$

where $\mathcal{S}(\cdot)$ denotes orientation-channel switching, $\mathcal{I}(\cdot)$ denotes interpolation in the orientation dimension, and s is the discrete orientation index under an N-fold rotation group. Here, N has the same meaning as in Equation (2). In this way, RiRoI Align transforms rotated proposals into fixed-size RoI features that are normalized not only spatially but also geometrically in the orientation dimension.

In our framework, the RoI Transformer and RiRoI Align are not introduced as the main source of methodological novelty, but as a necessary intermediate bridge between backbone-level equivariant feature extraction and head detection.

3.3. Dual-Branch Decoupled-Head

The non-shared weights of the fc-head make it more sensitive to spatial configurations, favoring classification, whereas the conv-head with shared weights better preserves spatial structures and is more suitable for regression. This distinction is particularly important for ORSOD, where angle regression and oriented boundary localization rely on preserving local geometric layouts after RoI alignment, while classification benefits from more global and rotation-normalized semantic representations. Motivated by the inherent inconsistency between classification and localization [37], we adopt a dual-head design in the ORSOD setting, where the fc-head focuses on semantic discrimination and the conv-head preserves fine-grained spatial information for precise oriented bounding box regression.

The bottom right panel of Figure 2 illustrates the dual-branch decoupled-head structure. The fc-head which consists of two fully connected layers following the standard FPN practice, outputs a 1024-dimensional vector. The conv-head comprises a stack of residual blocks [46], with the specific blocks corresponding to the solid and dashed arrows detailed in Figure 2. As illustrated in Figure 3a, the residual branch adopts a bottleneck structure consisting of 3×3, and 1×1 convolutions. The internal channel width is 256, and the channels are ultimately expanded by a factor of four. To maintain consistency with this dimension, the skip branch is projected using a 1×1 convolution. The outputs from both branches are then summed to perform fusion. As shown in Figure 3b, subsequent residual blocks maintain the same number of channels. Each residual block employs a sequence of 1×1, 3×3, and 1×1 convolutions, accompanied by an identity skip connection. Finally, a global average pooling layer aggregates the spatial dimensions into a 1024-dimensional feature vector, which serves as the final representation of the conv-head.

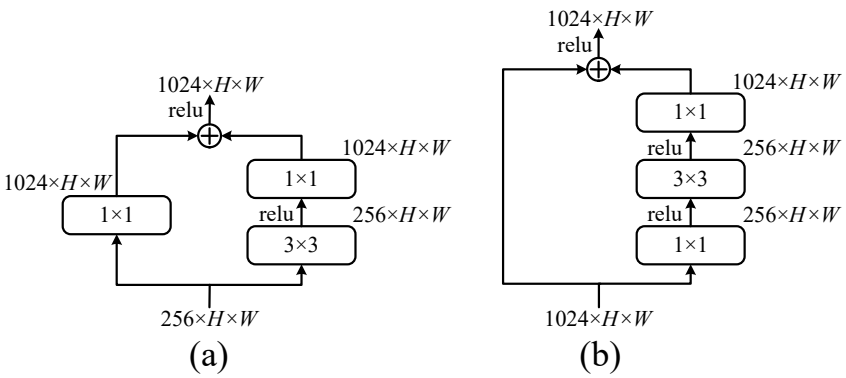


Figure 3. The two components of dual-branch decoupled-head: (**a**) residual block to increase the number of channels, (**b**) residual bottleneck block.

3.4. Total Loss

RedhDet is optimized in an end-to-end manner under a unified multi-task objective. The overall loss is formulated as:

$$\mathcal{L} = \lambda^{e}\mathcal{L}_{e} + \lambda^{reg}\mathcal{L}_{reg} + \lambda^{cls}\mathcal{L}_{cls} + \lambda^{rpn}\mathcal{L}_{rpn}, \tag{8}$$

where $\mathcal{L}_{e}$ denotes the equivariance loss of LSIEN and follows Equation (5). $\mathcal{L}_{reg}$ and $\mathcal{L}_{cls}$ denote the regression and classification losses of the detection head, respectively, while $\mathcal{L}_{rpn}$ is the loss of the proposal generation branch.

For box regression, the loss is defined as:

$$\mathcal{L}_{reg} = \frac{1}{N_{reg}} \sum_{i} \mathbb{I}(y_i > 0)\, L_{smooth}(\mathbf{b}_i, \mathbf{b}_i^{gt}), \tag{9}$$

where y_i denotes the class label of the i-th sample, $\mathbb{I}(\cdot)$ is the indicator function, and N_{reg} denotes the number of positive samples. Here, L_{smooth} is the Smooth L1 loss [47]:

$$L_{smooth}(\mathbf{b}_i, \mathbf{b}_i^{gt}) = \begin{cases} \dfrac{0.5(\mathbf{b}_i - \mathbf{b}_i^{gt})^2}{\beta}, & |\mathbf{b}_i - \mathbf{b}_i^{gt}| < \beta, \\ |\mathbf{b}_i - \mathbf{b}_i^{gt}| - 0.5\beta, & \text{otherwise}, \end{cases} \tag{10}$$

where $\mathbf{b}_i$ and $\mathbf{b}_i^{gt}$ denote the predicted box and its corresponding ground-truth target, respectively, and β is the hyperparameter controlling the transition point of the piecewise function.

For classification, we adopt the standard cross-entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{N_{cls}} \sum_{i} \sum_{c=0}^{C} y_{ic} \log(\hat{y}_{ic}), \tag{11}$$

where $\hat{y}_{ic}$ denotes the predicted probability that the i-th sample belongs to class c, y_{ic} is the corresponding one-hot ground-truth label, C is the number of object categories, and N_{cls} is the number of samples used for classification.

The RPN loss $\mathcal{L}_{rpn}$ consists of an objectness classification term and a proposal regression term. These two parts are computed in the same manner as the classification and regression losses defined above. In our experiments, we set $\lambda^{e} = 1.0$ and $\lambda^{rpn} = 1.0$. The remaining weights are set to $\lambda^{reg} = 2.5$ and $\lambda^{cls} = 2.0$, as in Wu et al. [37].

4. Experiments

4.1. Datasets

DOTA-1.0 [5] is a widely used benchmark for oriented object detection in aerial images. It contains 2806 images and 188,282 annotated instances from 15 ground object categories, including plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC). DOTA-1.0 provides an official split into training, validation, and test sets, with an approximate ratio of 3:1:2. We crop the original images into 1024 × 1024 patches with an overlap of 200 × 200. In the multi-scale experiments, the original images are resized to three scales {0.5,1.0,1.5}.

HRSC2016 [48] is a high-resolution optical remote sensing dataset specifically designed for ship detection, containing 1061 images and 2976 instances. The image sizes range approximately from 300 × 300 to 1500 × 900 pixels. The dataset includes 436 training

images, 181 validation images, and 444 test images. In our experiments, all images are resized to 800 × 800 pixels.

DIOR-R [6] provides OBBs derived from the DIOR dataset [49] for oriented object detection. It contains 23,463 remote sensing images and 192,518 oriented instances. This dataset covers 20 common categories, including airplane (APL), airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DAM), expressway toll station (ETS), expressway service area (ESA), golf field (GF), ground track field (GTF), harbor (HA), overpass (OP), ship (SH), stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM). The dataset includes 5862 training images, 5863 validation images, and 11,738 test images. The image size is fixed at 800 × 800 pixels.

4.2. Experiment Setup

Implementation Details: We utilize the pre-trained ResNet-50 from ImageNet [50] as backbone, with Feature Pyramid Networks (FPN) serving as the neck network. All experiments are conducted on a single RTX 3090 GPU (with 24 GB memory) using the MMRotate [51] toolbox, with batch size set to 2. The initial learning rate is set to 0.005, referring to ReDet [13]. The stochastic gradient descent (SGD) [52] is adopted as the optimizer. We use both training and validation sets for training and the test set for testing for all datasets. The models are trained for 12, 36, and 12 epochs on DOTA-1.0, HRSC2016, and DIOR-R, respectively.

Evaluation Metrics: Following standard evaluation protocols in object detection, mean Average Precision (mAP) is adopted as the primary metric, which is computed from the precision-recall relationship. Based on the Intersection over Union (IoU) criterion, each predicted bounding box is classified as either a true positive (TP) or a false positive (FP). A prediction is considered a TP if its IoU with the matched ground-truth box is higher than a predefined threshold; otherwise, it is considered an FP. False negatives (FNs) denote ground-truth instances that are missed by the detector. Accordingly, precision (P) and recall (R) are defined as follows:

$$\mathrm{P} = \frac{TP}{TP + FP}, \tag{12}$$

$$\mathrm{R} = \frac{TP}{TP + FN}. \tag{13}$$

To further reflect the balance between P and R, the F1-score is additionally introduced, which is defined as the harmonic mean of P and R:

$$\mathrm{F1} = \frac{2PR}{P + R}. \tag{14}$$

The Average Precision (AP) score is obtained by calculating the area beneath the PR curve:

$$\mathrm{AP} = \int_0^1 p(r)\,\mathrm{d}r, \tag{15}$$

where p denotes the precision rate and r denotes the recall rate. The mAP refers to the average of the AP scores for every single category.

$$\mathrm{mAP} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{AP}_k, \tag{16}$$

where K represents the number of object classes. In this work, detector accuracy is evaluated using mAP at IoU thresholds of 0.5 (mAP_{50}) and 0.75 (mAP_{75}). Unless otherwise specified, mAP refers to mAP_{50}. In addition, the HRSC2016 dataset is commonly reported with two evaluation settings, namely the PASCAL VOC2007 and VOC2012 metrics [53], whereas

the remaining datasets are evaluated under the PASCAL VOC2007 protocol. In this paper, mAP(07) denotes the mAP_{50} metric under VOC2007, and mAP(12) denotes the mAP_{50} metric under VOC2012.

4.3. Comparison with State-of-the-Art Methods

To evaluate the performance of the proposed method, representative and state-of-the-art ORSOD methods are selected for comparison. Specifically, recent representative methods are chosen from anchor-free [22,27,28,54,55], one-stage [9–11,21], and two-stage [2,3,7,12,18,56] oriented object detectors. In addition, recent rotation-equivariant feature extraction methods, including ReDet [13], FPNFormer [34], SREDet [35], FRED [36], RO^2-DETR [57], are also included for comparison.

4.3.1. Results on DOTA-1.0

We compare our proposed method with state-of-the-art approaches on DOTA-1.0 OBB task, and the results are presented in Table 1. The table reports the detection performance of all methods on the 15 categories of DOTA-1.0, along with mAP. All results except our own are drawn from published papers. Our proposed detector RedhDet, equipped with LSIEN based on ResNet-50, achieves a competitive result of 76.37% mAP. When multi-scale training and testing are adopted, RedhDet reaches 80.40% mAP, achieving the highest performance among the compared methods. It is noteworthy that compared to other rotation-equivariant methods such as FRED, FPNFormer and SREDet, our method outperforms them. Additionally, our approach achieves either the best or second-best results across the BD, BR, LV, TC, ST, SBF, HA, and HC categories.

Table 1. Comparison with state-of-the-art methods on the OBB-based task of the DOTA-1.0 dataset. * means multi-scale training and testing. R-101 denotes ResNet 101 (likewise for R-50 and R-152); D-53, and ReR-50 stand for DarkNet 53, and ReResNet 50, respectively. Best (bold) and second-best (underlined) results are highlighted.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (%) ↑
Anchor-free:																	
O-RepPoints [27]	R-101	89.53	84.07	59.86	71.76	79.95	80.03	87.33	90.84	87.54	85.23	59.15	66.37	75.23	73.75	57.23	76.52
GGHL * [22]	D-53	89.74	85.63	44.50	77.48	76.72	80.45	86.16	90.83	88.18	86.25	67.07	69.40	73.38	68.45	70.14	76.95
EARL * [54]	R-50	90.13	83.90	47.19	72.17	81.54	84.26	88.24	90.69	79.10	86.71	60.47	**72.21**	71.26	73.81	58.80	76.03
ARS-DETR [28]	R-50	86.97	75.56	48.32	69.20	77.92	77.94	87.69	90.50	77.31	82.86	60.28	64.58	74.88	71.76	66.62	74.16
FRED [36]	ReR-50	89.37	82.12	50.84	73.89	77.58	77.38	87.51	90.82	86.30	84.25	62.54	65.10	72.65	69.55	63.41	75.56
RO^2-DETR * [57]	R-50	89.40	84.81	57.11	75.54	78.77	82.74	87.31	90.87	84.66	85.39	62.01	69.97	75.02	71.90	72.14	79.42
PSSANet [55]	LRFormer	88.70	83.20	52.20	63.30	78.70	81.70	87.20	90.80	86.40	80.30	58.30	62.00	73.00	73.80	67.60	75.21
One-stage:																	
R^3Det * [9]	R-152	89.80	83.77	48.11	66.77	78.76	83.27	87.84	90.82	85.38	85.51	65.67	62.68	67.53	78.56	72.62	76.47
S^2A-Net * [10]	R-101	89.28	84.11	56.95	79.21	80.18	82.93	**89.21**	90.86	84.66	87.61	71.66	68.23	78.58	78.20	65.55	79.15
KFIoU * [11]	R-101	89.28	83.32	59.40	80.29	80.43	84.70	88.85	90.87	84.51	87.95	71.86	71.60	78.31	79.42	66.60	79.83
FPNFormer * [34]	R-50	89.10	82.41	56.15	81.24	**82.05**	85.16	88.74	90.89	85.78	87.07	68.93	68.47	77.93	78.65	68.35	79.39
SREDet * [35]	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
SDDLA * [21]	R-101	89.51	83.81	59.60	**81.92**	80.39	84.79	88.60	90.81	85.98	87.50	71.41	69.42	78.44	78.28	68.38	79.92
Two-stage:																	
RoI Trans. * [3]	R-101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet [2]	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
Gliding Vertex [7]	R-101	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
CSL * [8]	R-152	**90.25**	85.53	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	69.60	68.04	73.83	71.10	68.93	76.17
ReDet	ReR-50	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25
ReDet * [13]	ReR-50	88.81	82.48	**60.83**	80.82	78.34	**86.06**	88.31	90.87	**88.77**	87.03	68.65	66.90	**79.26**	79.71	**74.67**	80.10
O-R-CNN [12]	R-101	88.86	83.48	55.27	76.92	74.27	82.10	87.52	**90.90**	85.56	85.33	65.51	66.82	74.36	70.15	57.28	76.28
SCRDet++ * [18]	R-101	90.05	84.39	55.44	73.99	77.54	71.11	86.05	90.67	87.32	87.08	69.62	68.90	73.74	71.29	65.08	76.81
RSDB-Net * [56]	STC-R50	88.89	83.69	55.99	69.42	80.07	84.84	89.13	90.85	83.75	**88.15**	**76.97**	66.17	76.87	**80.48**	72.27	77.81
RedhDet (Ours)	LSIEN-R50	89.29	82.53	53.31	75.19	78.64	82.94	87.89	**90.90**	86.69	85.19	64.33	60.29	75.07	68.25	65.10	76.37
RedhDet * (Ours)	LSIEN-R50	89.44	86.24	59.87	79.62	79.09	85.43	88.32	90.87	86.88	88.06	72.29	68.78	78.75	78.33	73.97	**80.40**

The detection results for all 15 categories from our proposed method are visualized in Figure 4. The detection results exhibit fewer missed detections and false alarms, with bounding boxes tightly adhering to the ground truth and yielding high-confidence predictions. This demonstrates the precise detection capability of RedhDet for densely arranged and oriented objects. We chose the two-stage rotation-equivariant detector ReDet [13] to compare qualitative results with the RedhDet method. As shown in Figure 5, our approach reduces missed targets and false alarms.



Figure 4. Detection results visualization of our proposed method on the DOTA-1.0 dataset.

4.3.2. Results on HRSC2016

HRSC2016 contains numerous elongated ship instances with arbitrary orientations. We evaluate various advanced ORSOD methods on this dataset, and the results are summarized in Table 2. RedhDet built on the ResNet-101 backbone achieves the highest performance of 90.48% mAP(07) and 97.71% mAP(12). Figure 6 showcases challenging instance detection results, further elucidating the advantages of our method in detecting objects with large aspect ratios.

4.3.3. Results on DIOR-R

Table 3 presents a comparison between our method and several contemporary approaches on the DIOR-R dataset. Our proposed RedhDet achieves the highest overall performance of 66.44% mAP, and maintains state-of-the-art results across multiple categories. Figure 7 shows the PR curves of all categories on the DIOR-R dataset. Strong performance can be observed for categories such as BC, TC, GTF, SH, and ESA, whose curves maintain relatively high precision over a broad recall range. In contrast, categories such as DAM, BR, APO, and VE remain more challenging, with precision dropping more rapidly as recall increases. A more detailed analysis of the error distribution and inter-class confusion is provided by the normalized confusion matrix in Figure 8. It can be observed that APL, APO, BF, BR, DAM, HA, STO, TS, and VE are relatively often missed and thus classified as background. In addition, BR and OP show noticeable confusion, indicating that their geometric patterns are not always easy to distinguish in remote sensing images.

A similar but more severe phenomenon occurs between GTF and STA, suggesting that categories with similar spatial layouts or surrounding contexts remain challenging for accurate discrimination.

Table 2. Comparison with state-of-the-art methods on the HRSC2016 dataset. Best (bold) and second-best (underlined) results are highlighted.

Methods	Backbone	mAP(07) (%) ↑	mAP(12) (%) ↑
RoI Trans. [3]	R-101	86.20	—
Gliding Vertex [7]	R-101	88.20	—
CSL [8]	R-101	89.62	96.10
BBAVectors [26]	R-101	88.60	—
R^3Det [9]	R-101	89.26	96.01
DAL [19]	R-101	89.77	—
S^2A-Net [10]	R-101	90.17	95.01
ReDet [13]	R-50	<u>90.46</u>	<u>97.63</u>
O-RepPoints [27]	R-50	90.38	97.26
EARL [54]	R-101	88.80	92.10
SREDet [35]	R-101	89.80	—
SDDLA [21]	R-101	<u>90.46</u>	97.35
RSDB-Net [56]	STC-R50	90.10	—
RO^2-DETR [57]	R-101	90.44	97.47
PSSANet [55]	LRFormer	89.90	—
RedhDet (Ours)	LSIEN-R101	**90.48**	**97.71**

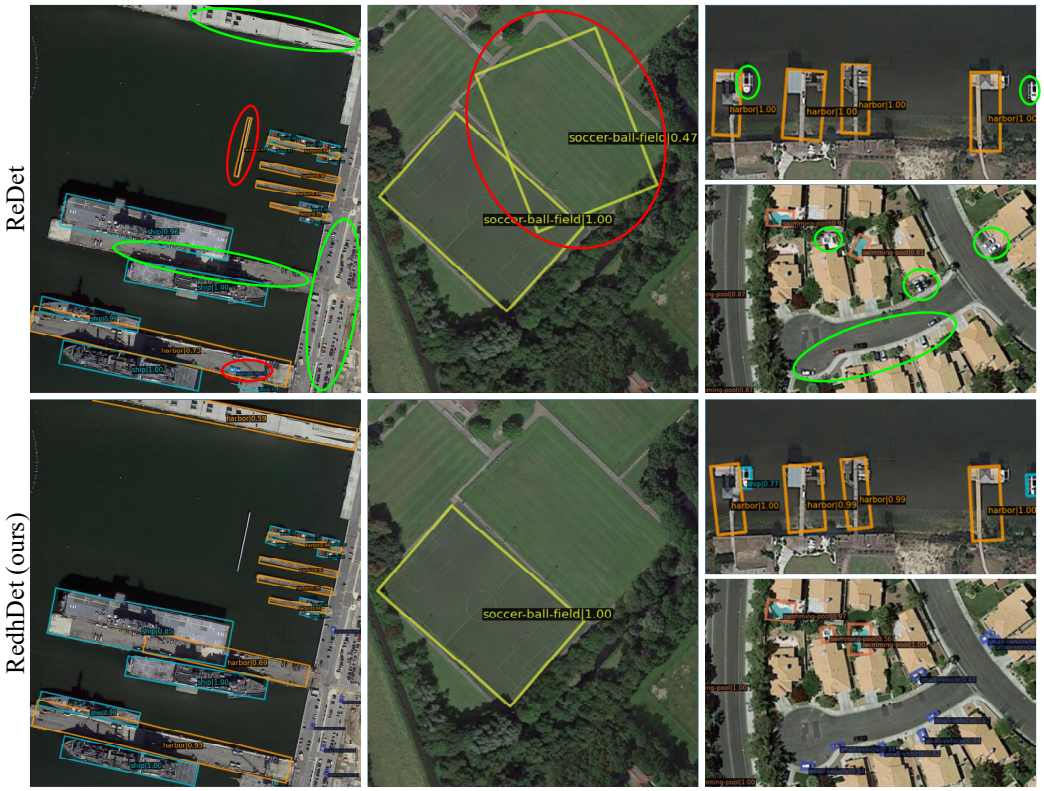


Figure 5. Qualitative comparisons between the proposed RedhDet and ReDet [13] on the DOTA-1.0 dataset. Red and green ellipses mark false alarm and missed targets results, respectively.

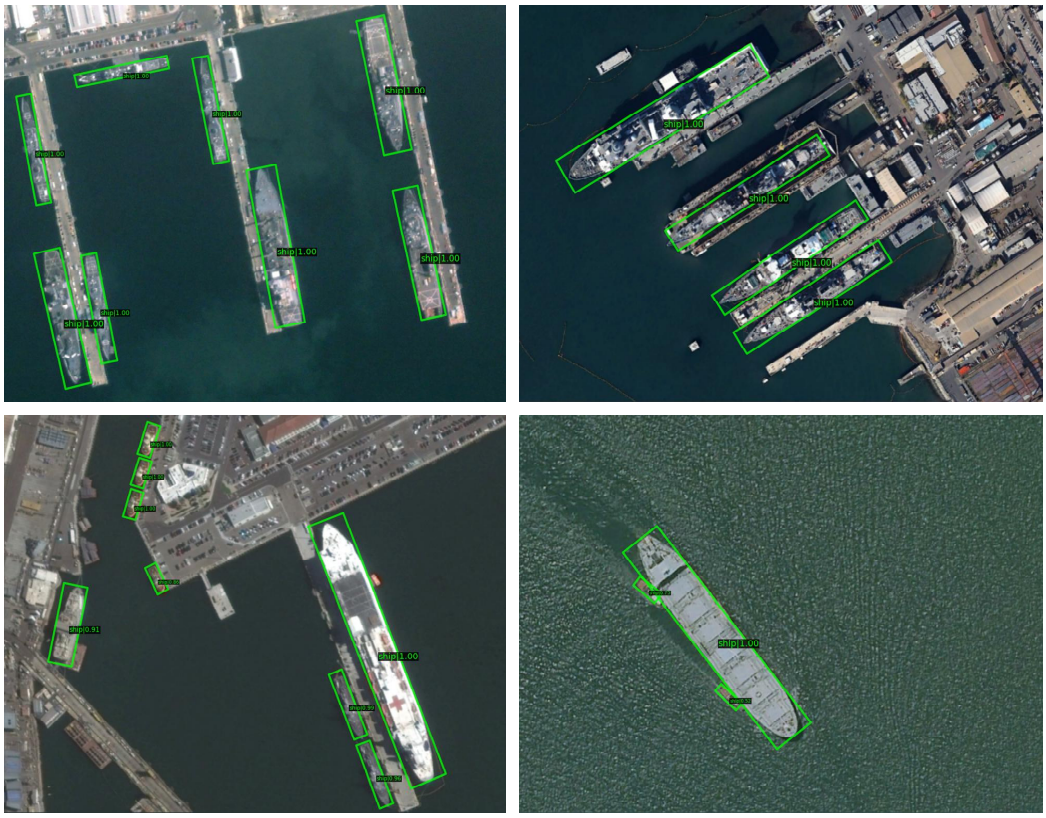


Figure 6. Detection results visualization of our proposed method on the HRSC2016 dataset.

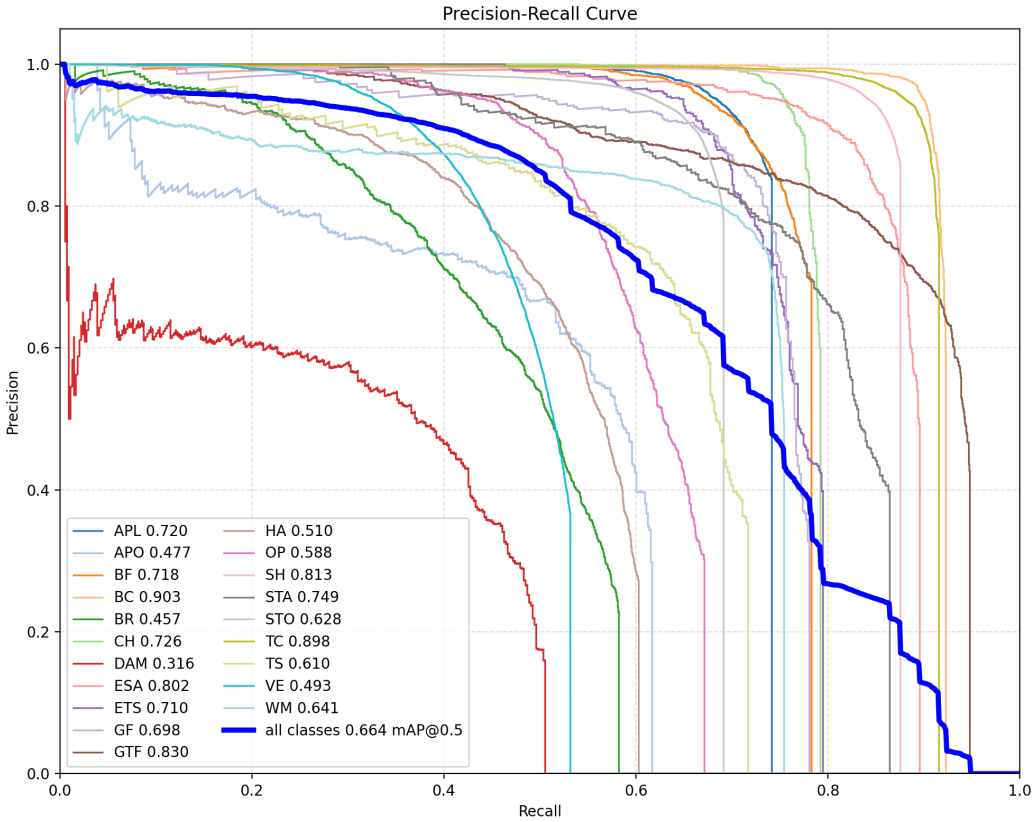


Figure 7. The PR curves of all categories on the DIOR-R dataset.

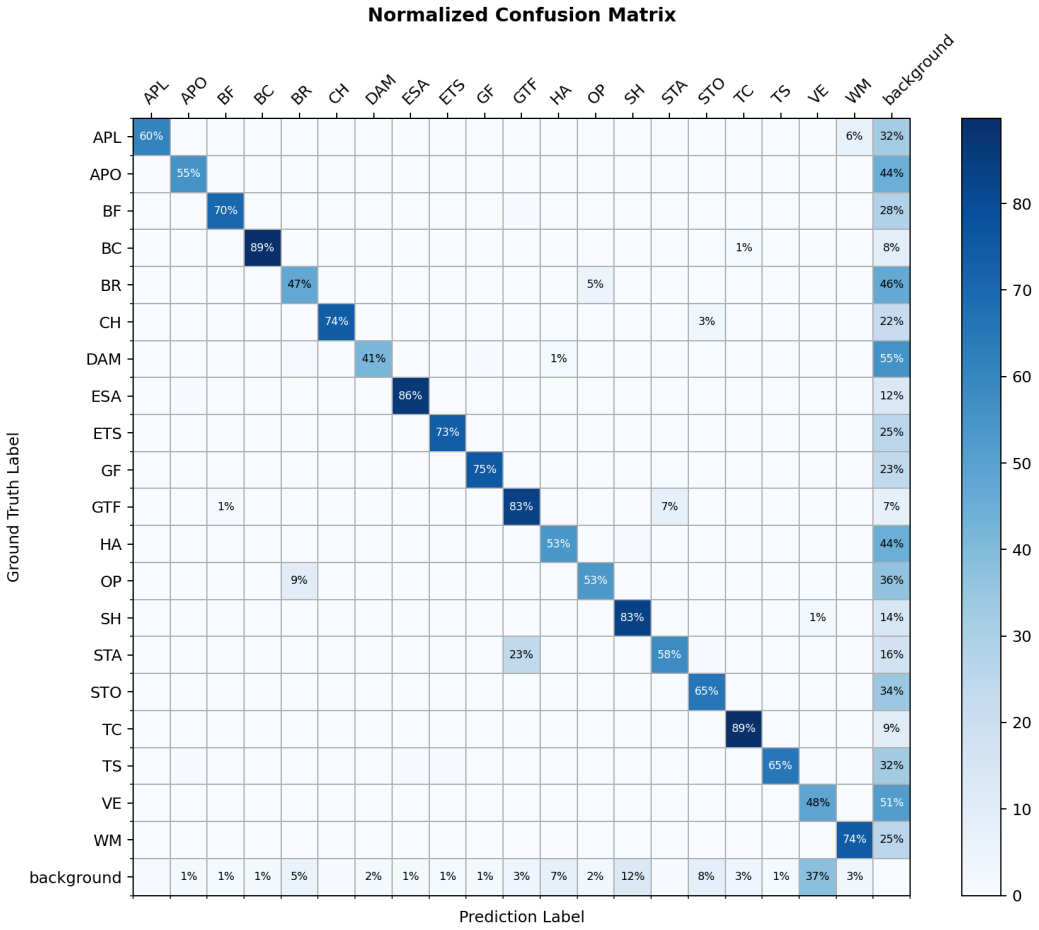


Figure 8. The confusion matrix on the DIOR-R dataset.

Table 3. Comparison with state-of-the-art methods on the DIOR-R dataset. Best (bold) and second-best (underlined) results are highlighted.

Methods	**RoI Trans. [3]**	**Gliding Vertex [7]**	**S^2A-Net [10]**	**ReDet [13]**	**AOPG [6]**	**O-RepPoints [27]**	**ARS-DETR [28]**	**ORSR [33]**	**RO^2-DETR [57]**	**RedhDet**
Backbone	R-50	R-50	R-50	ReR-50	R-50	R-50	R-50	ReR-50	R-50	LSIEN-R50
APL	63.34	65.35	67.98	**77.15**	62.39	67.80	68.00	63.16	69.78	71.79
APO	37.88	28.87	44.44	43.68	37.79	48.01	54.17	52.24	**54.77**	47.67
BF	71.78	74.96	71.63	71.66	71.62	**77.02**	74.43	71.54	74.49	71.80
BC	87.53	81.33	81.39	87.41	87.63	85.37	81.65	89.44	82.11	**90.35**
BR	40.68	33.88	42.66	43.70	40.90	38.55	41.13	39.47	42.30	**45.69**
CH	72.60	74.31	72.72	73.40	72.47	78.45	75.66	**80.72**	76.66	72.63
DAM	26.86	19.58	27.08	34.55	31.08	33.67	**34.89**	34.19	33.94	31.55
ETS	**78.71**	70.72	70.40	68.55	65.42	72.06	73.07	71.13	72.99	70.99
ESA	68.09	64.70	79.03	79.33	77.99	81.13	81.92	79.36	**82.01**	80.21
GF	68.96	72.30	75.56	69.57	73.20	76.00	76.10	70.94	**76.11**	69.81
GTF	82.74	78.68	81.02	76.23	81.94	79.89	78.62	81.68	78.56	**83.04**
HA	47.71	37.22	43.41	45.49	42.32	45.72	36.33	48.04	40.11	**51.02**
OP	55.61	49.64	56.45	**59.02**	54.45	54.72	55.41	57.55	56.06	58.82
SH	81.21	80.22	81.12	81.24	81.17	**85.13**	84.55	80.76	85.06	81.32
STA	**78.23**	69.26	68.00	64.46	72.69	76.04	70.09	65.56	73.11	74.93
STO	70.26	61.13	70.03	69.31	71.31	65.27	**72.23**	61.46	72.19	62.81
TC	81.61	81.49	87.07	81.42	81.49	85.38	81.14	88.85	84.10	**89.84**
TS	54.86	44.76	53.88	49.18	60.04	59.76	**61.52**	56.62	60.66	61.03
VE	43.27	47.71	51.12	48.24	52.38	48.02	50.57	46.23	**53.02**	49.25
WM	65.52	65.04	65.31	65.20	69.99	68.92	70.28	64.56	**71.11**	64.15
mAP (%) ↑	63.87	60.06	64.52	64.44	64.41	66.35	66.10	65.18	66.43	**66.44**

4.4. Ablation Studies

This section evaluates the effectiveness of the proposed method through a series of ablation studies conducted on DOTA-1.0, HRSC2016, and DIOR-R datasets. It is important

that RoI Transformer [3] serves as the baseline model and adopts the original ResNet-50 combined with FPN as its backbone.

4.4.1. Ablation of the Main Components

The contribution of each component is assessed through an ablation study, as shown in Table 4. All configurations in Table 4 are evaluated using the same dataset split, IoU thresholds, and AP computation protocol to ensure a consistent comparison. It confirms that LSIEN (with RiRoI) and the dual-head consistently improve upon the baseline performance. When both components are used jointly, the detector achieves its best performance, reaching 76.37% mAP_{50} on DOTA-1.0 and 90.21% mAP_{50} on HRSC2016. Furthermore, our approach attains gains of 2.09% mAP_{50} and 14.84% mAP_{75} on DOTA-1.0, as well as 14.12% mAP_{50} and 51.85% mAP_{75} on HRSC2016. These results demonstrate that each component of our method is effective, and the full detector significantly improves performance over the baseline. The substantial improvement in mAP_{75} on HRSC2016 can be primarily attributed to the improved localization accuracy for elongated ship targets. Due to their large aspect ratios, even small deviations in orientation or bounding box boundaries can lead to significant reductions in IoU, particularly under the stricter threshold of 0.75. The proposed LSIEN, combined with RiRoI Align, enhances the geometric consistency and rotation alignment of object features, while the adopted dual-head design further supports precise OBB regression. Their combination improves mAP_{75} to 80.45%, indicating that both components contribute to the substantial localization improvement.

Table 4. Ablation study of each structure in our proposed method on DOTA-1.0 and HRSC2016 datasets. The bold value represents the best result. RR denotes random rotation augmentation.

Dataset	LSIEN+RiRoI	Dual-Head	RR	mAP_{50} (%) ↑	mAP_{75} (%) ↑
DOTA-1.0	–	–	–	72.30	34.11
			✓	74.28	34.15
	✓		✓	76.12 (+1.84)	48.85 (+14.70)
		✓	✓	76.23 (+1.95)	48.19 (+14.04)
	✓	✓	✓	**76.37 (+2.09)**	**48.99 (+14.84)**
HRSC2016	–	–	–	74.84	28.55
			✓	76.09	28.60
	✓		✓	90.17 (+14.08)	79.83 (+51.23)
		✓	✓	90.16 (+14.07)	76.18 (+47.58)
	✓	✓	✓	**90.21 (+14.12)**	**80.45 (+51.85)**

4.4.2. Hyperparameter Selection

The hyperparameters investigated in this section are derived from the loss function in Equation (8). Following the best-performing settings reported by Wu et al. [37], $\lambda^{reg} = 2.5$ and $\lambda^{cls} = 2.0$ are adopted as the reference values. To investigate the sensitivity of these loss weights, λ^{reg} is varied over $\{2.0, 2.5, 3.0\}$ and λ^{cls} over $\{1.5, 2.0, 2.5\}$, with the corresponding values paired in ascending order. In addition, the equivariance loss weight λ^{e} is evaluated at $\{1.0, 2.0\}$, while λ^{rpn} is fixed at 1.0. This results in six parameter configurations as shown in Table 5. All configurations are trained using the combined training and validation sets, and their performance is evaluated on the DOTA-1.0 test set. The results indicate that the mAP performance fluctuates within the range of 75.53% to 76.37%, with the optimal performance achieved when $\lambda^{e} = 1$, $\lambda^{reg} = 2.5$, and $\lambda^{cls} = 2.0$.

Table 5. Analysis of the effects of different parameters on the DOTA-1.0 dataset. The bold value represents the best result.

λ^{e}	λ^{reg}	λ^{cls}	mAP (%) ↑
1.0	2.0	1.5	75.94
1.0	2.5	2.0	**76.37**
1.0	3.0	2.5	75.95
2.0	2.0	1.5	75.88
2.0	2.5	2.0	75.53
2.0	3.0	2.5	75.96

4.4.3. Effect of Rotation-Equivariant Backbone

To evaluate the performance differences of LSIEN under different rotation transformation groups, we conduct the experiments in Table 6 based on ReDet (without considering the impact of the dual-head). The results show that G_8 provides improvements of 0.24% mAP_{50}, 0.26% mAP_{75} on DOTA-1.0, and 0.16% mAP_{50}, 0.63% mAP_{75} on HRSC2016, compared with G_4. Although G_8 achieves slightly higher detection accuracy, its training time increases from approximately 9.1 h for G_4 to 14.3 h, representing a 57% increase due to the additional computation of rotated features. Therefore, G_4 offers a more efficient alternative when training resources are limited, whereas G_8 is preferred when detection accuracy is prioritized. To further verify the correctness of the ladder-shaped connections for loss computation and the removal of group pooling operation, Table 6 includes comparisons with IEN and a feature-deformed group pooling variant (LSIEN-GP). The results demonstrate that, under both G_4 and G_8 transformations, LSIEN consistently outperforms IEN and LSIEN-GP on both datasets.

Table 6. Performance comparison of different backbone under different rotation transformation groups on DOTA-1.0 and HRSC2016 datasets. The bold value represents the best result.

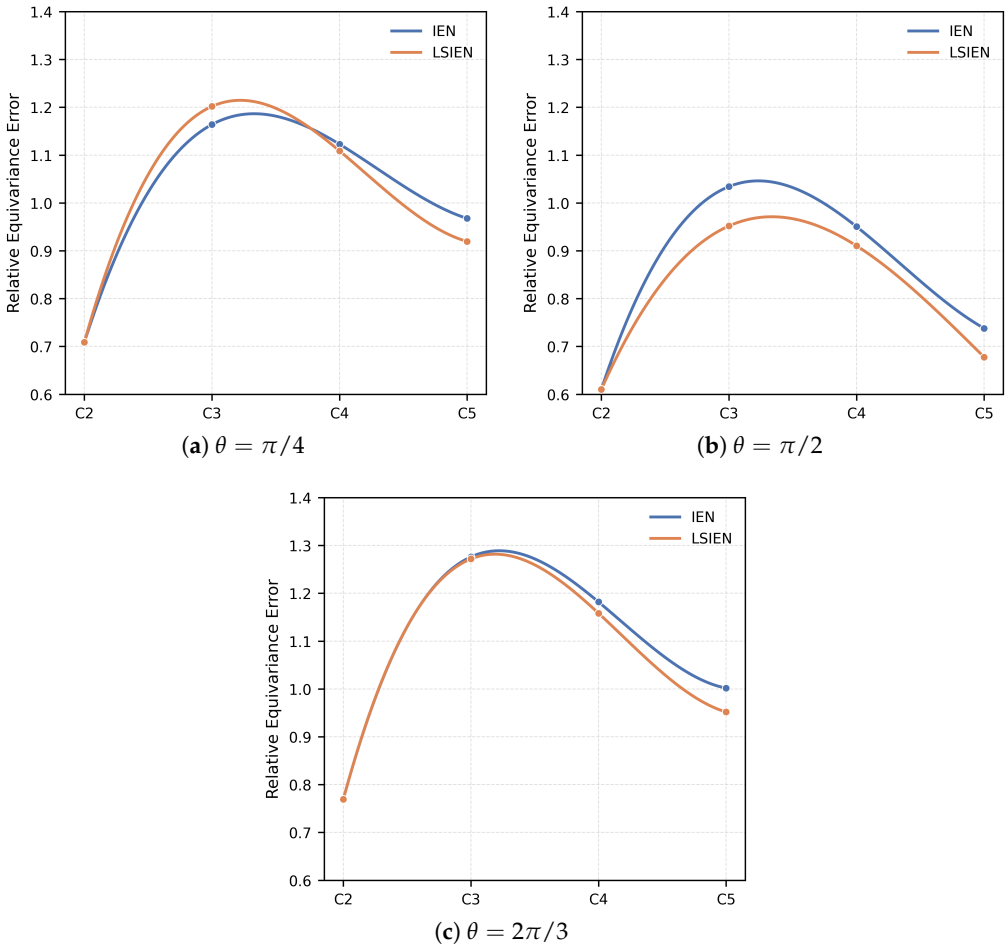
G_N	Backbone	DOTA-1.0		HRSC2016	
		mAP_{50} (%) ↑	mAP_{75} (%) ↑	mAP_{50} (%) ↑	mAP_{75} (%) ↑
G_4	IEN [16]	74.31	47.56	89.24	76.14
	LSIEN-GP	75.70	47.65	89.85	78.63
	LSIEN	**75.88**	**48.59**	**90.01**	**79.20**
G_8	IEN [16]	74.34	47.98	89.43	78.50
	LSIEN-GP	75.68	48.44	89.93	79.43
	LSIEN	**76.12**	**48.85**	**90.17**	**79.83**

To further validate whether the proposed LSIEN alleviates equivariance error accumulation, we conduct an additional layer-wise equivariance error analysis on the DOTA-1.0 test set. Specifically, for each backbone stage, the normalized equivariance error is averaged over all test images and defined as

$$E_i = \frac{1}{M}\sum_{m=1}^{M}\frac{\left\|f_i(\phi_g(x_m)) - \phi_g(f_i(x_m))\right\|_2}{\left\|\phi_g(f_i(x_m))\right\|_2 + \epsilon}, \tag{17}$$

where x_m denotes the m-th test image, M is the total number of test images, $f_i(\cdot)$ denotes the feature output of the i-th backbone stage, $\phi_g(\cdot)$ denotes a rotation transformation by an evaluation angle $\theta \in \{\pi/4, \pi/2, 2\pi/3\}$, and ϵ is a small constant for numerical stability. All models are evaluated in inference mode using their trained checkpoints, without any stochastic training operations. Figure 9 compares the layer-wise equivariance errors under $\pi/4$, $\pi/2$, and $2\pi/3$ rotations, covering both in-group and out-of-group transformations.

Across all three angles, the errors increase from C2 to C3 and subsequently decrease toward C5, with both methods exhibiting identical errors at C2. Compared with IEN, LSIEN consistently achieves lower errors at C4 and C5, although a slightly higher error is observed at C3 under $\pi/4$ rotation. Furthermore, the overall errors are highest under $2\pi/3$ rotation, and the improvement achieved by LSIEN is less pronounced than that under $\pi/2$, suggesting that the tested out-of-group rotation remains more challenging. These results indicate that the ladder-structured constraint provides more stable equivariant feature propagation and helps mitigate the accumulation of equivariance errors across backbone stages.



(**a**) $\theta = \pi/4$ (**b**) $\theta = \pi/2$

(**c**) $\theta = 2\pi/3$

Figure 9. Layer-wise comparison of relative equivariance errors between IEN and the proposed LSIEN on the DOTA-1.0 test set under different rotation angles. Lower values indicate better rotation-equivariant feature consistency.

4.4.4. Effect of Dual-Branch Head

To verify the suitability of the fc-head and the conv-head for classification and regression tasks, respectively, experiments are conducted on four head combinations. The DIOR-R dataset is selected for this study because it contains a larger number of categories and provides publicly available ground-truth annotations for the test set than the DOTA-1.0 dataset. Furthermore, to better evaluate classification performance and localization accuracy, the Macro-F1 is introduced in addition to mAP. The corresponding results are reported in Table 7. It can be observed that when employing the fc-head for classification and the conv-head for regression, the model achieves the best values on both metrics. This demonstrates the suitability of fully connected and convolutional heads for their respective tasks.

Table 7. Comparison of different classification and regression head combinations on the DIOR-R dataset. Macro-F1 is reported at IoU = 0.5 and score threshold = 0.3. The bold value represents the best result.

Classification Type	Regression Type	mAP (%) ↑	Macro-F1 ↑
conv-head	fc-head	65.81	69.77
conv-head	conv-head	65.90	68.93
fc-head	fc-head	65.92	69.52
fc-head	conv-head	**66.44**	**70.91**

4.4.5. Visualization Explanation

To intuitively examine the effectiveness of the proposed method in extracting rotation-equivariant features, we visualize the feature heatmaps of RoI Transformer (baseline) and RedhDet (ours) on the DOTA-1.0 dataset, as shown in Figure 10. It can be observed that feature heatmaps of RedhDet highlight regions that align more closely with the true object contours and exhibit a clearer foreground-background contrast. Moreover, after rotating the input image, the features extracted by the baseline detector display substantial deviations, whereas our method remains nearly rotation equivariance.

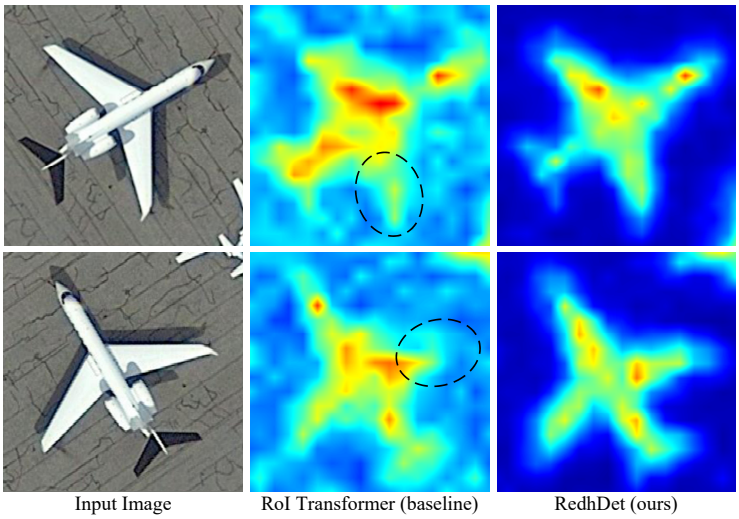


Figure 10. Visualization of feature heatmap of RoI Transformer (baseline) and RedhDet (ours) on the DOTA-1.0 dataset. Warmer colors indicate stronger feature responses. The dashed circles highlight regions where the baseline exhibits less consistent feature responses than RedhDet.

4.4.6. Computational Cost and Efficiency

To evaluate the computational efficiency of the proposed method, the number of parameters (Params), floating-point operations (FLOPs), and inference speed (FPS) are compared with those of RoI Transformer [3] and ReDet [13], as shown in Table 8. Params and FLOPs are measured using the MMRotate complexity analysis tool under a fixed input size, with FLOPs representing the estimated operations supported by the profiling tool. Compared with RoI Transformer, RedhDet improves mAP_{50} by 2.09 percentage points on DOTA-1.0, while maintaining the same reported Params (55.13 M) and FLOPs (225.29 G). Meanwhile, the FPS decreases slightly from 21.6 to 21.1, indicating a favorable accuracy-efficiency trade-off with limited inference overhead. Although ReDet requires fewer parameters and FLOPs, its measured FPS is lower than that of RedhDet, suggesting that theoretical computational complexity does not necessarily correspond directly to actual inference speed.

Table 8. Comparison of computational cost and inference efficiency among different detectors. The bold value represents the best result.

Methods	Params (M) ↓	FLOPs (G) ↓	FPS (img/s) ↑
RoI Trans. [3]	55.13	225.29	**21.6**
ReDet [13]	**31.64**	**80.22**	14.4
RedhDet (Ours)	55.13	225.29	21.1

5. Conclusions

In this paper, we propose a novel framework for ORSOD by unifying rotation-equivariant feature learning with dual-head decoupling. At the feature-learning level, the proposed Ladder-Structured Implicitly Equivariant Network (LSIEN) provides a new implicit rotation-equivariant backbone design that alleviates the major limitations of previous implicit-equivariant methods, including equivariance error accumulation, the loss of fine-grained details caused by group pooling, and constraints on input aspect ratios. At the prediction level, the introduced dual-branch decoupled head separates classification from OBB regression, enabling each task to benefit from a more suitable architectural form and more appropriate feature representations. Extensive experiments and ablation studies on DOTA-1.0, HRSC2016, and DIOR-R demonstrated the effectiveness of the proposed method. The results indicate that RedhDet is a practical and effective solution for remote sensing applications, with potential applicability to other rotation-equivariant vision tasks.

6. Patents

The work reported in this manuscript has resulted in a Chinese invention patent application entitled "A Rotation-Equivariant Dual-Branch Object Detection Method and System for Remote Sensing Images", Application No. 202610253798.3.

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

Funding: This research was funded in part by the National Key Research and Development Program of China under Grant No. 2025YFE0214901, in part by the National Natural Science Foundation of China under Grant No. T2600681, and in part by the Heilongjiang Provincial Natural Science Foundation under Grant No. QC2026F009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are publicly available. The DOTA-1.0 dataset can be accessed at https://captain-whu.github.io/DOTA/dataset.html (accessed on 1 September 2026), the HRSC2016 dataset at https://aistudio.baidu.com/datasetdetail/54106 (accessed on 1 September 2026), and the DIOR-R dataset at https://drive.google.com/drive/folders/1UdlgHk49iu6WpcJ5467iT-UqNPpx__CC (accessed on 1 September 2026). No new data were created in this study.

Acknowledgments: The authors would like to thank the open-source community for providing publicly available datasets and codebases that facilitated this research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ORSOD	oriented remote sensing object detection
HBB	horizontal bounding box
OBB	oriented bounding box
CNN	convolutional neural network
E2CNN	e(2)-equivariant convolutional neural network
RoI Align	region of interest align
IEN	implicitly equivariant network
LSIEN	Ladder-Structured Implicitly Equivariant Network
RedhDet	Rotation-equivariant Dual-head Detector
GP	group pooling
FPN	feature pyramid network
GPU	graphics processing unit
SGD	stochastic gradient descent
mAP	mean average precision
IoU	intersection over union
TP	true positive
FP	false positive
FN	false negative
AP	average precision
PR	precision–recall
RR	random rotation augmentation
Params	model parameters
FLOPs	floating-point operations
FPS	frames per second

References

1. Gui, S.; Song, S.; Qin, R.; Tang, Y. Remote sensing object detection in the deep learning era—A review. *Remote Sens.* **2024**, *16*, 327. [CrossRef]
2. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8231–8240. [CrossRef]
3. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858. [CrossRef]
4. Dai, L.; Liu, H.; Tang, H.; Wu, Z.; Song, P. AO2-DETR: Arbitrary-oriented object detection transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 2342–2356. [CrossRef]
5. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [CrossRef]
6. 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]
7. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1452–1459. [CrossRef] [PubMed]
8. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 677–694. [CrossRef]
9. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the 2021 AAAI Conference on Artificial Intelligence (AAAI), Virtual, 2–9 February 2021; pp. 3163–3171. [CrossRef]
10. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602511. [CrossRef]
11. Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; Tian, Q. The KFIoU loss for rotated object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023; pp. 1–18.

12. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3500–3509. [CrossRef]
13. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A rotation-equivariant detector for aerial object detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 20–25 June 2021; pp. 2785–2794. [CrossRef]
14. Weiler, M.; Cesa, G. General E(2)-equivariant steerable CNNs. In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 14334–14345.
15. Weissenbacher, M.; Agarwal, R.; Kawahara, Y. SiT: Symmetry-invariant transformers for generalisation in reinforcement learning. In Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024; pp. 52695–52719.
16. Khetan, N.; Arora, T.; ur Rehman, S.; Gupta, D.K. Implicitly rotation equivariant neural networks. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
17. Zhang, X.; Wu, Y.; Lin, D.; Wang, B.; Zhang, G.; Wang, X.; Sun, X.; Wan, K. RedhDet: A Rotation-Equivariant Dual-Head Detector for Oriented Remote Sensing Object Detection. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Harbin, China, 22–25 August 2026.
18. 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.* **2023**, *45*, 2384–2399. [CrossRef] [PubMed]
19. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic anchor learning for arbitrary-oriented object detection. In Proceedings of the 2021 AAAI Conference on Artificial Intelligence (AAAI), Virtual, 2–9 February 2021; pp. 2355–2363. [CrossRef]
20. Hou, L.; Lu, K.; Xue, J.; Li, Y. Shape-adaptive selection and measurement for oriented object detection. In Proceedings of the 2022 AAAI Conference on Artificial Intelligence (AAAI), Virtual, 22 February–1 March 2022; pp. 923–932. [CrossRef]
21. Zhang, X.; Wu, Y.; Zhang, G.; Yuan, Y.; Cheng, G.; Wu, Y. Shape-dependent dynamic label assignment for oriented remote sensing object detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 132–146. [CrossRef]
22. Huang, Z.; Li, W.; Xia, X.G.; Tao, R. A general Gaussian heatmap label assignment for arbitrary-oriented object detection. *IEEE Trans. Image Process.* **2022**, *31*, 1895–1910. [CrossRef] [PubMed]
23. Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; Yan, J. Dense label encoding for boundary discontinuity free rotation detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 20–25 June 2021; pp. 15819–15829. [CrossRef]
24. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking rotated object detection with gaussian wasserstein distance loss. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 11830–11841.
25. Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; Yan, J. Learning high-precision bounding box for rotated object detection via Kullback–Leibler divergence. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; pp. 18381–18394.
26. Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; Metaxas, D. Oriented object detection in aerial images with box boundary-aware vectors. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 2149–2158. [CrossRef]
27. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1819–1828. [CrossRef]
28. 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*, 1–15. [CrossRef]
29. Cohen, T.S.; Welling, M. Group Equivariant Convolutional Networks. In Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 20–22 June 2016; pp. 2990–2999.
30. Worrall, D.E.; Garbin, S.J.; Turmukhambetov, D.; Brostow, G.J. Harmonic Networks: Deep Translation and Rotation Equivariance. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5028–5037. [CrossRef]
31. 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 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 6589–6600. [CrossRef]
32. 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]
33. 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]
34. Tian, Y.; Zhang, M.; Li, J.; Li, Y.; Yang, H.; Li, W. FPNFormer: Rethink the method of processing the rotation-invariance and rotation-equivariance on arbitrary-oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5605610. [CrossRef]

35. 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. [CrossRef]
36. Lee, C.; Son, J.; Shon, H.; Jeon, Y.; Kim, J. FRED: Towards a full rotation-equivariance in aerial image object detection. In Proceedings of the 2024 AAAI Conference on Artificial Intelligence (AAAI), Vancouver, Canada, 20–27 February 2024; pp. 2883–2891. [CrossRef]
37. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10183–10192. [CrossRef]
38. Liu, C.; Xie, S.; Li, X.; Gao, J.; Xiao, W.; Fan, B.; Peng, Y. Mitigate the Classification Ambiguity via Localization-Classification Sequence in Object Detection. *Pattern Recognit.* **2023**, *138*, 109418. [CrossRef]
39. Zhao, Z.; He, C.; Zhao, G.; Zhou, J.; Hao, K. RA-YOLOX: Re-parameterization Align Decoupled Head and Novel Label Assignment Scheme Based on YOLOX. *Pattern Recognit.* **2023**, *140*, 109579. [CrossRef]
40. Zuo, F.; Liu, J.; Chen, Z.; Fu, M.; Wang, L. TGADHead: An Efficient and Accurate Task-Guided Attention-Decoupled Head for Single-Stage Object Detection. *Knowl.-Based Syst.* **2024**, *302*, 112349. [CrossRef]
41. Han, B.; He, L.; Yu, Y.; Lu, W.; Gao, X. General Deformable RoI Pooling and Semi-Decoupled Head for Object Detection. *IEEE Trans. Multimed.* **2024**, *26*, 9410–9422. [CrossRef]
42. Han, B.; He, L.; Ke, J.; Tang, C.; Gao, X. Weighted Parallel Decoupled Feature Pyramid Network for Object Detection. *Neurocomputing* **2024**, *593*, 127809. [CrossRef]
43. Chen, W.; Miao, S.; Wang, G.; Cheng, G. Recalibrating Features and Regression for Oriented Object Detection. *Remote Sens.* **2023**, *15*, 2134. [CrossRef]
44. Huang, Q.; Yao, R.; Lu, X.; Zhu, J.; Xiong, S.; Chen, Y. Oriented Object Detector with Gaussian Distribution Cost Label Assignment and Task-Decoupled Head. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5621916. [CrossRef]
45. He, S.; Yuan, R.; Yang, Z.; Liu, J. Multiscale Task-Decoupled Oriented SAR Ship Detection Network Based on Size-Aware Balanced Strategy. *Remote Sens.* **2025**, *17*, 2257. [CrossRef]
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
47. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
48. 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 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM), Porto, Portugal, 24–26 February 2017; pp. 324–331. [CrossRef]
49. 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]
50. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
51. 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 (ACM MM), Lisboa, Portugal, 10–14 October 2022; pp. 7331–7334. [CrossRef]
52. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT), Paris, France, 22–27 August 2010; pp. 177–186. [CrossRef]
53. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
54. Guan, J.; Xie, M.; Lin, Y.; He, G.; Feng, P. EARL: An elliptical distribution aided adaptive rotation label assignment for oriented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5619715. [CrossRef]
55. Deng, L.; Dong, R.; Zhou, T. PSSANet: Phase-sensitive and scale-adaptive anchor-free framework for geospatial object detection. *Opt. Laser Technol.* **2026**, *203*, 116391. [CrossRef]
56. Zhou, D.; Xiong, Y.; Yu, S.; Feng, P.; Liu, J.; Wu, N.; Dou, R.; Liu, L. RSDB-Net: A novel rotation-sensitive dual-branch network with enhanced local features for remote sensing ship detection. *Remote Sens.* **2025**, *17*, 3925. [CrossRef]
57. Dang, M.; Liu, G.; Kong, A.W.K.; Zheng, Z.; Luo, N.; Pan, R. RO^2-DETR: Rotation-equivariant oriented object detection transformer with 1D rotated convolution kernel. *ISPRS J. Photogramm. Remote Sens.* **2025**, *228*, 166–178. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.