remote sensing

MDPI

Article

ORCNN-X: Attention-Driven Multiscale Network for Detecting Small Objects in Complex Aerial Scenes

Yanfen Li [1], Hanxiang Wang [1], L. Minh Dang [2], Hyoung-Kyu Song [2] and Hyeonjoon Moon [3,*]

[1] School of Computer Science, Qufu Normal University, Rizhao 276826, China
[2] Department of Information and Communication Engineering, and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea; songhk@sejong.ac.kr (H.-K.S.)
[3] Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: hmoon@sejong.ac.kr

Abstract: Currently, object detection on remote sensing images has drawn significant attention due to its extensive applications, including environmental monitoring, urban planning, and disaster assessment. However, detecting objects in the aerial images captured by remote sensors presents unique challenges compared to natural images, such as low resolution, complex backgrounds, and variations in scale and angle. Prior object detection algorithms are limited in their ability to identify oriented small objects, especially in aerial images where small objects are usually obscured by background noise. To address the above limitations, a novel framework (ORCNN-X) was proposed for oriented small object detection in remote sensing images by improving the Oriented RCNN. The framework adopts a multiscale feature extraction network (ResNeSt+) with a dynamic attention module (DCSA) and an effective feature fusion mechanism (W-PAFPN) to enhance the model's perception ability and handle variations in scale and angle. The proposed framework is evaluated based on two public benchmark datasets, DOTA and HRSC2016. The experiments demonstrate its state-of-the-art performance in aspects of detection accuracy and speed. The presented model can also represent more objective spatial location information according to the feature visualization maps. Specifically, our model outperforms the baseline model by 1.43% mAP50 and 1.37% mAP_{12} on DOTA and HRSC2016 datasets, respectively.

Keywords: deep learning; remote sensing; object detection; attention module; multiscale feature extraction

check for updates

Citation: Li, Y.; Wang, H.; Dang, L.M.; Song, H.-K.; Moon, H. ORCNN-X: Attention-Driven Multiscale Network for Detecting Small Objects in Complex Aerial Scenes. *Remote Sens.* **2023**, *15*, 3497. https://doi.org/10.3390/rs15143497

Academic Editor: Konstantinos Topouzelis

Received: 15 June 2023
Revised: 7 July 2023
Accepted: 8 July 2023
Published: 11 July 2023

1. Introduction

Small object detection is an important and challenging subject in the field of remote sensing image analysis [1]. Compared to natural images, remote sensing images contain small objects with complex environments and large-scale variations. Moreover, objects can be oriented at different angles in remote sensing applications, which further increases the difficulty of small object detection [2].

Oriented small object detection in aerial images is a demanding task due to various factors. The challenges arise mainly because of the unique characteristics of the collected images, such as the presence of low resolution, noise, and occlusion [3]. Furthermore, small objects in these images often exhibit various orientations and aspect ratios, which makes their detection even more challenging.

In addition to the image-specific challenges, the detection of oriented small objects is further complicated by the limitations of traditional object detection algorithms. Traditional algorithms commonly rely on feature pyramids to fuse features of multiple scales directly and enhance the model's ability to detect objects of varying sizes. However, during the process of feature fusion and scale transformation, the feature information between different scales often gets lost or diluted, leading to a reduction in the model's feature expression capability [4,5]. This challenge is further compounded in remote sensing images, where small objects occupy only a small fraction of the image, and their discriminative features tend to be obscured by background noise [6].

To overcome the challenges of detecting small objects in remote sensing images, effective feature extraction and fusion methods are necessary to improve the model's perception ability and handle variations in scale and angle. Enhancing the model's ability to detect objects of various sizes and angles while mitigating the effects of scale and angle variations is critical.

In this research, we propose an innovative framework for oriented small object detection in remote sensing images. Our framework is based on an effective feature extraction network and an adaptive feature fusion mechanism to improve detection performance. We evaluate our framework on two public datasets and demonstrate that it achieves state-of-the-art performance on both datasets.

This work makes several significant contributions to addressing the limitations mentioned above.

1. An adaptive attention structure was constructed to enhance feature extraction ability. By designing and implementing this attention structure, the model is able to selectively focus on important features and effectively capture relevant information from the input data.
2. A creative feature extraction network is proposed for rotation-based small object detection, which leverages the multi-grid strategy to enhance the model's ability to capture and represent features of objects with different scales and orientations.
3. A novel feature fusion mechanism is introduced to address the issue of feature dilution in traditional feature pyramid networks. This mechanism aims to alleviate the loss of feature information by preserving more relevant features during multi-scale feature fusion.
4. The proposed framework achieves a promising performance on public datasets.

The rest of this article is arranged as follows. Section 2 presents a comprehensive overview of the research related to small object detection, feature extraction and feature fusion. In Section 3, we provide an explanation of the methodology and the corresponding flowchart of the proposed system. The data used to validate the performance of the presented framework are described in Section 4. In Section 5, a series of experiments are carried out to showcase the advancements achieved through this study. Finally, the article concludes by pointing out the current limitations and potential areas for future work in Section 6.

2. Related Work

In recent years, many studies have been conducted on small object detection in aerial images. Some researchers have focused on improving feature extraction methods, while others have explored various techniques to enhance feature fusion mechanisms. In this section, some recent papers in the relevant areas are reviewed and analyzed.

To address the limitations of small object detection, many studies have proposed effective feature extraction methods. In [7], the authors proposed a deep learning-based approach that utilizes a feature fusion module to extract multi-scale and multi-directional features for small object detection. However, it required a large number of computational resources and was computationally expensive. Similarly, in [8], the authors proposed an end-to-end network that uses residual blocks to refine features and a region proposal network to detect objects, but it had limitations in terms of scale variations. In [9], the authors proposed a method that uses a feature pyramid network to extract multi-scale features and a deformable convolutional network to enhance the model's feature representation. However, this method required a huge amount of training data and high computational cost. Furthermore, leveraging more efficient feature extraction networks such as ResNeSt [10] and ResNeXt [11] is considered an effective approach to enhance the accuracy of small object detection algorithms. However, there is still room for improvement in their capability to extract multi-scale features and spatial positional information.

Several studies have focused on feature extraction and fusion approaches for small object detection. For instance, the multi-level feature fusion networks proposed in [12,13] have the advantage of fusing features from multiple levels of a CNN (Convolutional Neural Network), which can enhance the ability of the model to detect small objects with varying sizes and scales. The networks achieved multi-level feature fusion via convolutional layers and pooling layers. However, this approach may suffer from the problem of feature dilution, which occurs when the feature information from different scales is merged. The improved FPN (Feature Pyramid Network) proposed in [14] overcomes the problem of feature dilution by using bottom-up and top-down pathways to enhance feature expression and fusion. This approach can help the model to detect small objects with high accuracy and robustness. However, the improved FPN still has limitations in handling scale variations, as it cannot dynamically allocate the weights of features of diverse scales. The feature fusion process is still based on fixed weight coefficients, which may not be optimal for small object detection in the whole image, with large variations in scale and orientation.

In this work, we propose a novel framework to detect oriented small objects in aerial images. Our presented framework addresses the weaknesses of previous studies by introducing an effective feature extraction network and an adaptively feature fusion mechanism. Specifically, an effective multiscale network with a dynamic attention module is designed to improve the feature extraction ability, and a novel feature fusion mechanism is introduced to alleviate the feature dilution issue in the multi-scale feature fusion process. Moreover, our framework can dynamically allocate the contribution of the features with different scales, which is critical for small object detection in aerial images with large variations in scale and orientation.

3. Methodology

Figure 1 depicts the overall architecture of our proposed rotation-based small object detection algorithm (ORCNN-X), which is based on ORCNN (Oriented Region-based Convolutional Neural Network) [15]. It consists of a feature extraction module (a), a feature pyramid module (b), ORPN (Oriented RPN) (c), and a prediction head (d). The feature extraction module is responsible for distinguishing between object and background in the input image and extracting the semantic information of the target object. Four different scales of feature maps are extracted from the backbone network to address objects of different sizes in object detection. The feature pyramid module includes four branches that process feature maps of distinct scales from the feature extractor and then fuse them by upsampling to obtain more comprehensive feature information.

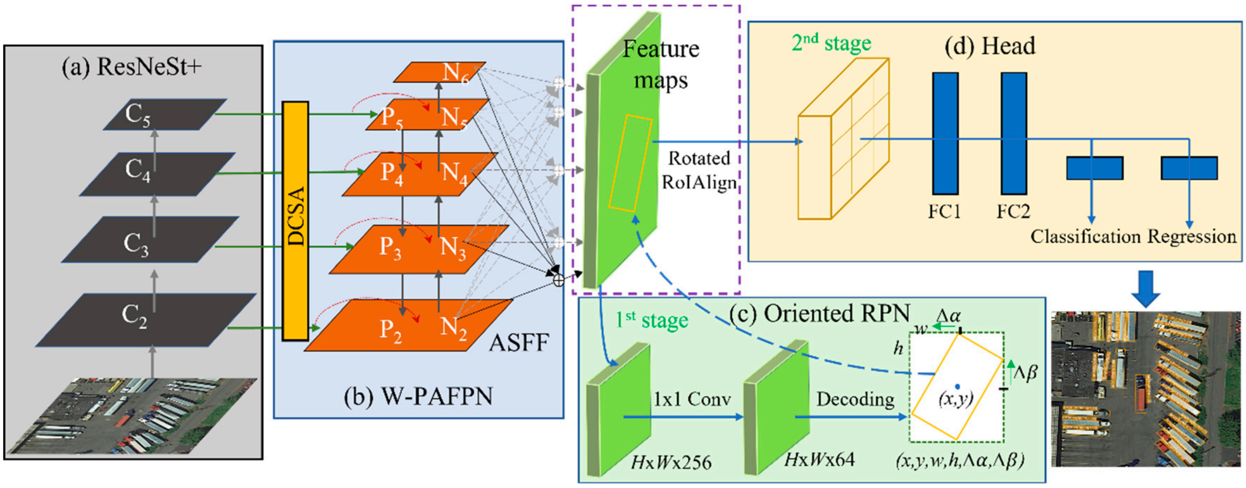


Figure 1. Flowchart of the proposed rotation-based small object detection framework.

After multiple information extraction and fusion, the feature pyramid module outputs five different feature maps containing various semantic information to the ORPN structure. ORPN uses anchor boxes with various aspect ratios and orientations to generate proposals for objects with rotated bounding boxes in the first stage. It generates anchors with predefined angles instead of horizontal anchors to cover the possible orientation range of objects, which helps accurately detect objects with rotated bounding boxes. Then, the Oriented Head uses a combination of classification and regression layers to predict the class label and the offset of the oriented bounding box for each proposal. The regression layer provides the offsets of the four corners of the oriented bounding box relative to the anchor box, and the classification layer predicts the class label probabilities for each proposal. In the following sections, we describe the analysis and improvements made to the feature extractor and feature pyramid.

3.1. Effective Feature Extraction Structure

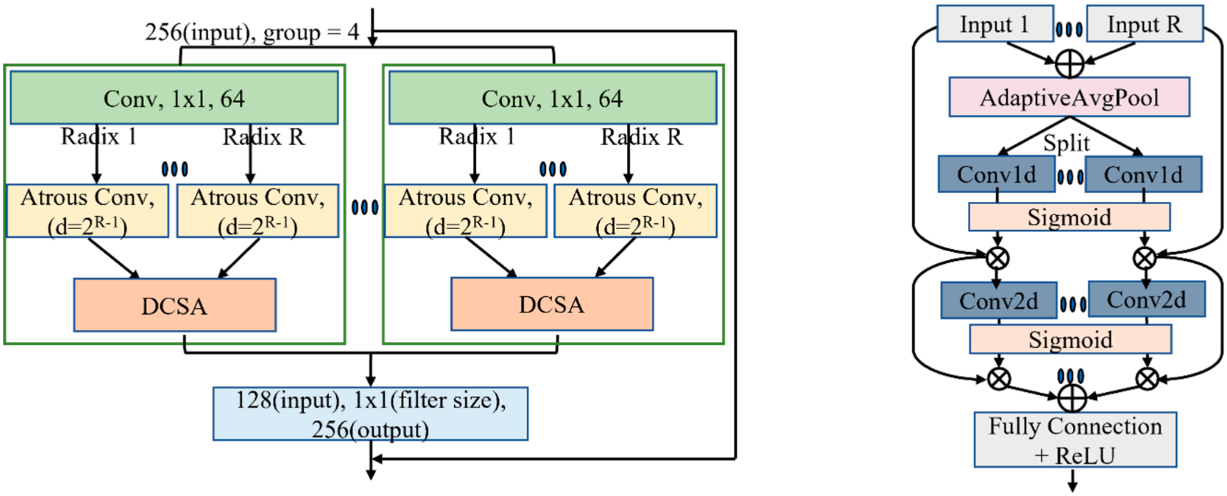
In deep learning, the feature extraction module is one of the most crucial parts of a model, as it directly affects the model's performance. A good feature extraction module can extract the most useful features from the data, making the model more accurate and faster in completing various computer vision tasks (image classification, object detection, and semantic segmentation). For example, the feature extraction module is responsible for transforming low-level features, such as edges and textures, into higher-level semantic features that better describe the image's content. In recent years, residual networks have been extensively used in diverse tasks and have gained remarkable performance. In this study, we conducted a thorough analysis of the deep residual module ResNeSt [9] and proposed an effective multi-scale residual network.

3.1.1. ResNeSt

ResNeSt (Introduced in 2020) has emerged as a prominent deep neural network architecture that has achieved remarkable success in image classification tasks, surpassing many other models in performance on several benchmark datasets, such as ImageNet [16], COCO [17], and CIFAR [18]. Thus, it was selected as a basic extractor model to replace the backbone in Oriented RCNN. ResNeSt is an enhanced version of ResNet [19] that adopts a multi-branch architecture to increase the model's capacity and an attention mechanism to selectively highlight important features in the object regions. It employs two hyperparameters, namely cardinality (K) and radius (R), to embody the multi-branch idea. The input features are partitioned into K groups, and each group is further split into R subsets. Each subset is processed by a layer of 1×1 convolution and a layer of 3×3 convolution. These layers apply mathematical operations to the input features in order to extract and transform information that can be used by the network for further processing. In this approach, the feature maps generated by each subset within a group are combined and passed through a split attention module. This module then selectively emphasizes relevant features within each channel, effectively re-weighting the importance of different channels to improve overall feature representation. The split attention module employs a softmax function to compute attention weights for each subset, and these weights are used to scale the input features from each subset before combining them to form new feature maps. This process helps to emphasize informative features and suppress less useful ones, leading to improved performance in object detection tasks.

3.1.2. ResNeSt+

ResNeSt has been utilized for object detection tasks and achieved satisfactory performance, but still suffers from false positives and false negatives when dealing with high-density multi-object scenes with varying sizes, such as aerial images and drone images. To improve the model's recognition ability for multi-scale objects in complex backgrounds, we made two improvements to ResNeSt. Firstly, we used the multi-Grid strategy to strengthen the model's feature extraction capability. As shown in Figure 2a, each group was assigned 64 channels when the input channel is set to 256. Each 3 × 3 convolution operation within a group was replaced by a multi-Grid convolution module, in which each convolution branch used a different dilation rate to increase the model's sensitivity to objects of different scales. The dilation rate of each branch was calculated as $2^{\mathrm{R}-1}$. R is the index of the branch. For example, the dilation rate of the first branch is 1. Secondly, we proposed a DCSA (Dynamic Channel-Spatial Attention) module to replace split attention in order to further ameliorate the feature extraction ability of the backbone.



(a) ResNeSt+ block

(b) Dynamic Channel-Spatial Attention (DCSA)

Figure 2. (**a**) The architecture of ResNeSt+ block and (**b**) the structure of the proposed Dynamic Channel-Spatial Attention.

The structure of DCSA is shown in Figure 2b. The DCSA consists of two sections: a DCA (dynamic channel attention) module and a DSA (dynamic spatial attention) module. The DCA module adaptively recalibrates channel-wise feature responses by learning channel attention weights. The channel attention module is inspired by ECANet (Efficient channel attention) [20], which uses the idea of one-dimensional convolution to reduce computational complexity. Firstly, the average pooling operation reduces each channel of the feature map to a single value, capturing the global correlations between channels. Then, the one-dimensional convolution performs convolution on each channel in different branches based on global correlations, capturing local correlations. Finally, the sigmoid activation function normalizes the output of the multiple branch convolutions, and uses the attention coefficient as the weight of the channel to perform a weighted sum on the feature maps of different branches, obtaining the feature map with enhanced channels. The calculation can be expressed as:

$$Y_i^B = f(\omega a^B)\cdot X_i,\ a \in \mathbb{R}^C,\ X \in \mathbb{R}^{C\times H\times W} \tag{1}$$

where X_i denotes the ith channel of the input feature map, ω indicates the parameters of the one-dimensional convolution operation, f is the sigmoid activation function, and Y_i refers to the output of that channel. With the one-dimensional convolution operation, we can learn a channel weight that can adaptively weight each channel's feature value to enhance useful features and suppress noise points during feature processing. B represents the index of the branch, which ranges from 1 to R. $C,\ H,\ W$, respectively, refer to the channel, height, and weight of the feature tensor. a is the average value of features over each channel, which is calculated as shown in Equation (2).

Equation (3) represents the calculation of attention weight for a single branch. In the equation, $\mathbf{z}$ represents the input feature tensor, and $\mathbf{W}^*$ is the learned weight matrix.

$$a_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{c,i,j} \tag{2}$$

$$f(\mathbf{z}^B)_i = \frac{1}{1 + \exp(-(\mathbf{W}^* \cdot \mathbf{z}^B)_i)}, i = 1, 2, \ldots, c \tag{3}$$

The output feature maps are fed into the DSA module, which improves the informative spatial regions by assigning them higher weights. The DSA module achieves its functionality by two-dimensional convolutional layers and a sigmoid activation function. DCA and DSA work together to learn the most important features and suppress irrelevant features, making the backbone more robust to various scales and aspect ratios of objects, especially for small object detection. Afterwards, a fully connected layer is used to transform the channel numbers of different branches to guarantee that all feature maps have the same channel number. The introduction of the DCSA structure and incorporation of the multi-Grid convolution module significantly enhanced the performance of ResNeSt+. Compared with prior studies, the proposed ResNeSt+ gained considerable improvement in terms of computation complexity (parameters and FLOPs) and detection accuracy (mAP, AP50, and AP75).

3.2. Feature Fusion Mechanism

The PAFPN (Path Aggregation Pyramid Fusion) model [13] is a type of feature pyramid network used in object detection tasks. As an upgraded version of FPN [21], PAFPN does not simply select a single feature map from different levels of the feature pyramid. Instead, it aggregates information about the feature pyramid at all levels through path aggregation, thereby improving the accuracy and stability of the features. It consists of a bottom-up pathway and a top-down pathway as illustrated in Figure 3a. In the bottom-up pathway, features of different scales are extracted from the input image using a CNN backbone. The features are then passed via a series of convolution layers to generate a feature pyramid. In the top-down pathway, the feature pyramid is processed in a reverse manner to generate a set of feature maps with the same spatial resolution as the original input image. These feature maps are then aggregated using a fusion strategy to yield the final detection results.

The fusion strategy in PAFPN is based on path aggregation and pyramid fusion. The path aggregation is performed by connecting the feature maps at different levels of the feature pyramid using lateral connections. This allows the high-level features to be combined with the low-level features to improve the detection results of small objects. The pyramid fusion is performed by using a weighted sum value to combine the features from different levels of the pyramid. The weights are learned during the training process and are used to manage the contribution of each level of the pyramid.

However, the PAFPN model has some limitations when it comes to small object detection. Since the feature areas of such objects are small, the PAFPN model divides the image into multiple scales through a feature pyramid, which can lead to small objects being ignored or misclassified during feature extraction. Additionally, multiple fusions can dilute important features, as feature fusion reduces the clarity of feature maps. Diluted features cannot provide sufficient information to detect small objects. Therefore, the feature fusion mechanism of the PAFPN model needs to be optimized and adjusted to improve its performance.

Figure 3b shows the structure of our proposed W-PAFPN (Weighted PAFPN) model, which has three main improvements: 1. The proposed DCSA attention module is added between ResNeSt+ and PAFPN to receive four different-sized feature maps and to control the contribution of features from different stages. 2. The lateral shortcut (red dashed line) is added to establish a direct connection between high-resolution and low-resolution feature maps, allowing information to flow more smoothly between different levels and improving the model's performance. 3. ASFF (Adaptively Spatial Feature Fusion) [22] is used to further process the output features. Through ASFF, feature maps of different resolutions can be effectively fused without being forced to concatenate them together. ASFF can preserve high-resolution feature information while effectively capturing contextual information from lower-resolution feature maps. Combining ASFF and PAFPN can further improve the model's feature fusion effect and detection performance, even in complicated background situations. Additionally, since both methods are based on feature fusion, combining them does not lead to a significant computation cost.

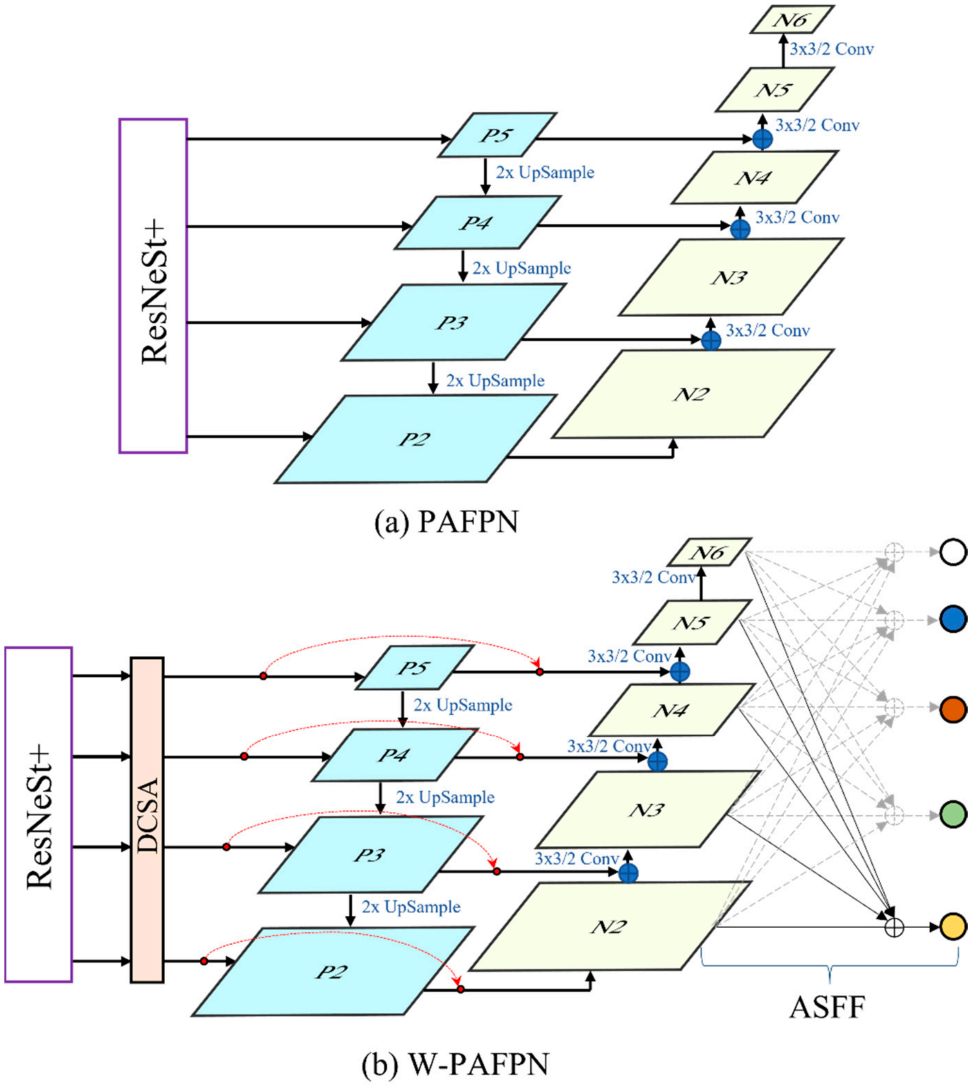


Figure 3. (**a**) The structure of PAFPN and (**b**) the proposed W-PAFPN. 'Conv' is convolutional layer. 'ASFF' refers to adaptively spatial feature fusion.

Assuming there are K feature maps with sizes of $H_k \times W_k \times C_k$, where H_k and W_k represent the height and width of the feature map, respectively, for the K-th feature map, ASFF uses a weight coefficient vector α_k of size $H_k \times W_k$ to represent the importance of each position. This weight coefficient vector is obtained by a branch network and can be adaptively trained for different tasks. Let $x_{i,j}^{(k)}$ represent the feature vector at the (i,j) position of the K-th feature map, $y_{i,j}$ represent the weighted result at the corresponding position, and $\alpha_{i,j}^{(B)}$ represent the weight coefficient at position (i,j) of the feature maps in B-th branch. Then, the calculation equation of ASFF can be defined as:

$$y_{i,j} = \sum_{k=1}^{k} \alpha_{i,j}^{(B)} x_{i,j}^{(k)} \tag{4}$$

4. Dataset

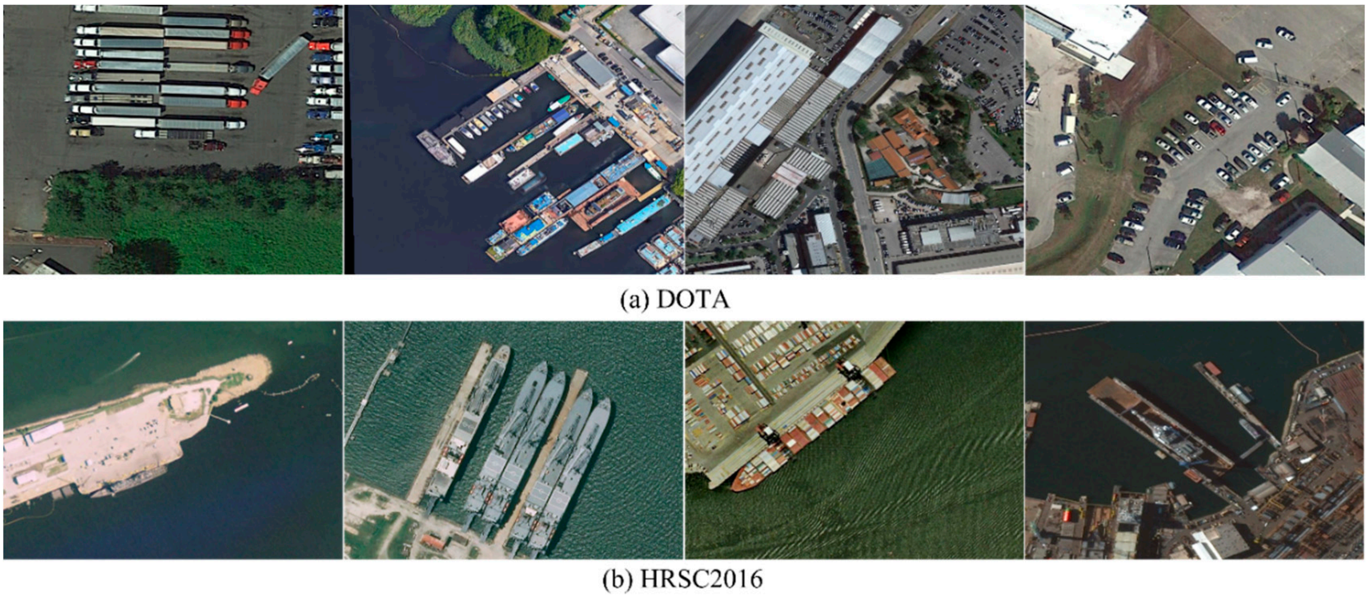
The datasets investigated in this research are DOTA [23] and HRSC2016 [24]. DOTA is a large-scale data for aerial object detection. It consists of 2806 high-resolution aerial images with diverse scenes and objects, such as airports, ships, and vehicles. The dataset covers a total area of 15 square kilometers and includes over 188,000 annotated instances of 15 object categories. Table 1 displays the abbreviations of 15 categories. The dataset consists of a training set, a validation set, and a testing set, with 1411, 458, and 938 images, respectively. The image sizes range from 800 × 800 pixels to 4000 × 4000 pixels, with a mean size of 2000 × 2000 pixels. The annotations are provided in both text and XML formats.

Table 1. The category names and corresponding abbreviations in DOTA dataset.

Classes	Abbreviations	Classes	Abbreviations	Classes	Abbreviations
Bridge	BR	Small vehicle	SV	Basketball court	BC
Harbor	HA	Large vehicle	LV	Soccer-ball field	SB
Ship	SH	Baseball diamond	BD	Roundabout	RA
Plane	PL	Ground track field	TF	Swimming pool	SP
Helicopter	HC	Tennis court	TC	Storage tank	ST

To increase the diversity and richness of the dataset, as well as to avoid cases where some targets are ignored or segmented into two parts, the DOTA training and testing sets are cropped and sliced into different scales using Python tools. The processed dataset contains approximately 160,000 images.

HRSC2016 dataset is a publicly available remote sensing dataset for ship detection and identification, which was collected by the optical and SAR (Synthetic Aperture Radar) sensors mounted on satellites. The dataset has a total of 1070 SAR images with corresponding annotations. The annotations are offered in the form of rectangular bounding boxes, which indicate the locations of ships in the images. The dataset is divided into two parts: the training set contains 626 images, and the testing set contains 444 images. Some sample images from the two datasets are shown in Figure 4.


(a) DOTA

(b) HRSC2016

Figure 4. Sample images of DOTA and HRSC2016.

5. Experimental Results

For the experimental setup, we utilized a powerful server with four Tesla V100 GPUs, which is a widely used hardware configuration for deep learning tasks. In addition, the server was equipped with an Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz processor and 256 GB RAM, providing sufficient computational resources for training and evaluation. To implement our proposed framework, we chose PyTorch as the deep learning framework due to its popularity and ease of use. The experiments were performed on the Ubuntu 18.04 operating system. To ensure the reliability of our experimental results, all the experiments were performed using the same hardware and software environment. This approach enables us to conduct a fair comparison of the results and eliminate the impact of hardware or software discrepancies.

5.1. Features Extraction

In this section, we designed experiments to prove the feature extraction capability of the proposed multiscale backbone model. Table 2 shows a comparison of different backbones and frameworks for the task of oriented object detection. The evaluated metrics include the Param (number of parameters) and FLOPs (floating-point operations) required for each model, as well as the mAP (mean average precision) at different IoU (intersection over union) thresholds (mAP, AP50 and AP75). Four different backbone networks were evaluated, namely ResNet, ResNeXt, ResNeSt, and ResNeSt+ (the proposed backbone). All models use the FPN neck and the Oriented RCNN framework with multi-scale training.

Table 2. Oriented object detection performances of different backbone models on DOTA dataset. Note: Params is short for parameters, FLOPs stands for Floating Point Operations per Second.

Backbones	Neck	Framework	Params	FLOPs	mAP (%)	AP50 (%)	AP75 (%)
ResNet	FPN	Oriented RCNN	60.13 M	282.88 G	51.89	80.63	56.93
ResNeXt			59.79 M	286.58 G	52.06	80.72	57.01
ResNeSt			44.93 M	227.65 G	52.25	81.06	57.22
ResNeSt+ (Proposed)			**44.78 M**	**227.42 G**	**52.30**	**81.27**	**57.34**

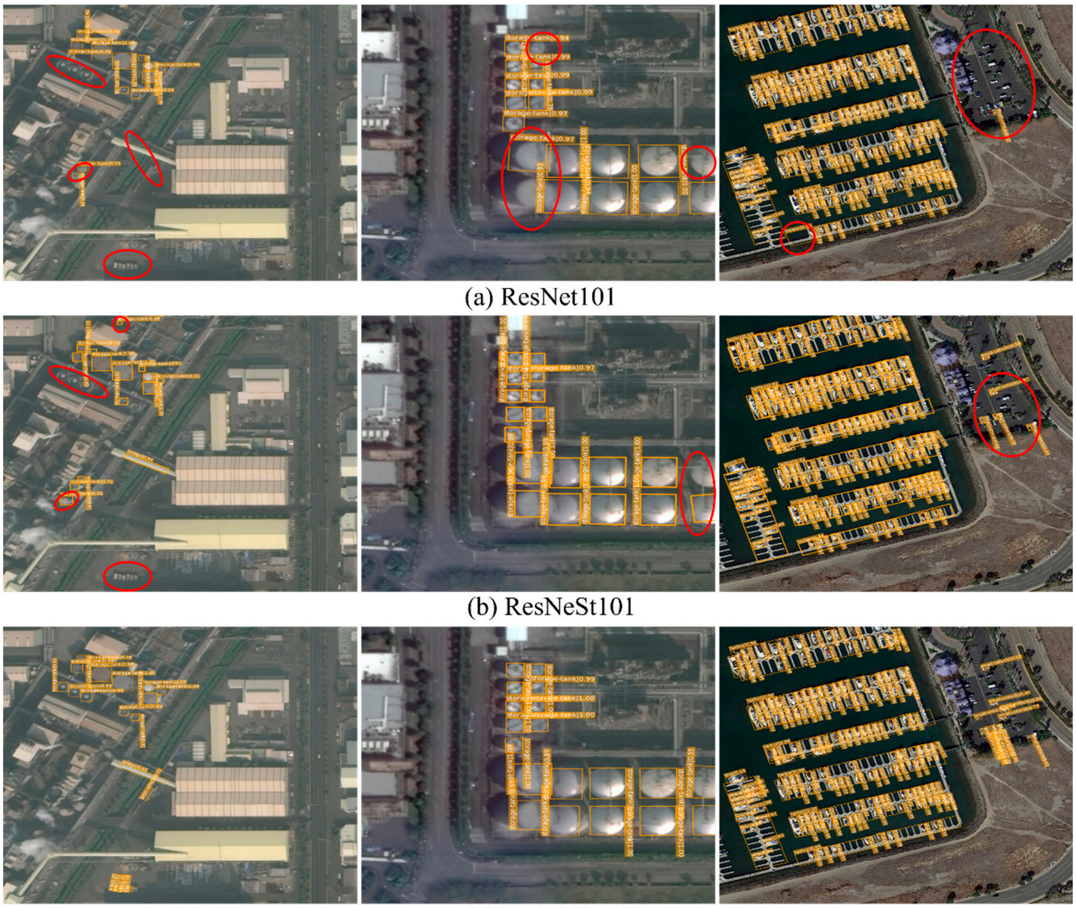
As shown in the table, the proposed ResNeSt+ backbone achieved the highest AP50 of 81.27% on the DOTA testing set, which is 0.64% higher than the baseline model. It also had the lowest number of parameters and FLOPs compared to the other backbone networks, making it more computationally efficient. The ResNeSt backbone achieved the second-best mAP scores, with a mAP of 52.25%. The ResNet and ResNeXt backbones had lower mAP scores than ResNeSt and ResNeSt+ but required a similar amount of parameters and FLOPs. Overall, the results suggest that ResNeSt+ is a promising backbone network for oriented object detection tasks due to its high performance and computational efficiency.

Moreover, to validate the feature extraction capability of the proposed model on small objects, a test dataset consisting only of small objects was created by extracting images from the DOTA validation set. The dataset contains 10,000 patch images with a size of 1024 × 1024. Two models are trained on the training dataset. Table 3 presents the performance comparison between two models, ResNet and ResNeSt+, on the new test dataset. The evaluation metric used is mAP. ResNet achieves a mAP score of 77.8%, while ResNeSt+, which is the proposed model, achieves a higher score of 78.9%. The table also shows the mAP scores for each of the 15 object categories, including Baseball diamond (BD), Harbor (HA), Storage tank (ST), etc. ResNeSt+ outperforms ResNet in most categories, indicating the better feature extraction capability of the proposed model on small objects.

Table 3. Performance comparison of ResNet and proposed ResNeSt+ on small objects detection using our DOTA testing set. '*' based on the Oriented RCNN framework with FPN.

Models	BR	HA	SH	PL	HC	SV	LV	BD	TF	TC	BC	SB	RA	SP	ST	mAP
ResNet *	62.7	80.2	81.9	89.4	66.1	73.7	83.8	83.3	78.6	90.2	85.2	70.0	65.3	74.9	82.4	**77.8**
ResNeSt+ * (Proposed)	64.4	81.3	81.7	90.2	67.7	75.4	84.5	84.7	79.3	90.1	87.6	71.9	66.4	75.8	83.2	**78.9**

Figure 5 demonstrates the visual effects of the two models on the small object testing set. Compared to ResNet, our proposed ResNeSt+ model extracts more discriminative features to support the detector in recognizing more objects. From the figure (in red circles), it can be observed that the detector based on ResNet tends to overlook the edge features of the targets, especially when the input image is of poor clarity. This can result in the detector confusing the targets with the background or predicting inaccurate bounding boxes. Although the ResNeSt model pays attention to the edge information, it still cannot offer a precise detection result. In contrast, the proposed ResNeSt+ utilizes multi-scale object features more effectively through the use of multi-receptive field convolution and dynamic attention mechanisms, which further enhances the detector's sensitivity to small objects.



(a) ResNet101

(b) ResNeSt101

(c) ResNeSt+101

Figure 5. Visualization results comparison of ResNet, ResNeSt, and proposed ResNeSt+ on small objects detection using our DOTA testing set. The red circles indicate the differences among the results of the three models.

5.2. Feature Fusion

To further improve the detector's performance, a novel dynamic feature pyramid network called W-PAFPN is introduced to replace FPN. W-PAFPN incorporates several modules, including DCSA, Lateral Shortcuts, and ASFF. To verify the influence of each component on the model, we designed an ablation study on two distinct datasets, HRSC2016 and DOTA. The experimental results are summarized in Table 4. All models adopt ResNeSt+ 101 as backbone, and different combinations of neck structures are evaluated. FPN is used as the baseline neck structure. The evaluated metrics are mAP07 and mAP12 for the HRSC2016 dataset, and AP50 for the DOTA dataset. The results demonstrate that PAFPN outperforms FPN. Moreover, the incorporation of DCSA and lateral shortcut components further enhances the performance of PAFPN. Additionally, the inclusion of ASFF on top of these structures yields the best results on both datasets. Finally, the proposed W-PAFPN achieves remarkable performance, achieving 98.97% mAP12 on HRSC2016 dataset, and 81.97% AP50 on DOTA testing set.

Table 4. Ablation study of W-PAFPN on HRSC2016 and DOTA datasets.

Backbone	Neck					HRSC2016		DOTA
	FPN [21]	PAFPN [13]	DCSA	Lateral Shortcut	ASFF [22]	mAP07	mAP12	AP50
ResNeSt+ 101	√					90.61	97.95	81.27
		√				90.69	98.17	81.38
		√	√			90.77	98.46	81.63
		√	√	√		90.80	98.63	81.70
		√	√	√	√	**90.84**	**98.97**	**81.97**

To visually demonstrate the effectiveness of the designed feature fusion structure in capturing informative features, we extracted the feature maps from three different feature pyramid model and generated the saliency maps as displayed in Figure 6. The results verify that the proposed modules can extract more accurate details of the objects compared to FPN and PAFPN, which supports the detector in discriminating object categories and rotation angles. Additionally, the DCSA and ASFF modules alleviate the feature dilution problem during the process of feature propagation and fusion at different scales, further enhancing the feature representation capability of the model. These improvements contribute to the better performance of the proposed approach on the small object detection task.

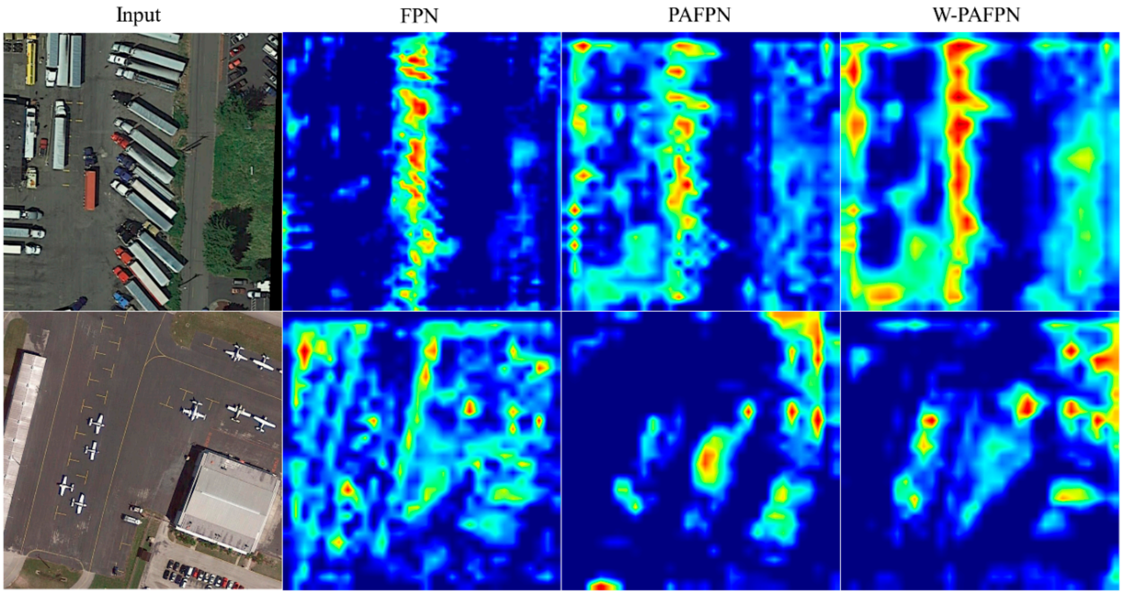
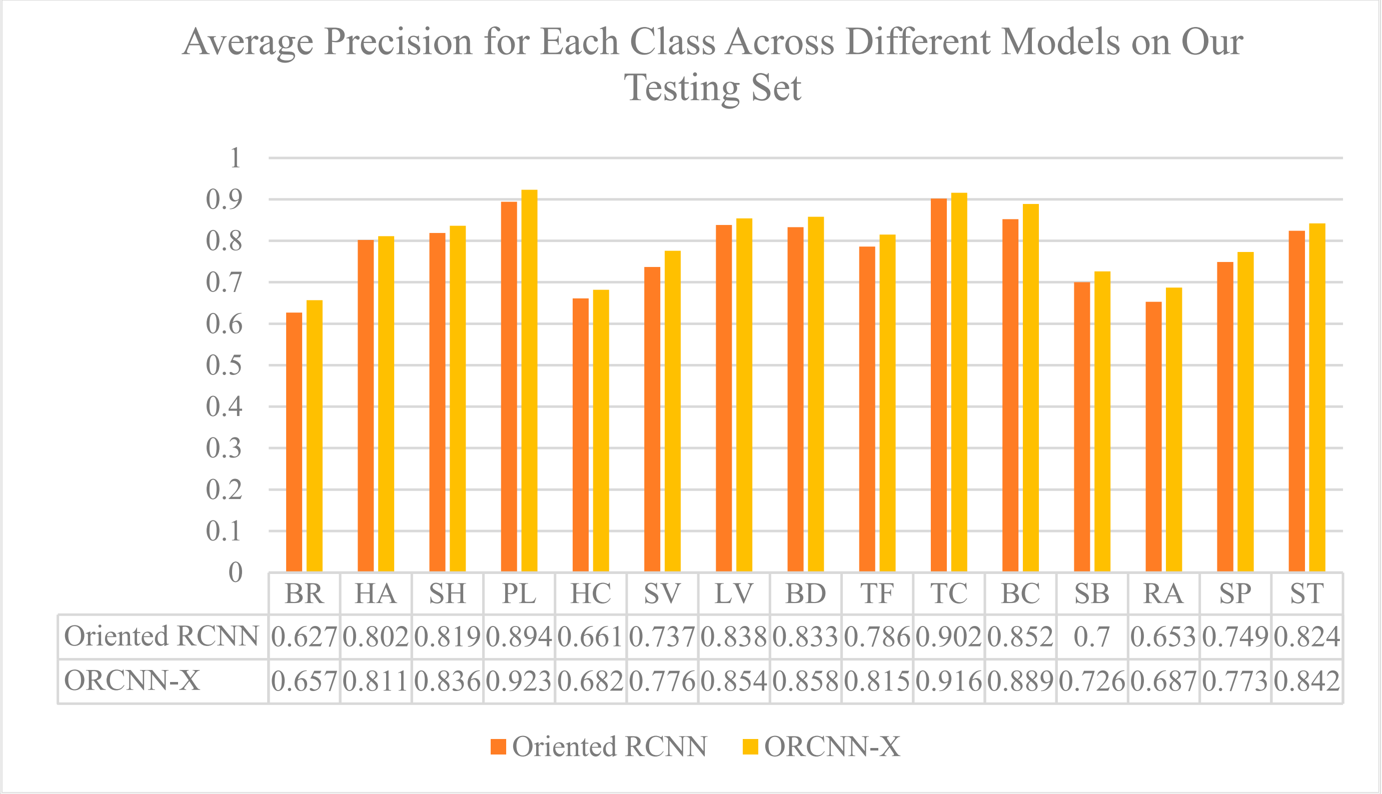


Figure 6. Comparison of visualized results among three feature pyramid modules (FPN, PAFPN and W-PAFPN).

Figure 7 shows the AP (average precision) for each class in our DOTA testing set using two different models, Oriented RCNN and ORCNN-X. The range of AP scores starts at 0 and goes up to 1, where a higher score means better performance in detecting objects of that class. Overall, ORCNN-X outperforms Oriented RCNN in terms of AP scores for all classes. Specifically, ORCNN-X achieved an average improvement of 2.5% in AP scores compared to Oriented RCNN, with the largest improvement seen in the "SV" class (3.9%). These results verify the effectiveness of the proposed ORCNN-X framework in detecting objects in aerial images, particularly for images with small, oriented objects.



	BR	HA	SH	PL	HC	SV	LV	BD	TF	TC	BC	SB	RA	SP	ST
Oriented RCNN	0.627	0.802	0.819	0.894	0.661	0.737	0.838	0.833	0.786	0.902	0.852	0.7	0.653	0.749	0.824
ORCNN-X	0.657	0.811	0.836	0.923	0.682	0.776	0.854	0.858	0.815	0.916	0.889	0.726	0.687	0.773	0.842

Figure 7. The average precision values for each class of objects in our DOTA dataset for two models: Oriented RCNN and ORCNN-X.

5.3. Comparison with Previous Work

In this section, a fair comparison is made between recently proposed oriented object detection models and our proposed model on two challenging datasets, including model accuracy and speed. In this comparison, all models were trained fairly, and then tested using the same parameter settings. Table 5 presents the testing results of different models on the DOTA testing dataset, including the AP for each of the 15 object categories in the dataset. The first row indicates the depth of the feature extractor and the names of the 15 categories. The ORCNN-X model is our proposed model. The results show that the ORCNN-X exceeds other models in terms of AP for most classes and achieves the highest AP50 score. In addition, we demonstrate the performance of the presented model using a multi-scale training approach in the table. The experimental results suggest that multi-scale training can considerably improve the detection performance of ORCNN-X, and the final AP50 score reached 81.97% with a feature extraction model depth of 101.

Table 5. Comparison of our method with other state-of-the-art techniques on the DOTA dataset. 'Dep.' is the depth of feature extraction network. 'ms' is multi-scale training.

Method	**Dep.**	**PL**	**BD**	**BR**	**TF**	**SV**	**LV**	**SH**	**TC**	**BC**	**ST**	**SB**	**RA**	**HA**	**SP**	**HC**	**AP50**
R3Det [25]	101	88.76	80.09	50.91	67.27	76.23	80.39	86.72	90.78	84.68	83.24	61.98	61.35	66.91	70.63	53.94	73.79
S^2ANet [26]	50	89.11	82.84	48.37	71.11	78.11	78.39	87.25	90.83	84.90	85.64	60.36	62.60	65.26	69.13	57.94	74.12
RoI Transformer [27]	101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
SCRDet [28]	101	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	87.94	86.86	65.02	66.68	66.25	68.24	65.21	72.61
Gliding Vertex [15,29]	101	89.64	85.00	52.26	77.34	73.01	73.14	85.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
Mask OBB [30]	50	89.61	85.09	51.85	72.90	75.28	73.23	85.57	90.37	82.08	85.05	55.73	68.39	71.61	69.87	66.33	74.86
ReDet [31]	50	88.79	82.64	53.97	74.00	78.13	84.06	88.04	90.89	87.78	85.75	61.76	60.39	75.96	68.07	63.59	76.25
Oriented RepPoints [32]	50	87.02	83.17	54.13	71.16	80.18	78.40	87.28	90.90	85.97	86.25	59.90	70.49	73.53	72.27	58.97	75.97
Oriented RCNN [15]	50	89.46	82.12	54.78	70.86	78.93	83.00	88.20	90.90	87.50	84.68	63.97	67.69	74.94	68.84	52.28	75.87
Oriented RCNN [15]	101	88.86	83.48	55.27	76.92	74.27	82.10	87.52	90.90	85.56	85.33	65.51	66.82	74.36	70.15	57.28	76.28
Ours																	
ORCNN-X	50	88.18	84.2	55.7	75	78.31	83.11	87.76	89.87	85.48	85.61	67.97	66.43	76.77	69.81	65.12	77.29
ORCNN-X	101	88.9	84.12	56.37	74.64	79.29	**83.94**	**88.26**	89.59	86.19	86.35	68.4	66.91	76.52	70.63	65.58	77.71
ORCNN-X (ms)	50	**90.58**	88.2	62.72	80	80.31	83.11	88.06	91.87	**87.48**	87.74	72.97	**74.43**	80.77	80.81	75.12	81.61
ORCNN-X (ms)	101	90.41	**88.74**	**63.4**	**80.51**	**80.67**	83.59	88.11	**92.35**	87.27	**88.39**	**73.88**	73.82	**81.57**	**81.56**	**75.23**	**81.97**

Table 6 presents the mAP07 and mAP12 results of different object detection methods on the HRSC2016 testing set. The methods compared include R3Det [25], S^2ANet [26], Rotated RPN [33], Oriented RCNN [30], ReDet [31], Oriented RepPoints [32], and the proposed ORCNN-X. The "Dep." column indicates the number of layers in the backbone model used in each method. The table shows that ORCNN-X achieves the best performance among all approaches, with mAP07 of 90.84% and mAP12 of 98.97% using a backbone network with 101 layers, which are respectively 0.34% and 1.37% higher than Oriented RCNN.

Table 6. Comparison of our method with other state-of-the-art techniques on the HRSC2016 dataset.

Method	**Dep.**	**mAP07**	**mAP12**
R3Det [25]	101	89.26	96.01
S^2ANet [26]	101	90.17	95.01
Rotated RPN [33]	101	79.08	85.64
ReDet [31]	101	90.46	97.63
Oriented RepPoints [32]	50	90.38	97.26
Point RCNN [34]	50	90.53	98.53
Oriented RCNN [15]	50	90.40	96.50
Oriented RCNN [15]	101	90.50	97.60
ORCNN-X	50	90.76	97.72
ORCNN-X	101	**90.84**	**98.97**

In Table 7, we compared AP50 and FPS (frames per second) of different models for oriented object detection on the DOTA testing dataset. The models include RoI Transformer, Faster RCNN-O, RetinaNet-O, S2Net, Oriented RCNN, and our proposed ORCNN-X. Among these models, ORCNN-X achieves the highest AP50 of 77.3%, outperforming other models by a significant margin. S2Net and Oriented RCNN follow closely with AP50s of 74.1% and 75.9%, respectively. RoI Transformer, Faster RCNN-O, and RetinaNet-O exhibit relatively lower AP50s of 74.6%, 69.1%, and 68.4%, respectively.

Table 7. Comparison of AP50 and FPS for Various Oriented Object Detection Models. '*' refers to the results from Aerial Detection.

	RoI Transformer *	**Faster RCNN-O ***	**RetinaNet-O ***	**S2Net [26]**	**Oriented RCNN [15]**	**ORCNN-X**
AP50	74.6	69.1	68.4	74.1	75.9	77.3
FPS	11.3	14.5	15.9	15.4	15.0	16.1

In terms of FPS, ORCNN-X achieves the highest speed of 16.1, followed by RetinaNet-O with an FPS of 15.9. Oriented RCNN exhibits a relatively lower FPS of 15.0, while RoI Transformer, Faster RCNN-O, and S2Net exhibit even lower FPSs of 11.3, 14.5, and 15.4, respectively. Overall, the results demonstrate that ORCNN-X achieves state-of-the-art performance on oriented object detection in aerial images while maintaining high inference speed. Furthermore, some visualization results of ORCNN-X on two datasets are shown in Figures 8 and 9.

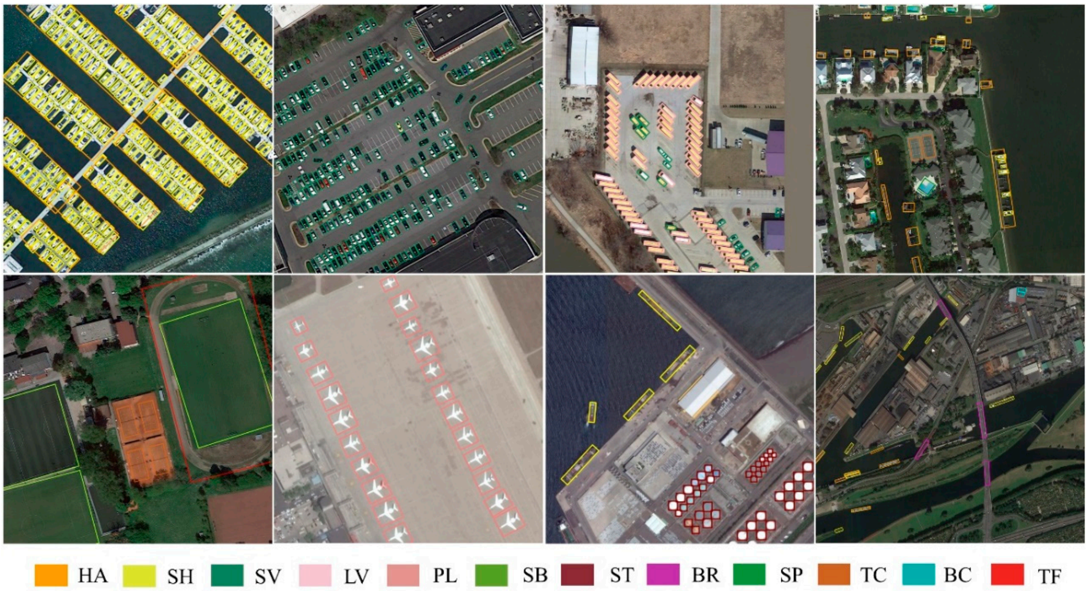


Figure 8. Visualization results on DOTA testing set using ORCNN-X with a backbone of depth 50.

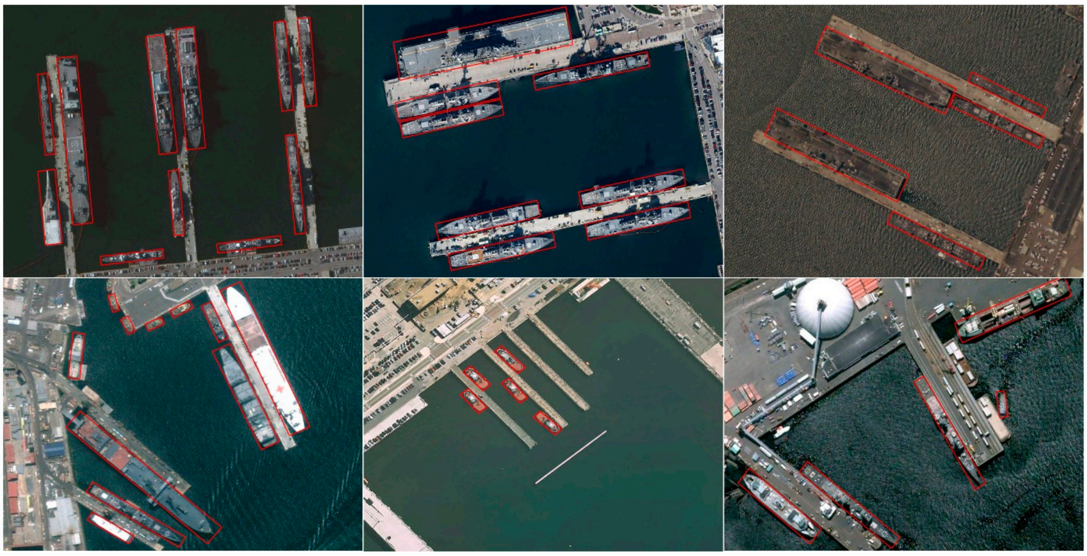


Figure 9. Visualization results on HRSC2016 testing set using ORCNN-X with a backbone of depth 50.

6. Conclusions

In this study, we proposed an innovative framework (ORCNN-X) for oriented small object detection in remote sensing images, which is based on Oriented RCNN. The framework addressed the unique challenges of detecting objects in aerial images, such as low resolution, complex backgrounds, and variations in scale and angle. It incorporated a multiscale feature extraction network with a dynamic attention module and an effective feature fusion mechanism to strengthen the model's perception ability and handle variations in scale and angle. We evaluated our proposed framework on two public datasets, DOTA and HRSC2016, and demonstrated its state-of-the-art performance. The experimental results showed that our model outperformed the baseline model by a significant margin, achieving 1.43% AP50 and 1.37% mAP12 improvement on DOTA and HRSC2016 datasets, respectively. Our proposed framework provides a promising approach for oriented small object detection in aerial images. In addition, it has potential and practical significance in urban planning, environmental monitoring, and disaster assessment.

Since this research employed multiple receptive fields to capture fine-grained details and contextual information for achieving high-precision small object detection, the sizes of these receptive fields were pre-defined and fixed. Therefore, it may lead to variations in the effectiveness of feature extraction for different data. Moreover, the non-dynamic receptive field selection method will also affect the model's inference speed. In future work, more effort should be put into extracting multi-scale contextual information. Further improvements in the accuracy and speed of small object detection models will be sought through in-depth exploration of algorithm design for dynamic receptive fields selection. The use of transfer learning will also be explored to further improve the model's generalization ability on different datasets. Additionally, the investigation of other data augmentation techniques, such as geometric transformations, will be conducted to boost the model's robustness to variations in orientation and scale.

Author Contributions: Conceptualization, Y.L. and H.W.; methodology, Y.L.; validation, L.M.D.; investigation, H.-K.S.; writing—original draft preparation, Y.L.; writing—review and editing, H.W.; supervision, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (2021R1F1A1046339) and by a grant(20212020900150) from "Development and Demonstration of Technology for Customers Bigdata-based Energy Management in the Field of Heat Supply Chain" funded by Ministry of Trade, Industry and Energy of Korean government.

Data Availability Statement: Not applicable.

Conflicts of Interest: There is no conflict of interest.

References

1. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
2. Hou, L.; Lu, K.; Xue, J. Refined one-stage oriented object detection method for remote sensing images. *IEEE Trans. Image Process.* **2022**, *31*, 1545–1558. [CrossRef] [PubMed]
3. Li, J.; Huang, X.; Tu, L.; Zhang, T.; Wang, L. A review of building detection from very high resolution optical remote sensing images. *GIScience Remote Sens.* **2022**, *59*, 1199–1225. [CrossRef]
4. Li, Y.; Wang, H.; Dang, L.M.; Song, H.K.; Moon, H. Attention-guided multiscale neural network for defect detection in sewer pipelines. *Comput.-Aided Civ. Infrastruct. Eng.* **2023**. [CrossRef]
5. Nguyen, T.N.; Nguyen-Xuan, H.; Lee, J. A novel data-driven nonlinear solver for solid mechanics using time series forecasting. *Finite Elem. Anal. Des.* **2020**, *171*, 103377. [CrossRef]
6. Ma, W.; Li, N.; Zhu, H.; Jiao, L.; Tang, X.; Guo, Y.; Hou, B. Feature split–merge–enhancement network for remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
7. Peng, L.; Zhang, T.; Huang, S.; Pu, T.; Liu, Y.; Lv, Y.; Zheng, Y.; Peng, Z. Infrared small-target detection based on multi-directional multi-scale high-boost response. *Opt. Rev.* **2019**, *26*, 568–582. [CrossRef]

8. Rabbi, J.; Ray, N.; Schubert, M.; Chowdhury, S.; Chao, D. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. *Remote Sens.* **2020**, *12*, 1432. [CrossRef]
9. Guo, H.; Bai, H.; Yuan, Y.; Qin, W. Fully deformable convolutional network for ship detection in remote sensing imagery. *Remote Sens.* **2022**, *14*, 1850. [CrossRef]
10. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R. Resnest: Split-attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 2736–2746.
11. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
12. Zhang, Q.; Zhang, H.; Lu, X.; Han, X. Anchor-free small object detection algorithm based on multi-scale feature fusion. In Proceedings of the 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Chengdu, China, 19–21 August 2022; pp. 370–374.
13. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
14. Wang, X.; Wang, C. MSFM: Multi-Scale Fusion Module for Object Detection. *arXiv* **2020**.
15. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 3520–3529.
16. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
17. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part V 13, pp. 740–755.
18. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; Technical Report TR-2009; University of Toronto: Toronto, ON, Canada, 2009.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part IV 14. pp. 630–645.
20. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
21. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. *arXiv* **2019**, arXiv:1911.09516.
23. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
24. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
25. Pan, X.; Ren, Y.; Sheng, K.; Dong, W.; Yuan, H.; Guo, X.; Ma, C.; Xu, C. Dynamic refinement network for oriented and densely packed object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11207–11216.
26. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
27. Ding, J.; Xue, N.; Long, Y.; Xia, G.-S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2849–2858.
28. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. *IEEE/CVF Int. Conf. Comput. Vis.* **2019**, *27*, 8232–8241.
29. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.-S.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 1452–1459. [CrossRef] [PubMed]
30. Wang, J.; Ding, J.; Guo, H.; Cheng, W.; Pan, T.; Yang, W. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sens.* **2019**, *11*, 2930. [CrossRef]
31. Han, J.; Ding, J.; Xue, N.; Xia, G.-S. Redet: A rotation-equivariant detector for aerial object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2786–2795.
32. 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, 19–24 June 2022; pp. 1829–1838.

33. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
34. Zhou, Q.; Yu, C. Point rcnn: An angle-free framework for rotated object detection. *Remote Sens.* **2022**, *14*, 2605. [CrossRef]

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