

Article

Weakly Supervised Oriented Object Detection in Remote Sensing via Geometry-Aware Enhancement Network

Yufei Zhu [†] , Jianzhi Hong ^{*,†} and Taoyang Wang ^{ID} 

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China; 2023302131281@whu.edu.cn (Y.Z.); wangtaoyang@whu.edu.cn (T.W.)

* Correspondence: hongjianzhi@whu.edu.cn

[†] These authors contributed equally to this work.

Highlights

What are the main findings?

- The proposed DAE-ResNet and AG-FPN modules significantly enhance feature extraction capabilities for objects with extreme aspect ratios and diverse orientations.
- The adaptive geometric perception loss (AGL) effectively improves the detection accuracy of few-shot categories, demonstrating superior performance on DOTA-v1.0, HRSC, and RSAR datasets.

What are the implications of the main findings?

- The proposed method significantly reduces the reliance on costly oriented bounding box annotations while maintaining high detection precision in complex scenes.
- This study demonstrates the critical role of geometric priors (e.g., angle and shape alignment) in alleviating sample imbalance and improving weakly supervised learning robustness.

Abstract

In remote sensing image oriented object detection tasks, weakly supervised learning methods based on horizontal bounding boxes have attracted much attention due to their lower annotation costs compared to fully supervised methods. However, remote sensing images, characterized by complex backgrounds, exhibit a wide range of target scales and diverse geometric characteristics across target categories. Existing methods exhibit inadequate exploitation of background and angular information under weak supervision, resulting in compromised perception of dense and high-aspect-ratio targets. Neglecting the imbalance in angle estimation samples further leads to excessively low detection accuracy for few-shot categories. To address the aforementioned issues, this paper proposes a Geometry-Aware Enhancement Network (WSOOD-GAEN) for weakly supervised oriented object detection tasks. First, in the backbone network stage, a channel-space deformable attention module (DAE-ResNet) was constructed. Through deformable sampling and screening of key regions, feature extraction has both morphological adaptability to complex shapes and semantic discriminability of key features in complex backgrounds. Secondly, in the feature pyramid stage, an Angle-Guided Feature Pyramid Network (AG-FPN) is proposed. This module dynamically applies rotation transformation to the sampling offsets of deformable convolutions, thereby enhancing the feature representation of objects with different orientations and scales. Furthermore, an adaptive geometric perception loss (AGL) was designed. Based on the geometric characteristics of different categories, it automatically learns differentiated rotation and flip consistency weights, thereby improving the prediction accuracy



Academic Editors: Jianming Hu, Xiyang Zhi, Yongqiang Mao and Longfei Ren

Received: 14 March 2026

Revised: 14 April 2026

Accepted: 18 April 2026

Published: 21 April 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

of small sample categories. Experiments on the DOTA-v1.0, HRSC, and RSAR datasets validate our approach. Specifically, under the AP75 evaluation metric, the proposed method outperforms existing weakly supervised methods by 1.51%, 9.86%, and 3.28%, respectively.

Keywords: oriented object detection; weak supervision; remote sensing; feature alignment; adaptive weighting

1. Introduction

With the rapid development of remote sensing technology, object detection in high-resolution images has become essential for tasks such as urban monitoring and military reconnaissance [1,2]. In recent years, using Convolutional Neural Networks (CNN) for feature extraction has greatly improved general object detection performance [3–5]. However, applying these methods directly to remote sensing fields faces significant challenges. Unlike natural images, objects in remote sensing images often show large scale variations and are densely distributed [6–8]. Moreover, targets like bridges, ships, and vehicles usually appear in arbitrary directions and have very large aspect ratios. In such complex scenes, traditional horizontal bounding boxes (HBox) often include substantial background noise, lowering the quality of feature extraction, and struggle to handle crowded objects [9–11]. For example, in ports or airports, the high overlap between horizontal boxes may cause non-maximum suppression (NMS) to mistake real targets for duplicates. In contrast, oriented object detection (OBB) adds an angle parameter to create tighter boxes that fit the object's shape more accurately, effectively separating dense targets. As a result, OBB provides more accurate localization and has become a key technique for high-precision remote sensing analysis [12].

However, most existing remote sensing benchmark datasets are annotated with HBox. While this approach is more cost-effective than using rotated boxes [13], it has resulted in a significant gap between the currently available training dataset and the growing demand for oriented object detection [14]. Weakly supervised oriented object detection (WS-OOD) provides an effective solution to this issue. Weakly supervised oriented object detection based on HBox aims to reduce the cost of manual annotation by using coarser-grained labels for training compared to fully supervised methods, while preserving, to the greatest extent possible, the accuracy of fully supervised methods [15]. This method does not rely on precisely labeled rotated bounding boxes (RBox), solving the problem of time-consuming and laborious oriented bounding box annotation in dense or extremely small-scale remote sensing object scenes. Furthermore, it can utilize existing large-scale horizontal labeling datasets without the need for secondary labeling, making it a more efficient alternative.

Furthermore, as shown in Figure 1, in contrast to other forms of weakly supervised labeling (such as image-level labels and point annotations), horizontal bounding box-based supervision strikes a superior balance between labeling cost and model performance. As it provides more precise spatial location and scale information, this method offers stronger geometric priors, thereby reducing the uncertainty in inferring rotated bounding boxes despite the lack of angular information. Therefore, using HBox as weak supervision signals offers greater practicality and robustness than other weak labeling methods, making it the optimal choice for achieving high-performance oriented object detection while ensuring low cost.

Weakly supervised object detection methods based on HBox can be broadly categorized into three types based on the weak label derivation process: segmentation-driven methods, region proposal-based methods, and transformation consistency-based meth-

ods. Compared to segmentation-driven methods which are limited by mask quality, and proposal-based methods which lack adaptability to objects with large aspect ratios, transformation consistency-based methods exhibit distinct advantages: avoiding complex intermediate mask generation processes and proposal presets. By constructing a weakly supervised–self-supervised network and imposing geometric consistency constraints between views, end-to-end angle adaptive learning is achieved, resulting in regression efficiency and accuracy far exceeding other methods.

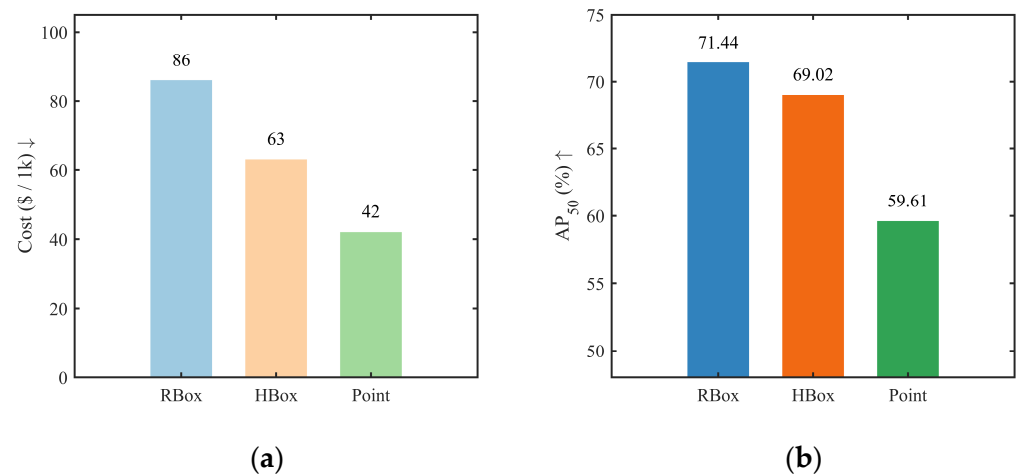


Figure 1. Cost and AP₅₀ comparison across RBox supervision, HBox supervision, and point supervision. (a) Annotation cost comparison across different supervision types, where \$/1k denotes the cost per thousand images. (b) The best AP₅₀ achieved in our reproduced results for each supervision type.

As a representative method based on transformation consistency, H2RBox [16] achieves performance close to that of fully supervised methods. It follows the idea of self-supervised learning, generating rotated and flipped transformed views based on the same image and learning consistency among the prediction results of the three views, thus reconstructing angle information even without angle labels. The improved H2RBox-v2 [17] introduces prior knowledge of object symmetry, allowing the self-supervised branch to learn angles independently. Simultaneously, a new angle encoding method is used to transform the angle regression problem into a continuous phase prediction problem, effectively solving the boundary discontinuity problem caused by angle periodicity. Building on this, subsequent researchers optimized the angle encoding method, proposing the Unit Circle Angle Resolver (UCR) and incorporating a unit circle loss to further enhance the model's ability to perceive subtle angle changes [18]. These methods optimize the model's learning of angle and scale information in complex scenes and alleviate the problem of missing angular supervision to some extent. However, in-depth analysis of the results reveals (as shown in Figure 2) that existing methods still face the following three major problems:

- (1) Remote sensing images are predominantly captured from high-altitude overhead views, covering extensive areas with complex backgrounds where targets are frequently densely distributed—this renders feature extraction significantly more challenging than in general images [19,20]. Furthermore, the feature distribution under weak supervision signals is overly dispersed, and relying exclusively on backpropagation of the regression branch makes it arduous for the network to autonomously focus on the key semantic regions of objects. Because traditional backbone design lacks proactive geometric adaptation and semantic filtering of key features, it is susceptible to background clutter in complex backgrounds [21]. As shown in Figure 2a, the prediction box of the ship is not accurate enough and introduces substantial irrelevant background regions.

- (2) Remote sensing images encompass extensive regions, with target scales exhibiting significantly broader variations than those of typical images. Moreover, remote sensing targets are often elongated geometries, requiring feature representations to aggregate along the target's principal axis for accurate identification. Existing methods, in the feature extraction stage, employ conventional convolutional operations, which fail to extract directional information and fail to have sampling points align with object edges, thus limiting the directional representation capabilities of the features [22–24]. As shown in Figure 2b, the feature responses of the port and bridge in the heat map exhibit scattered distributions and fail to be focused around the main axis of the target.
- (3) Remote sensing datasets contain various targets with significantly different geometric characteristics and an imbalanced distribution. However, existing methods employ a fixed consistency loss weight across all categories, and the weight ratios of the flip consistency loss and rotation consistency loss are fixed, leading to lower detection accuracy for some few-shot categories. As shown in Figure 2c, the detection accuracy for few-shot categories is low.

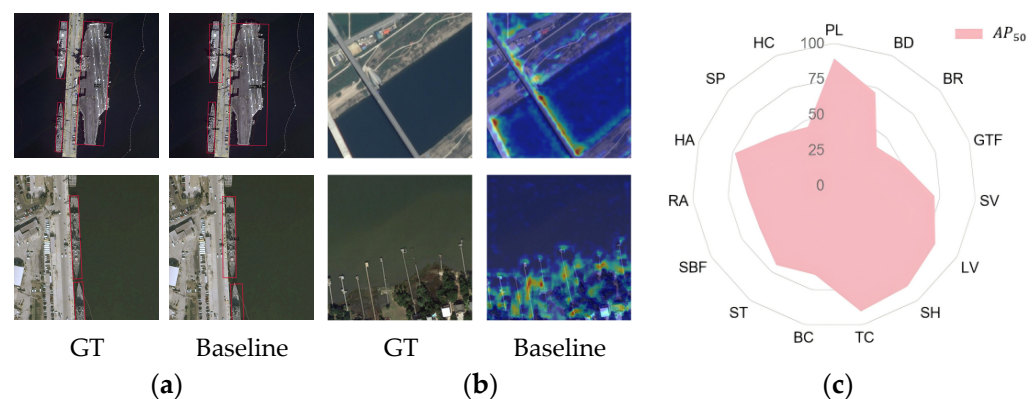


Figure 2. Problems found in visualization. (a) The detection bounding box contains a lot of irrelevant backgrounds. (b) The heat map exhibits scattered distributions. (c) Differences in AP₅₀ accuracy of various object categories in the DOTA-v1.0; and PL, BD, etc., are abbreviations for object categories in the DOTA-v1.0.

To address the aforementioned issues, this paper proposes the WSOOD-GAEN method based on the concepts of multi-scale angle guidance and adaptive form perception, optimizing it from three levels:

At the backbone level, a channel-spatial deformable attention module (DAE-ResNet) was designed to address the feature dispersion problem arising from the shape variability of oriented targets and cluttered backgrounds in remote sensing images. By incorporating deformable convolutions [25], an adaptive deformation mechanism is introduced into the backbone network, enabling it to dynamically align with the geometric contours of oriented objects. Channel and spatial attention calculations are then performed on the aligned features of the oriented objects, allowing the model to automatically focus on the object itself and filter out irrelevant background noise. This design produces features that are better aligned with object geometry while suppressing background noise.

At the feature pyramid level (Neck), to address the limitations of existing weakly supervised methods in orientation representation for multi-scale, high-aspect-ratio remote sensing targets, we propose the Angle-Guided Feature Pyramid Network (AG-FPN). AG-FPN introduces an explicit geometric guidance mechanism, explicitly rotating the offset field of sampling points by predicting dense pixel-level angle maps at the feature layer, thus endowing the feature pyramid with rotational equivariance. Utilizing this prior,

multi-scale features can be aggregated along the principal axis of the object, enhancing the representation capability for oriented objects with high aspect ratios.

At the loss function level, to address the challenges of geometric diversity across target categories and low detection accuracy for few-shot classes in remote sensing datasets, this paper proposes Adaptive Geometry-Aware Loss (AGL). Inspired by the adaptive weighting mechanism of focal loss, AGL assigns learnable weights (including rotation loss weights and flip loss weights) to each category during training, automatically adapting to the geometric characteristics of each class through gradient backpropagation. This improves the prediction accuracy for few-shot class objects. The contributions of this paper are summarized as follows:

- (1) This paper proposes the WSOOD-GAEN method. Compared with other HBox-based weakly supervised methods, this method optimizes feature extraction and loss function from the perspective of fine excavation of orientation information, enhances the semantic discriminativeness of key features in complex backgrounds, and enables feature extraction to more accurately perceive target contours, thereby improving the model's learning ability for high-aspect-ratio objects and small sample objects in remote sensing images.
- (2) The proposed DAE-ResNet is designed to focus on the semantics of key object regions, enabling the backbone to possess both geometric adaptability and semantic discriminativeness between foreground and background, thus solving the problem of feature extraction being easily interfered with by background noise under weak supervision signals.
- (3) We propose AG-FPN, which introduces multi-scale pixel-level angle guidance in the FPN stage to provide angle prior knowledge for subsequent learning in the detection head, thereby achieving higher accuracy detection and positioning of high-aspect-ratio objects.
- (4) We propose AGL, which optimizes geometric consistency loss through class-adaptive weights, allowing the network to automatically adapt to the geometric characteristics of different classes, improving learning efficiency compared to fixed weight strategies.
- (5) Compared with other HBox-based weakly supervised methods, WSOOD-GAEN achieves improvements of 1.5% and 1.21% in AP50 and AP75 on the DOTA-v1.0 [26], 0.69% and 9.86% on the HRSC [27], and 4.27% and 3.35% on the RSAR [18], respectively. These results significantly outperform existing methods, validating the effectiveness of our proposed optimizations.

The remaining chapters are structured as follows. Section 2 introduces related works on oriented object detection and weakly supervised methods. Section 3 elaborates on our WSOOD-GAEN method. Section 4 presents experimental results and analysis. Section 5 summarizes the entire paper.

2. Related Works

Based on the level of supervision required for model training, oriented object detection can currently be divided into fully supervised, semi-supervised, and weakly supervised paradigms. In Earth observation tasks, acquiring large-scale and precise oriented bounding box annotations for massive high-resolution satellite imagery is highly labor-intensive and time-consuming. Therefore, the latter two paradigms can effectively alleviate the heavy annotation burden associated with remote sensing datasets. Unlike weakly supervised methods, semi-supervised learning trains the model using a small number of oriented bounding box annotations and a large amount of unlabeled data [28,29], and still relies on some high-cost bounding box annotations. Therefore, this paper focuses on the more challenging weakly supervised paradigm, which operates without any angle priors.

2.1. Full-Supervision Oriented Object Detection in Remote Sensing

Oriented object detection aims to predict bounding boxes (x, y, w, h, θ) containing orientation information [20]. While fully supervised methods have made progress in areas such as remote sensing imagery [30,31] and scene text detection [32], early approaches directly regressed these parameters, facing numerous challenges in practical applications. Among these challenges, the angular periodicity caused by the arbitrary orientations of remote sensing objects can lead to boundary discontinuity and training divergence; in dense remote sensing scenarios, traditional IoU loss suffers from gradient vanishing for non-overlapping boxes and insensitivity to subtle angular deviations; furthermore, the axis-aligned features of standard convolutions struggle to adapt to the complex rotations of remote sensing targets. Consequently, related research has primarily focused on the following three directions:

In terms of angle encoding strategies, to resolve the periodic ambiguity caused by the arbitrary orientations of remote sensing targets, CSL transforms angle regression into a classification problem by discretizing angles into multiple intervals and introduces circular smooth labels to resolve the boundary discontinuity problem [33]. Building on CSL, DCL introduces dynamic classification labels and utilizes dense coded labels to further improve the prediction accuracy for subtle angle variations in remote sensing targets [34]. In addition, approaches such as denoising-based methods [35] have explored enhancing the feature sensitivity to the angular transformation of remote sensing targets through denoising reconstruction.

Regarding detection architectures, various approaches have been developed to accommodate the arbitrary orientations of remote sensing objects [36–38]. Two-stage methods, such as RoI Transformer [39], align horizontal candidate boxes to rotated targets through spatial transformation. Single-stage methods, such as S2A-Net [40] and R3Det, design feature alignment and refinement modules, effectively alleviating the mismatch between features and anchor boxes for objects with large aspect ratios (e.g., bridges and ships in ports) in remote sensing images. More recently, LSKNet [41] uses a large selective kernel network for dynamic adjustment to better model the ranging context of various objects in remote sensing scenes.

In the design of loss functions, to address the gradient vanishing problem of non-overlapping boxes in dense remote sensing object detection, early methods like PIoU [42] and SCRDet [43] provide continuous gradients by approximating the Intersection over Union (IoU). Rotated IoU loss directly optimizes the IoU through a differentiable computation graph, achieving consistency between the training objective and evaluation metrics. Gaussian modeling-based methods such as GWD [44] and KLD [45] model the rotated boxes as a two-dimensional Gaussian distribution, optimizing bounding box regression in highly dense remote sensing scenes. CircumIoU [46] introduces a circumscribed rectangle constraint, further enhancing regression robustness for objects with high aspect ratios in remote sensing images.

These methods highlight the importance of explicit geometric modeling for oriented object detection in remote sensing. However, their heavy reliance on costly, fine-grained annotations makes them impractical for the massive volumes of data generated by Earth observation systems. Therefore, the key to closing the performance gap between weakly and fully supervised methods lies in adapting these geometric concepts for feature extraction in complex remote sensing scenes, using only HBox as supervision.

2.2. Weak-Supervision Oriented Object Detection in Remote Sensing

To reduce annotation costs, weakly supervised oriented object detection methods have been proposed. These methods learn to predict oriented boxes from coarser-grained

annotations. Based on the richness of the supervision information, existing weakly supervised methods can be mainly divided into point-supervised methods and horizontal box supervision methods:

Point-supervised methods. Early approaches like P2BNet [47] and RepPoints [48] explored the application of point set representation in object detection. Tailored for remote sensing scenes, PointOBB [49] combines weak and self-supervision within a Multiple Instance Learning (MIL) framework, optimizing the model using generated proposals. Point2RBox [50] combines simple geometric priors for object size and rotation with point annotations for classification, enabling joint learning. Furthermore, Point2RBox-v2 [51] leverages the mutual constraints between different instances in the same scene, designing Gaussian overlap and related loss functions to further improve size prediction accuracy. However, due to the scarcity of directly usable datasets, points must still be manually annotated for each instance. Moreover, the limited scale and orientation information provided by such annotations often leads to suboptimal detection accuracy.

Horizontal box-supervised methods. The idea behind the horizontal bounding box-supervised instance segmentation algorithm introduced in Section 1 is to obtain a mask from the horizontal bounding box, and then generate directional bounding boxes based on the mask, such as BoxInst [52] and BoxLevelSet [53]. However, the computational efficiency of the models is limited, and the prediction accuracy is insufficient. Secondly, there are proposal-based or geometric prior-based methods. These algorithms usually generate a large number of candidate boxes with different angles inside the HBox and use geometric constraints or classification confidence to select the best orientation box. However, when facing objects with large aspect ratios, the pre-set anchor boxes are prone to misalignment, resulting in a large deviation in the final selection results. The H2RBox series of methods introduced below solves these problems well.

The H2RBox series of methods is based on transform consistency. H2RBox is a pioneering work that introduces self-supervised learning into horizontal bounding box-supervised oriented detection. By generating multiple views of the same image (original, flipped, and rotated) and imposing consistency constraints between the predicted angles of different views, H2RBox can learn rich angular information in the absence of oriented bounding box annotations. H2RBox-v2 extends the method by introducing flip consistency and rotation consistency losses, leveraging the object's symmetry prior, resulting in a significant improvement in angle prediction performance. Furthermore, subsequent research proposed a Unit Cycle Resolver (UCR) [18], which further enhances performance by mapping angles to points on a unit circle and introducing a corresponding distribution loss. A comprehensive, large-scale rotated SAR object detection dataset was also constructed based on the prediction results and manual refinement [18]. Aside from optimizing angle encoding, other parallel directions based on H2RBox-v2 explore bounding box representations and scale matching. For example, AFWS [54] decouples parameters using concentric circles, and ABBSP0 [55] handles scale variations with symmetric priors.

Nonetheless, the methods described above still have limitations. Feature extraction lacks geometric adaptability, relying on axis-aligned convolutions methods, resulting in extracted features that are often corrupted by complex background noise inherent to remote sensing images. Secondly, the learning of angle information is limited to the regression branch of the detection head, neglecting the potential of angle-guided feature learning in the feature pyramid stage. Finally, the geometric constraint strategy is oversimplified: the consistency loss function uses a uniform weight configuration for all categories, failing to consider the geometric differences between different objects. These limitations undermine the orientation-awareness of feature representations. Consequently, the model struggles to

accurately localize and recognize multi-category oriented objects in complex real-world remote sensing scenarios, such as varying illumination and cloud occlusion.

3. Materials and Methods

This paper proposes WSOOD-GAEN (Weakly Supervised Oriented Object Detection via Fine-Grained Orientation Perception Network) to address the aforementioned limitations. It addresses these issues through three innovations: (1) to overcome the limitation of feature extraction not adapting to object deformation, DAE-ResNet is proposed, which deeply integrates deformable convolution and dual attention mechanisms in the backbone network, endowing the feature extraction process with dynamic geometric adaptability and semantic discriminative ability; (2) to address the lack of directional guidance in multi-scale feature fusion, AG-FPN is proposed, which predicts dense angle maps in the feature pyramid stage and explicitly rotates the sampling offset of the convolution, achieving rotational-equivariant feature alignment at the feature level; (3) to address the limitation of geometric constraint strategies ignoring class differences, AGL is proposed, which automatically learns rotation and flip consistency loss weights adapted to the geometric characteristics of each class through gradient backpropagation, achieving refined regulation of the supervision signal. Compared with existing methods, WSOOD-GAEN achieves multi-level angle guidance from the backbone network to the detection head, providing a more comprehensive solution for horizontal bounding box-supervised oriented object detection.

3.1. Overview of the WSOOD-GAEN Method

In weakly supervised oriented object detection methods, the H2RBox method pioneered the learning of rotated boxes from horizontal box supervision. Our method adopts the overall framework of H2RBox-v2 while incorporating significant improvements. The overall framework, from data augmentation to feature extraction, detection head stage, and loss calculation, is further divided into weakly supervised and self-supervised branches, corresponding to a scale regression branch based on horizontal bounding box supervision and an angle learning branch based on self-supervised geometric constraints.

Specifically, in the data augmentation stage, the original input image and its rotated and flipped views are provided to facilitate subsequent learning of rotation consistency. In the feature extraction stage, we use an improved DAE-ResNet to extract multi-scale features in the backbone network (details in Section 3.2). In the feature pyramid network, we propose AG-FPN, which adds an angle guidance module to the standard FPN (details in Section 3.3). For the detection head, we adopt an anchor-free FCOS-based architecture [56], comprising classification, regression and centrality branches. The loss calculation is divided into two branches: weakly supervised and self-supervised. The calculation process in the weakly supervised branch is the same as that in H2RBox-v2. The self-supervised branch calculates the angle learning loss, accurately locating the object's orientation by incorporating rotation consistency constraints and flip consistency constraints, and also includes the unit circle loss mentioned earlier. In this branch, we added Adaptive Geometry-Aware Loss (AGL) to learn adaptive weights for each category; details are in Section 3.4. The framework is shown in Figure 3.

3.2. Deformable Attention Enhancement ResNet

The standard ResNet backbone network uses a standard 3×3 convolution in the second layer, which lacks adaptability to geometric changes in rotated objects and has significant room for optimization in information utilization. DAE-ResNet improves upon the ResNet Bottleneck structure by changing the standard Bottleneck to:

$$F_2 = \text{RELU}(\text{BN}(\text{DCN}_{3 \times 3}(F_1))) \quad (1)$$

$$F_4 = \text{ChannelAttention}(F_3) \quad (2)$$

$$F_5 = \text{SpatialAttention}(F_4) \quad (3)$$

The second 3×3 convolution is replaced with a deformable convolution, and channel attention and spatial attention are applied sequentially before the residual connection. The first and third layers remain unchanged, as shown in Figure 4.

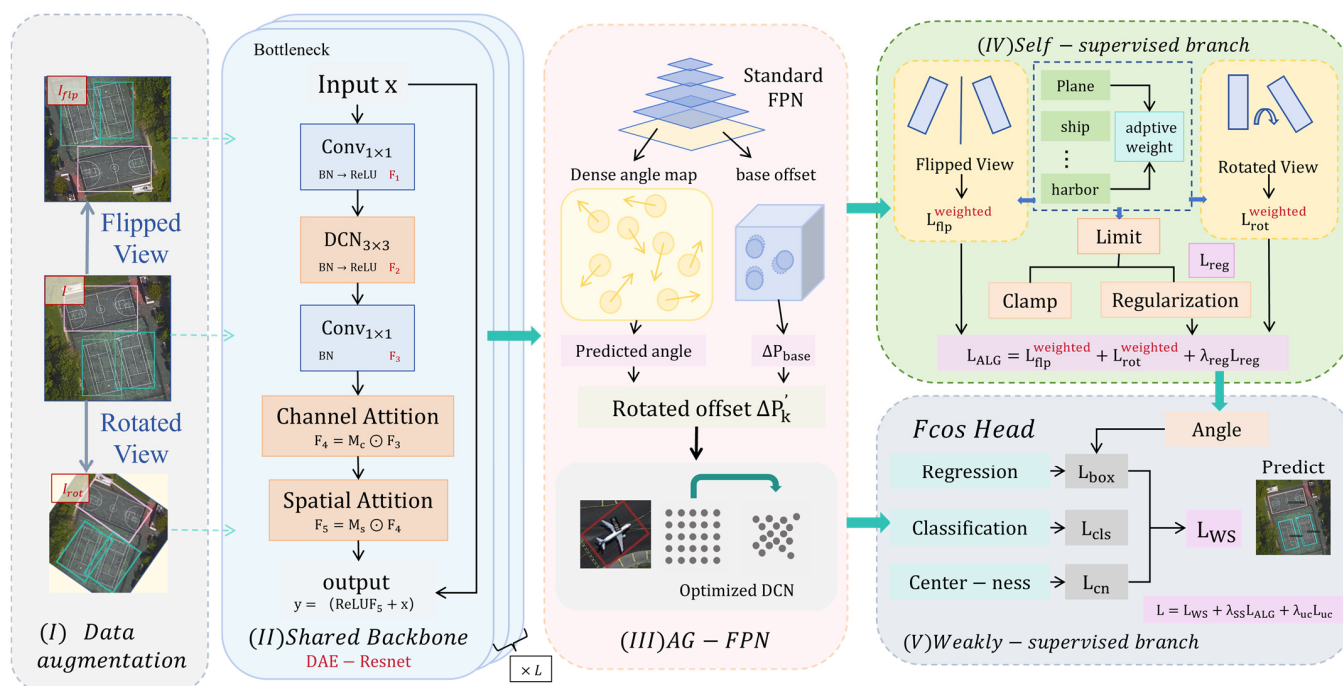


Figure 3. Framework of the WSOOD-GAEN method.

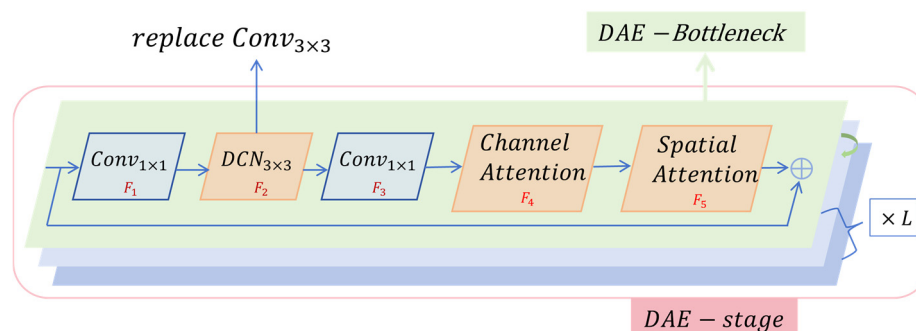


Figure 4. Construction Framework of DAE.

Specifically, the weights for channel attention and spatial attention are calculated as follows:

$$M_{c1} = \sigma(W_2 \cdot \text{RELU}(W_1 \cdot z^{avg}) + W_2 \cdot \text{RELU}(W_1 \cdot z^{max})) \quad (4)$$

$$M_{c2} = \sigma(\text{Conv}_{7 \times 7}(F^{avg}; F^{max})) \quad (5)$$

In this model, z^{avg} and z^{max} represent global average pooling and max pooling, respectively; F^{avg} and F^{max} represent the average and maximum values along the channel dimension; $\sigma(\cdot)$ is the sigmoid activation function; and $[\cdot; \cdot]$ represents channel concatenation, resulting in $2 \times H \times W$ features. The first formula assigns weights between 0 and 1 to each channel, suppressing irrelevant channels and enhancing useful ones. The second

formula goes further, assigning weights to each spatial location to highlight object regions and suppress background regions. In each step, the weighted modulated features can be obtained using $F' = M_c \odot F$.

Compared to the fixed sampling positions of ResNet's convolutional kernels, the proposed deformable attention enhancement strategy mitigates the lack of geometric adaptability to rotated objects by incorporating explicit semantic guidance into deformable convolutions. This allows the backbone to simultaneously possess adaptability to diverse morphologies (adapting to object deformation through deformable convolution) and semantic discriminability (focusing on features of key regions through attention).

3.3. Angle-Guided Feature Pyramid Network

Standard FPN [57] uses a uniform feature extraction strategy for objects in all orientations, without utilizing the object's orientation information. Although simply optimizing the convolution sampling method can allow the learning of the spatial pose of a rotated object, this learning is implicit and lacks robust generalizability. Since it can only directly learn the offset, it lacks explicit directional constraints, often requiring massive amounts of data to converge, and it is difficult to guarantee the rotational equivariance of feature sampling. Our core innovation lies in explicitly predicting the angle at each position. This prediction is optimized in an end-to-end manner, and the sampling offset of the DCN [25,58] is rotated based on the result, thereby achieving rotation-equivariant feature extraction. Therefore, it is easier to learn and can achieve the same effect with less data. AG-FPN adds angle guidance modules to the P3, P4, and P5 layers based on standard FPN. Each module includes the following steps: (I) predicting dense angle maps; (II) predicting the base DCN offset; (III) rotating the offset based on the angle; (IV) applying modulated deformable convolution; (V) feature fusion and gating (as illustrated in Figure 5). In general, it can be divided into two main stages:

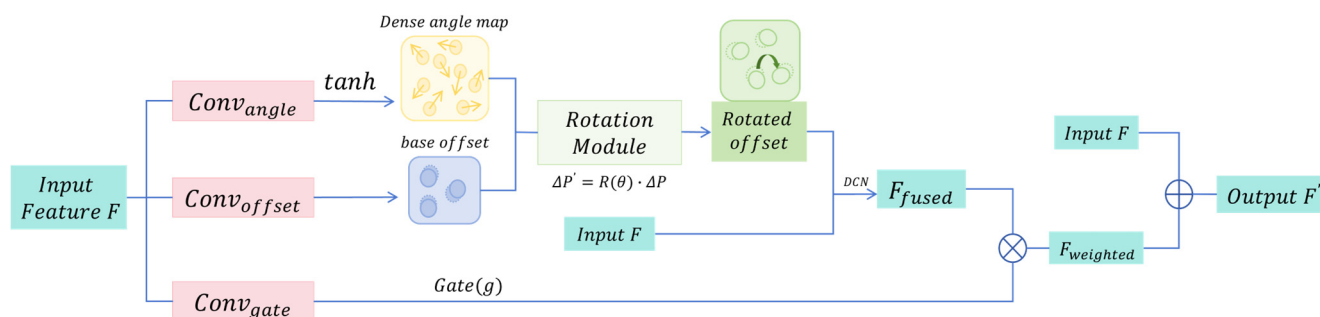


Figure 5. Overall process of the AG-FPN method.

Dense Angle Map and Base Offset Prediction: For a feature $F \in \mathbb{R}^{C \times H \times W}$ at a certain level of FPN, we predict the angle of each spatial location and the base offset of DCN. Notably, the angle map serves as a deformable offset rotation/alignment auxiliary variable. We do not apply an explicit angle regression loss; rather, it is implicitly optimized via the overall detection loss. This setup naturally conforms to the $\text{le}90$ angle definition and guarantees no discontinuities at the boundaries. For the latter, we initialize the offset and mask to 0, which ensures that AG-FPN is equivalent to standard FPN in the early stages of training, thus guaranteeing the stability of training and avoiding feature destruction caused by random offsets in the initial stage. The specific calculation formula is as follows:

$$\theta_{map} = \pi \cdot \tanh \left(\text{Conv}_{angle}(F) \right) \quad (6)$$

$$\Delta P_{base} = \text{Conv}_{offset}(F) \in \mathbb{R}^{18 \times H \times W} \quad (18 = 9 \times 2) \quad (7)$$

Here, the 18 channels represent 9 sampling points with 2 coordinates each. For every spatial location (h, w) , the 9 sampling point offsets along the x -axis and y -axis are defined as:

$$[\Delta P_1, \Delta P_2, \dots, \Delta P_9], \Delta P_k = (\Delta x_k, \Delta y_k) \quad (8)$$

Angle-guided offset rotation: This is the core innovation of AG-FPN. We rotate the offset at this position based on the predicted angle $\theta(h, w)$. This explicit rotation mechanism ensures rotational equivariance; if the input image rotates by θ degrees, the sampling point will also rotate by θ degrees, accordingly. The specific formulas are as follows:

$$\Delta x'_k = \cos(\theta_{map}) \odot \Delta x_k - \sin(\theta_{map}) \odot \Delta y_k \quad (9)$$

$$\Delta y'_k = \sin(\theta_{map}) \odot \Delta x_k + \cos(\theta_{map}) \odot \Delta y_k \quad (10)$$

where $\odot$ represents element-wise multiplication. The result of the offset rotation is that when the object angle is 0° , the offset remains unchanged. When the object angle is 45° , the offset rotation is 45° to align the sampling point with the object edge. Finally, to adaptively regulate the contribution of the alignment feature and improve computational efficiency, we introduce a gating mechanism [59]:

$$g = \sigma(\text{Conv}_{\text{gate}}(F)) \in [0, 1]^{C \times H \times W} \quad (11)$$

$$F' = F + g \odot F_{\text{fused}} \quad (12)$$

where F_{fused} is the fused feature. Acknowledging that not all positions require angular alignment, the formula regulates the alignment feature based on the value of g . When $g \approx 1$, the alignment feature is fully utilized, and when $g \approx 0$, the original feature is maintained.

3.4. Adaptive Geometry-Aware Loss

The original consistency loss function disregards the rotational characteristics of the objects themselves, using instead uniform weights. Our Adaptive Geometry-Aware Loss (AGL), however, employs different metrics for each object, defining two sets of learnable weight parameters to optimize the calculation of flip and rotation losses, as shown in Figure 6. The calculation process is performed in two branches: self-supervised branch loss function and weakly supervised branch loss function.

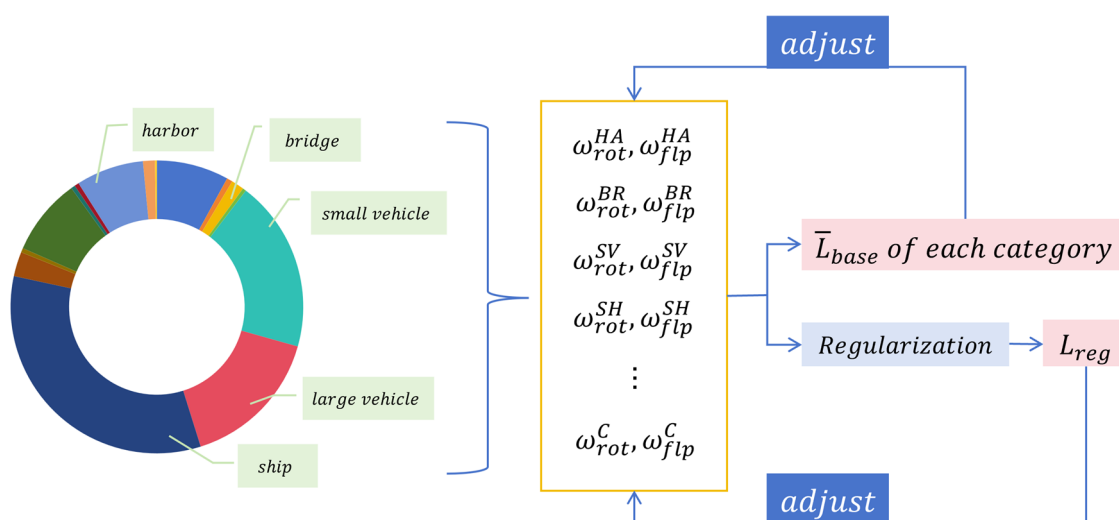


Figure 6. Visualization of the weight balancing process. The pie chart shows the relative proportions of various types of objects in DOTA-v1.0 dataset. HA, BR, SV, and SH are the abbreviations for harbor, bridge, small vehicle, and ship.

Self-supervised branch loss function. In H2RBox-v2, the self-supervised loss L_{SS} is defined as a linear weighted sum of the flip consistency loss and rotation consistency loss, i.e., $L_{SS} = \lambda L_{flp} + L_{rot}$ [17]. Our innovative AGL adds weight adjustments for different object categories. During training, the network automatically learns the weights for each category. First, we define two sets of weighted learning parameters for C categories:

$$\omega_{rot} = [\omega_{rot}^{(1)}, \omega_{rot}^{(2)}, \dots, \omega_{rot}^{(C)}] \in \mathbb{R}^C \quad (13)$$

$$\omega_{flp} = [\omega_{flp}^{(1)}, \omega_{flp}^{(2)}, \dots, \omega_{flp}^{(C)}] \in \mathbb{R}^C \quad (14)$$

Initially, ω_{rot} and ω_{flp} can be set to 1.0. However, to prevent extreme weight divergence due to sample balance, we apply weight clamping:

$$\tilde{\omega}_{rot}^{(c)} = \text{clamp}(\omega_{rot}^{(c)}, \omega_{min}, \omega_{max}) \quad (15)$$

$$\tilde{\omega}_{flp}^{(c)} = \text{clamp}(\omega_{flp}^{(c)}, \omega_{min}, \omega_{max}) \quad (16)$$

where ω_{min} and ω_{max} are the minimum and maximum values of the truncation function clamp. The specific threshold settings will be analyzed in detail in the parameter sensitivity analysis in Section 4.4. The rotation loss $L_{rot}(i)$ for each object category is multiplied by the weight $\omega_{rot}^{(c_i)}$ to obtain the weighted value $L_{rot}^{weighted}(i)$. The total weighted rotation loss and weighted flip loss are then summed and averaged over each category. Furthermore, we have a regularization loss function, which is a key factor in adjusting the loss for each object category.

$$L_{reg} = \lambda_{reg} \left[\frac{1}{C} \sum_{c=1}^C (\omega_{rot}^{(c)} - 1)^2 + \frac{1}{C} \sum_{c=1}^C (\omega_{flp}^{(c)} - 1)^2 \right] \quad (17)$$

The AGL function is calculated as:

$$L_{ALG} = L_{rot}^{weighted} + L_{flp}^{weighted} + L_{reg} \quad (18)$$

In Equation (18), $L_{rot}^{weighted} + L_{flp}^{weighted}$ can be expanded as follows:

$$\frac{1}{N} \sum_{i=1}^N [\omega_{rot}^{(c_i)} \ell_{rot}(i) + \omega_{flp}^{(c_i)} \ell_{flp}(i)] \quad (19)$$

The gradient formula for the weight ω_c of category c can well elucidate how ω_c is learned and adjusted:

$$\frac{\partial L_{AGL}}{\partial \omega_c} = \frac{1}{N} \sum_{i:c_i=c} L_{base}(i) + 2L_{reg}(\omega_c - 1) \quad (20)$$

The above equation is derived directly by differentiating L_{AGL} with respect to ω_c . Here, L_{base} serves as a unified notation for ℓ_{rot} and ℓ_{flp} . Specifically, when computing the partial derivative regarding the rotation weights, ω_c corresponds to $\omega_{rot}^{(c)}$ and L_{base} becomes ℓ_{rot} . The gradient equation intuitively explains the adaptive adjustment mechanism of the weights: the first term is the data-driven term; when the average base loss L_{base} of a certain class is high (indicating that the class is difficult to learn), the gradient is positive, driving the weights to increase to strengthen attention; the second term is a regularization term, used to constrain the degree of weight deviation and prevent training instability.

The situation is different when the dataset has only one target class. This is because AGL operates through two different mechanisms: class-level rebalancing for multi-class scenarios, and L_2 regularization applicable to all datasets to achieve stable multi-objective

optimization. Therefore, for the single-class dataset HRSC, the first mechanism becomes ineffective, and the final performance improvement comes solely from the second. The regularization prevents extreme weight configurations and enhances the model's generalization ability through balanced learning.

Weakly supervised branch loss function. Similar to H2RBox-v2, we use HBox to label the output of the supervised detection head. This module is mainly responsible for learning scale information. The total loss function in the weakly supervised stage is:

$$L_{WS} = \lambda_{cls}L_{cls} + \lambda_{box}L_{box} + \lambda_{ctr}L_{ctr} \quad (21)$$

where L_{cls} , L_{box} , and L_{ctr} use the same formulations as H2RBox-v2. In our implementation, L_{cls} is the sigmoid focal loss, L_{box} is an IoU-based regression loss, and L_{ctr} is the sigmoid cross-entropy loss for centerness. Our Adaptive Geometric Perception Loss focuses on learning the weights λ_{cls} , λ_{box} , and λ_{ctr} during training, without changing the definitions of these losses. Based on this, the total loss function is:

$$L_{total} = L_{WS} + \lambda_{SS}L_{ALG} + \lambda_{uc}L_{uc} \quad (22)$$

where λ_{SS} is the weight of the self-supervised branch, generally set to 1 by default, and L_{uc} is the unit-circle-based loss used to supervise angle prediction (via the UCResolver module). In the revised manuscript, we state its explicit definition as follows:

$$L_{uc} = \left| \frac{n}{2} - \sum_{i=1}^n m_i^2 \right| + \sigma(n) \left| \sum_{i=1}^n m_i \right| \quad (23)$$

where m_i is the encoded value, and $\sigma(n)$ is a segmentation function, taking the value 1 when n is 3 and 0 when n is 2.

4. Results

4.1. Datasets and Settings

The datasets used in this paper are the DOTA-v1.0 [26], HRSC [27], and RSAR [18] datasets. The DOTA-v1.0 and HRSC datasets contain optical imagery, while the RSAR dataset contains radar imagery.

The DOTA-v1.0 dataset is a large-scale dataset specifically designed for object detection in aerial images, particularly suitable for research and evaluation of oriented object detection algorithms. It consists of 2806 images, encompassing fifteen object categories. This dataset features diverse scenes, including rotation-invariant objects such as storage tanks and baseball fields, as well as highly orientation-sensitive objects such as ships and bridges. The training set contains 1411 images, the validation set contains 458 images, and the test set contains 937 images, representing 1/2, 1/6, and 1/3 of the total, respectively. The images used for training and validation are identical to the default allocation in the dataset.

The HRSC dataset is a high-resolution remote sensing ship detection dataset with a total of 2976 instances, divided into four main classes (warships, civilian ships, aircraft carriers, and submarines) and 19 subclasses. The training set contains 436 images (approximately 1207 instances), the validation set contains 181 images (approximately 541 instances), and the test set contains 444 images (approximately 1228 instances).

The RSAR dataset is a comprehensive, multi-class, large-scale oriented SAR object detection dataset, containing 95,842 SAR images and 183,534 labeled instances, covering six typical SAR object categories. It was constructed via the H2RBox-v2+UCR weakly supervised method, based on predictions from the SARDet-100K dataset, and manually refined.

Regarding implementation details, the DOTA-v1.0 and RSAR datasets are trained for 12 epochs, while the HRSC dataset is configured for 72 epochs. The training batch size was

set to four, the learning rate for the DOTA-v1.0 dataset was set to 10^{-4} , and the learning rate for the RSAR and HRSC datasets was set to 5×10^{-5} . The experimental setup is based on CUDA 11.6, PyTorch 1.13.0, and MMRotate [60] 1.0.0rc1. The hardware configuration includes a GeForce RTX 3090 GPU. All models were trained using the AdamW optimizer (implemented via torch.optim.AdamW).

4.2. Comparative Experiment Results and Analysis

Table 1 presents the quantitative results for AP₅₀, AP₇₅, and mAP accuracy metrics on the DOTA-v1.0, HRSC, and RSAR datasets.

The comparative results on the DOTA-v1.0 and HRSC datasets validate the method's effectiveness in optical remote sensing scenarios. On the DOTA-v1.0 dataset, compared to H2RBox-v2+UCR method, the WSOOD-GAEN method improves the AP₅₀ metric by 1.51%, along with improvements of 1.21% and 1.26% on the AP₇₅ and mAP metrics, respectively. These results signify high-precision localization capabilities. Furthermore, its AP₅₀ accuracy is only 0.88% lower than the fully supervised OrientedFormer [61]. This performance, which closely approaches that of fully supervised methods, demonstrates the effectiveness of the proposed angle adaptation method and consistency loss weight adjustment.

In the HRSC dataset results, H2RBox [15] yields suboptimal performance, lagging far behind other methods, primarily due to its lack of effective angle guidance, making it difficult to obtain accurate angle information under small sample conditions. The improved H2RBox-v2 [16], however, learns the angle from the object's inherent symmetry, resulting in a significant performance improvement. On the AP₅₀ metric, the performance of all methods tends to be consistent. However, on the AP₇₅ metric, which better reflects high positioning accuracy, our method shows a significant advantage, exceeding other weak supervision methods by more than 9.86%. Furthermore, on the mAP metric, our method is only 0.8% behind the fully supervised method Oriented R-CNN [54]. This indicates that WSOOD-GAEN can generate high-quality prediction boxes and has strong robustness on small-scale datasets with high-aspect-ratio objects.

Table 1. The performance of our method and previous approaches on the DOTA-v1.0, HRSC, and RSAR.

Method	Resolver	Supervised	DOTA-v1.0			HRSC			RSAR		
			AP ₅₀	AP ₇₅	mAP	AP ₅₀	AP ₇₅	mAP	AP ₅₀	AP ₇₅	mAP
Oriented R-CNN [62]	-	RBox	70.51	41.72	41.20	89.87	78.02	62.12	70.05	34.11	36.22
OrientedFormer [61]	-	RBox	69.90	42.28	41.28	90.11	82.63	69.27	73.18	38.54	39.69
YOIOv11 [63]	-	RBox	82.84	59.83	67.42	97.73	95.32	88.04	86.97	53.69	60.55
FCOS [56]	-	RBox	71.44	41.76	41.81	89.26	77.47	63.21	66.66	31.45	34.22
H2RBox [16]	-	HBox	64.46	26.91	32.40	2.61	0.11	0.42	60.05	11.63	22.56
H2RBox-v2 [17]	PSC [64]	HBox	65.17	34.52	36.01	89.70	63.55	56.24	65.87	24.09	31.07
ABBSPO [55]	PSC	HBox	68.60	34.82	37.54	89.25	64.64	54.23	68.26	25.73	32.98
H2RBox-v2	UCR [18]	HBox	67.51	39.04	38.85	89.33	62.62	54.60	70.38	25.82	34.07
WSOOD-GAEN	UCR	HBox	69.02	40.25	40.11	90.02	74.50	61.32	74.65	29.17	35.74

Furthermore, detection results in the RSAR dataset demonstrate that feature-aligned convolution also has good optimization effects when dealing with SAR images containing speckle noise and sensitive to geometric deformation, improving the model's adaptability in multimodal remote sensing. In the RSAR dataset detection results, WSOOD-GAEN outperforms other weak supervision methods method by 4.27% on the AP₅₀ metric, significantly leading not only other weakly supervised methods but also surpassing some fully

supervised methods (such as outperforming FCOS [56] on the mAP metric), and the evaluation results indicate that many weakly supervised methods achieve AP₅₀ scores comparable to those of fully supervised approaches, with the proposed method even outperforming most of the latter. This notable performance boost on the RSAR dataset compared to DOTA-v1.0 stems from the fact that RSAR contains only six categories and frequently features images with sparse instances, yielding more pronounced gains under a lenient IoU threshold. Conversely, fully supervised baselines retain a clear advantage on the AP₇₅ metric, which is a stricter measure of high-precision localization. This performance gap highlights a key area where weakly supervised frameworks require further optimization.

Tables 2 and 3 show the AP₅₀ accuracy comparison of various methods on the DOTA-v1.0 and RSAR datasets for each category. The detection results demonstrate that WSOOD-GAEN has superior performance in multi-category scenarios. Quantitative results indicate that WSOOD-GAEN outperforms existing weakly supervised methods in accuracy across the majority of categories on both datasets, and even reaches the performance level of fully supervised methods in some categories. Benefiting from the angle-aware enhancement of the feature extraction network and the introduction of Adaptive Geometric Awareness Loss (AGL), the detection model can adaptively allocate consistency constraint weights according to the geometric variations in different object categories under rotation and flip transformations in multi-category scenarios. Consequently, it addresses the issue of fixed weights in traditional methods and significantly improves the detection accuracy for few-shot categories. In addition, AP₅₀ on elongated categories (bridge/ship/harbor) increases from 40.4% to 43.8%, 87.8% to 88.4%, and 73.4% to 73.9% on DOTA-v1.0, and from 53.4% to 55.2%, 79.1% to 86.0%, and 65.5% to 75.0% on RSAR, respectively. These gains indicate that AG-FPN better preserves geometric cues in multi-scale features, which benefits objects with extreme aspect ratios under HBox-based supervision.

Table 2. Quantitative results of each category on the DOTA-v1.0.

Method		PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	AP ₅₀
Oriented R-CNN [62]	RB	89.4	77.1	50.2	73.2	64.0	84.8	89.2	90.7	71.7	62.4	66.3	65.5	73.3	58.6	41.4	70.51
OrientedFormer [61]		89.7	72.5	49.9	64.5	67.1	84.1	86.9	90.7	65.9	73.1	49.8	64.0	76.5	60.9	53.3	69.90
YOIOv11 [63]		96.2	90.7	69.0	84.6	81.1	85.5	90.2	93.0	77.5	78.6	69.8	75.9	90.2	84.5	75.6	82.84
FCOS [56]		88.8	76.7	49.7	63.9	77.8	79.7	87.1	90.3	76.7	80.5	59.6	65.1	65.7	68.5	41.6	71.44
H2RBox [16]	HB	89.3	72.8	28.8	56.2	70.1	68.2	79.5	90.6	69.3	68.5	56.9	65.4	51.7	57.0	41.2	64.46
H2RBox-v2(PSC) [17]		88.7	66.8	38.7	52.6	69.7	77.9	87.8	90.5	64.6	67.1	51.4	60.4	68.2	55.2	37.2	65.17
ABBSP0 [55]		89.1	74.1	46.6	59.4	71.2	78.0	88.3	90.5	69.1	68.1	57.8	62.8	70.0	55.9	48.2	68.60
H2RBox-v2(UCR) [18]		89.3	71.4	40.4	50.6	71.0	82.2	87.8	90.6	64.3	69.2	60.4	61.6	73.4	53.2	45.2	67.51
WSOOD-GAEN		89.3	71.1	43.8	54.0	72.0	81.4	88.4	90.7	71.6	74.1	62.2	61.5	73.9	50.7	49.9	69.02

Notes: RB and HB represent the rotated bounding box supervision and the horizontal bounding box supervision, respectively. PL, BD, BR, etc., are abbreviations for the object class names in the DOTA-v1.0 dataset.

Figures 7–9 show visualizations of the three datasets. In the DOTA-v1.0 dataset, WSOOD-GAEN improves upon the scale regression bias problem common in weakly supervised methods. Taking the detection results of orientation-sensitive objects such as harbors and planes as examples, the predicted bounding boxes generated by our method closely match the ground truth annotations, reducing false negatives and false positives caused by background noise. In the HRSC dataset, since it only includes ship categories, the accuracy requirements for the angle and boundary prediction of high-aspect-ratio objects are high. Therefore, relying solely on prediction scores is insufficient; fine-grained details

must also be considered. Visualization results show that WSOOD-GAEN is more refined in boundary localization, and the generated predicted bounding boxes tightly envelop the ship's outline. In the RSAR dataset, despite the learning challenge of low signal-to-noise ratio in radar images, WSOOD-GAEN still achieves robust detection results. Compared to other weak supervision methods, WSOOD-GAEN exhibits almost no missed detections and maintains accurate boundary regression even with blurred edges.

Table 3. Quantitative results of each category on the RSAR.

Method		Ship	Aircraft	Car	Tank	Bridge	Harbor	AP ₅₀
Oriented R-CNN [62]	RB	86.1	87.7	89.7	38.4	56.7	61.6	70.05
OrientedFormer [61]		87.0	91.4	90.3	33.7	65.0	71.5	73.18
H2RBox [16]	HB	47.1	90.7	85.7	46.5	42.9	47.4	60.05
H2RBox-v2 (PSC) [17]		79.2	90.6	89.7	34.0	39.3	62.2	65.87
ABBSPO [55]		79.2	90.4	89.9	30.4	54.2	64.8	68.26
H2RBox-v2 (UCR) [18]		79.1	90.3	90.0	43.5	53.4	65.5	70.38
WSOOD-GAEN		86.0	90.3	90.1	50.9	55.2	75.0	74.65

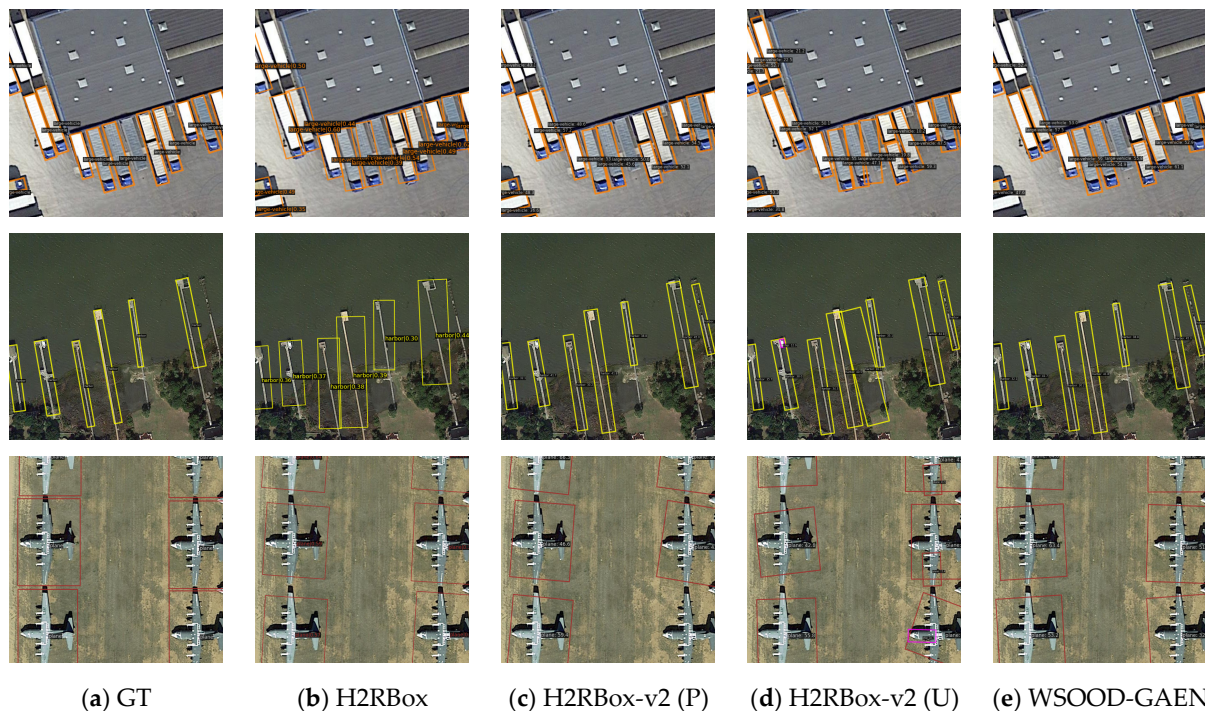


Figure 7. Comparative results of DOTA-v1.0 dataset visualization. (a) shows the original annotations of the dataset; (b–d) show the visualization annotations reproduced by other weak supervision methods; and (e) shows the detection results of the proposed method.

Overall, H2RBox suffers from weak angle learning capabilities due to a lack of effective geometric constraints, resulting in predictions that bias towards horizontal orientations and deviate significantly from the ground truth annotations. H2RBox-v2, by introducing symmetry consistency constraints, significantly improves angle prediction accuracy and scale regression. H2RBox-v2+UCR further enhances detection accuracy by optimizing angle encoding and introducing unit circle loss, but the tightness of the predicted bounding boxes remains insufficient [18]. In contrast, WSOOD-GAEN, with its synergistic effect of multi-scale angle guidance and adaptive geometric perception loss, improves accuracy in both orientation determination and scale regression. This establishes our method as a robust framework for weakly supervised oriented object detection, capable of tackling

the localization difficulties inherent to densely arranged and arbitrarily oriented targets in complex remote sensing imagery.

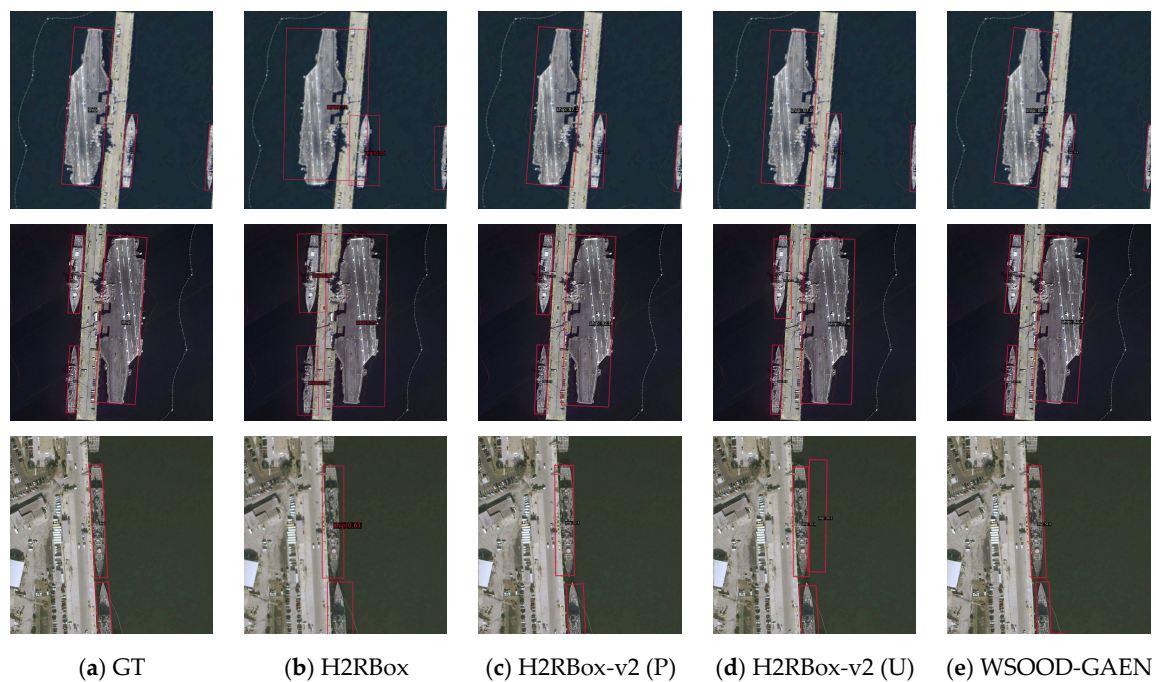


Figure 8. Comparative results of HRSC dataset visualization. The explanations for (a–e) are the same as above. In (c), (P) refers to the PSC angle encoding method used by the H2RBox-v2 method, and in (d), (U) refers to the UCR angle encoding method used by the improved H2RBox-v2 method.

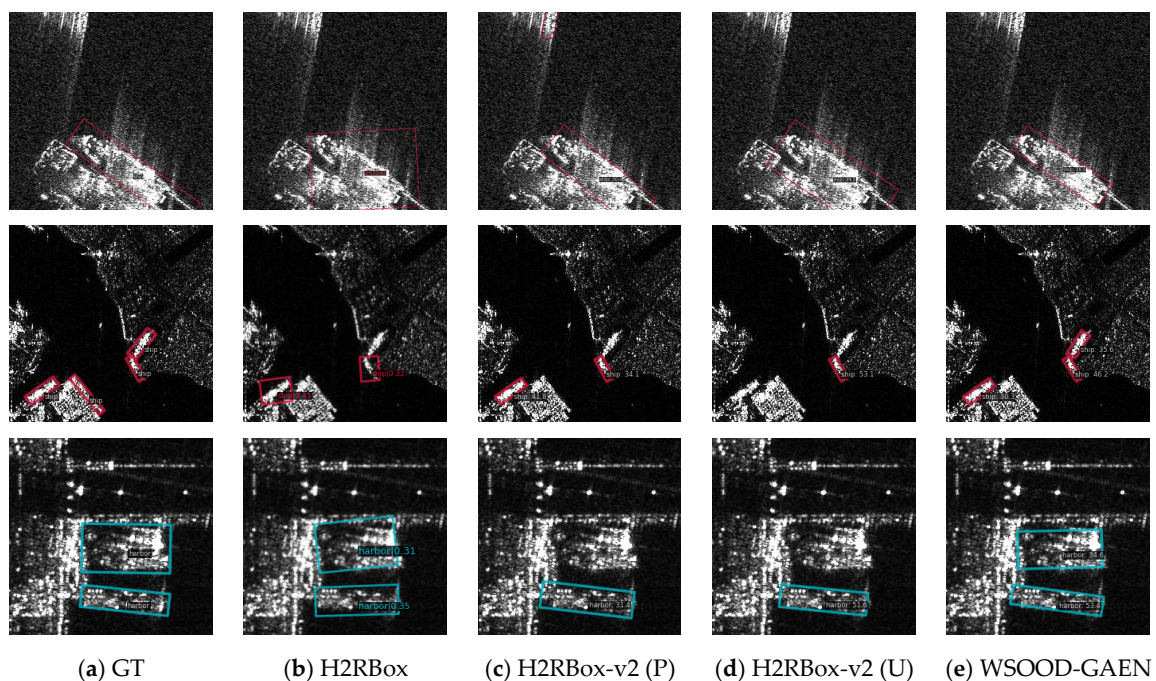


Figure 9. Comparative results of RSAR dataset visualization. The explanations for (a–e) are the same as above.

4.3. Ablation Study Results and Analysis

The data in Table 4 shows that, with the sequential addition of the AGL, DAE-ResNet, and AG-FPN modules, the detection accuracy of the three datasets generally exhibits an upward trend. The improvement is most significant after the introduction of the DAE-ResNet

module. This indicates that the three modules comprehensively optimize the network from three dimensions: key region feature enhancement, multi-scale orientation guidance, and adaptive loss constraints, respectively, contributing to the overall performance improvement. Furthermore, even on the HRSC dataset containing only a single class, the addition of AGL still brings performance improvements. This is because AGL not only distinguishes between class differences but also rebalances the ratio of rotational consistency to flip consistency weights within the same class, thus learning the geometric properties of specific objects more precisely.

Table 4. Ablation results of DAE-ResNet, AG-FPN module, and AG Loss.

Module			DOTA-v1.0		HRSC		RSAR	
AGL	DAE-ResNet	AG-FPN	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅	AP ₅₀	AP ₇₅
-	-	-	67.51	39.04	89.33	62.62	70.38	25.82
✓	-	-	68.06	35.58	89.58	64.59	71.81	27.50
✓	✓	-	68.62	39.63	89.90	72.73	72.52	27.73
✓	✓	✓	69.02	40.25	90.02	74.50	74.65	29.17

Table 5 summarizes the efficiency changes when we progressively add our components on DOTA v1.0. As expected, the additional prediction and alignment branches increase the parameter count and reduce the throughput from 21.2 to 17.5 FPS. We also observe that FLOPs does not strictly correlate with latency in our setting: after switching to the DAE ResNet backbone, FLOPs drops from 206.8 to 187.0 G, but FPS slightly decreases from 19.5 to 18.9. This is likely because deformable sampling is less latency-friendly on common hardware, involving irregular memory access as well as extra indexing and interpolation. Enabling AG-FPN further adds multi-scale fusion and dense angle-related computations, which increases memory traffic and leads to additional slowdown. Although these modules introduce extra runtime cost, they consistently improve accuracy in our main experiments, yielding a reasonable accuracy–efficiency trade-off.

Table 5. Efficiency impact of progressively adding the proposed components on DOTA-v1.0 compared with the H2RBox-v2+UCR baseline (FPS, FLOPs, Params).

Module			FPS	FLOPs (G)	Parameters (M)
AGL	DAE-ResNet	AG-FPN			
-	-	-	21.2	206.821	32.150
✓	-	-	19.5	206.821	32.150
✓	✓	-	18.9	186.997	33.552
✓	✓	✓	17.5	195.760	36.547

Figure 10 shows the feature response maps of the ablation experiment of the WSOOD-GAEN key module. The top two rows of images mainly analyze the role of the DAE-ResNet module, while the bottom two rows mainly analyze the role of the AG-FPN module. First, regarding the feature distribution of the backbone stage output, a relatively dense and noisy scene was selected. The baseline feature distribution was chaotic, with some high-activation regions incorrectly focused on the background. After adding the DAE-ResNet module, the features began to accurately focus on objects such as vehicles and ships. Finally, with the integration of AG-FPN, the two worked together to make the features more concentrated and more closely aligned with the object orientations. Next, regarding the feature distribution of the FPN stage output, a scene containing high-aspect-ratio objects (such as bridges and ports) was selected. Other weak supervision methods' feature

distribution was scattered, and adjacent regions showed adhesion. After adding the DAE-ResNet module, the distribution became more concentrated. Finally, the AG-FPN module was added, resulting in a substantial improvement. The features exhibit distinct orientation awareness, and the distribution became continuous and closely aligned with the axes of the rotated objects. Qualitatively, the visualization of feature heatmaps provides intuitive evidence for these improvements. It reveals that the DAE-ResNet module effectively suppresses background noise while enhancing semantic focus. Moreover, it validates that the AG-FPN module generates high-quality orientation guidance essential for subsequent learning stages.

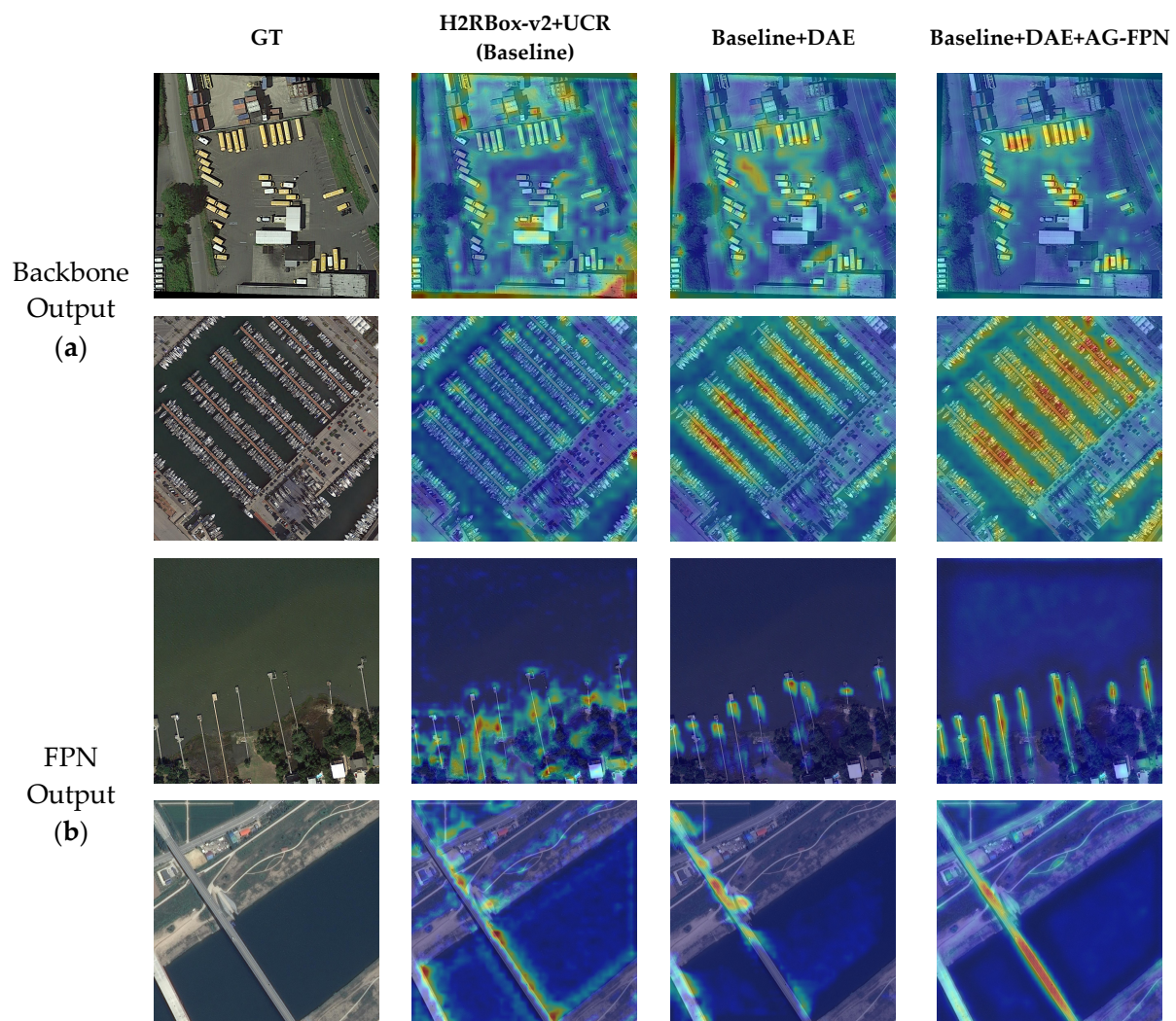


Figure 10. Feature response maps of key module ablation experiments for WSOOD-GAEN. (a) The top two rows of images show the feature distribution of the last layer of the backbone. (b) The bottom two rows show the feature distribution of the P_3 layer in FPN. DAE is an abbreviation for DAE-ResNet.

Furthermore, Table 6 presents the ablation study results of the AG-FPN module across different FPN levels. The results demonstrate that the model achieves the highest detection accuracy (69.02%) when the module is applied in combination with P_3 , P_4 , and P_5 levels. In terms of the FPS metric, adding the P_5 level has a negligible impact on the model's inference speed (the FPS only marginally decreases from 18.4 to 17.5). However, further incorporating the P_6 and P_7 levels not only compromises detection accuracy but also leads to a substantial decrease in inference speed. Simultaneously, computational metrics including FLOPs (G)

and Params (M) exhibit a gradual increase as more pyramid levels are incorporated. In summary, applying the AG-FPN module solely to P_3 , P_4 , and P_5 achieves the optimal trade-off between accuracy and speed. This indicates that selectively introducing the AG-FPN module, rather than uniformly applying it across all levels, can enhance detection accuracy with only a marginal impact on inference speed. This is because, on the one hand, compared to applying it exclusively to P_3 and P_4 , introducing the P_5 level enhances the network's feature extraction capability for medium-to-large objects; on the other hand, due to the extremely low spatial resolution and excessively large receptive fields of deep features, extending the module to these deeper levels introduces geometric noise and complex background interference.

Table 6. Ablation study of AG-FPN across different FPN levels on DOTA-v1.0.

FPN Levels					FPS	FLOPs (G)	Params (M)	DOTA-v1.0
P_3	P_4	P_5	P_6	P_7				AP ₅₀
✓	✓	-	-	-	18.4	195.342	35.549	68.35
✓	✓	✓	-	-	17.5	195.760	36.547	69.02
✓	✓	✓	✓	-	16.3	195.864	37.545	68.54
✓	✓	✓	✓	✓	15.2	195.890	38.543	67.58

4.4. Parameter Sensitivity Analysis

Regarding the formulation in Section 3.4, the parameters, clamping range and λ_{reg} , play a crucial role.

Selection of the clamping range. As shown in Table 7, experiments were conducted with ranges from [0.9, 1.1] to [0.65, 1.35]. As the range expands, the model's accuracy initially increases and then decreases. Theoretically, a range that is too narrow limits the model's adaptability; a range equal to 1 is equivalent to the original method. Conversely, a range that is too wide may lead to large weight variations, disrupting training stability and also degrading prediction accuracy. Experiments show that [0.75, 1.25] is the optimal truncation range.

Table 7. Parameter sensitivity analysis of the clamping range.

Clamping Range		λ_{reg}	DOTA-v1.0
Max	Min		AP ₅₀
1.1	0.9	0.1	67.41
1.15	0.85	0.1	67.68
1.25	0.75	0.1	68.06
1.3	0.7	0.1	65.82
1.35	0.65	0.1	67.35

Regularization coefficient λ_{reg} . Theoretically, λ_{reg} controls the sparsity and smoothness of the weight distribution: when the coefficient is too large, the learned weights will excessively shrink to the initial value of 1, causing the model to degenerate into the baseline method; conversely, when the coefficient is close to 0, the weights will lose constraint, and the distribution will become chaotic. Based on the experimental findings in Table 8, we can infer that a moderate regularization strength is necessary to achieve optimal performance. This hypothesis is corroborated in the quantitative results in Table 8: when the clamping range is fixed at [0.75, 1.25], setting λ_{reg} to 0.1 yields the best detection accuracy.

Table 8. Parameter sensitivity analysis of regularization loss weights.

λ_{reg}	DOTA-v1.0
	AP ₅₀
0.01	66.52
0.05	66.94
0.1	68.06
0.15	67.12
0.2	67.30

5. Discussion

Experiments on the RSAR dataset show that our method outperforms most state-of-the-art fully supervised methods on the AP₅₀ metric, confirming the effectiveness of the proposed weakly supervised strategy in general detection tasks. However, when measured by the more stringent AP₇₅ metric, fully supervised methods still maintain an advantage—indicating that there is still room for improvement in high-precision localization (especially target boundary and orientation depiction against complex backgrounds), and future works should focus on enhancing this capability.

Furthermore, point-supervised oriented object detection methods have shown great potential in recent years. Given the dual value of point supervision in both research and practical applications, combining point annotations with existing horizontal bounding box labels can provide supplementary localization cues for the model, helping to learn more accurate target outlines even with limited annotation budgets. Exploring hybrid supervision strategies combining weak labels and point-level annotations may become a new direction for improving the reliability and efficiency of detection frameworks, potentially accelerating the development of better oriented object detection systems.

6. Conclusions

This paper proposes a geometric perception enhancement network named WSOOD-GAEN for weakly supervised oriented object detection in remote sensing images, leveraging HBox as supervision. The method addresses three key problems of existing weakly supervised frameworks in remote sensing image oriented object detection: high susceptibility to background noise interference, significant deviations in feature extraction for high-aspect-ratio targets, and insufficient detection accuracy for few-shot categories. Through targeted optimization strategies, WSOOD-GAEN achieves detection accuracy comparable to fully supervised methods across multiple metrics. The core innovations of this paper are as follows: in the backbone stage, a DAE-ResNet backbone network is constructed. This module introduces a dual attention mechanism and couples it with deformable convolutions, endowing the network with geometric adaptability and semantic discrimination capabilities for key regions. In the feature pyramid stage, we propose AG-FPN, which achieves accurate alignment of multi-scale features with object boundaries by explicitly predicting dense angle maps and rotating the feature sampling offset field accordingly, enhancing the representation of high-aspect-ratio rotated objects. An AGL function is designed, which improves the model's learning ability for limited-sample objects by adaptively learning rotation and flip consistency weights for different categories. Experimental results on three public remote sensing datasets—DOTA-v1.0, HRSC, and RSAR—demonstrate that the WSOOD-GAEN method outperforms previous weakly supervised oriented object detection methods in terms of accuracy across AP₅₀, AP₇₅, and mAP, validating the method's effectiveness.

Author Contributions: Conceptualization, J.H.; Methodology, Y.Z.; Validation, Y.Z.; Investigation, Y.Z.; Resources, J.H.; Data curation, Y.Z.; Writing—original draft, Y.Z.; Writing—review & editing,

J.H.; Visualization, Y.Z.; Supervision, J.H.; Project administration, T.W.; Funding acquisition, T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Foundation Strengthening Fund Project (2023-JCJQ-JJ-0343) and the Fundamental Research Funds for the Central Universities (2042022dx0001).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Zheng, S.D.; Wu, Z.B.; Xu, Y.; Liu, Q.; Wei, Z.H. More accurate constraints for self-supervised learning in remote sensing images-based object detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 12303–12314. [\[CrossRef\]](#)
2. Wu, G.P.; Liu, L.D.; Liu, Z.X.; Liu, Y.; Gao, T. Task alignment interaction and cross-scale guided enhancement for remote sensing object detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 4007305. [\[CrossRef\]](#)
3. Feng, X.X.; Han, J.W.; Yao, X.W.; Cheng, G. Progressive contextual instance refinement for weakly supervised object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8002–8012. [\[CrossRef\]](#)
4. Lei, M.Q.; Liu, X.Y. SOLO-Net: A sparser but wiser method for small object detection in remote-sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *21*, 6001005. [\[CrossRef\]](#)
5. Zhang, X.; Li, Y.; Li, F.; Jiang, H.Z.; Wang, Y.H.; Zhang, L.; Zheng, L.; Ding, Z.G. Ship-Go: SAR ship images inpainting via instance-to-image generative diffusion models. *ISPRS J. Photogramm. Remote Sens.* **2024**, *207*, 203–217. [\[CrossRef\]](#)
6. Yu, D.W.; Ji, S.P. A new spatial-oriented object detection framework for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4407416. [\[CrossRef\]](#)
7. Li, Z.K.; Liu, W.; Xie, Z.J.; Kang, X.D.; Duan, P.H.; Li, S.T. FAA-Det: Feature augmentation and alignment for anchor-free oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5539411. [\[CrossRef\]](#)
8. Han, Y.Q.; Liu, H.P.; Wang, Y.F.; Liu, C.L. A comprehensive review for typical applications based upon unmanned aerial vehicle platform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9654–9666. [\[CrossRef\]](#)
9. Yue, T.X.; Zhang, Y.M.; Wang, J.; Xu, Y.B.; Liu, P.Y. A weak supervision learning paradigm for oriented ship detection in SAR image. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5207812. [\[CrossRef\]](#)
10. Tan, Z.Q.; Wu, C. Point-based weakly semisupervised oriented vehicle detection in optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 15635–15650. [\[CrossRef\]](#)
11. Wang, Q.; Sun, W. A Dual-Branch Perception Network for High-Precision Oriented Object Detection in Remote Sensing. *Remote Sens.* **2026**, *18*, 839. [\[CrossRef\]](#)
12. Zhang, M.; Ouyang, Y.; Yang, M.; Guo, J.; Li, Y. ORPSD: Outer Rectangular Projection-Based Representation for Oriented Ship Detection in SAR Images. *Remote Sens.* **2025**, *17*, 1511. [\[CrossRef\]](#)
13. Fasana, C.; Pasini, S.; Milani, F.; Fraternali, P. Weakly Supervised Object Detection for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 5362. [\[CrossRef\]](#)
14. Yang, Y.; Zhang, Y.M. Weakly supervised learning for oriented ship detection in SAR images using multiscale feature enhancement and angle encoding. *IEEE Geosci. Remote Sens. Lett.* **2025**, *23*, 4001005. [\[CrossRef\]](#)
15. Liu, P.; Huang, B.; Jin, T.; Long, H. Low-Shot Weakly Supervised Object Detection for Remote Sensing Images via Part Domination-Based Active Learning and Enhanced Fine-Tuning. *Remote Sens.* **2025**, *17*, 1155. [\[CrossRef\]](#)
16. Yang, X.; Zhang, G.; Li, W.; Wang, X.; Zhou, Y.; Yan, J. H2RBox: Horizontal box annotation is all you need for oriented object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023; pp. 1–18.
17. Yu, Y.; Yang, X.; Li, Q.Y.; Zhou, Y.; Zhang, G.F.; Da, F.P.; Yan, J.C. H2RBox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, 10–16 December 2023; pp. 71832–71856.
18. Zhang, X.; Yang, X.; Li, Y.X.; Yang, J.; Cheng, M.M.; Li, X. RSAR: Restricted state angle resolver and rotated SAR benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 7416–7426.
19. Sun, X.; Wang, B.; Wang, Z.R.; Li, H.; Li, H.C.; Fu, K. Research progress on few-shot learning for remote sensing image interpretation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2387–2402. [\[CrossRef\]](#)
20. Wang, X.; Han, C.; Huang, L.; Nie, T.; Liu, X.; Liu, H.; Li, M. AG-Yolo: Attention-Guided Yolo for Efficient Remote Sensing Oriented Object Detection. *Remote Sens.* **2025**, *17*, 1027. [\[CrossRef\]](#)

21. Yang, Z.G.; Liu, Y.M.; Gao, Z.H.; Wen, G.W.; Zhang, W.E.; Xiao, Y.H. Deep convolutional feature enhancement for remote sensing object detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6017105. [[CrossRef](#)]
22. Xie, X.; You, Z.H.; Chen, S.B.; Huang, L.L.; Tang, J.; Luo, B. Feature enhancement and alignment for oriented object detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 778–787. [[CrossRef](#)]
23. Chen, Y.; Jiang, W.; Wang, Y. FAMHE-Net: Multi-Scale Feature Augmentation and Mixture of Heterogeneous Experts for Oriented Object Detection. *Remote Sens.* **2025**, *17*, 205. [[CrossRef](#)]
24. 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](#)]
25. Zhu, X.Z.; Hu, H.; Lin, S.; Dai, J.F. Deformable ConvNets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9308–9316.
26. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.B.; Datcu, M.; Pelillo, M.; Zhang, L.P. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
27. Liu, Z.K.; Yuan, L.; Weng, L.B.; Yang, Y.P. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM), Porto, Portugal, 24–26 February 2017; pp. 324–331.
28. Hua, W.; Liang, D.K.; Li, J.Y.; Liu, X.L.; Zou, Z.K.; Ye, X.Q.; Bai, X. SOOD: Towards semi-supervised oriented object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 15558–15567.
29. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [[CrossRef](#)]
30. Li, K.; Wan, G.; Cheng, G.; Meng, L.Q.; Han, J.W. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [[CrossRef](#)]
31. Cheng, G.; Han, J.W.; Lu, X.Q. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [[CrossRef](#)]
32. Liao, M.H.; Shi, B.G.; Bai, X. TextBoxes++: A single-shot oriented scene text detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [[CrossRef](#)]
33. Yang, X.; Yan, J.C. Arbitrary-Oriented object detection with circular smooth label. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual, 23–28 August 2020; pp. 677–694.
34. Yang, X.; Hou, L.P.; Zhou, Y.; Wang, W.T.; Yan, J.C. Dense label encoding for boundary discontinuity free rotation detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 15814–15824.
35. Yang, X.; Yan, J.C.; Feng, Z.M.; He, T. R3Det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual, 2–9 February 2021; pp. 3163–3171.
36. Ran, Q.; Wang, Q.; Zhao, B.Y.; Wu, Y.F.; Pu, S.L.; Li, Z.J. Lightweight oriented object detection using multiscale context and enhanced channel attention in remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5786–5795. [[CrossRef](#)]
37. Yu, C.; Shin, O.-S.; Shin, Y. YOSDet: A YOLO-Based Oriented Ship Detector in SAR Imagery. *Remote Sens.* **2026**, *18*, 645. [[CrossRef](#)]
38. Yu, X.; Xiang, W.; Yu, L.; Han, K.; Yang, Y. SODE-Net: A Slender Rotating Object Detection Network Based on Spatial Orthogonality and Decoupled Encoding. *Remote Sens.* **2025**, *17*, 3042. [[CrossRef](#)]
39. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q.K. Learning RoI Transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2844–2853.
40. Han, J.M.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [[CrossRef](#)]
41. Li, Y.X.; Hou, Q.B.; Zheng, Z.H.; Cheng, M.M.; Yang, J.; Li, X. Large selective kernel network for remote sensing object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 16748–16759.
42. Chen, Z.M.; Chen, K.; Lin, W.Y.; See, J.; Yu, H.; Ke, Y.; Yang, C. PIoU Loss: Towards accurate oriented object detection in complex environments. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual, 23–28 August 2020; pp. 195–211.
43. Yang, X.; Yang, J.R.; Yan, J.C.; Zhang, Y.; Zhang, T.F.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8231–8240.
44. Yang, X.; Yan, J.C.; Ming, Q.; Wang, W.T.; Zhang, X.P.; Tian, Q. Rethinking rotated object detection with gaussian wasserstein distance loss. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 11830–11841.
45. Yang, X.; Yang, X.J.; Yang, J.R.; Ming, Q.; Wang, W.T.; Tian, Q.; Yan, J.C. Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; pp. 29707–29718.

46. Zheng, Z.H.; Wang, P.; Liu, W.; Li, J.Z.; Ye, R.G.; Ren, D.W. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
47. Chen, P.F.; Yu, X.H.; Han, X.M.; Hassan, N.; Wang, K.; Li, J.C.; Zhao, J.; Shi, H.; Han, Z.J.; Ye, Q.X. Point-to-box network for accurate object detection via single point supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 51–67.
48. Yang, Z.; Liu, S.H.; Hu, H.; Wang, L.; Lin, S. RepPoints: Point Set Representation for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9656–9665.
49. Lu, J.W.; Yang, X.; Yu, Y.; Li, Q.Y.; Van, J.C.; Li, Y.S. PointOBB: Learning oriented object detection via single point supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16730–16740.
50. Yu, Y.; Yang, X.; Li, Q.Y.; Dai, F.P.; Dai, J.F.; Qiao, Y.; Yan, J.C. Point2RBox: Combine knowledge from synthetic visual patterns for end-to-end oriented object detection with single point supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16783–16793.
51. Yu, Y.; Ren, B.T.; Zhang, P.Y.; Liu, M.X.; Luo, J.W.; Zhang, S.F.; Da, F.P.; Yan, J.C.; Yang, X. Point2RBox-v2: Rethinking point-supervised oriented object detection with spatial layout among instances. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 19283–19293.
52. Tian, Z.; Shen, C.; Wang, X.; Chen, H. BoxInst: High-performance instance segmentation with box annotations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 5443–5452.
53. Li, W.T.; Liu, W.Y.; Zhu, J.K.; Cui, M.M.; Hua, X.S.; Zhang, L. Box-Supervised Instance Segmentation with Level Set Evolution. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 1–18.
54. Lu, J.Y.; Hu, Q.L.; Zhu, R.F.; Wei, Y.L.; Li, T. AFWS: Angle-Free Weakly Supervised Rotating Object Detection for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3485590. [[CrossRef](#)]
55. Lee, W.; Chang, H.; Moon, J.; Lee, J.; Kim, M. ABBSP: Adaptive bounding box scaling and symmetric prior based orientation prediction for detecting aerial image objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 8848–8858.
56. Tian, Z.; Shen, C.H.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9626–9635.
57. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.M.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
58. Dai, J.F.; Qi, H.Z.; Xiong, Y.W.; Li, Y.; Zhang, G.D.; Hu, H.; Wei, Y.C. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
59. Li, X.; Wang, W.H.; Hu, X.L.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
60. Zhou, Y.; Yang, X.; Zhang, G.F.; Wang, J.B.; Liu, Y.Y.; Hou, L.P.; Jiang, X.; Liu, X.Z.; Yan, J.C.; Lyu, C.Q.; et al. MMRotate: A rotated object detection benchmark using PyTorch. In Proceedings of the ACM International Conference on Multimedia (MM), Lisbon, Portugal, 10–14 October 2022; pp. 7331–7334.
61. Zhao, J.Q.; Ding, Z.Y.; Zhou, Y.; Zhu, H.C.; Du, W.L.; Yao, R.; El Saddik, A. OrientedFormer: An End-to-End Transformer-Based Oriented Object Detector in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5640816. [[CrossRef](#)]
62. Xie, X.X.; Cheng, G.; Wang, J.B.; Yao, X.W.; Han, J.W. Oriented R-CNN for object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 3500–3509.
63. Ultralytics. Ultralytics YOLOv11. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 20 October 2025).
64. Yu, Y.; Da, F.P. Phase-Shifting Coder: Predicting accurate orientation in oriented object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 13354–13363.

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.