drones

MDPI

Article

Research on Bidirectional Multi-Span Feature Pyramid and Key Feature Capture Object Detection Network

Heng Zhang, Faming Shao *, Xiaohui He, Dewei Zhao, Zihan Zhang and Tao Zhang

College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China; hengzhang4216@aeu.edu.cn (H.Z.); gcbhxh@aeu.edu.cn (X.H.); lq821321@aeu.edu.cn (D.Z.); zzh2023@aeu.edu.cn (Z.Z.); 15951598133@163.com (T.Z.)

* Correspondence: shaofaming@aeu.edu.cn; Tel.: +86-185-4985-4591

Abstract: UAV remote sensing (RS) image object detection is a very valuable and challenging technology. This article discusses the importance of key features and proposes an object detection network (URSNet) based on a bidirectional multi-span feature pyramid and key feature capture mechanism. Firstly, a bidirectional multi-span feature pyramid (BMSFPN) is constructed. In the process of bidirectional sampling, bicubic interpolation and cross layer fusion are used to filter out image noise and enhance the details of object features. Secondly, the designed feature polarization module (FPM) uses the internal polarization attention mechanism to build a powerful feature representation for classification and regression tasks, making it easier for the network to capture the key object features with more semantic discrimination. In addition, the anchor rotation alignment module (ARAM) further refines the preset anchor frame based on the key regression features extracted by FPM to obtain high-quality rotation anchors with a high matching degree and rich positioning visual information. Finally, the dynamic anchor optimization module (DAOM) is used to improve the ability of feature alignment and positive and negative sample discrimination of the model so that the model can dynamically select the candidate anchor to capture the key regression features so as to further eliminate the deviation between the classification and regression. URSNet has conducted comprehensive ablation and SOTA comparative experiments on challenging RS datasets such as DOTA-V2.0, DIOR and RSOD. The optimal experimental results (87.19% mAP, 108.2 FPS) show that URSNet has efficient and reliable detection performance.

Keywords: UAV RS images; object detection; bidirectional multi-span feature pyramid; key feature capture mechanism

Citation: Zhang, H.; Shao, F.; He, X.; Zhao, D.; Zhang, Z.; Zhang, T. Research on Bidirectional Multi-Span Feature Pyramid and Key Feature Capture Object Detection Network. *Drones* **2024**, *8*, 189. https://doi.org/10.3390/drones8050189

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 10 March 2024
Revised: 7 May 2024
Accepted: 7 May 2024
Published: 9 May 2024

1. Introduction

In recent years, object detection technology in drone RS images based on deep learning has developed rapidly. It utilizes cameras or sensors mounted on drones to capture ground images, and then identifies specific objects on the ground through image processing. This technology is widely applied in various fields, such as agricultural monitoring [1], environmental protection [2], urban planning [3], disaster assessment [4], intelligence reconnaissance [5], and more, providing significant support and contributions to the development of human society. However, despite its rapid growth, it also faces a series of challenges.

First, drones may be subject to interference from weather conditions and limitations of their own equipment when collecting RS images, leading to noise in the captured images and affecting the accuracy of object detection and recognition [6]. Second, due to differences in the height and angle of drone photography, the same type of object may exhibit different sizes and shapes, while the size differences between different types of objects are even more pronounced. This imposes higher requirements for object detection algorithms [7]. Third, unlike traditional images captured from a fixed perspective, objects in drone RS images

may appear in any direction. This diversity in direction poses additional challenges for detection [8]. Some typical examples of the above challenges are shown in Figure 1.

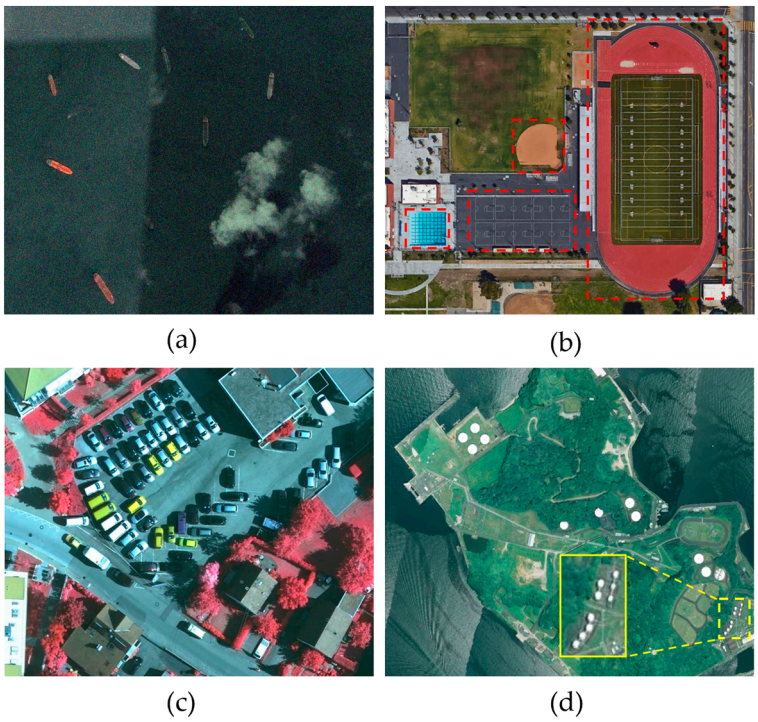
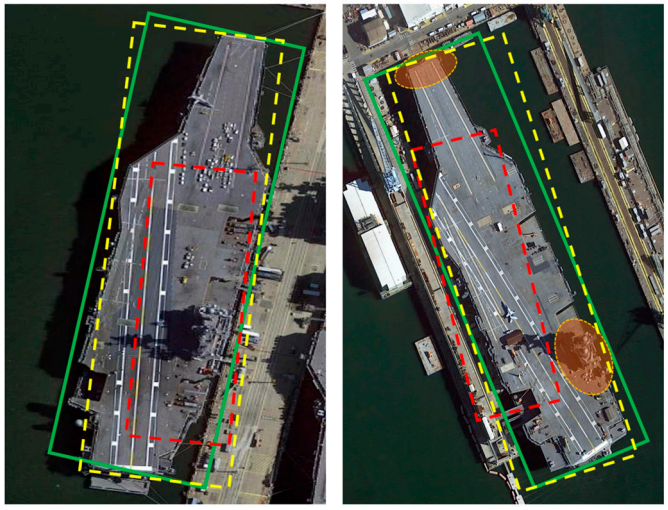


Figure 1. Several challenging examples. (**a**) The image contains noise, clouds, and light and shadow effects. (**b**) There is a significant variation in the scale of the objects in the image. (**c**) The objects are distributed in arbitrary directions in the image. (**d**) The image contains extremely challenging small objects.

Addressing the issue of noise in RS images, article [9] proposes a two-stage approach to separately filter out Gaussian noise and salt-and-pepper noise. It introduces dilated convolutions into a DnCNN (Denoising Convolutional Neural Network) for initial noise reduction. Additionally, the median of the filtering window is improved for secondary noise reduction. Lossy compression often leads to the generation of noise during image transmission. To address this, Kovalenko et al. [10] adjust the noise variance in three-channel images to predict a denoising threshold known as OPP. Subsequently, they use BPG (Better Portable Graphics) to obtain high-quality RS images. To preserve more object details in the images, Wang et al. [11] utilize wavelet technology to decompose the noise in RS images and then reconstruct the denoised images using wavelet techniques. To address the impact of varying object scales and complex backgrounds in RS images, Lin et al. [12] proposed a Multi-Scale Context Network (MSCNet). This network effectively addresses the issue of low precision through a multi-scale context feature extraction module and a pyramid aggregation mechanism. Detecting small-scale objects has always been a challenging task. Zhang et al. [13] focused on detecting small vehicles and improved the network's ability to discern features of small objects by modifying the loss function in YOLOv3. Similarly, aiming at complex backgrounds and small object detection, Article [14] employed an advanced architecture combining convolutional neural networks and transformers. This architecture utilizes a Cross-Shaped Window Transformer (CSWin) to build powerful feature representations, thereby enhancing the detection capabilities for small objects. The arbitrary orientation of objects in RS images makes it difficult for detection models to accurately locate and classify them, especially under the influence of complex backgrounds. The authors of [15] propose an arbitrary-oriented detection method that integrates an attention mechanism within the RCNN-like framework to highlight useful features, enabling the model to possess state-of-the-art detection capabilities. To effectively

handle randomly oriented objects in RS images, Shamsolmoali et al. [16] use a Rotation Equivariant Feature Image Pyramid Network (REFIPN) to efficiently extract features of widely distributed objects and spatially determine their locations and angles. The arbitrary orientation of objects means that traditional horizontal bounding boxes cannot guarantee accurate predictions by the model. To address this, Shi et al. [17] integrate a search framework (NAS-FPN) into a dense detector (RetinaNet) based on angle classification to capture target motion information and trajectories.

Furthermore, through research, we have discovered that key features are of crucial importance for the tasks of object classification and regression in UAV RS images [18]. As shown in Figure 2, although the ship objects in both images are accurately located, the key features of the object in Figure 2b are not accurately captured by the detection model, resulting in a classification error. Therefore, the crucial role of key features provides important insights for the construction of our subsequent methods.



(a) (b)

Figure 2. In both (**a**,**b**), the green boxes represent the ground truth boxes, the red boxes correspond to the prior anchors, and the yellow boxes depict the predicted boxes obtained through regression based on the prior anchors. It is evident from the figures that the object positions in both (**a**,**b**) are accurately located. Nevertheless, the classification error in (**b**) arises due to the prior anchor's inability to capture crucial features, such as the island and bow of the ship.

The above section enumerates and discusses common methods for addressing several typical challenges in object detection for drone RS images. It is evident that these methods are designed to tackle individual issues independently. However, in reality, the challenges discussed often occur simultaneously. To address this, the present study introduces an efficient and comprehensive detection model, URSNet, to overcome the aforementioned difficulties.

Specifically, first, the BMSFPN, which is composed of a bidirectional path, is proposed to gradually filter out noise in the image and smooth the edge details of the object. Second, the designed FPM enhances the key feature regions of the detection target through a polarization attention mechanism, enabling the construction of robust feature representations for both classification and regression. Additionally, given the arbitrary distribution of objects in RS images, we utilize ARAM to construct rotationally aligned anchors that match the key features of the target, facilitating the model's extraction of accurate localization information. Finally, the designed DAOM optimizes the label assignment of training samples, effectively addressing the inconsistency in confidence between classification and regression, and enabling the model to achieve precise classification and regression performance. The superior experimental results of URSNet on large and challenging RS datasets validate its efficient performance. The contributions of this paper can be summarized as follows:

- BMSFPN is proposed to address the issues of noise interference and the loss of small object detail features during the feature extraction process. It utilizes bicubic interpolation and cross-layer connections in a bidirectional sampling path to gradually filter out noise present in the feature layers, while simultaneously weighting and enhancing the edges and texture details of the object.
- To mitigate the interference between classification and regression when sharing features, FPM is employed to decouple the input features from the upper level and construct robust feature representations specifically for each individual task.
- The designed ARAM further obtains high-quality initial candidate anchors that are more aligned with arbitrarily directed objects through rotation alignment, based on the key feature regions constructed by FPM. The refined anchor regions provide the model with more accurate visual information for regression.
- DAOM addresses the issue of the confidence mismatch between classification and regression during the training phase through a matching degree strategy. It primarily optimizes the assignment of sample labels to enable the model to dynamically select high-quality anchor samples with critical regression feature capture capabilities. These positive anchor samples, after training, ensure the model's ability to accurately locate objects.

The rest of this article is organized as follows: Section 2 summarizes the current research status of object detection in RS images; Section 3 elaborates on the proposed method and its details; Section 4 implements experiments and analyzes the relevant results; and Section 5 summarizes the achievements and limitations of this research.

2. Related Works

This section provides a comprehensive overview of the related work on object detection technology in drone RS images. Below, we will specifically elaborate on several closely related aspects of this study, including RS object detection, RS image denoising technology, and key feature roles of RS objects.

2.1. RS Object Detection

In recent years, RS image object detection has made significant progress driven by deep learning techniques. Traditional RS image object detection methods include manually designed image feature methods (such as color and texture features) [19] and classifier-based methods (such as SVM and Decision Trees) [20]. While these methods can achieve object detection to a certain extent, their performance is limited by image quality and complex scenes. In contrast, deep learning-based object detection methods, such as Convolutional Neural Networks (CNNs) [21], Region-based Convolutional Neural Networks (R-CNNs) [22], and YOLO [23], can better adapt to RS object detection tasks in different scenarios. For example, Article [24] proposed an RS object detection method based on deep learning. The authors introduced DenseNet and SE into the original backbone network Darknet-53 of YOLOv3, significantly improving the model's feature extraction capability. Small objects have very few pixels, making their features difficult to extract. To address this issue, Teng et al. [25] innovated a small object detection model (GLNet) that collects global contextual information through the Multi-Scale Perception (MSP) module and Clip-LSTM encoding, providing crucial assistance for the model to detect small objects. Yu et al. [26] focused on the issues of occlusion and overlap of objects in RS images, emphasizing the strategy of large-scale proposal bounding boxes, and constructed a novel spatial adaptive detector (RSADet). Zhao et al. [27] introduced non-striding convolution and an attention mechanism into YOLOv7 to improve the feature extraction capability for small targets, and optimized the fusion process of deep information for small targets using the Lion optimizer. In article [28], a joint motion mechanism based on a three-degree-of-freedom (DOF) framework was designed for drones in complex motion patterns to achieve real-time active tracking of targets. Lai et al. [29] proposed a background subtraction method to detect moving targets and used the Mask R-CNN model to identify target types. Article [30]

provides an overview of the 2022 L4S competition aimed at overcoming the challenges of detecting landslides in remote sensing images. The top-ranked team improved the image patch size to overcome the weak representation of small landslides and achieved a high-performance landslide detection capability by emphasizing self-operation. Ye et al. [31] designed a detection model with an Adaptive Attention Fusion Mechanism (AAFM) to address the sensitivity of RS multi-scale targets. This model balances the proportion of multi-scale targets through a stitcher and introduces spatial and channel attention models to enhance feature information. Compared to the current SOTA detectors, this model improves accuracy and robustness.

2.2. RS Image Denoising Technology

Noise interference in RS images can lead to problems such as blurred object edges and loss of details, which in turn affect the accuracy and stability of object detection. Removing noise can enhance the recognizability of objects and improve the results of object detection and recognition. In [32], the authors used ResNet and DenseNet to generate an adversarial network (RRDGAN) and employed total variation (TV) regularization for high-quality denoising and ultra-high-resolution image reconstruction. To address the shortcomings of the BM3D algorithm in removing strong noise, Chen et al. [33] studied the similarity between object edges and utilized an edge search strategy to match local image blocks, resulting in excellent denoising effects. The presence of noise in RS images makes supervised deep neural network training inefficient. To address this, Xie et al. [34] proposed an unsupervised training method for specialized noise removal. They constructed a noisy image dataset and improved the deep image prior (DIP) method, allowing the DIP model to be fully trained and achieve powerful denoising capabilities using non-local regularization. Article [35] proposes a Global-to-Local Scale-Aware Network (GLSANet), which aims to improve the performance of multi-scale target detection in RS images by reducing complex background interference and suppressing noise through the Global Semantic Information Interaction Module (GSIIM), optimizing the feature pyramid, and introducing the Local Attention Pyramid (LAP).

2.3. Key Feature Roles of RS Objects

The key features in RS images play a pivotal role in achieving high-performance object classification and regression. By fully leveraging these critical features for object detection, we can reduce false positives and false negatives, thus improving the precision and robustness of the detection process. For instance, to address the challenges posed by multi-scale objects, Lin et al. [36] focused on analyzing the critical distribution characteristics of objects and proposed an efficient detection model based on prominent object features. This model utilizes search operators to extract critical information about object features, resulting in superior algorithm performance. Article [37] introduces a network that addresses the shortcomings of feature pyramids and label allocation. This network comprises the Aware Feature Pyramid Network (AFPN) and the Group Allocation Strategy (GAS). These components are designed to capture high-level critical features from the feature pyramid, enhancing the model's capabilities in classification and localization. Liu et al. [38] discussed the limitations of anchor-free detectors in detecting objects with arbitrary orientations and proposed a novel detection network called CBDA-Net, which is based on the Center-Boundary Dual Attention (CBDA) mechanism. This network primarily leverages attention mechanisms to extract critical features of object centers and boundaries, facilitating rapid object localization. Ghorbanzadeh et al. [39] employed optical data from the Rapid Eye satellite to extract and select crucial training patches from satellite imagery. After training a CNN model, they used the mIOU strategy to enhance the accuracy of landslide detection. Remote sensing ship targets possess characteristics of an arbitrary orientation and dense arrangement, posing significant challenges for target detection tasks. To address this, Article [40] proposed a dynamic adjustment learning (DAL) strategy based

on binary-coded learning (BCL) to improve the ability to capture key features and enhance the accuracy of angle prediction.

3. Overview of the Proposed Methods

The detection process of the proposed URSNet is depicted in Figure 3. Its backbone framework employs ResNet-101 [41] to accomplish initial feature extraction and optimization. Subsequently, BMSFPN is utilized to enhance and denoise the object features. In the top-down process, bicubic interpolation is applied to mitigate the negative impact of noise on the feature images. In the bottom-up process, cross-layer and cross-node connections are leveraged to fuse multi-scale features, enabling the network to improve its ability to extract features from small objects. Then, FPM decouples the upper-level input features into task-specific sensitive features, providing more useful information for individual tasks. Subsequently, ARAM generates high-quality initial candidate anchors based on the sensitive features extracted by FPM, and further rotates them to create anchors that better match the sensitive features. Finally, DAOM is employed to optimize the label assignment approach during the training phase, enabling the model to dynamically select high-quality candidate anchors and thereby eliminate biases between classification and regression. Detailed descriptions of each component are provided in the following subsections.

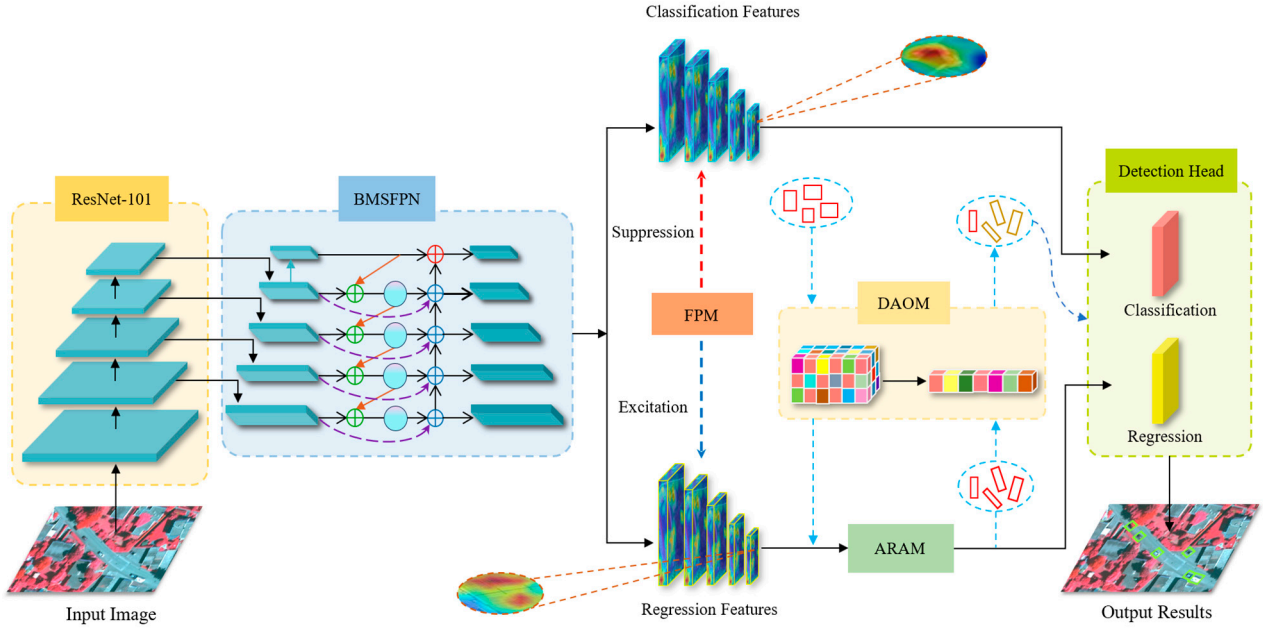


Figure 3. The overall structure and process of the proposed URSNet.

3.1. Bidirectional Multi-Span Feature Pyramid Network—BMSFPN

The upper-level backbone network, ResNet-101, has two types of negative impacts on the accuracy of object detection during the feature extraction process. Firstly, as the depth of feature extraction increases, the resolution of the image gradually decreases, making it difficult to effectively preserve detailed information such as the edges and textures of object parts. Secondly, noise generated by the performance of the image acquisition sensor is amplified during feature mapping, further affecting the detection accuracy.

Figure 4 demonstrates the detection results for two types of objects with different scales before and after processing with Gaussian noise [42]. The detection models used in this comparison are two SOTA models: YOLOv7 [43] and Swin-Transformer [44]. As can be seen in the figure, after introducing noise, both models exhibit a decrease in detection accuracy for large objects, such as a plane. For small objects, like cars, YOLOv7 mistakenly detects them as harbors, while Swin-Transformer fails to detect the cars at all. Overall,

noise introduces significant interference to image quality, having a notable impact on the accuracy and reliability of model detection.

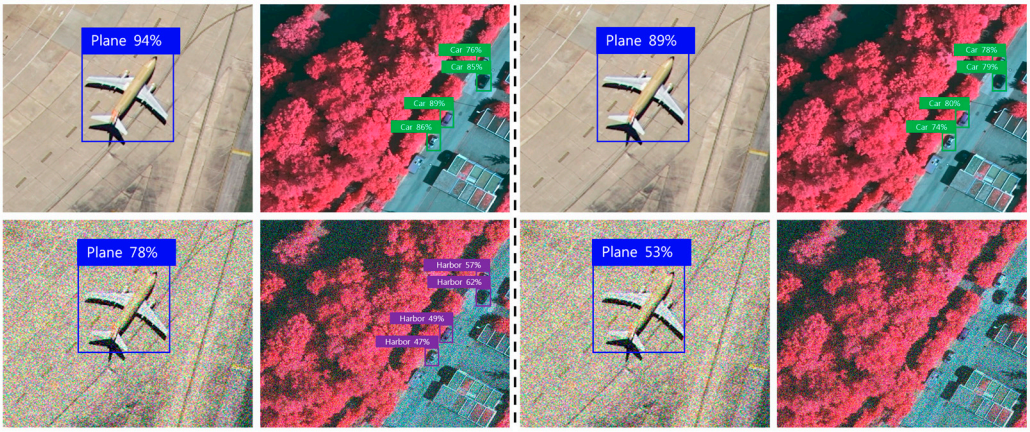


Figure 4. (**a**,**b**) The detection results for a plane and cars using YOLOv7 and Swin-Transformer, respectively. In both (**a**,**b**), the top images are the original ones, while the bottom images are those processed with Gaussian noise.

To enhance the model's ability to address the aforementioned challenges, this section proposes a Bidirectional Multi-span Feature Pyramid Network (BMSFPN), as shown in Figure 5. The operations depicted in Figure 5 can be divided into three stages. In the first stage, the initial input feature $P_1 \sim P_4$ is passed through 1×1 convolutional layer and *Swish* activation function to obtain the subsequent input feature map $P_1^{in} \sim P_4^{in}$. In the second stage, $P_1^{in} \sim P_5^{in}$ is processed through 1×1 convolutional layer and 3×3 max pooling layer to obtain a new input feature P_5^{in}, with a stride of 2. In the third stage, $P_1^{in} \sim P_5^{in}$, obtained from the preceding operations, is used as the final feature input for the bidirectional multi-span feature pyramid structure. This stage represents the core functionality of the BMSFPN. Specifically, the functionality of this structure is mainly realized by the following two components:

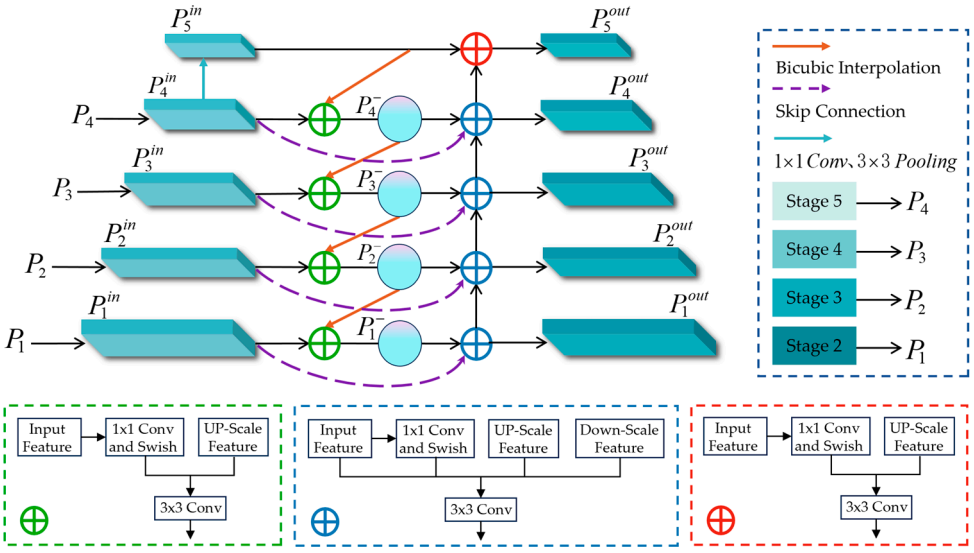


Figure 5. Structure and process of BMSFPN.

Firstly, during upsampling, the quartic bicubic interpolation method [45] is employed to precisely smooth the feature images. This process primarily utilizes the continuous relationship between pixels to enhance image resolution and filter out noise. Specifically,

bicubic interpolation is first used to upsample the input feature P_5^{in}. Then, the upsampled features are added with P_4^{in} in the channel dimension to obtain feature-rich P_4^-. Finally, bicubic interpolation is once again applied to upsample P_4^-, and the upsampled features are added with P_3^{in} in the channel dimension to derive the final feature P_3^-. By repeating this process, new P_2^- and P_1^- features can be obtained. After top-down upsampling, high-level and low-level features are fused, thereby improving the model's perception and anti-interference capabilities.

Secondly, in the downsampling process, multiple cross-node connections are added to the $P_1 \sim P_4$ path, and multi-feature propagation is utilized to prevent the loss of target information during feature extraction. Additionally, cross-scale connections are employed to propagate features from shallow to deep layers, fusing features of different scales. Specifically, through a skip connection, features P_1 and P_1^- are added together along the channel dimension to obtain the P_1^{out} feature. Subsequently, 3×3 max pooling with a stride of 2 is performed on feature P_1^{out}, and the result is added with the skip-connected features P_2 and P_2^- along the channel dimension to obtain P_2^{out}. Similarly, this approach is used to derive features P_3^{out} and P_4^{out}. Finally, cross-scale connections are used to concatenate the input feature P_5^{in} with P_4^{out} along the channel dimension to obtain the P_5^{out} feature. Below, we illustrate the specific calculation formulas for P_4^- and P_4^{out} in BMSFPN, as shown in Equations (1) and (2).

$$P_4^- = conv\left[\frac{\omega_1 \cdot P_4^{in} + \omega_2 \cdot BI\left(P_5^{in}\right)}{\omega_1 + \omega_2 + \beta}\right] \tag{1}$$

$$P_4^{out} = conv\left[\frac{\omega_1' \cdot P_4 + \omega_2' \cdot P_4^- + \omega_3' \cdot Resize\left(P_3^{out}\right)}{\omega_1' + \omega_2' + \omega_3' + \beta}\right] \tag{2}$$

where P_i and P_i^{in} represent the initial and final input features in BMSFPN, respectively. P_i^- represents the fusion feature after bicubic interpolation, and P_i^{out} represents the output feature. ω_1 and ω_2 represent the learnable weights in upsampling, and ω_1', ω_2' and ω_3' represent the learnable weights in downsampling. β is taken as 0.0001. The BI function represents the upsampling operation, which is implemented by bicubic interpolation. The function *Resize* represents maximum pooling (stripe = 2, k = 3×3). By adjusting the size of feature mapping, it makes each feature layer keep the same dimension.

3.2. Feature Polarization Module—FPM

Based on the noise reduction and feature detail preservation capabilities of the BMSFPN described in Section 3.1, this section focuses on addressing the issue of incompatible shared features between the regression and classification tasks. This is primarily achieved by extracting key target features to enhance the model's ability for accurate classification and regression.

Currently, most visual detection models rely on shared features for both classification and regression tasks. However, incompatibility often arises between these two tasks, leading to a decline in detection performance to a certain extent [46]. To eliminate the feature interference between the two tasks and assist the model in effectively extracting key features for different tasks, this section proposes a Feature Polarization Module (FPM). The structure of the FPM is illustrated in Figure 6.

Firstly, feature pyramid networks (FPN) are constructed for classification and regression tasks, respectively. Secondly, a well-structured polarization attention mechanism is designed to enhance the representation capabilities of various types of features. Finally, we utilize a polarization function to generate discriminative features for different task branches. Specifically, for classification, we tend to select global features with high responses to reduce interference from complex backgrounds. For regression, we pay more attention to the edge features of the object and suppress irrelevant regions with high activation.

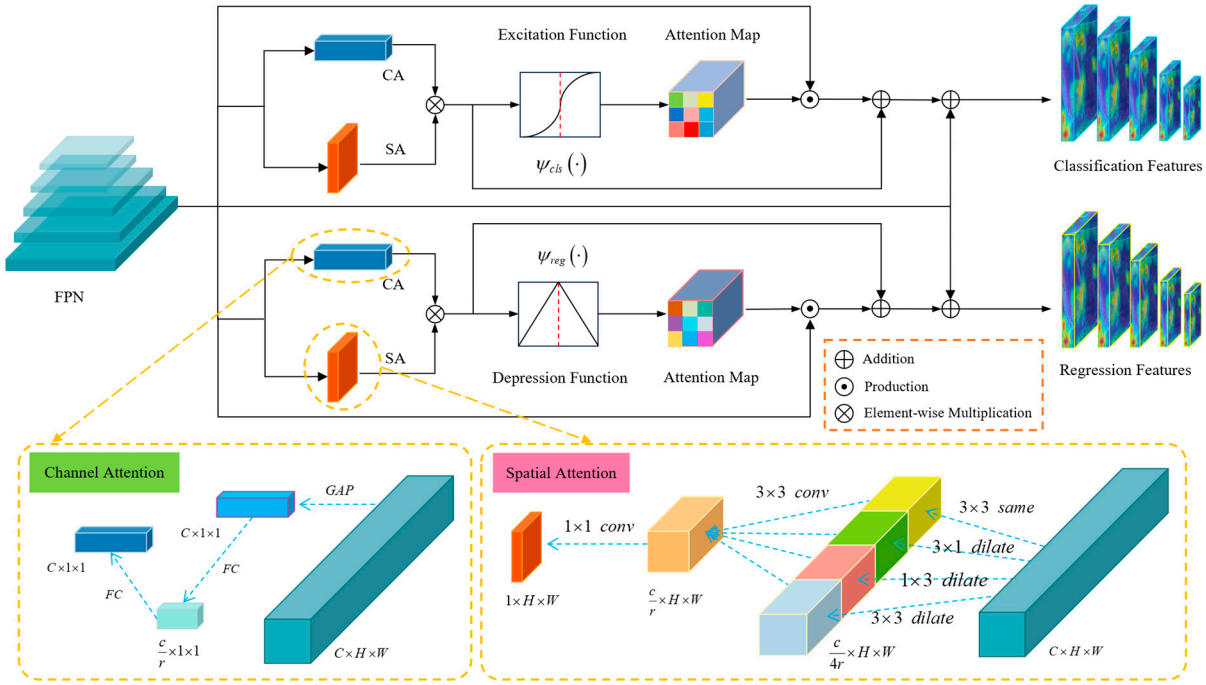


Figure 6. The structure and flow of the FPM. CA and SA in the figure represent channel attention and spatial attention, respectively.

Assuming the given input feature is $f \in \mathbb{R}^{C\times H\times W}$, the construction of the key feature f' is detailed below, as shown in Equations (3) and (4).

$$M = M_c(f) \otimes M_s(f) \tag{3}$$

$$f' = M + \psi(\sigma(M)) \odot f + f \tag{4}$$

where $\otimes$ denotes the tensor product symbol and $\odot$ represents the element-wise multiplication symbol. σ refers to the activation function *Sigmoid*. Firstly, during the process of input feature convolution, the channel attention M_c and spatial attention M_s are extracted. Here, channel attention serves to extract channel relationships from the feature layer. The weights of each channel are extracted using both maximum pooling and fully connected methods. The calculation formula for M_c is as follows:

$$M_c(f) = \sigma(W_1(W_0(f_{gap}))) \tag{5}$$

where $W_0 \in \mathbb{R}^{C/r\times C}$ and $W_1 \in \mathbb{R}^{C\times C/r}$ represent the computational weights in the fully connected layers, f_{gap} is the result of applying maximum pooling to the input features f, and σ denotes the activation function *Sigmoid*.

Spatial attention M_s is primarily used to model the global dependencies among pixels in the input image. The specific calculation process is shown in Equation (6).

$$M_s(f) = \sigma\left(c^{3\times3}\left(cat\left(\left(c^{3\times3}, c_d^{1\times3}, c_d^{3\times1}, c_d^{3\times3}\right)(f)\right)\right)\right) \tag{6}$$

where $c^{3\times3}$ denotes the convolutional operation using the filters of 3×3. $c_d^{1\times3}, c_d^{3\times1}, c_d^{3\times3}$ represent dilated convolutions with kernel sizes of 1×3, 3×1, and 3×3, respectively. *cat* represents the concatenation of features. In this section, we enlarge the receptive field of the convolutional kernels through dilated convolutions. Additionally, to accurately detect elongated objects, we employ convolutional kernels with different proportions.

Secondly, we multiply the channel attention M_c with the spatial attention M_s to obtain a task-specific key response M. Subsequently, a comprehensive feature representation is

constructed through a task-specific polarization function $\psi(\cdot)$. The curve of this function is illustrated in Figure 6. Specifically, for the classification branch, we desire the features to focus more on high-response regions while ignoring unnecessary parts that are only used for object localization or produce interference signals. To achieve this, we can employ the following excitation function to fulfill the classification functionality:

$$\psi_{cls}(x) = \frac{1}{1 + e^{-\eta(x-0.5)}} \tag{7}$$

where η serves as the intensity modulation factor during feature activation. In this experiment, $\eta = 15$ is set accordingly. The high-response regions corresponding to key features in classification tasks are sufficient to assist the network in achieving accurate classification, thus eliminating the need to collect excessive additional information.

For the regression branch, as the key features are generally distributed along the object boundaries, we expect the feature map to focus more on the object contours and contextual information. Next, we process the input features through the following *depression* function to achieve the aforementioned functionality:

$$\psi_{reg}(x) = \begin{cases} x & \text{if } x < 0.5 \\ 1 - x & \text{otherwise} \end{cases} \tag{8}$$

In regression tasks, a strong response generated from a small area along the object's edge cannot effectively determine the entire object's location. The *depression* function in Equation (8) suppresses similar high-response regions in the regression features, encouraging the model to actively seek more potential visual information for precise localization.

Finally, by combining the designed polarized attention-weighted features with the BMSFPN described in Section 3.1, the model is able to extract key features of the object more effectively. Figure 7 provides a visual representation of the FPM results. As can be seen in the figure, the key features required for regression are fully extracted by the FPM, which facilitates the network's better location of the object boundaries and improves positioning accuracy. The extracted key features for classification are mainly concentrated in the most recognizable parts of the target region, contributing to the improvement of target classification accuracy.

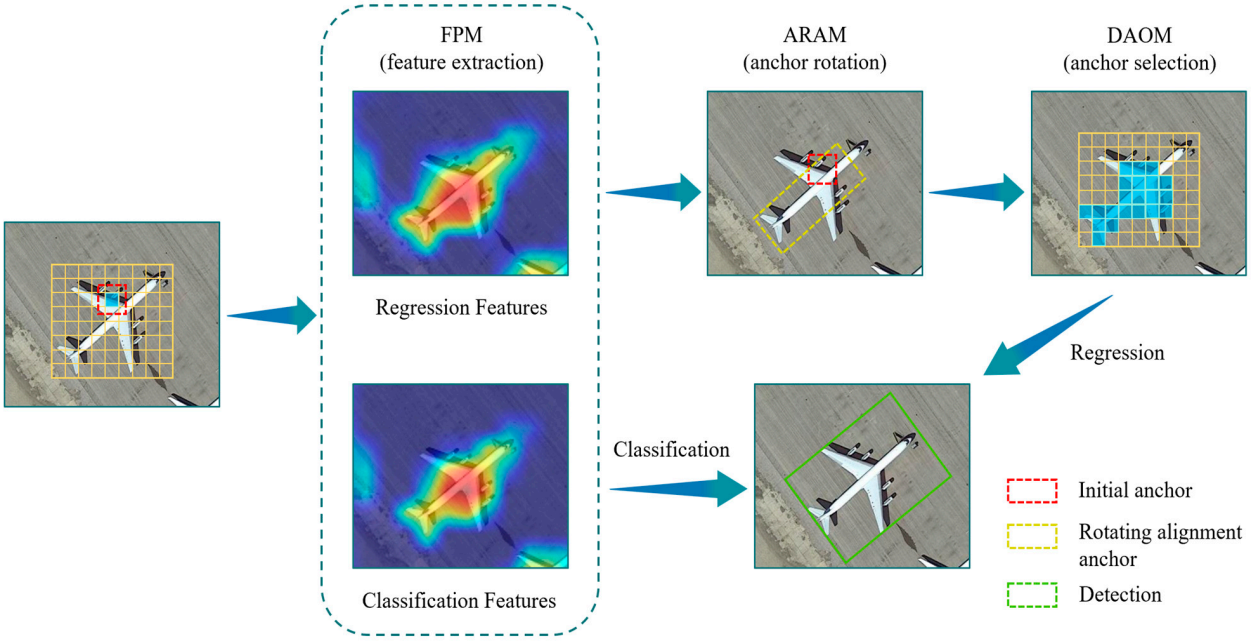


Figure 7. Visualization of the URSNet detection process. The blue parts in the figure represent high-quality candidate anchor centers.

3.3. Anchor Rotation Alignment Module—ARAM

In current anchor-based object detection models, regression tasks are typically performed on predefined dense anchors. However, due to the multi-scale and multi-directional variations exhibited by UAV RS objects, alignment issues arise between the anchors and rotated objects, making it difficult to achieve accurate positioning. To address this issue, this section proposes an Anchor Rotation Alignment Module (ARAM). This module generates high-quality initial candidate anchors based on the sensitive regression features extracted by the FPM in Section 3.2. Through rotation, it further obtains rotated anchors that better match the sensitive regression features. In the rotated anchor regions, the model is able to capture boundary and visual feature information that is conducive to precise positioning.

The structure of the ARAM is illustrated in Figure 8. Firstly, horizontal anchors are set at each location on the regression feature map and represented as (x, y, w, h). Here, (x, y) denotes the coordinates of the center, while w and h represent the preset width and height of the horizontal anchors, respectively. Secondly, the ARAM regresses the new additional rotation angle θ and the preset anchor offset to obtain refined rotated anchors, which can be specifically represented as (x, y, w, h, θ). Finally, the ARAM generates accurate rotated bounding boxes that align with the true object boxes. Specifically, the offset $t^r = (t_x, t_y, t_w, t_h, t_\theta)$ during the rotation of the anchor boxes is calculated using Equation (9):

$$\begin{aligned} t_x^r &= (x - x^a)/w^a,\ t_y^r = (y - y^a)/h^a \\ t_w^r &= log(w/w^a),\ t_h^r = log(h/h^a) \\ t_\theta^r &= tan(\theta - \theta^a) \end{aligned} \tag{9}$$

where x, y, w, h and θ are used for box refinement, and x^a, y^a, w^a, h^a and θ^a are used for anchor refinement.

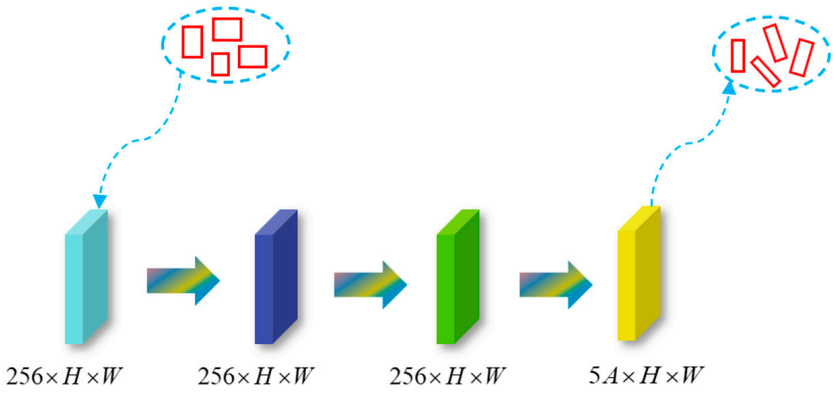


Figure 8. The structure and flow of the ARAM.

Compared to the traditional anchor setting approach, our method only presupposes one anchor (A = 1) at each location on the feature map, enabling the model to achieve more efficient performance. Combined with the special design of the ARAM, we can eliminate the cumbersome setting of hyperparameters such as anchor angles and aspect ratios.

Figure 7 provides a visual representation of anchor refinement. Based on the sensitive regression features extracted by the FPM, the preset square anchors undergo rotation alignment under the guidance of the ARAM to generate accurate candidate bounding boxes.

3.4. Dynamic Anchor Optimization Module—DAOM

Sections 3.2 and 3.3 have introduced the sensitive feature extraction and anchor rotation alignment components of URSNet. However, during the training process, we observed a discrepancy between classification and regression, where a high classification confidence score does not necessarily guarantee high accuracy in target regression. Therefore, this section discusses the procedures used in response to this issue.

During the training process, the detector typically assigns labels based on the IoU between the anchor and the ground truth box, and then selects positive anchor samples [39]. Here, for ease of representation, we use IoU_{in} and IoU_{out} to denote the IoU between the anchor and the ground truth box, and the IoU between the predicted box and the ground truth box, respectively. In general, the more semantic information a selected positive anchor sample has, the more favorable it is for object regression. However, from a statistical perspective of confidence scores, even though there is a strong correlation between the classification confidence and IoU_{in} overall (as shown in Figure 9a), a high IoU_{in} does not guarantee high-precision localization of the anchor, as can be seen in Figure 9b. There is only a weak correlation between the classification score and IoU_{out} (i.e., the object regression capability of the predicted box). We believe that this deviation is due to the unreasonable selection of training anchor samples based on IoU_{in} and the lack of precise alignment between the localization anchor and key object features.

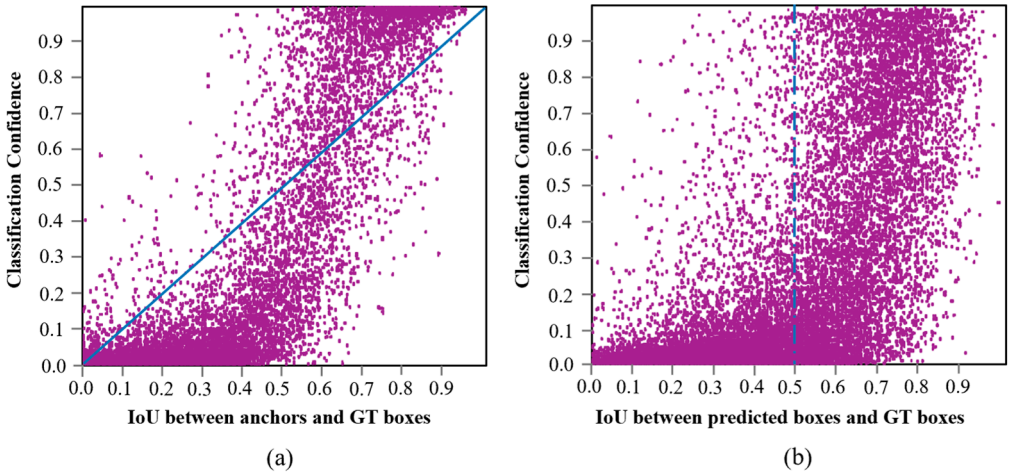


Figure 9. Analysis of the relationship between classification confidence scores and IoU. (**a**) A positive distribution between confidence scores and anchor localization capability; the graph in (**b**) does not demonstrate a strong positive relationship between confidence scores and predicted boxes.

To address the above issues, we propose a Dynamic Anchor Optimization Module (DAOM). In the training phase, this method enables the model to dynamically select anchor samples with critical regression feature capturing capabilities, further assisting the network in improving the accuracy of object localization. Specifically, we introduce a matching degree as a guiding criterion for selecting training samples, as defined in Equation (10):

$$MD = \alpha \cdot IoU_{in} + (1 - \alpha) \cdot IoU_{out} - \mu^{\gamma} \tag{10}$$

where α and γ represent different weighted hyperparameters before and after regression, respectively. MD measures the spatial alignment capability and regression feature alignment capability of the initial anchor through IoU_{in} and IoU_{out}, respectively. It can be seen that the higher IoU_{out} is, the better the predefined anchor can capture key feature information for object regression and perform localization. However, due to uncertainty, some anchor samples with a high IoU_{in} but low IoU_{out} may be falsely identified as negative samples even though they are of high quality [47]. To address this issue, we introduce a penalty term μ into the matching degree metric to reduce the impact of uncertainty. The definition of μ is as follows:

$$\mu = |IoU_{in} - IoU_{out}| \tag{11}$$

We evaluate the erroneous anchor samples based on the change in IoU before and after regression, and apply a distrust penalty to such uncertain samples using μ. By suppressing uncertainty, reasonable training samples can be selected during the regression process. In our experiments, we set a matching degree threshold, where anchor samples with a

matching degree higher than 0.6 are considered positive samples, and otherwise, they are considered negative samples.

The introduction of the matching degree further enhances the model's capabilities in feature alignment and positive–negative sample selection, enabling the accurate division of classification and regression features. As can be seen from the visualization in Figure 7, the designed DAOM dynamically selects candidate anchors that capture key regression features. These high-quality anchors ensure that the model possesses precise object localization capabilities after regression, further mitigating the discrepancy between classification and regression.

4. Experiment

In this section, we first analyze the dataset to provide a good data-driven foundation for model experimentation. Second, we outline the evaluation criteria and implementation details during the experimental process. Finally, we conduct ablation experiments and comparisons with multiple SOTA models to quantitatively and qualitatively verify the effectiveness of the proposed method for object detection in drone RS images.

4.1. Dataset Preparation

To comprehensively evaluate the detection performance of the proposed URSNet method from multiple perspectives, we specifically selected four RS datasets of different scales and types: DOTA-V2.0 [48], RSOD [49], DIOR [50], and UCAS-AOD [51]. Below, we will introduce them in detail.

The DOTA-V2.0 dataset is the latest upgrade of the DOTA series. It contains a larger number of high-resolution images (11,268 images) with sizes ranging from 800×800 to $20,000 \times 20,000$ pixels, and includes a more diverse set of object categories (18 categories). Additionally, the dataset boasts a significant number of instances, totaling 1,793,658. The DOTA-V2.0 dataset is divided into training, testing, and validation sets in a ratio of 6:3:1.

DIOR, proposed by Northwestern Polytechnical University, is a large-scale benchmark RS image dataset. It consists of 20 object categories and 23,463 images, containing 192,388 object instances. All images have a size of 800×800 pixels and are annotated using both HBB and OBB annotation methods. The images in this dataset were collected during different seasons and weather conditions, and certain data augmentation techniques were applied.

RSOD, an open object detection dataset released by Wuhan University, is designed for the detection of aircraft, oil tanks, stadiums, and overpasses in RS images. It adopts the HBB data annotation format and contains a total of 6950 object instances across 976 images. Specifically, there are 446 images of aircraft, 189 of stadiums, 176 of overpasses, and 165 of oil tanks. The image sizes are 512×512 or 1083×923 pixels.

The UCAS-AOD dataset includes two types of objects: airplanes and cars. It contains a certain number of challenging samples (negative examples) and comprises a total of 1000 airplane images and 510 car images, with 14,596 relevant instances. The image sizes are either 1280×659 pixels or 1372×941 pixels. This dataset adopts the HBB method for image annotation.

The objects in these RS datasets are artificially designed and possess unique edge and texture characteristics compared to natural objects. To clearly understand the distribution of instances within the datasets and facilitate the subsequent analysis and evaluation of model detection performance, we have summarized the basic data information of the four datasets in Table 1 and presented the percentages of instance counts in Figure 10.

Figure 10. This figure reflects the percentages of instance counts in the four datasets: DOTA-V2.0, RSOD, DIOR, and UCAS-AOD. It can be observed that DOTA-V2.0 has rich instance categories and a large number of instances, which is conducive to the training of the model proposed in this paper.

Table 1. Information on the four RS datasets: DOTA-V2.0, RSOD, DIOR, and UCAS-AOD.

Dataset	Object Category	Number of Images	Annotation Method	Image Size
DOTA-V2.0 [48]	18	11,268	OBB	800×800 to $20,000 \times 20,000$
RSOD [49]	4	976	HBB	512×512, 1083×923
DIOR [50]	20	23,463	HBB, OBB	800×800
UCAS-AOD [51]	2	1510	HBB	1280×659, 1372×941

4.2. Implementation Details

The backbone network of the proposed model URSNet in this paper is ResNet-101 [39]. To accelerate the training process during model execution, we pre-trained it on Google Open Images-V4 [52] for 120 epochs and fine-tuned it on the training set of DOTA-V2.0. Additionally, we employed data augmentation techniques such as Gaussian noise, HSV jittering, and rotation to enhance some of the data and improve their richness.

In this section, we establish the experimental environment and uniformly set some important parameters for the experiments. Unless otherwise specified, these settings will be used by default. Specific details are presented in Tables 2 and 3.

Table 2. Environment setup.

Environment	Configuration
OS	Windows 11
CPU	Intel Core i7-10700k
GPU	NVIDIA GeForce RTX 4060Ti
Python	3.8.12
PyTorch	1.9.1
Torchvision	0.9.1
OpenCV-Python	4.5.5.64

Table 3. Parameter settings.

Input Size	Optimizer	Learning Rate	Momentum	Batch Size	Weight Decay	Training Epoch
416 × 416	SGD	0.0001	0.937	32	0.0005	1200

4.3. Evaluation Metrics

In order to effectively quantify the experimental data and reasonably evaluate the accuracy, speed and efficiency of each detection model, we selected the evaluation criteria listed below based on the actual situation of this paper.

Precision (P) refers to the proportion of positive predictions that are actually positive among all predictions made by the model. Recall (R) measures the proportion of positive instances that are correctly predicted as positive among all positive instances. Additionally, they can also be used to calculate the F1 score, which comprehensively considers the balance between precision and recall. Their definitions are shown in Equations (12)–(14):

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

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

$$F1\ score = \frac{2 \times P \times R}{P + R} \tag{14}$$

where TP represents the number of true positives, FP represents the number of false positives, and FN represents the number of false negatives.

AP represents the average prediction accuracy of a model for a specific object category across different confidence thresholds. mAP is a commonly used evaluation metric in multi-class object detection tasks. It comprehensively assesses the performance of a model across multiple categories by averaging the AP values of each category. The definitions of AP and mAP are as follows:

$$AP = \int_0^1 P(r)dr = \sum_{k=0}^{n} P(k)R(k)\mathrm{m}AP = \frac{1}{C}\sum_{i=1}^{c} AP(i) \tag{15}$$

$$\mathrm{m}AP = \frac{1}{C}\sum_{i=1}^{c} AP(i) \tag{16}$$

where $P(r)$ is the precision at a certain recall rate in $0 \sim 1$, C is the total number of object categories in the detection dataset, and $AP(i)$ represents the average precision of a specific object category in the dataset.

FPS stands for frames per second, indicating the number of images processed by the model per second, which is used to measure the inference speed of the model. The definition of FPS is shown in Equation (17):

$$FPS = 1000/(Pre + Inf + Nms) \tag{17}$$

Params represents the number of parameters in the model, which is related to the complexity and storage requirements of the model. Its definition is as follows:

$$Params \sim O\left(\sum_{i=1}^{D} K_i^2 \cdot Channel_i^{input} \cdot Channel_i^{output}\right) \tag{18}$$

where D refers to the total number of layers in the algorithm, K represents the size of the convolutional kernel, and $Channel_i^{input}$ and $Channel_i^{output}$ indicate the numbers of channels in the convolutional process.

FLOPs represents the number of floating-point operations executed by a model during the inference stage, which is closely related to the computational complexity and inference time of the model. The definition of FLOPs is as follows:

$$FLOPs_{Net} \sim O\left(\sum_{i=1}^{D} M_i^2 \cdot K_i^2 \cdot Channel_i^{input} \cdot Channel_i^{output}\right) \tag{19}$$

where M represents the size of the feature map obtained during the convolution process.

4.4. Experimental Results and Analysis

4.4.1. Ablation Experiments

In this section, we mainly conduct ablation experiments to verify the rationality and effectiveness of the proposed module in combination with the baseline ResNet-101. The DOTA-V2.0 dataset provides the data foundation for this experiment. To ensure the scientific nature of the experimental process, we proceed from two perspectives:

(1) Quantitative Analysis. We use AP_S, AP_M, and AP_L to represent the average detection results of small, medium, and large objects in the dataset, respectively. Additionally, mAP@0.5 and mAP@0.5 : 0.95 are used to represent the average AP value of all object categories when the IoU threshold is set to 0.5 and the average mAP value of all object categories with a step size of 0.05 when the threshold ranges from 0.5 to 0.95. Among these, mAP@0.5 : 0.95 is the most effective metric in assessing the combination of the proposed module with the baseline framework.

Table 4 demonstrates the detection results for the integration of BMSFPN, FPM, ARAM, and DAOM with the baseline framework, with the optimal results highlighted in bold. Specifically, the integration of the FPM, ARAM, and DAOM results in significant improvements in AP_M and AP_L, outperforming the worst results by 9.18% and 8.54%, respectively. Similarly, mAP@0.5 and mAP@0.5 : 0.95 increase by 12.88% and 9.89%, respectively. These positive outcomes are attributed to the decoupling and refinement of classification and regression features achieved by these three modules, simplifying the model's handling of remotely sensed objects with critical features and variable orientations. The combination of BMSFPN, FPM, ARAM, and DAOM yields the best overall performance, with AP_S, AP_M, and AP_L reaching 65.29%, 85.30%, and 88.17%, respectively, and mAP@0.5 and mAP@0.5 : 0.95 achieving 75.03% and 66.93%, respectively. This demonstrates that the four modules designed in this paper can collaborate effectively with the baseline framework to comprehensively address issues such as noise and the rotational distribution in drone RS image object detection.

Table 4. Ablation results for each module combined with the baseline.

Baseline	BMSFPN	FPM	ARAM	DAOM	Dataset	AP_S	AP_M	AP_L	mAP@0.5	mAP@0.5:0.95
✓	✓					50.13	68.41	77.36	53.94	51.30
✓		✓				45.26	67.02	75.72	52.15	50.13
✓	✓	✓			DOTA	53.70	65.97	78.05	60.47	54.40
✓		✓	✓	✓	-V2.0	52.99	75.15	84.26	65.03	60.02
✓	✓	✓	✓			60.41	82.36	86.73	72.19	62.35
✓	✓	✓	✓	✓		**65.29**	**85.30**	**88.17**	**75.03**	**66.93**

To further validate the efficiency and superiority of the combinations of each module with the baseline framework in detecting objects of various scales, we analyzed the precision and recall rates of the various combinations presented in Table 4 and utilized P-R curves for comparison and verification. The combinations in Table 4 are named in sequential order from I to VI, such as I (Baseline + BMSFPN), II (Baseline + FPM), and so on.

The P-R curves are presented in Figure 11. As can be seen in the figure, VI (the proposed URSNet in this paper) occupies the optimal position in the detection of various object sizes. For large- and medium-sized objects, the precision of VI is slightly higher than that of the second-ranked V, but both significantly outperform the third-ranked IV. This demonstrates that after image filtering, decoupling of key features, and anchor box refinement, URSNet can generally ensure efficient precision. For small objects, VI exhibits the most significant advantage, indicating that the enhancement of classification and localization capabilities for small objects in URSNet is strengthened by highlighting object detail textures through BMSFPN, as well as the rotation and optimization of anchor boxes through the FPM. Furthermore, the distribution of the P-R curves for all object sizes is consistent with the data presented in Table 4, comprehensively reflecting the unique contributions of each module and their indispensability, as well as validating the superiority of the proposed URSNet in this paper.

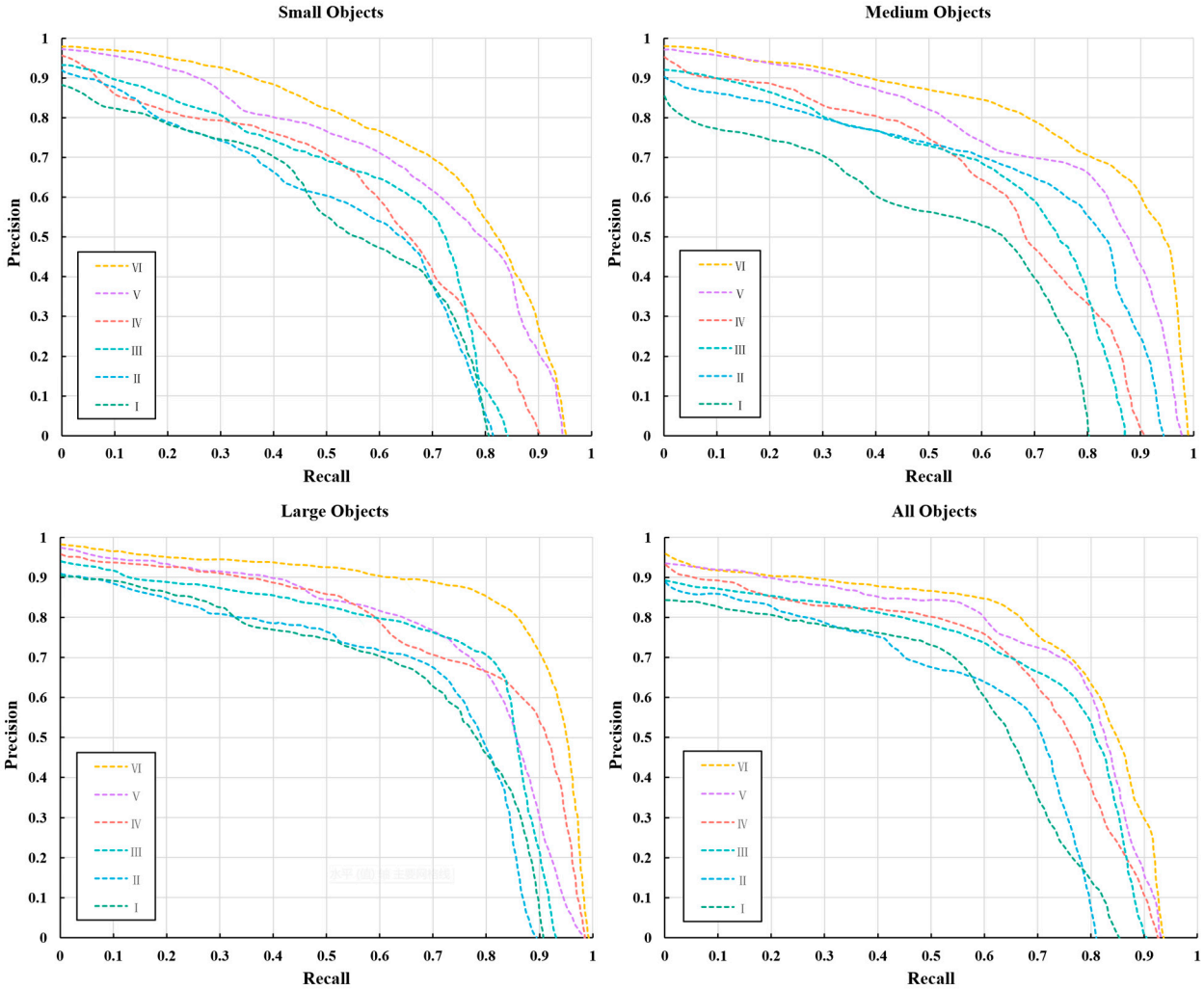


Figure 11. Table 4 presents the precision and recall statistics for various combinations of modules (I–VI) on the DOTA-V2.0 dataset for different types of objects. The resulting P-R curves are displayed above.

Based on the above analysis, we will now conduct experimental validation of the rationality and scientific basis for selecting the ResNet-101 [41] backbone network using the DOTA-V2.0 dataset. According to the models utilized by most researchers [53–58], we have chosen several advanced neural network frameworks for discussion, namely ResNet-50 [59], VGG-16 [60], LSKNet [61], Swin-Trans [44], and DLA-34 [62]. Figure 12 illustrates the P-R curves for these six types of backbone networks. As can be seen in the figure, ResNet-101 exhibits superior precision and recall performance compared to the other networks. In the low recall region, the precision of ResNet-101 is significantly higher than the other networks, indicating its excellent performance in handling complex backgrounds and distinguishing similar targets. In the high recall region, ResNet-101 still maintains high precision, demonstrating its strong generalization ability and robustness to noise and interference. Therefore, through an experimental performance analysis, it can be verified that using ResNet-101 as the backbone network is feasible.

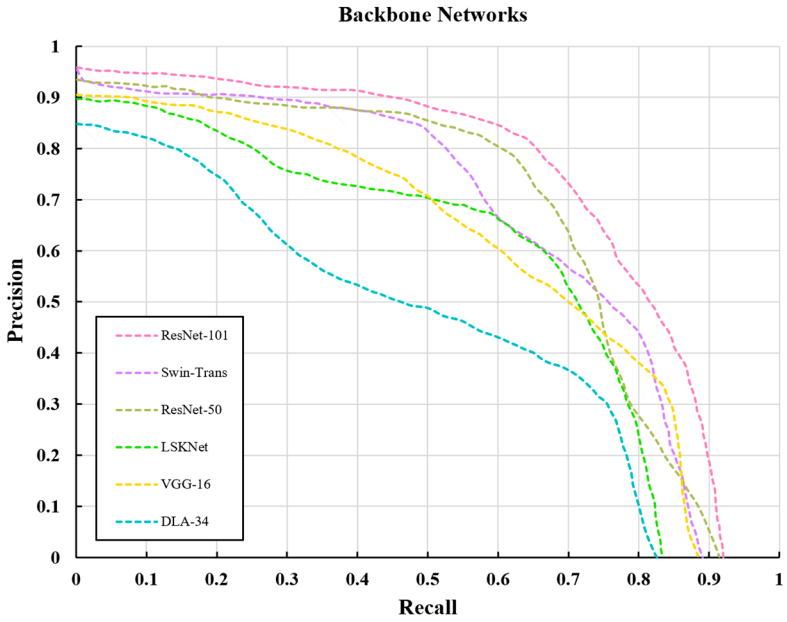


Figure 12. The above figure shows the P-R curves for six types of backbone networks: ResNet-101, ResNet-50, VGG-16, LSKNet, Swin-Trans, and DLA-34.

Additionally, Table 5 presents the evaluation results for various baseline frameworks. It can be observed that the ResNet-101 baseline selected in this paper generally exhibits the best performance. It achieves an F1-Score of 85.72 and a top-1 accuracy of 82.95%, ranking first with an 8.85% advantage over the top-1 accuracy score of DLA-34. Therefore, it can be validated that the selection of ResNet-101 as the baseline framework for the model in this paper is reasonable.

Table 5. Multiple evaluation results for each baseline framework on the dataset DOTA-V2.0.

Baselines	F1-Score	Params (M)	Flops (G)	FPS	Top-1 Accuracy (%)	Excess over DLA-34 (%)
DLA-34 [62]	74.49	7.10	0.58	50.00	74.10	
H-104 [63]	80.13	11.40	3.70	51.60	76.84	+2.74
Swin-Trans [44]	75.36	28.00	4.50	**73.77**	79.60	+5.50
LSKNet [61]	77.90	30.98	17.39	61.80	81.30	+7.20
VGG-16 [60]	80.67	13.84	15.47	60.00	76.03	+1.93
ResNet-50 [59]	83.19	25.60	3.86	55.14	80.10	+6.00
ResNet-101 [41]	**85.72**	44.60	7.80	70.20	**82.95**	**+8.85**

(2) Qualitative Analysis. To observe the ablation results more intuitively, we present them in a visual form in Figure 13. Specifically, the three top-ranked methods (IV, V, and VI) from Table 4 are selected for detection in five challenging scenarios of drone RS images.

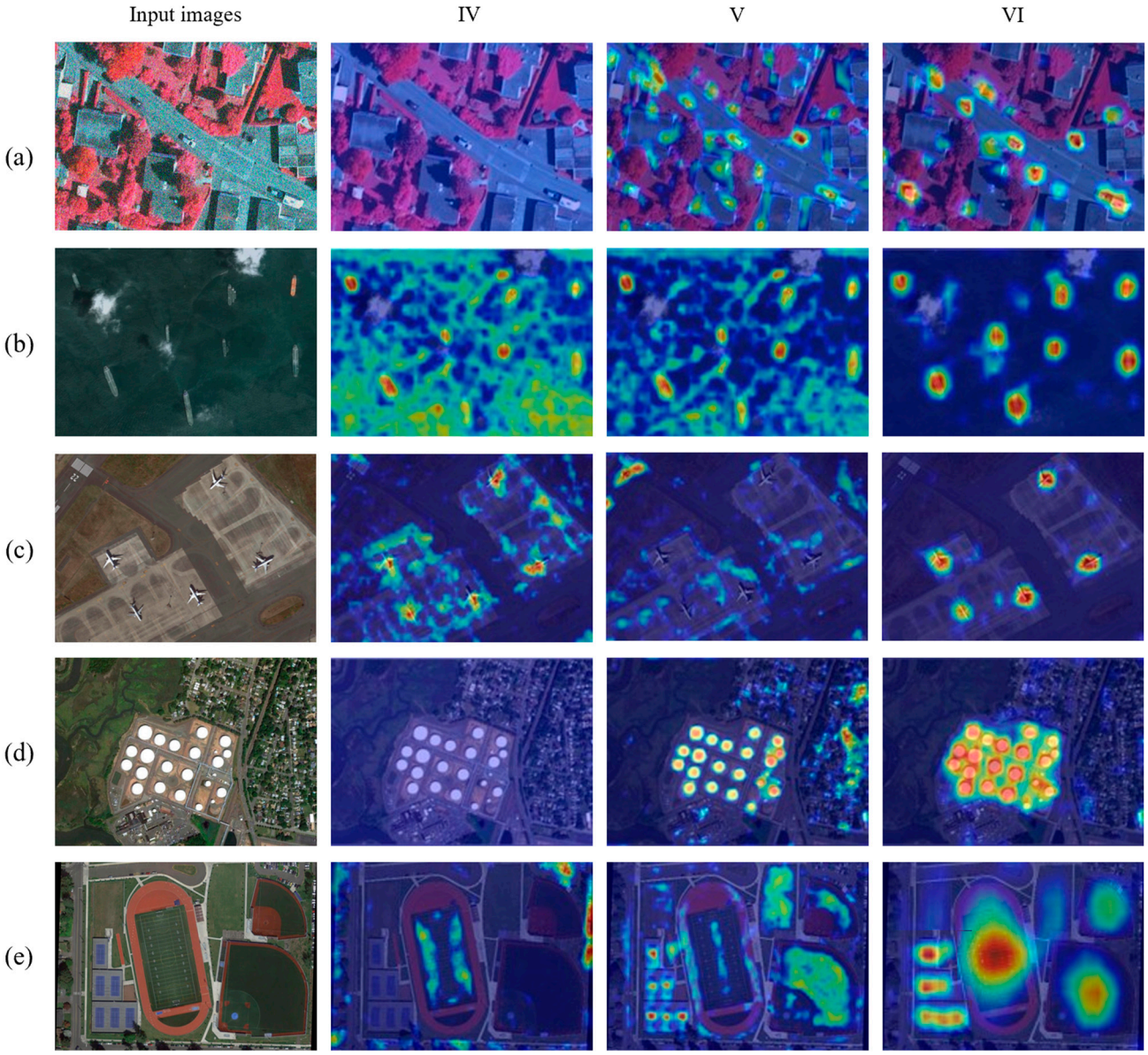


Figure 13. The heatmaps are utilized to visualize the detection results for three combinations of methods on challenging images containing objects. Specifically, (**a**) represents an image with noise; (**b**) depicts an image with clouds and fog; (**c**) shows an image with objects distributed randomly; (**d**) illustrates an image with densely arranged objects; (**e**) presents an image with objects of varying scales.

Observing the output results, the heatmap of our method VI covers the largest number of object regions, maintaining relatively accurate capture capabilities even in complex noisy images. Additionally, for difficult objects with large scale differences, dense distributions, and rotational characteristics, VI utilizes the FPM, ARAM, and DAOM to enhance the expression of object boundary features and anchoring regression capabilities, extracting more feature information compared to V and IV and resulting in more accurate detections. Therefore, these results demonstrate that the strategy of combining the four modules proposed in this paper with ResNet-101 is reasonable and efficient.

4.4.2. Comparison Experiments with SOTA Methods

To fully demonstrate the unique advantages of the proposed URSNet in the task of object detection in drone RS images, it is necessary to conduct a comprehensive performance comparison experiment with similar SOTA algorithms. Therefore, based on the stage division of deep learning detection algorithms, we selected over twenty of the most advanced and classic object detection models for drone RS images from three categories: anchor-free, single-stage, and two-stage algorithms. These include DRN [64], O^2-DNet [65], AOPG [66], CenterMap [67], S^2ANet [68], and AO^2-DETR [69], among others.

This experiment was conducted on four RS datasets: DOTA-V2.0, RSOD, DIOR, and UCAS-AOD. The following is a specific explanation:

(1) The detection results for the DOTA-V2.0 dataset are presented in Table 6. The data in the table are calculated and evaluated strictly according to the AP and mAP standards in MS COCO [70]. For ease of expression, we have simplified the names of various objects in the dataset, such as swimming pool (SWP), helicopter (HC), bridge (BE), large vehicle (LVE), ship (SP), plane (PE), soccer ball field (SBF), basketball court (BC), airport (AT), container crane (CCE), ground track field (GTF), small vehicle (SV), harbor (HB), baseball diamond (BDD), tennis court (TCT), roundabout (RT), storage tank (ST), and helipad (HD). Observing the data in the table, we can find that our model URSNet has the highest mAP score (84.03%), which is 2.75 percentage points higher than the second-ranked LSKNet-S. This indicates that URSNet has the optimal overall detection performance for various objects in DOTA-V2.0. Additionally, in terms of AP performance, SGR-Net, which has an advanced architecture of Swin-Trans, achieves the highest detection result for the small object HB (72.04%), but falls behind URSNet in detecting SP, ST, and HC. This suggests that URSNet still has relatively good performance in small object detection. For elongated objects such as SWP and BE, URSNet has a unique advantage due to the carefully designed spatial attention convolution kernel in the FPM. For medium and large objects with significant scale and edge features, such as BC, TCT, and GTF, both URSNet and AO^2-DETR achieve over 85%, with URSNet outperforming AO^2-DETR by 0.71%, 2.16%, and 0.40%, respectively.

In addition, based on the data in Table 6, we will further analyze the target detection performance of different SOTA models in challenging scenarios in the subsequent results visualization section.

Table 6. Comparison of results from the URSNet method proposed in this article with various SOTA methods using the DOTA-V2.0 dataset. Top-1 and top-2 results are highlighted in red and green, respectively.

Method	Backbone	Object Categories (*AP* (%))																	*mAP* (%)
		SWP	HC	BE	LVE	SP	PE	SBF	BC	AT	CCE	GTF	SV	HB	BDD	TCT	RT	ST	HD	
Anchor-free Methods:																				
DRN [64]	H-104	69.43	57.86	45.21	75.73	56.85	90.74	55.45	80.18	78.63	80.47	65.16	73.25	70.33	73.35	90.54	65.33	80.81	86.87	72.01
O^2-DNet [65]	H-104	66.98	61.03	47.65	73.06	74.62	88.31	80.93	82.28	65.52	73.17	66.27	72.32	58.21	81.14	90.66	60.17	80.06	60.48	71.27
CenterNet-O [71]	DLA-34	56.74	57.77	28.60	67.00	64.75	83.06	83.00	79.05	69.33	75.53	58.60	39.67	56.50	67.00	90.83	53.10	74.54	80.37	65.86
Oriented Rep. [72]	ResNet-50	76.35	53.26	58.85	83.03	68.32	90.53	65.84	80.51	78.94	63.09	79.70	78.90	67.23	75.07	90.86	69.35	73.11	76.41	73.86
CFA-Net [73]	ResNet-101	72.64	65.48	54.86	78.27	63.41	86.29	56.88	82.40	80.33	59.94	84.17	80.90	70.14	81.82	90.71	80.15	81.04	57.93	73.74
SGR-Net [74]	Swin-Trans	77.40	63.16	64.07	80.17	59.13	88.54	77.15	83.46	78.04	69.53	75.57	80.90	72.04	67.65	90.83	78.00	78.54	75.80	75.56
Two-stage Methods:																				
AOPG [66]	ResNet-50	78.85	73.29	70.74	84.73	57.95	89.96	69.07	84.60	76.48	75.80	83.38	78.90	70.72	84.87	90.78	83.21	81.07	65.53	77.78
CenterMap [67]	ResNet-50	73.31	57.70	53.32	69.07	78.98	89.34	45.63	78.86	69.90	80.41	60.78	76.67	63.05	60.74	87.83	79.32	63.45	78.59	70.39
Faster RCNN-O [75]	ResNet-50	69.41	65.48	43.87	79.52	67.55	85.73	48.68	79.86	68.77	68.77	58.89	69.54	67.32	80.69	90.85	60.09	74.39	83.17	70.14
LSKNet-S [61]	LSKNet	83.55	79.86	67.75	87.69	78.06	88.52	85.34	85.97	78.58	80.01	84.03	83.31	68.88	89.65	90.06	86.11	68.70	77.01	81.28
LSKNet-S * [61]	LSKNet	85.41	79.64	64.94	87.04	78.76	89.05	83.72	80.75	76.31	68.74	83.57	81.29	71.79	72.37	90.56	70.64	76.06	84.13	79.15
RoI-Trans [76]	ResNet101	67.84	60.04	56.64	75.46	66.75	88.63	54.62	84.43	74.80	78.37	72.33	79.90	70.68	83.42	86.74	67.53	72.16	70.07	72.80
RC1&RC2 [77]	VGG-16	66.72	74.38	58.78	76.01	69.93	85.53	79.66	79.60	70.65	81.21	69.09	76.31	69.52	85.96	89.90	80.51	80.93	75.47	76.12
Single-stage Methods:																				
S^2ANet [68]	ResNet-50	68.94	59.34	46.28	76.89	77.33	87.11	80.15	84.50	71.46	74.51	88.19	79.46	64.28	86.34	90.52	62.31	80.46	80.04	75.45
AO^2-DETR [69]	ResNet-50	80.81	77.69	59.92	83.15	67.53	91.59	83.24	87.19	78.77	80.00	85.51	81.46	69.64	91.53	90.31	80.04	78.95	83.72	80.61
RetinaNet-O [78]	ResNet-50	69.92	60.94	49.56	71.45	69.12	88.36	77.91	82.10	80.18	74.13	82.76	75.49	62.48	81.52	91.24	81.53	75.41	78.04	75.11
RRD [79]	VGG-16	74.25	57.91	47.24	70.61	60.76	80.79	84.05	84.42	78.51	81.00	67.18	58.19	70.14	90.42	90.85	72.41	80.59	75.13	73.58
R^3Det [80]	ResNet-101	69.25	67.46	58.24	78.54	78.42	89.15	62.08	85.13	80.43	68.14	87.43	75.16	67.49	70.16	90.42	83.15	73.15	83.64	75.97
R^3Det-DCL [81]	ResNet-101	69.54	66.19	56.94	82.43	67.51	89.55	73.54	84.23	75.81	80.07	68.14	76.44	68.54	82.56	90.48	64.59	75.77	85.90	75.46
RetinaNet-R [78]	ResNet-101	76.59	74.17	71.12	80.14	76.58	85.62	80.61	80.64	66.97	78.41	70.15	80.10	48.16	76.58	87.47	78.64	79.31	80.15	76.19
URSNet (ours)	ResNet-101	86.27	75.24	75.47	88.22	79.90	90.12	87.81	87.90	80.22	81.48	88.59	84.15	70.42	91.04	92.47	84.51	82.16	86.63	84.03

* O represents a model framework with a directional bounding box detection capability.

(2) The detection results for the RSOD dataset are presented in Table 7. It can be observed that the proposed URSNet method achieves the best AP and mAP scores for the four types of objects in the RSOD dataset, surpassing the powerful YOLOv7 [43] and Vision-Trans [82]. Specifically, the Anchor Rotation Alignment Module (ARAM) and Feature Polarization Module (FPM) in URSNet are highly effective in handling aircraft with a multi-directional distribution and overpasses with significant differences in length and width.

Table 7. The detection results from multi class SOTA models for the RSOD dataset. The bold result is the best.

Method	Aircraft (%)	Oil tank (%)	Overpass (%)	Playground (%)	*mAP* (%)
CFA-Net [72]	63.92	57.83	80.04	89.30	72.77
SGR-Net [75]	72.05	78.01	83.25	90.13	80.86
LSKNet-S [61]	55.63	63.92	82.91	90.12	73.14
RoI-Trans [76]	78.39	71.77	84.62	92.45	81.81
RC1&RC2 [77]	65.07	70.62	78.33	86.53	75.14
YOLOv7 [43]	87.60	78.35	83.61	88.01	84.39
Vision-Trans [82]	85.73	70.02	86.35	90.36	83.12
URSNet (ours)	**87.59**	**79.16**	**88.41**	**93.58**	**87.19**

(3) The evaluation results for the DIOR dataset are presented in Table 8. The results demonstrate that our method, URSNet, achieves an mAP score of 85.13% and a processing speed of 108.20 FPS, indicating its superior performance in both detection accuracy and image processing speed. This is attributed to the advanced modular architecture design of URSNet. However, it is worth noting that URSNet does not achieve the highest level in terms of model parameters (Params) and computational complexity (FLOPs), which suggests that further improvements are needed in model lightweighting and hardware resource allocation to achieve more efficient performance in future work.

Table 8. The evaluation results from multiple SOTA methods for the DIOR dataset. The optimal results are bolded.

Method	Backbone	FPS	Params (M)	FLOPs (G)	*mAP* (%)
FCOS [83]	ResNet-50	51.50	32.10	38.60	81.01
YOLOX [84]	Modified CSP V5	57.80	99.10	–	80.43
SAR-Net [85]	ResNet-50	–	42.60	–	74.46
RetinaNet-R [78]	ResNet-101	51.80	36.30	40.10	69.37
YOLOv5 [86]	CSPDarkNet53	87.70	7.00	15.80	72.31
LRTrans [87]	ViT	75.80	**3.07**	**9.60**	83.69
URSNet (ours)	ResNet-101	**108.20**	36.33	50.41	**85.13**

(4) The performance evaluation results for the UCAS-AOD dataset are presented in Table 9. The UCAS-AOD dataset contains only two types of objects: cars and planes. However, these two types of objects exhibit characteristics such as an arbitrary orientation, small scale, and dense distribution, making them complex targets that can be used to further validate the effectiveness of the proposed method in this paper. As can be seen in the table, most SOTA models have lower detection accuracy for planes than for cars. Additionally, URSNet outperforms the second-ranked YOLOv7 by 4.05%, 0.15%, and 1.77% in the AP and mAP metrics for both targets. This suggests that for cars with regular edges, most models can effectively extract key feature information for classification and localization. However, when faced with planes with complex boundary information, most models lack efficient key information extraction and localization refinement capabilities, resulting in lower detection accuracy. In contrast, URSNet maintains high accuracy due to the design of BMSFPN and DAOM.

Table 9. Performance evaluation results from several types of SOTA models for the UCAS-AOD dataset. The added part of the optimal result is highlighted in bold.

Method	Cars (%)	Planes (%)	*mAP*
O^2-DNet [65]	78.93	59.43	57.27
Oriented Rep. [72]	80.05	74.01	70.83
YOLOv7 [43]	87.30	84.31	74.09
CenterMap [67]	75.03	80.02	63.11
Faster RCNN-O [75]	66.81	79.50	53.58
S^2ANet [68]	83.65	75.74	68.03
R^3Det-DCL [81]	86.59	83.92	73.16
URSNet (ours)	**+4.05**	**+0.15**	**+1.77**

(5) Visualization of experimental results.

Figure 14 illustrates the visualized detection results from our proposed method URSNet for the large-scale RS dataset DOTA-V2.0. As can be seen in the figure, targets with multi-directional variations such as planes, ships, large vehicles, and small vehicles all exhibit high detection results. This demonstrates that our designed ARAM and DAOM can fully utilize the key features of such targets for anchor rotation refinement, and further improve the classification and localization performance of URSNet for these targets through label assignment.

Furthermore, for medium and large-scale objects such as a baseball diamond, tennis court, and ground track field, URSNet maintains a superior detection level, achieving accuracy rates above 70%. For densely distributed small objects like storage tanks and harbors, URSNet ensures accuracy rates of over 60%. This is attributed to the powerful feature extraction and detail representation capabilities of the designed BMSFPN and FPM.

Figure 15 demonstrates the visualization results from URSNet for the DIOR dataset. Since DIOR contains 20 different categories of objects, the generalization and robustness of URSNet have been thoroughly tested. For elongated objects such as bridges, airports, and swimming pools, their aspect ratios vary significantly, making it difficult for conventional models to extract effective feature information. However, the unique design of the spatial attention convolution kernel in the FPM proposed in this paper effectively overcomes this issue. As can be seen in the figure, URSNet exhibits excellent performance for such objects.

Furthermore, for targets such as golf courses with blurred backgrounds, small-scale chimneys, multi-scale ships, and planes, URSNet effectively reduces the misalignment between predicted and actual bounding boxes by smoothing out redundant background details and dynamically optimizing the target anchor boxes. This enhancement in both classification and localization capabilities results in generally impressive detection performance.

Figure 16 demonstrates the detection results from several advanced SOTA models for the RSOD dataset. Based on Table 7, we visualize the detection results from our proposed URSNet, along with Vision-Trans, YOLOv7, and RoI-Trans, for four types of objects: aircraft, oil tank, overpass, and playground. It can be observed that URSNet detects the oil tank, which has a large size but a small scale, and the overpass, which is located in a variable background, more accurately. At the same time, aircraft and playground are also well detected. The efficient performance of URSNet further validates its advantages and reliability.

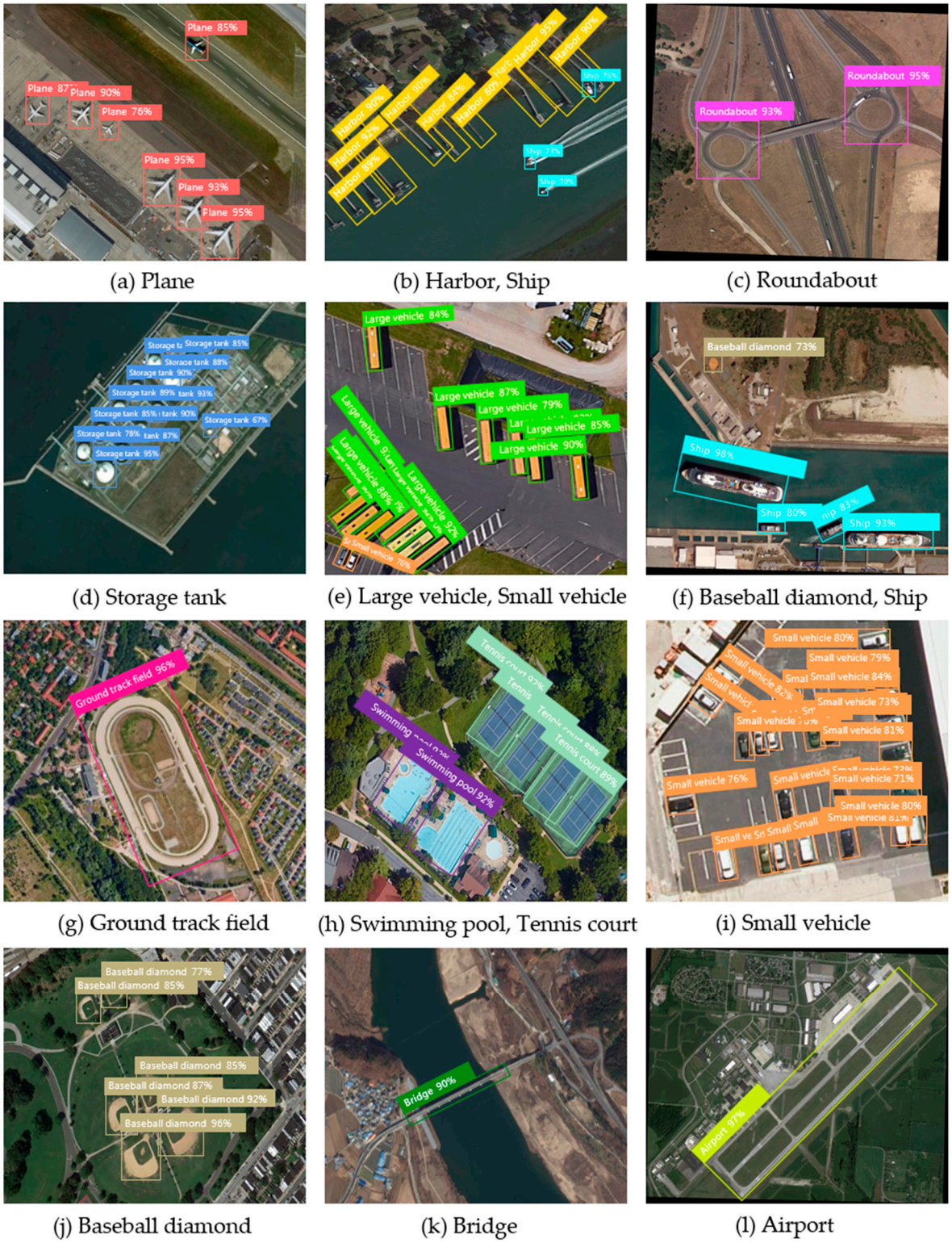


Figure 14. (**a**–**l**) The visual detection results of the proposed URSNet method for the DOTA-V2.0 dataset. We have selected some typical examples for presentation.

Figure 15. (**a**–**l**) The visualization results of the proposed URSNet method for the DIOR dataset. We have selected some typical examples for presentation.

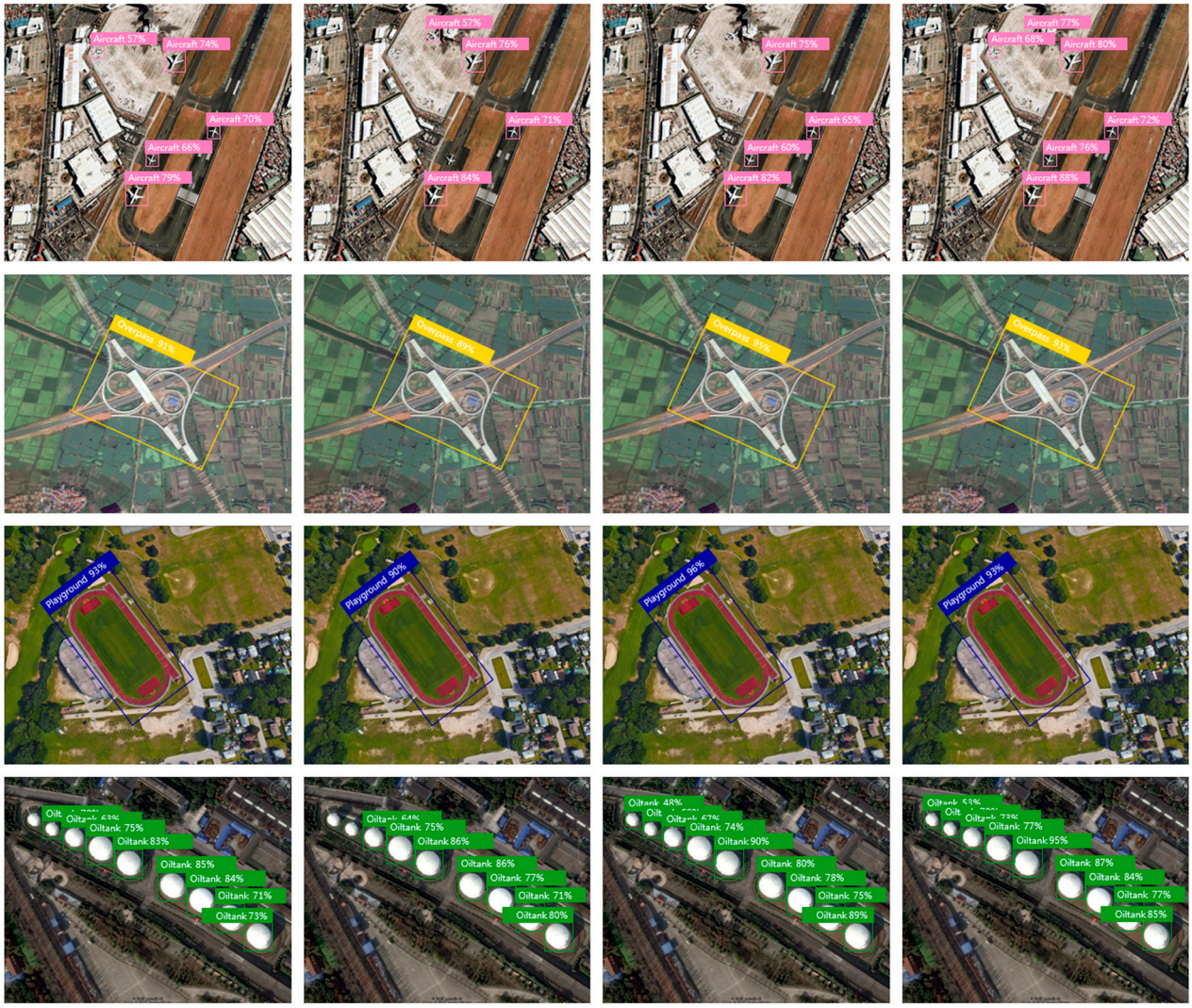


Figure 16. (**a**–**d**) The detection results from RoI-Trans, YOLOv7, Vision-Trans, and our proposed method URSNet for the RSOD dataset. It can be observed that URSNet exhibits the optimal detection performance. It successfully detects all four categories of objects in RSOD and achieves the highest prediction accuracy score.

Figure 17 demonstrates the detection results from URSNet, YOLOv7, R^3Det-DCL, and Oriented Rep. for the UCAS-AOD dataset. As can be seen, for densely arranged cars and planes on similar backgrounds, our proposed method detects all targets with relatively high accuracy, while the other three models exhibit varying degrees of missed detections and poorer accuracy. This illustrates that the advanced architecture of the FPM designed for URSNet effectively highlights the key features of the targets, which are then precisely captured and optimized by the ARAM and DAOM. The efficient performance of URSNet further validates its reliability and applicability in handling small targets in complex scenarios.

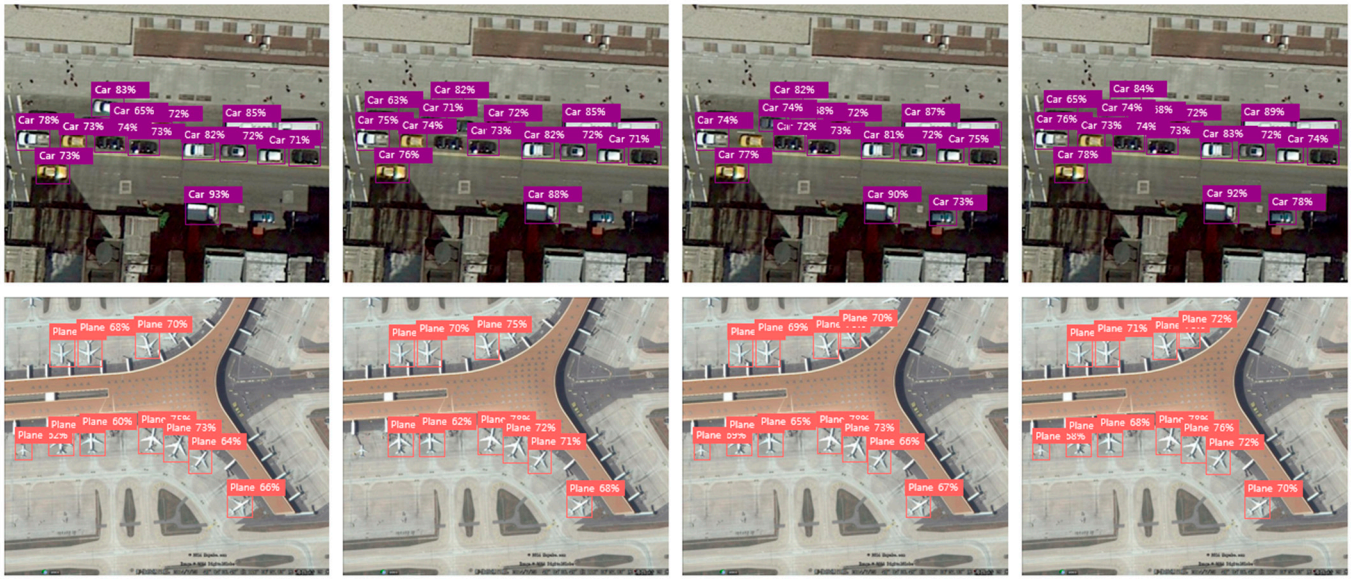


Figure 17. (**a**–**d**) The detection results from URSNet, YOLOv7, R^{3}Det-DCL, and Oriented Rep. for the car and plane targets in the UCAS-AOD dataset. It can be observed that URSNet achieves a higher detection rate and accuracy compared to the other three methods.

5. Conclusions

Based on the current development status and research achievements of object detection technology in the field of drone RS imagery, this paper proposes a target detection network (URSNet) that incorporates a bidirectional multi-span feature pyramid and a key feature capture mechanism to address the challenging issues of noise interference, significant differences in target scale, and arbitrary directional distributions present in RS images. Firstly, BMSFPN is constructed. During the top-down and bottom-up sampling processes, bicubic interpolation, feature weighting, and cross-layer fusion are employed to filter out image noise and enhance the detailed features of the targets. Secondly, our designed FPM constructs robust feature representations for both classification and regression tasks, making it easier for the network to capture key target features with high semantic discrimination. Additionally, the ARAM is introduced to further refine the preset anchor boxes, resulting in high-quality rotated anchors that better match the key regression features. These refined anchor regions provide the model with accurate visual information for localization. Finally, the DAOM is utilized to enhance the model's feature alignment and positive–negative sample discrimination capabilities, enabling the model to dynamically select candidate anchors that capture key regression features and further eliminating the discrepancy between classification and regression. We conducted comprehensive ablation studies and SOTA comparison experiments on challenging RS datasets such as DOTA-V2.0, DIOR, and RSOD. The comparison revealed that URSNet achieves superior experimental results (87.19% mAP and 108.2 FPS) across multiple datasets, indicating that URSNet is effective in addressing the challenges posed by complex drone RS images.

In the future, we will continue to enrich the multi-type object recognition capabilities of URSNet and construct a more comprehensive RS dataset to enhance the robustness and generalization of the model, making it more adaptable to real-world scenarios. Additionally, due to the complexity of our method's structure, it does not demonstrate superiority in terms of model parameters (Params) and computational complexity (FLOPs). This reminds us that further improvements are needed in model lightweighting and hardware resource allocation in subsequent work.

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

Funding: This research was funded by the National Natural Science Foundation of China (grant number: 61671470).

Data Availability Statement: The links to the public datasets used in this study are as follows: https://captain-whu.github.io/DOTA/dataset.html (accessed on 10 February 2024), https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset (accessed on 10 February 2024), https://github.com/ming71/UCAS-AOD-benchmark (accessed on 10 February 2024), and http://www.escience.cn/people/JunweiHan/DIOR.html (accessed on 10 February 2024).

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

Abbreviations

The full names of abbreviations in the manuscript are as follows:

UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once
IoU	Intersection over Union
FPN	Feature Pyramid Network
SOTA	State-of-the-Art
OBB	Oriented Bounding Boxes
HBB	Horizontal Bounding Boxes
AP	Average Precision
mAP	mean Average Precision
FPS	Frames Per Second
Params	Parameters
FLOPs	Floating Point Operations

References

1. Yu, S.; Xie, X.; Du, P.; Wang, X.; Yang, S.; Liu, C.; Nazri, F.M. A Method for Rapidly Determining the Seismic Performance of Buildings Based on Remote-Sensing Imagery and Its Application. *Adv. Civ. Eng.* **2022**, *2022 Pt 13*, 5760913. [CrossRef]
2. Chen, W.; Wang, H.; Li, H.; Li, Q.; Yang, Y.; Yang, K. Real-Time Garbage Object Detection with Data Augmentation and Feature Fusion Using SUAV Low-Altitude Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3074415. [CrossRef]
3. Hu, C.; Li, X.; Jia, S.; Zhang, F.; Xiao, C.; Ruan, M.; Thrasher, J. UPDExplainer: An interpretable transformer-based framework for urban physical disorder detection using street view imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *204*, 209–222. [CrossRef]
4. Cheng, C.S.; Behzadan, A.H.; Noshadravan, A. Uncertainty-aware convolutional neural network for explainable artificial intelligence-assisted disaster damage assessment. *Struct. Control Health Monit.* **2022**, *29*, e3019. [CrossRef]
5. Wang, B.J.; Hsu, C.B.; Lee, J.C.; Chuang, S.J.; Chen, C.H.; Tu, T.M. Military Target Detection in Remote Sensing Imagery Based on YOLOv4-Faster. *J. Imaging Sci. Technol.* **2022**, *66*, 040405-1–040405-9. [CrossRef]
6. Vijayakumar, S.; Santhi, V. Speckle noise reduction in SAR images using type-II neuro-fuzzy approach. *Int. J. Adv. Intell. Paradig.* **2022**, *23*, 276–293. [CrossRef]
7. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote Sensing Image Super-resolution and Object Detection: Benchmark and State of the Art. *arXiv* **2021**. [CrossRef]
8. Bo, Z.; Luyuan, Y. Improved YOLOv5 in Remote Sensing Slender and Rotating Target Detection. In Proceedings of the 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 25–27 February 2022; pp. 918–923. [CrossRef]
9. Zhou, Y.; Ren, C.; Zhang, S.; Xue, X.; Liu, Y.; Lu, J.; Ding, C. A Second-Order Method for Removing Mixed Noise from Remote Sensing Images. *Sensors* **2023**, *23*, 7543. [CrossRef] [PubMed]
10. Kovalenko, B.; Lukin, V.; Vozel, B. BPG-Based Lossy Compression of Three-Channel Noisy Images with Prediction of Optimal Operation Existence and Its Parameters. *Remote Sens.* **2023**, *15*, 1669. [CrossRef]
11. Wang, C.; Zhang, Y.; Wang, X.; Song, J.I. An Effective Strip Noise Removal Method for Remote Sensing Image. *J. Geod. Geoinf. Sci.* **2022**, *5*, 72–85. [CrossRef]
12. Lin, Y.; Sun, H.; Liu, N.; Bian, Y.; Cen, J.; Zhou, H. A lightweight multi-scale context network for salient object detection in optical remote sensing images. *arXiv* **2022**. [CrossRef]

13. Zhang, Z.; Liu, Y.; Liu, T.; Lin, Z.; Wang, S. DAGN: A Real-Time UAV Remote Sensing Image Vehicle Detection Framework. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1884–1888. [CrossRef]
14. Lu, W.; Lan, C.; Niu, C.; Liu, W.; Lyu, L.; Shi, Q.; Wang, S. A CNN-Transformer Hybrid Model Based on CSWin Transformer for UAV Image Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1211–1231. [CrossRef]
15. Zou, F.; Xiao, W.; Ji, W.; He, K.; Li, K. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. *Neural Comput. Appl.* **2020**, *32*, 14549–14562. [CrossRef]
16. Shamsolmoali, P.; Zareapoor, M.; Chanussot, J.; Zhou, H.; Yang, J. Rotation Equivariant Feature Image Pyramid Network for Object Detection in Optical Remote Sensing Imagery. *arXiv* **2021**. [CrossRef]
17. Shi, P.; Zhao, Z.; Fan, X.; Yan, X.; Xin, Y. Remote Sensing Image Object Detection Based on Angle Classification. *IEEE Access* **2021**, *9*, 118696–118707. [CrossRef]
18. Nie, H.; Fu, Z.; Tang, B.H.; Li, Z.; Chen, S. A Multiscale Unsupervised Orientation Estimation Method with Transformers for Remote Sensing Image Matching. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3234531. [CrossRef]
19. Kaur, R.; Singh, S. A comprehensive review of object detection with deep learning. *Digit. Signal Process.* **2022**, *132*, 103812. [CrossRef]
20. Sutradhar, P.; Tarefder, P.K.; Prodan, I.; Saddi, M.S.; Rozario, V.S. Multi-Modal Case Study on MRI Brain Tumor Detection Using Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbor, Temporal Convolution & Transfer Learning. *Am. Int. Univ. -Bangladesh* **2021**, *20*, 107–117. [CrossRef]
21. Rastogi, A.; Jain, R. Deep Learning Applications: An Overview. *J. Adv. Robot.* **2022**, 9.
22. Jabir, B.; Rabhi, L.; Noureddine, F. RNN- and CNN-based weed detection for crop improvement: An overview. *Food Raw Mater.* **2021**, *9*, 387–396. [CrossRef]
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
24. Yng, T.; Li, J. Remote Sensing Image Object Detection Based on Improved YOLOv3 in Deep Learning Environment. *J. Circuits Syst. Comput.* **2023**, *32*, 23502651. [CrossRef]
25. Teng, Z.; Duan, Y.; Liu, Y.; Zhang, B.; Fan, J. Global to Local: Clip-LSTM-Based Object Detection from Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3064840. [CrossRef]
26. Yu, D.; Ji, S. A New Spatial-Oriented Object Detection Framework for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3127232. [CrossRef]
27. Zhao, D.; Shao, F.; Liu, Q.; Yang, L.; Zhang, H.; Zhang, Z. A Small Object Detection Method for Drone-Captured Images Based on Improved YOLOv7. *Remote Sens.* **2024**, *16*, 1002. [CrossRef]
28. Hansen, J.G.; de Figueiredo, R.P. Active Object Detection and Tracking Using Gimbal Mechanisms for Autonomous Drone Applications. *Drones* **2024**, *8*, 55. [CrossRef]
29. Lai, Y.-C.; Lin, T.-Y. Vision-Based Mid-Air Object Detection and Avoidance Approach for Small Unmanned Aerial Vehicles with Deep Learning and Risk Assessment. *Remote Sens.* **2024**, *16*, 756. [CrossRef]
30. Ghorbanzadeh, O.; Xu, Y.; Zhao, H.; Wang, J.; Zhong, Y.; Zhao, D.; Zang, Q.; Wang, S.; Zhang, F.; Shi, Y.; et al. The Outcome of the 2022 Landslide4Sense Competition: Advanced Landslide Detection from Multisource Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9927–9942. [CrossRef]
31. Ye, Y.; Ren, X.; Zhu, B.; Tang, T.; Tan, X.; Gui, Y.; Yao, Q. An Adaptive Attention Fusion Mechanism Convolutional Network for Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 516. [CrossRef]
32. Feng, X.; Zhang, W.; Su, X.; Xu, Z. Optical Remote Sensing Image Denoising and Super-Resolution Reconstructing Using Optimized Generative Network in Wavelet Transform Domain. *Remote Sens.* **2021**, *13*, 1858. [CrossRef]
33. Chen, J.; Li, H.; Chen, T.; Hu, B.; Liu, S. A Denoising Method of Remote Sensing Images Based on Improved BM3D. In Proceedings of the CSAE 2020: The 4th International Conference on Computer Science and Application Engineering, Sanya, China, 20–22 October 2020. [CrossRef]
34. Xie, Z.; Liu, L.; Luo, Z.; Huang, J. Image Denoising Using Nonlocal Regularized Deep Image Prior. *Symmetry* **2021**, *13*, 2114. [CrossRef]
35. Gao, T.; Niu, Q.; Zhang, J.; Chen, T.; Mei, S.; Jubair, A. Global to Local: A Scale-Aware Network for Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
36. Lin, B.; Yang, X.; Wang, J.; Wang, Y.; Wang, K.; Zhang, X. A Robust Space Target Detection Algorithm Based on Target Characteristics. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3080319. [CrossRef]
37. Cheng, G.; He, M.; Hong, H.; Yao, X.; Qian, X.; Guo, L. Guiding Clean Features for Object Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3104112. [CrossRef]
38. Liu, S.; Zhang, L.; Lu, H.; He, Y. Center-Boundary Dual Attention for Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3069056. [CrossRef]
39. Ghorbanzadeh, O.; Blaschke, T. Optimizing Sample Patches Selection of CNN to Improve the mIOU on Landslide Detection. In Proceedings of the International Conference on Geographical Information Systems Theory, Applications and Management, Heraklion, Greece, 3–5 May 2019. [CrossRef]
40. Pan, C.; Li, R.; Liu, W.; Lu, W.; Niu, C.; Bao, Q. Remote Sensing Image Ship Detection Based on Dynamic Adjusting Labels Strategy. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3268330. [CrossRef]

41. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
42. Huynh, K.T.; Nguyen, H. Drunkenness detection using a CNN with adding Gaussian noise and blur in the thermal infrared images. *Int. J. Intell. Inf. Database Syst.* **2022**, *15*, 398–419. [CrossRef]
43. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
44. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
45. Han, G.; Chen, Y.; Wu, T.; Li, H.; Luo, J. Adaptive AFM imaging based on object detection using compressive sensing. *Micron* **2022**, *154*, 103197. [CrossRef]
46. Song, G.; Liu, Y.; Wang, X. Revisiting the sibling head in object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11563–11572.
47. Ming, Q.; Zhou, Z.; Miao, L.; Zhang, H.; Li, L. Dynamic anchor learning for arbitrary-oriented object detection. *arXiv* **2020**, arXiv:2012.04150. [CrossRef]
48. Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M.Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7778–7796. [CrossRef] [PubMed]
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. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In Proceedings of the ICIP, Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.
51. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens.* **2015**, *36*, 618–644. [CrossRef]
52. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.R.R.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [CrossRef]
53. Xu, Z.; Sun, K.; Mao, J. Research on ResNet101 Network Chemical Reagent Label Image Classification Based on Transfer Learning. In Proceedings of the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Weihai, China, 14–16 October 2020. [CrossRef]
54. Jiang, M.; Jing, C.; Chen, L.; Wang, Y.; Liu, S. An application study on multimodal fake news detection based on Albert? ResNet50 Model. *Multimed. Tools Appl.* **2024**, *83*, 8689–8706. [CrossRef]
55. Balachandran, G.; Krishnan, J.V.G. Moving scene-based video segmentation using fast convolutional neural network integration of VGG-16 net deep learning architecture. *Int. J. Model. Simul. Sci. Comput.* **2023**, *14*, 23410143. [CrossRef]
56. Sharsha, A.; Matsun, A. Innovative Horizons in Aerial Imagery: LSKNet Meets DiffusionDet for Advanced Object Detection. *arXiv* **2023**, arXiv:2311.12956.
57. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3144165. [CrossRef]
58. Zhou, G.; Chen, L.; Wu, F. LaneAF: Robust Multi-Lane Detection with Affinity Fields. *arXiv* **2022**. [CrossRef]
59. Sharma, A.K.; Nandal, A.; Dhaka, A.; Polat, K.; Alwadie, R.; Alenezi, F.; Alhudhaif, A. HOG transformation based feature extraction framework in modified Resnet50 model for brain tumor detection. *Biomed. Signal Process. Control.* **2023**, *84 Pt 1*, 104737. [CrossRef]
60. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
61. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. *arXiv* **2023**, arXiv:2303.09030.
62. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. *arXiv* **2017**. [CrossRef]
63. Yang, J.; Liu, Q.; Zhang, K. Stacked hourglass network for robust facial landmark localisation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 79–87.
64. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.W.; Ma, C.; Xu, C. Dynamic Refinement Network for Oriented and Densely Packed 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. 11204–11213.
65. Wei, H.; Zhang, Y.; Chang, Z.; Li, H.; Wang, H.; Sun, X. Oriented objects as pairs of middle lines. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 268–279. [CrossRef]
66. 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*, 5618111. [CrossRef]
67. Wang, J.; Yang, W.; Li, H.C.; Zhang, H.; Xia, G.S. Learning center probability map for detecting objects in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4307–4323. [CrossRef]
68. Han, J.; Ding, J.; Li, J.; Xia, G. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 5602511. [CrossRef]
69. Dai, L.; Liu, H.; Tang, H.; Wu, Z.; Song, P. Ao2-detr: Arbitrary-oriented object detection transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2342–2356. [CrossRef]

70. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Doll'ar, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
71. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
72. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1829–1838.
73. Guo, Z.; Liu, C.; Zhang, X.; Jiao, J.; Ji, X.; Ye, Q. Beyond Bounding-Box: Convex-hull Feature Adaptation for Oriented and Densely Packed Object Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8788–8797.
74. Mai, S.; You, Y.; Feng, Y. SGR: An Improved Point-Based Method for Remote Sensing Object Detection via Dual-Domain Alignment Saliency-Guided RepPoints. *Remote Sens.* **2024**, *16*, 250. [CrossRef]
75. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
76. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q.K.; Soc, I.C. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2844–2853.
77. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017; Volume 2, pp. 324–331.
78. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
79. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.-s.; Bai, X. Rotation-sensitive regression for oriented scene text detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918.
80. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv* **2019**, arXiv:1908.05612. [CrossRef]
81. Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; Yan, J. Dense label encoding for boundary discontinuity free rotation detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 15819–15829.
82. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
83. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. *arXiv* **2019**. [CrossRef]
84. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**. [CrossRef]
85. Lin, H.; Liu, Z.; Cheang, C.; Fu, Y.; Guo, G.; Xue, X. SAR-Net: Shape Alignment and Recovery Network for Category-level 6D Object Pose and Size Estimation. *arXiv* **2021**. [CrossRef]
86. Dong, X.; Yan, S.; Duan, C. A lightweight vehicles detection network model based on YOLOv5. *Eng. Appl. Artif. Intell. Int. J. Intell. Real-Time Autom.* **2022**, *113*, 113. [CrossRef]
87. Feng, K.; Lun, L.; Wang, X.; Cui, X. LRTransDet: A Real-Time SAR Ship-Detection Network with Lightweight ViT and Multi-Scale Feature Fusion. *Remote Sens.* **2023**, *15*, 5309. [CrossRef]