Article

MFCANet: Multiscale Feature Context Aggregation Network for Oriented Object Detection in Remote-Sensing Images

Honghui Jiang , Tingting Luo , Hu Peng , Guozheng Zhang *

Posted Date: 11 December 2023

doi: 10.20944/preprints202312.0665.v1

Keywords: object detection; complex backgrounds; remote sensing images; context information; multiscale feature fusion

Article

MFCANet: Multiscale Feature Context Aggregation Network for Oriented Object Detection in Remote-Sensing Images

Honghui Jiang [1,†,‡], Tingting Luo [2,‡], Hu Peng [3,‡] and Guozheng Zhang [1,*,†,‡]

[1] Anhui Technical College of Mechanical and Electrical Engineering, Wuhu 241003; 0120200036@ahcme.edu.cn
[2] State Gride Wuhu Power Supply Company, Wuhu 230061; luott@aliyun.com
[3] Hefei University of Technology, Hefei 230009; hpeng@hfut.edu.cn
[*] Correspondence: jenuel@163.com
[†] Current address:16 Wenjin West Road, Yijiang District, Wuhu City, Anhui Province.
[‡] These authors contributed equally to this work.

Abstract: Detecting rotated objects in remote sensing images poses a substantial challenge. These images usually cover a wide field of view, containing diverse and complex backgrounds with variously sized ground objects densely distributed. Therefore, identifying objects of interest in remote-sensing images is challenging. While integrating Convolutional Neural Networks (CNNs) and Transformer networks has exhibited progress in detecting rotated objects, there is still scope for enhancing feature extraction and fusion. To address this issue, we propose a feature extraction module, a feature context aggregation module, and a multi-scale feature fusion module. Initially, we substitute the Spatial Pyramid Pooling Bottleneck (SPPFBottleneck) with a new module aimed at extracting multi-scale features, thereby enhancing the detection of small objects in complex backgrounds. Next, we develop a novel module for the multi-scale fusion of contextual information within feature maps, extracting valuable information. Finally, we combine the original features with the fused ones to prevent the loss of specific features in the fusion process. We refer to our newly proposed model as the "Multi-scale Feature Context Aggregation Network" (MFCANet). We assess our approach on three challenging remote sensing datasets: MAR20, SRSDD, and HRSC. Comprehensive experimental results show that our method surpasses baseline models by 2.13%, 10.28%, and 1.46% mAP on the MAR20, SRSDD, and HRSC datasets, respectively.

Keywords: object detection; complex backgrounds; remote sensing images; context information; multiscale feature fusion

1. Introduction

Object detection in remote sensing imagery [1–5] holds a significant role in identifying and pinpointing the exact locations of objects within an image. Its diverse applications span environmental monitoring, military applications, national security, transportation, forestry, and oil and gas activity detection. Remote-sensing images originate from varied sources like aerial, satellite, and unmanned aerial vehicle platforms. However, the complexities in remote sensing imagery—comprising intricate backgrounds, arbitrary object orientations, varying object densities, and differences in object size ratios—pose substantial challenges for small object detection. Rotating bounding boxes, contrary to conventional horizontal ones, notably decrease background overlap and provide more accurate object boundary delineation. Consequently, there's an increasing need for research in rotating object detection within remote sensing imagery.

In remote sensing images, the same object can vary significantly in appearance based on the background, resulting in notable intra-class variability. This is especially prominent in fine-grained remote sensing images where object class distinctions are less apparent. Leveraging feature information to the fullest becomes crucial for effective detection in such instances. Multi-level feature pyramid networks are commonly used to address the challenge of object scale variations in remote-sensing images. The Feature Pyramid Network (FPN) [19] framework comprises higher-level feature maps with richer semantic information but smaller scales, making them less efficient in detecting small objects. On the contrary, lower-level feature maps have larger scales but lack distinctive object representations. To bridge this discrepancy, FPN incorporates a top-down lateral connection structure, facilitating semantic information flow from higher to lower-level features, thereby enabling object detection across various scales. As a result, extensive research is dedicated to further enhancing FPN to better suit the demands of object detection in remote sensing images.

The DCFPN [25] employs densely connected multi-path dilated layers to encompass objects of diverse sizes in remote sensing environments, enabling dense and accurate extraction of multi-scale information, thereby boosting the detection prowess for varying-sized objects. LFPN [26] accounts for both low-frequency and high-frequency features, utilizing trainable Laplacian operators to extract high-frequency object features from Laplacian pathways. It introduces an attention mechanism within the feature pyramid network to emphasize more distinct multi-scale object features. SPH-YOLOv5 [12] integrates an attention mechanism into FPN, aiding in capturing semantic information among features to highlight crucial spatial features and suppress redundant ones. Info-FPN [23] introduces a PixelShuffle-based lateral connection module (PSM) to fully preserve channel information within the feature pyramid. Simultaneously, to mitigate confusion arising from feature misalignment, it proposes a feature alignment module (FAM). FAM uses template matching and learns feature offsets during feature fusion to achieve alignment. However, existing FPN-based methods often neglect the drawbacks of the feature pyramid network structure, inadequately exploiting original feature information and encountering performance issues due to attention mechanisms. These limitations lead to reduced feature representation capacity, particularly noticeable when dealing with objects exhibiting significant scale variations in remote sensing images.

In summary, The ***The existing challenges*** are as follows: (1) Most SPPBottleneck modules lack the capability to capture both coarse spatial information and fine-grained feature details, constraining further detection of the target of interest. (2) The mutual fusion among features at different layers within the feature pyramid is not thorough enough, leaving room for further improvement in the extraction and enhancement of the fused feature information. Additionally, a majority of feature pyramid networks do not fully exploit original features, yet these original features play a crucial role in reinforcing feature fusion, enhancing residual functions, and ensuring stable gradient propagation during backpropagation.

In this paper, we propose robust solutions to overcome the mentioned challenges. Leveraging the RTMDet model as our baseline, we substitute the SPPBottleneck module with a Focused Feature Context Aggregation Module (FFCA Module). This module effectively captures coarse spatial information and fine-grained feature details at different scales, gathering intricate details across various target scales, thereby enhancing the model's perception of the targets. Additionally, we design a multi-scale feature fusion feature pyramid to integrate spatial context within feature maps, maximizing the amalgamation of feature information across layers, and consequently enhancing the model's representational capacity. These solutions seamlessly integrate into object detectors, enhancing detection performance without increasing training complexities. To summarize, our contributions are outlined below:

- In the backbone network, we utilize a multi-level feature fusion mechanism to acquire features of different scales. Subsequently, context information is selectively extracted from local to global levels at varying granularities, resulting in feature maps equal in size to the input features. Finally, these feature maps are injected into the original features to obtain relevant information about the objects of interest without altering their size.
- We design a feature aggregation module that assigns varying attention across multiple dimensions to the fused feature map information, thereby improving performance in capturing rich contextual information and consequently enhancing pixel-level attention towards objects of interest.

- Within the feature pyramid, we efficiently harness original feature information to process multi-scale features more effectively by introducing a multi-scale fusion pyramid network. This network connects original features and fused features while shortening the information transmission paths, extending from large-scale features to fused small-scale features, and enabling the module to optimally utilize features at each stage.
- We introduce a novel object network and conduct extensive experiments on three challenging datasets: MAR20, SRSDD, and HRSC, affirming the effectiveness of our approach. The experimental results demonstrate outstanding performance.

2. Releated Work

2.1. Object Detection in General Scenarios

Over the past decade, computer vision technology has rapidly advanced due to the continual iteration of large-scale annotated datasets, which has further propelled advancements in object detection tasks. These methodologies can be broadly classified into two major categories: those based on convolutional neural networks and those leveraging attention mechanisms.

Within CNN models, there exist both single-stage detection models (such as SSD [15], RetinaNet [16], R^2ANet [17], the YOLO series [10–14], RTMDet [18], among others) and two-stage models (R-CNN [6], Fast R-CNN [7], Faster R-CNN [8], R-FCN [9], and so forth). These models have shown significant achievements. However, downsampling operations during processing in CNN-based models may render extremely small targets undetectable. To address the challenge of detecting small targets, the introduction of FPN and its variants [27,28] aimed to improve their detection. However, this introduction brought new challenges, including increased computational complexity, the necessity for parameter adjustments within FPN, and the potential for incomplete feature map matching due to introduced cross-level connections, resulting in inaccurate predictions at boundaries. Some researchers have optimized feature spatial pooling modules and achieved certain results [13,69–71]. However, they have not fully considered the impact of feature context information on detection results.

Moreover, some researchers have introduced attention mechanisms into CNNs [12,29–31], which to some extent enhance the accuracy of object detection. Methods combining attention with convolution capture both static and dynamic contextual information in images, possessing self-attention learning capabilities while incorporating contextual information. Furthermore, certain researchers have transformed temporal information into the frequency domain through techniques like wavelet and Fourier transforms [21,26], subsequently extracting frequency domain features that have yielded promising results. Various approaches have been proposed from different perspectives, involving the design of a series of channel weight-solving methods to adaptively learn the importance of each channel and weight each channel feature map [32–34], all of which have demonstrated favorable results.

In recent years, Transformer-based models [35–37,67] have shown promising results in the field of object detection. The Vision Transformer (ViT) [35]demonstrated that Transformers can be applied to computer vision with minimal modifications and achieve excellent performance. The DETR [36] model provides end-to-end object detection without the need for post-processing steps like non-maximum suppression (NMS) or prior knowledge and constraints such as anchors. It can be parallelized and achieves results comparable to Faster R-CNN, with better performance on large objects. However, DETR, which utilizes CNN for feature extraction and dimension reduction before applying Transformers, still faces challenges in small object detection. To build a comprehensive Transformer-based model, the Swin Transformer [37] adopts a strategy inspired by the favorable properties of CNN networks. It divides the image into patches and further subdivides them into multiple windows. Within each window, it calculates self-attention among patches and then computes global self-attention through a sliding window mechanism. This approach overcomes the memory and computational limitations of Transformers when dealing with large images. Additionally, the Swin-Transformer exhibits strong scalability and performs well on large-scale datasets. Nevertheless, it still requires relatively high computational costs compared to traditional neural networks and has certain limitations related to input image size, which needs adjustments based on window size and model architecture.

2.2. Object Detection in Remote Sensing Scenarios

Deep learning methods are presently extensively utilized for object detection in remote sensing imagery. A variety of CNN-based approaches for remote sensing object detection have emerged, showing promising results.

To address multiscale detection challenges arising from different object sizes in remote sensing imagery, mSODANet [38] utilizes parallel dilated convolutions to investigate a hierarchical dilation network. This network facilitates contextual information learning for diverse object types across various scales and fields of view. The Super-Yolo model [39] integrates multimodal data and incorporates auxiliary super-resolution learning to address high-resolution object detection of multiscale objects, balancing detection accuracy and computational cost. MFAF [40] proposes a method for adaptive multiscale feature fusion, using multiscale feature integration modules and spatial attention weight modules to create a feature fusion module, facilitating adaptable fusion of multiscale features. MDCT [28] introduces a single-stage object detection model that relies on multi-kernel dilated convolution blocks and Transformer blocks, improving intrinsic and neighboring spatial features of small objects. ANSDA [41] utilizes NASFPN for feature extraction and incorporates context enhancement modules and channel attention modules, enhancing the feature extraction capabilities for shallow-level features and small object semantics. ORCNN-X [27] incorporates a dynamic attention module and an efficient feature fusion mechanism into a multiscale feature extraction network, improving the model's perception capabilities to handle scale and orientation variations. DCFPN [25] develops a Dense Context Feature Pyramid Network (DCFPN) and uses Gaussian loss for rotation object detection. It utilizes dense multi-path dilated layers to accurately extract multiscale information and addresses boundary regression discontinuity via the Gaussian loss function, resulting in favorable performance. ESRTMDet [42] designs a lightweight embedded feature map super-resolution module and embeds it into PAFPN to enhance and magnify the backbone's output features, facilitating small object detection. HFAN [43] introduces an adjacent feature alignment module to integrate adjacent features in the feature map using a non-parametric alignment strategy, improving detection performance. YOLO-DCTI [44] addresses the challenge of globally modeling pixel-level information for small objects by designing a context transformer framework and embedding it into the detection head. SPH-Yolo [12] incorporates the Swin-Transformer into PAFPN to more effectively detect objects of various scales.

Additionally, researchers are exploring anchor-free mechanisms as alternatives to anchor-based methods for rotation object detection. AOPG [45] generates coarse-oriented boxes without anchors using a coarse localization module and further refines them into high-quality-oriented proposals. FCOS [46] introduces a fully convolutional single-stage object detector that addresses object detection through per-pixel prediction, entirely avoiding the complex computations associated with predefined anchor boxes. CLU [47] presents an unsupervised object detection approach, leveraging self-supervised models' characteristics to identify objects without supervision. H2RBOX [48] utilizes weakly supervised training with horizontal bounding box annotations for rotation box object detection. Specifically, it employs weakly supervised learning and self-supervised learning to predict object angles by exploiting consistency between two different views, yielding promising outcomes.

In remote sensing images, both sparse and dense small objects hold significant proportions, posing substantial challenges for feature extraction networks. Convolutional Neural Networks (CNNs) typically capture local information well due to their ability to extract features with translational invariance. However, they often lack in extracting contextual information from these features.

Conversely, attention mechanisms are adept at global modeling, enabling the acquisition of contextual information for feature maps. Thus, combining these two approaches can harness their individual strengths and produce features more suitable for detection. Based on these insights, we propose an enhanced PAFPN-based single-stage object detection model, leveraging the groundwork of RTMDet. We aim for our work to contribute to the advancement of object detection in remote sensing imagery.

3. Methodology

3.1. Basic Rotated Detection Method as Baseline

Previous approaches overlooked the detection of rotated bounding boxes, relying commonly on horizontal bounding boxes for object delineation [49,50]. However, remote-sensing images frequently contain numerous objects with complex backgrounds. The use of traditional horizontal bounding boxes includes background information that hinders precise object localization. In contrast, rotated bounding boxes facilitate precise object localization while minimizing background interference. Moreover, rotated bounding boxes have minimal overlap, ensuring distinct object delineation within them. Hence, it's crucial to explore and implement more precise representations of rotated bounding boxes for object detection in remote sensing images. The typical definition of rotated bounding boxes (RBB) is as follows:

$$(X, Y, W, H, \theta), \tag{1}$$

Here, $\theta \in [-\pi/2, \pi/2]$ represents the clockwise rotation angle from the image's X-direction to the bounding box's X-direction in its relative coordinate system. We use the long-edge-based format [51], requiring the width w to be greater than the height h. We utilize the one-stage rotation object detector RTMDet [18] to detect sparse and dense objects with complex backgrounds in remote sensing images. RTMDet is an improved version derived from YOLOX [52], having a comparable overall macro-architecture to the YOLO series. The complete model structure is depicted in Figure 1.

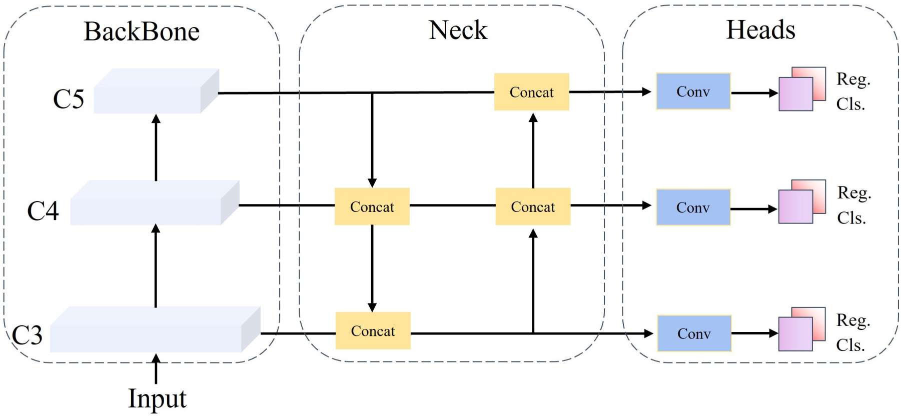

Figure 1. The fundamental macro-architecture comprises three segments: the backbone, neck, and heads. Input images are processed through the backbone network to extract features, resulting in three sets of feature maps at varying scales. The neck section utilizes PAFPN for the bidirectional merging of these multi-scale feature maps before passing them to the head. In the head component, predictions encompass various aspects, including object category counts, boundary regression, and detected target rotation angles, derived from the input features.

Specifically, RTMDet comprises CSPNeXt, CSPNeXtPAFPN, and SepBNHead, which share convolutional weights while performing batch normalization independently. Furthermore, it takes cues from ConvNeXt [53] and RepLKNet [54], improving feature extraction by incorporating large kernel convolutions within the Basic Block. Additionally, the authors adopt a dynamic SimOTA approach for detecting rotated objects, employing DistanceAnglePointCoder for Bbox encoding and

doi:10.20944/preprints202312.0665.v1

decoding. RTMDet introduces a Dynamic Soft Label Assigner to execute a dynamic label-matching strategy. This method primarily employs prior position information loss, sample regression loss, and sample classification loss, incorporating soft processing on these losses to fine-tune parameters for the optimal dynamic matching effect. Upon summing these three losses to derive the final cost matrix, SimOTA is utilized to ascertain the quantity of matched samples for each ground truth (GT) and thereby establish the final samples.

3.2. Focused Feature Context Aggregation Module

The FFCA Module is mathematically described as follows:

$$Y = f(X) \odot (\sum_{i=0}^{3} h_i(X^{'}) \odot g_i(X^{''})) \tag{2}$$

Spatial feature pooling and its variations [69–71] are commonly employed in the backbone network for extracting multi-scale features in target detection. However, these methods have not adequately addressed the aggregation of contextual features relevant to the specific feature of interest, which is crucial for target localization and regression prediction. To address this, we have devised a novel method to acquire semantic information related to focal features, which we term the Focal Feature Context Aggregation Module. The specific workflow is depicted in Figure 2. This module initially adjusts the channel dimensions of the input features through convolution and then divides the features into three parts: A1, A2, and A3. Subsequently, A3 is further segmented into four channels: A31, A32, A33, and A34. Next, A2 undergoes convolution using kernels of sizes 1, 3, 5, and 7, followed by averaging operations. The resulting outcomes are tensor-operated with A31, A32, A33, and A34, followed by summation and activation functions. The resultant is multiplied with A1 and then refined using a 1x1 convolution to extract specific target features along with their corresponding contextual information.

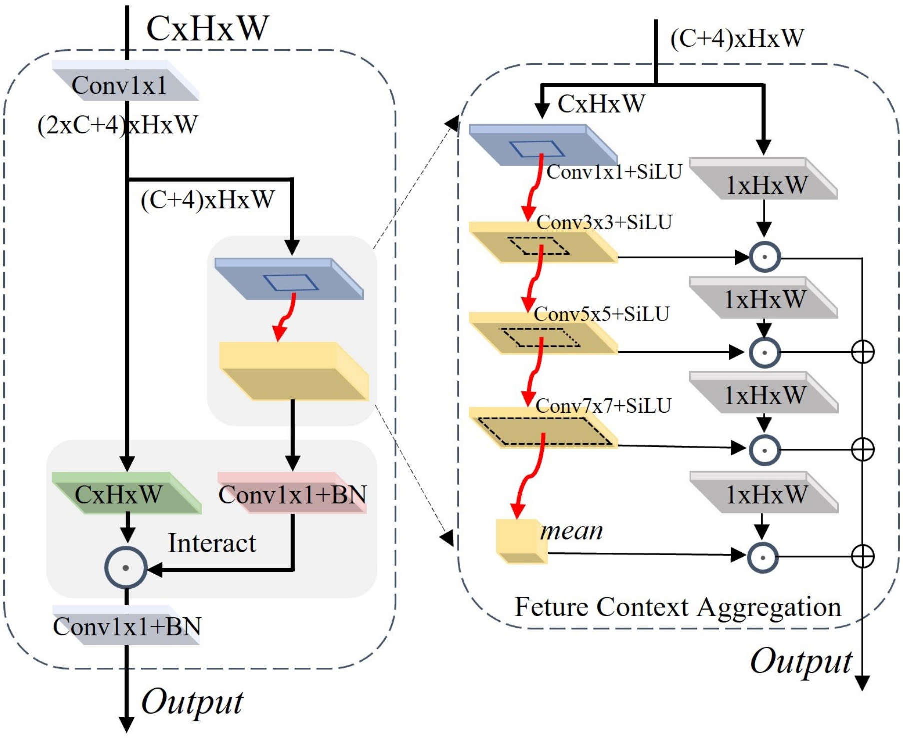

Figure 2. The FFCA module is specifically designed to acquire multi-scale focal feature context information. C, H, and W respectively denote the channel, height, and width of the feature map. ’mean’ represents the tensor’s mean operation, ⊙ denotes tensor multiplication, and ⊕ signifies tensor addition.

In this equation, $f(\cdot)$ represents the focusing function used to extract results conforming to the Feature Context Aggregation from the original features. h_i is the feature context extraction function for the i^{th} layer, $g_i(\cdot)$ represents the gating function for the j^{th} layer, $\odot$ denoting the tensor multiplication operation, and $\sum$ representing the tensor summation operation. The variables X, $X^{'}$, and $X^{''}$ respectively represent the sections of the original features used for querying, extracting context information, and gate selection.

3.3. Multiscale Feature Fusion Module

Objects in remote sensing images often exhibit significant size variations, requiring neural networks' feature maps to encompass diverse receptive field scales for comprehensive object feature extraction. PAFPN [20] initially extracts feature maps at various scales through a bottom-up approach and subsequently performs upsampling using a top-down structure. It then integrates the downsampled and upsampled outcomes via lateral connections, producing feature maps at higher pyramid levels to incorporate enriched semantic information. However, the PAFPN model encounters challenges in detecting objects with complex backgrounds. Object features with complex backgrounds within this model are confined to small regions, potentially leading to their oversight or misclassification during the image partitioning into multiple scales using feature pyramids. Additionally, multiple fusions can diminish vital features, reducing feature map clarity, and thus hindering effective object detection. Optimizing and adjusting the PAFPN model's feature fusion mechanism becomes essential to improve its performance.

Figure 3 illustrates our proposed model architecture. The model incorporates two levels of lateral skip connections, merging the original feature information with the intermediate and final results. This creates direct connections between the original features and the fused feature maps, effectively utilizing the original feature maps' characteristics to improve model performance. Furthermore, the integrated residual structures maintain essential information throughout the fusion process, preventing the loss of crucial details and mitigating gradient vanishing issues. As this approach relies on feature fusion, the combination does not inherently increase computational costs. The entire process is outlined as follows.

$$\begin{cases} P_3 = f_3(g_3(g_4(C_4, C_5), C_3) + C_3) + C_3 \\ P_4 = f_4(g_4(f_4(g_4(C_4, C_5)) + C_4, P_3 - C_3)) + C_4 \\ P_5 = f_5(g_5(C_5, P_4 - C_4) + C_5) + C_5 \end{cases} \tag{3}$$

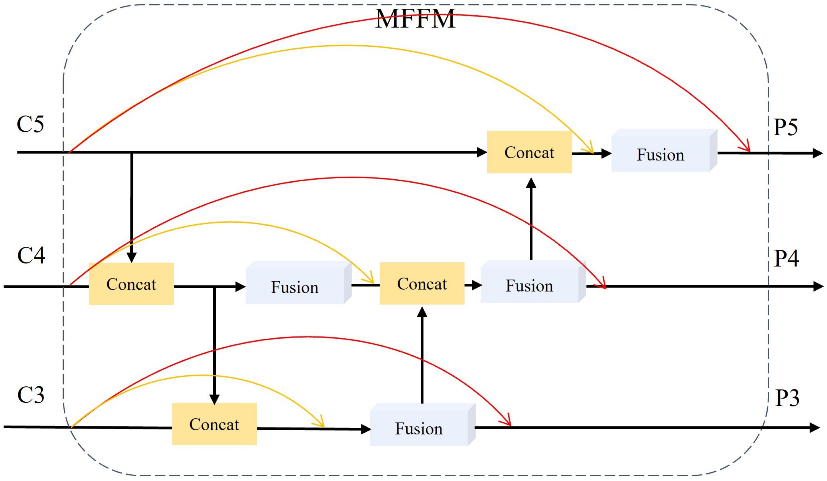

Figure 3. The Multiscale Feature Fusion Network integrates intermediate and final outputs from PAFPN with the original output features using a red solid line as a residual connection. The fusion of intermediate-level information with deep-layer information is denoted by a deep yellow dashed line, employing a 1x1 convolutional kernel for channel dimension adjustment. The Fusion module, inherent in the baseline, is used for merging the concatenated features.

In this context, C3, C4, and C5 represent the features extracted by the backbone network, while P3, P4, and P5 correspond to the fused feature outcomes. The function $f_i(\cdot)$ denotes the fusion of the merged results, and function $g_i(\cdot)$ signifies channel-wise concatenation. The subscript i denotes the respective layers, ranging in values from 3 to 5.

3.4. Feature Context Information Enhancement Module

Conventional convolutional methods use a fixed kernel independent of input samples. Dynamic convolution integrates attention mechanisms and deformable convolutions to improve the model's perception of temporal and spatial information. ODConv [72], utilizing a parallel strategy, incorporates a multi-dimensional attention mechanism, allowing flexible attention learning across four dimensions of convolutional kernel space. Illustrated in Figure 4, ODConv assigns varying attention values to convolutional parameters in spatial positions, diverse input channels, convolution filters, and 'n' overall convolutional kernels. These attention types complement each other, enabling different convolutional operations based on position, channel, filter, and kernel, thereby enhancing contextual information capture. Consequently, ODConv significantly improves convolution operations' feature extraction capability. Building upon ODConv's strong performance, we developed the ODCLayer depicted in Figure 4 and applied it in the 'neck' section of our model for feature fusion within PAFPN.

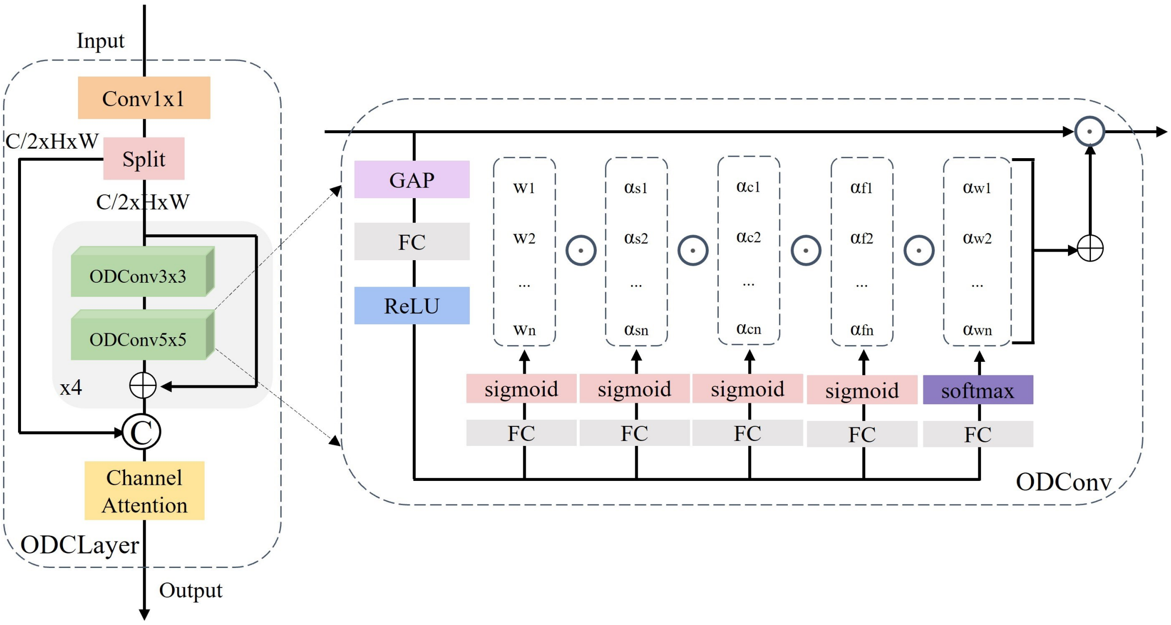

Figure 4. The ODCLayer initiates by integrating input features through a 1-sized convolutional kernel. These integrated features are split into two segments. Successively, ODConv modules with kernel sizes of 3 and 5 are concatenated in series while preserving residual connections. This sequence repeats four times. Later, the other segment concatenates along the channel dimension. Finally, channel attention mechanisms assign diverse weights to distinct channels.

ODCLayer is mathematically described as follows:

$$Y = h(f(c_1(X))\mathbb{C}c_2(X)) \tag{4}$$

The equation contains notations: $h(\cdot)$ for channel attention weighting, $f(\cdot)$ representing ODConv operation with four layers, each with a depth of 4, utilizing convolutional kernels of sizes 3 and 5. $c_1(\cdot)$ and $c_2(\cdot)$ represent ordinary convolution with a kernel size of 1, and $\mathbb{C}$ denotes tensor concatenation across the channel dimension.

3.5. MFCANet

Figure 5 illustrates the overall architecture of our proposed multi-scale feature context aggregation network, constructed upon RTMDet. It consists of a feature extraction module, a feature pyramid module, and prediction heads. The backbone network extracts features at three different scales for handling objects of diverse sizes in object detection. We replaced the SPPFBottleneck with the FFCA Module to enhance feature extraction at varied scales. Additionally, we integrated original and output features using PAFPN. A new ODCLayer was designed, employing ODConv with various convolutional kernels to capture information representing real features at diverse scales.

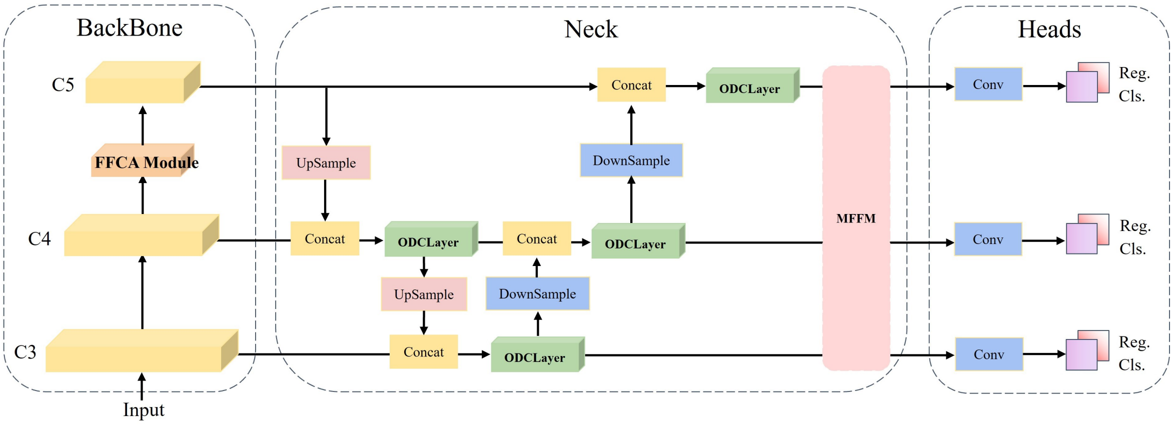

Figure 5. The architectural components of MFCANet consist of crucial modules. Initially, we employ the FFCA module to replace the SPPFBottleneck in the backbone network, capturing multi-scale feature context information related to focal targets. Subsequently, utilizing the MFFM module enhances the utilization of original features, minimizing the loss of specific feature information during fusion processes. Finally, leveraging our designed ODCLayer maximizes the enhancement of cross-layer feature integration and extraction, considering information across various feature dimensions. Our improvements notably enhance the model's detection capability within the context of remote sensing applications.

4. Experiments

This section assesses the effectiveness of our proposed model by training and testing it on three widely used datasets: MAR20, SRSDD, and HRSC. We present a comprehensive overview of our experiments, covering experimental design, parameter configurations, comparisons with state-of-the-art (SOTA) models, and experimental outcomes. Additionally, we conducted an ablation study on the MAR20 dataset to demonstrate the effectiveness of each module. Our software environment includes CUDA 11.8, Python 3.8.10, PyTorch 2.0, mmdetection3.1.0, and mmrotate1.x. The hardware setup includes an Intel(R) Xeon(R) Platinum 8350C @ 2.60GHz, NVIDIA GeForce RTX 3090, and 80GB of memory. Configuration files follow the default settings of mmrotate, with a linear decay in learning rate for the first 1000 iterations, followed by a cosine decay at $max_epoch/2$. All experiments are assessed using DotaMetric. The AdamW optimizer is utilized with a base learning rate of 0.00025, a momentum of 0.9, and a weight decay of 0.05 for all experiments. Random seeds for both the numpy library and tensors are set to 42.

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

The ***MAR20*** [64] dataset stands as the largest publicly available dataset for recognizing military aircraft targets in remote sensing images. It includes 3842 images featuring 20 distinct military aircraft models, totaling 22341 instances. Most images have a resolution of 800×800 pixels. These instances were gathered from 60 military airfields situated in countries like the United States, Russia, and others, using Google Earth imagery. The MAR20 dataset comprises a specific array of 20 aircraft models, including six Russian aircraft such as the SU-35 fighter, TU-160 bomber, TU-22 bomber, TU-95 bomber, SU-34 fighter-bomber, and SU-24 fighter bomber. The remaining 14 models belong to the United States, including the C-130 transport plane, C-17 transport plane, C-5 transport plane, F16 fighter, E-3 AWACS (Airborne Warning and Control System) aircraft, B-52 bomber, P-3C anti-submarine warfare aircraft, B-1B bomber, E-8 Joint Surveillance Target Attack Radar System (Joint STARS) aircraft, F-15 fighter, KC-135 aerial refueling aircraft, F-22 fighter, F/A-18 fighter-attack aircraft, and KC-10 aerial refueling aircraft. These aircraft model types are labeled A1 to A20. The training set contains 1331 images and 7870 instances, while the test set includes 2511 images and 14471 instances.

The ***SRSDD*** [73] dataset is a high-resolution Synthetic Aperture Radar (SAR) dataset designed for ship detection, characterized by complex backgrounds and notable interference. The original SAR images are in spotlight mode, displaying HH and VV polarization. Annotations within the dataset employ rotated bounding boxes, specifically suitable for detecting objects within rotational frames. It consists of 666 smaller patches extracted from 30 China High-Resolution Gaofen-3 SAR panoramic images at a 1-meter resolution, with each patch containing 1024×1024 pixels. The dataset includes 2884 ship instances distributed among six distinct categories: Container, Dredger, Ore-oil, LawEnforce, Cell-Container, and Fishing, containing 89, 263, 166, 25, 2053, and 288 instances, respectively. Most images in the dataset capture coastal areas, featuring intricate background interferences, which pose substantial challenges for detection.

HRSC [65] is a widely utilized benchmark for arbitrary-oriented object detection. It consists of 1061 images ranging in size from 300×300 to 1500×900. The training set comprises 436 images, the validation set has 181 images, and the rest are designated for testing. Regarding evaluation metrics, we utilize COCO-style mean average precision (mAP) along with average precision scores at 0.5 and 0.75 IoU thresholds (AP50 and AP75) for HRSC.

4.1.2. Evaluation Metrics

Various commonly used metrics for Remote Sensing Object Detection (RSOD) are used in the experiment to evaluate the proposed model's effectiveness. Average Precision (AP) is utilized in this paper as the performance metric for object detection models. The calculation formula for AP is:

$$\begin{cases} P = \frac{TP}{TP+FP} \\ r = \frac{TP}{TP+FN} \\ AP = \int_0^1 p(r)\,dr \end{cases} \tag{5}$$

TP represents correctly classified targets, FP signifies background identifications as targets, and FN indicates object identifications misclassified as background. Precision (p) is the ratio of correctly identified targets to all detected results, while Recall (r) is the ratio of correctly identified targets to the true values of all targets. The area under the curve with p on the vertical axis, r on the horizontal axis, and the coordinate axes represents the AP value. AP considers both precision and recall, where a higher value suggests better detection accuracy. The mean Average Precision (mAP) for each class is calculated with the formula below:

$$mAP = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 P_i(R_i)\,dR_i \tag{6}$$

Here, N represents the number of object categories. mAP@0.5 indicates the mean average precision for all classes at an Intersection over the Union (IoU) threshold of 0.5. mAP@0.5:0.95 denotes the average mAP calculated across IoU thresholds from 0.5 to 0.95.

4.2. Implementation details

We perform experiments utilizing RTMDet [18] within the MMRotate toolbox [56]. Our experiments adopt the configuration from RTMDet, employing CSPNetXtBlock as the backbone network and CSPNetXt-PAFPN as the neck. Throughout the model training phase, we utilize diverse data augmentation techniques like random flipping, rotation, scale variation, and padding. Scale variation augmentation is similarly applied in the testing and inference phases. In comparative experiments, we uphold consistent hyperparameter settings during training to ensure a fair comparison with other SOTA methods.

The MAR20 dataset is divided into patches of 800x800 pixels with a 200-pixel overlap between contiguous patches. During the training, validation, and testing phases of the SRSDD and HRSC datasets, we resize the images to 1024x1024 and 800x800 pixels, respectively, using data augmentation techniques without cropping. We use the training subset for training purposes and the test subset for validation and inference. The training duration comprises 36 epochs for the MAR20 dataset, 144 epochs for the SRSDD dataset, and 108 epochs for the HRSC dataset to derive the inference model.

4.3. Comparisons with SOTA

We compare our proposed method with other SOTA approaches using the MAR20, SRSDD, and HRSC datasets. As indicated in the table, our method exhibits superior performance compared to the SOTA approaches without excessive details.

4.3.1. Results on MAR20

MAR20 is a detailed dataset specifically created for detecting military aircraft, covering a broad spectrum of target sizes. It comprises remote sensing images captured in diverse climatic conditions, various seasons, and under differing lighting conditions. Due to the modules we designed that combine both convolutional and attention characteristics, our model efficiently extracts features and aggregates feature contexts, obtaining high-quality feature maps. This enables effective category recognition and precise learning of object bounding boxes, resulting in significantly higher accuracy than the current SOTA. We have chosen various object categories at different scales and scenes where objects are arranged densely and sparsely against different backgrounds for visualization. The detection results are illustrated in the Figure 6. It can be observed from the figures that the proposed method accurately detects densely arranged objects. Table 1 presents the specific performance metrics for each object category. For individual categories like A11, A13, and A14, there is considerable room for improvement in detection results due to the limited number of training instances for each class, which is fewer than 200. Similarly, some small object categories (such as A15 and A20) face challenges in accurate detection owing to their small size, with approximately 70% of instances having pixel values less than 100 pixels. Additionally, the similarity between the A13 and A15 classes, both representing aircraft, further complicates accurate detection. The same conclusion can be obtained by analyzing the mAP. Overall, our approach outperforms most categories and achieves an outstanding performance of 85.96%.

In the MAR20 dataset, we selected two images from the test set for showcasing feature heatmaps, and the feature heatmaps of the baseline model and MFCANet at scales P3, P4, and P5 are visualized in Figure 7. Observing the images, it's evident that the baseline CSPPAFPN model lacks sufficient feature extraction for the targeted objects. The heatmap points for features are relatively small, and there are instances of misalignment, with certain features undetected (such as the plane in the third column of the image). Conversely, our approach significantly enhances feature extraction, resulting in more prominent, clearer-shaped, and accurately positioned extracted features. This showcases the exceptional feature acquisition capability of our method, excelling in target differentiation, background noise suppression, and optimized feature extraction.

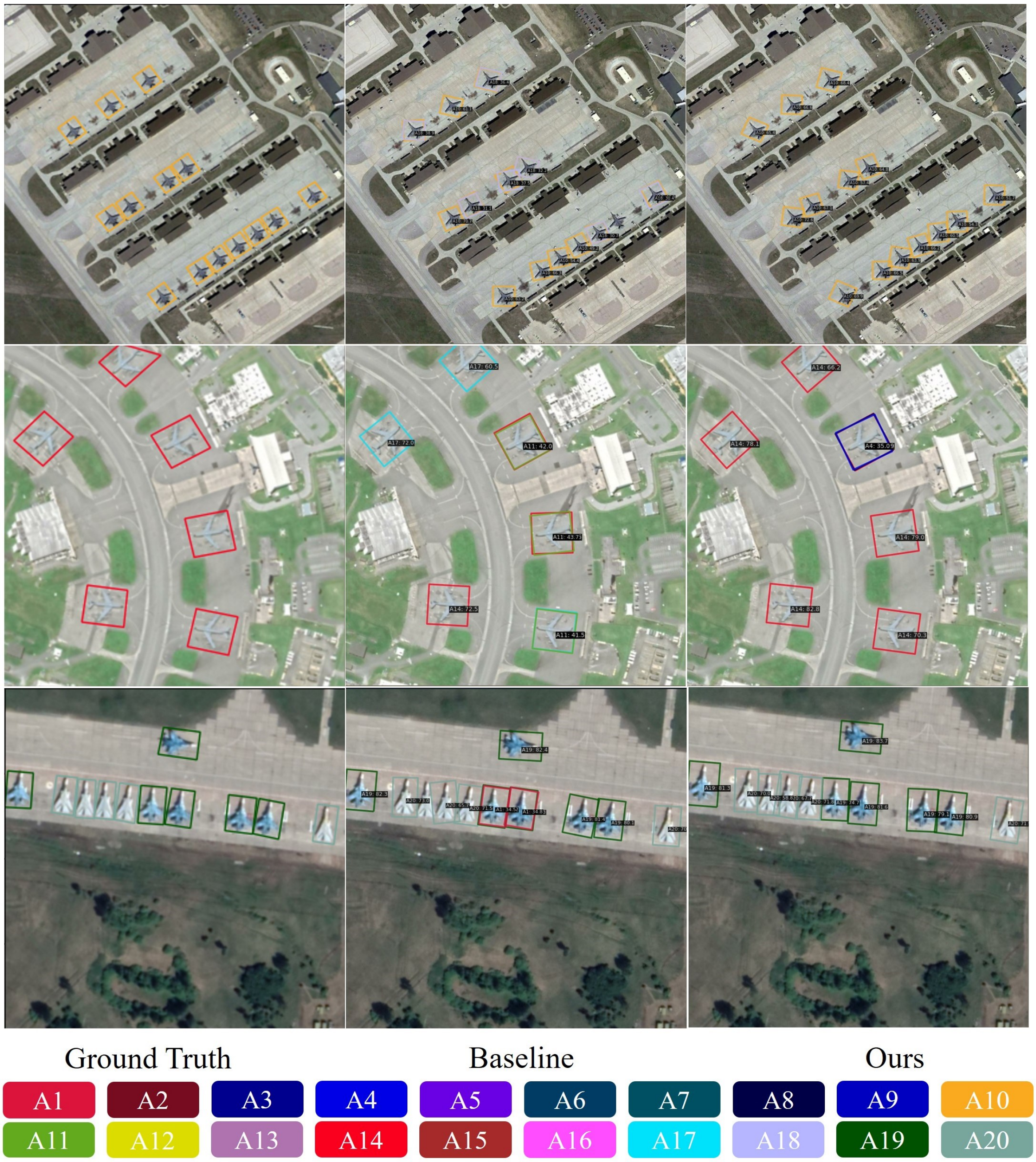

Figure 6. The depicted image demonstrates the outcomes derived from our proposed approach on the MAR20 dataset, encompassing 20 distinct categories. The initial column portrays the dataset's authentic annotations, while the second column displays the baseline results, and the third column exhibits our method's outcomes. Each row corresponds to three sets of results for a single image. The rectangular boxes labeled A1 to A20 at the bottom signify the distinct colors representing respective category bounding boxes.

Table 1. Detection Accuracy of Different Detection Methods on the MAR20 Dataset. The numerical value in black bold represents the maximum.

Method	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	mAP
$S^2A - Net$ [64]	82.6	81.6	86.2	80.8	76.9	90.0	84.7	85.7	88.7	90.8	81.7	86.1	69.6	82.3	47.7	88.1	90.2	62.0	83.6	79.8	81.1
Faster R-CNN [64]	85.0	81.6	87.5	70.7	79.6	90.6	89.7	89.8	90.4	91.0	85.5	88.1	63.4	88.3	42.4	**88.9**	90.5	62.2	78.3	77.7	81.4
Oriented R-CNN [64]	86.1	81.7	**88.1**	69.6	75.6	89.9	90.5	89.5	89.8	**90.9**	**87.6**	88.4	67.5	88.5	46.3	88.3	**90.6**	70.5	78.7	**80.3**	81.9
RoI Trans [64]	85.4	81.5	87.6	78.3	80.5	90.5	90.2	87.6	87.9	**90.9**	**85.9**	**89.3**	67.2	88.2	47.9	89.1	90.5	74.6	81.3	80.0	82.7
RTMDet [18]	**87.7**	**84.0**	82.5	77.4	77.7	**90.7**	90.5	**90.0**	90.5	90.6	84.5	87.7	69.2	86.9	71.7	85.7	90.5	82.9	81.5	74.4	83.83
Ours	86.7	83.5	83.0	**84.5**	**81.2**	90.5	**90.9**	89.4	**90.8**	90.7	85.7	88.3	**78.1**	**88.9**	**76.1**	88.2	90.4	**88.5**	**83.8**	79.8	**85.96**

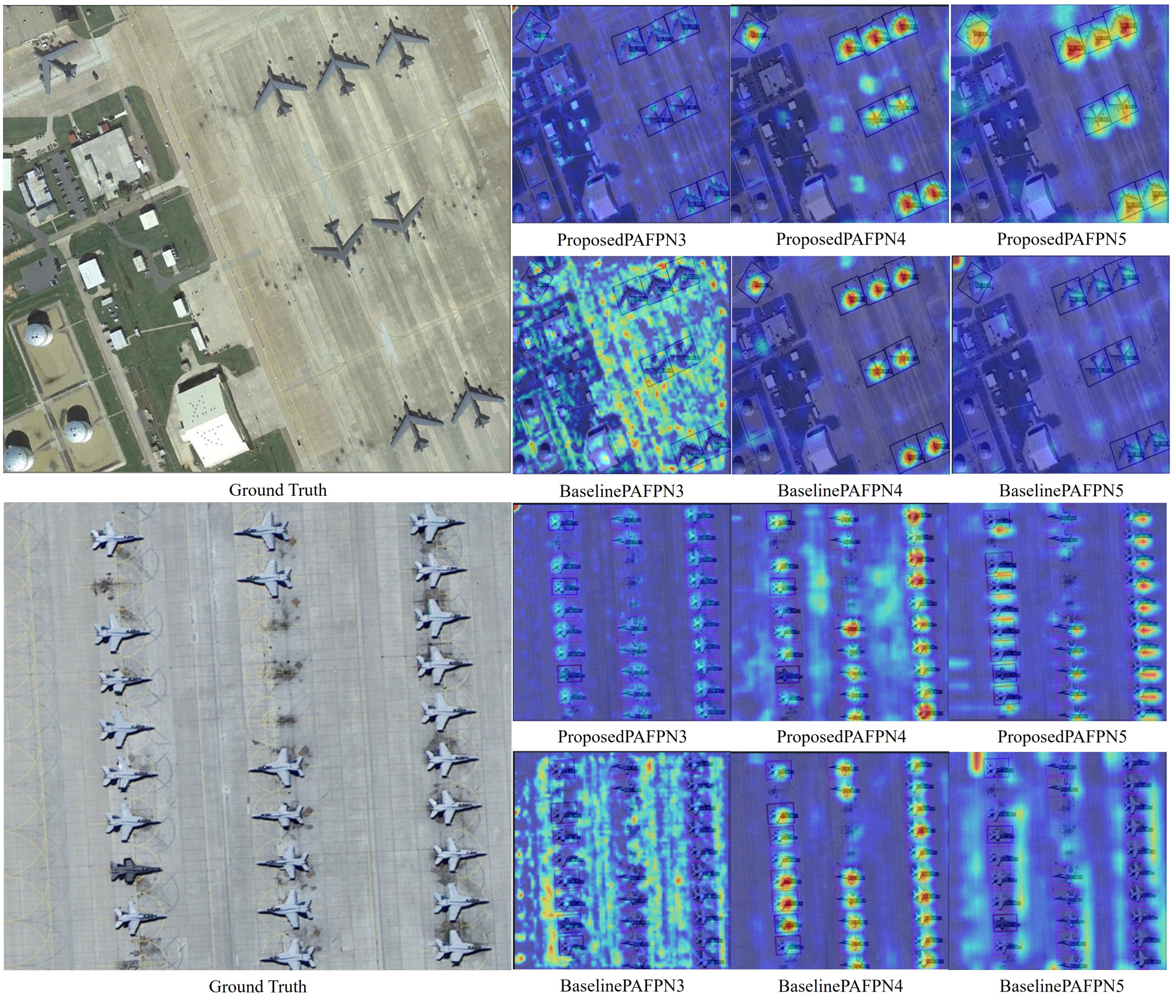

Figure 7. Each image's top row represents the output results of our method, while the second row showcases the baseline's output results. The first column corresponds to the real image, and the subsequent columns, from the second to the fourth, display the output features from the P3, P4, and P5 levels of the pyramid. Blue denotes background, while red and yellow indicate highlighted responses of that specific feature part.

The following conclusions can be drawn from the experimental results: Compared to the baseline, our network effectively captures intricate features of smaller targets within complex backgrounds, enabling precise identification of fine-grained objects and mitigating classification errors. This illustrates that our network thoroughly considers feature and contextual information extraction, effectively eliminating background noise interference. During the feature fusion phase, the network enhances target features, enabling better discrimination of subtle differences within categories, consequently yielding superior results compared to the baseline. However, our network still encounters certain issues. For instance, in scenarios involving more ambiguous images with complex background noise, our model exhibits instances of missed detections and classification errors.

4.3.2. Results on SRSDD

The SRSDD serves as a dataset for detecting rotated objects amidst intricate backgrounds. There is a substantial variation in the quantities among different categories within this dataset, leading to a pronounced issue of data imbalance. Simultaneously, the dataset's complex background contains considerable noise, posing significant challenges for detection. The majority of algorithms exhibit relatively low detection results, as evident from Table 2. Our model has been compared against various state-of-the-art approaches on the SRSDD dataset, demonstrating superior performance with a 10.28% enhancement over the baseline. Specifically, our model achieves the best results in two categories: Ore-oil vessels, characterized by distinct features facilitating easier detection across algorithms, and Law-enforce vessels, which are scarce and usually poorly detected by most algorithms. Our model's improvement in this category stems from its ability to capture distinct features and contextual information specific to Law-enforce vessels, enhancing accuracy due to their scarcity. Container vessels, often overlapping with onshore targets, pose significant interference, while their similarity to fishing vessels complicates their detection amidst high noise levels. Addressing this challenge remains a focal point for our future work. Overall, our method demonstrates commendable performance across most categories, achieving a notable overall accuracy of 66.28%. Nonetheless, there persist issues in our network, such as missed detections when numerous vessels are in proximity and classification errors for vessels with less distinct features. Figure 8 showcases a segment of the detection outcomes, highlighting the proposed method's adeptness in accurately detecting objects within complex backgrounds despite these aforementioned challenges.

Table 2. Detection Accuracy of Different Detection Methods on the SRSDD Dataset. We utilize B1 to B6 to represent the six categories: Ore-oil, Fishing, Law-enforce, Dredger, Cell-Container, and Container. The numerical value in black bold represents the maximum.

Method	B1	B2	B3	B4	B5	B6	mAP
R-RetinaNet [75]	30.4	11.5	2.1	67.7	35.8	48.9	32.73
R^3Det [60]	44.6	18.3	1.1	54.3	43.0	73.5	39.12
BBAVeectors [76]	54.3	21.0	1.1	82.2	34.8	78.5	45.33
R-FCOS [77]	54.9	25.1	5.5	**83.0**	47.4	81.1	49.49
Glid Vertex [78]	43.4	34.6	27.3	71.3	52.8	79.6	51.50
FR-O [74]	55.6	30.9	27.3	77.8	46.7	**85.3**	53.93
ROI [59]	61.4	32.9	27.3	79.4	48.9	76.4	54.38
RTMDet(baseline)	59.4	40.0	27.3	80.5	**76.5**	52.3	56.00
RBFA-Net [79]	59.4	**41.5**	73.5	77.2	57.4	71.6	63.42
Ours	**66.2**	31.4	**94.8**	81.8	73.0	50.5	**66.28**

From Figure 8, it's evident that our model detects targets more accurately compared to the baseline. Within the same image, MFCANet can detect and correctly classify nearshore vessels amid complex coastal backgrounds. This capability stems from MFCANet utilizing the FFCA Module to extract rich contextual feature information. Subsequently, the Feature Context Information Enhancement Module amalgamates and enhances multiscale features, significantly boosting the model's ability to focus on global information. Simultaneously, it's observable from the figure that our network still exhibits instances of misclassification and missed detections. Nevertheless, despite these limitations, our model surpasses the current state-of-the-art. We aim to address these issues of missed detections and misclassification by refining our network for optimal performance.

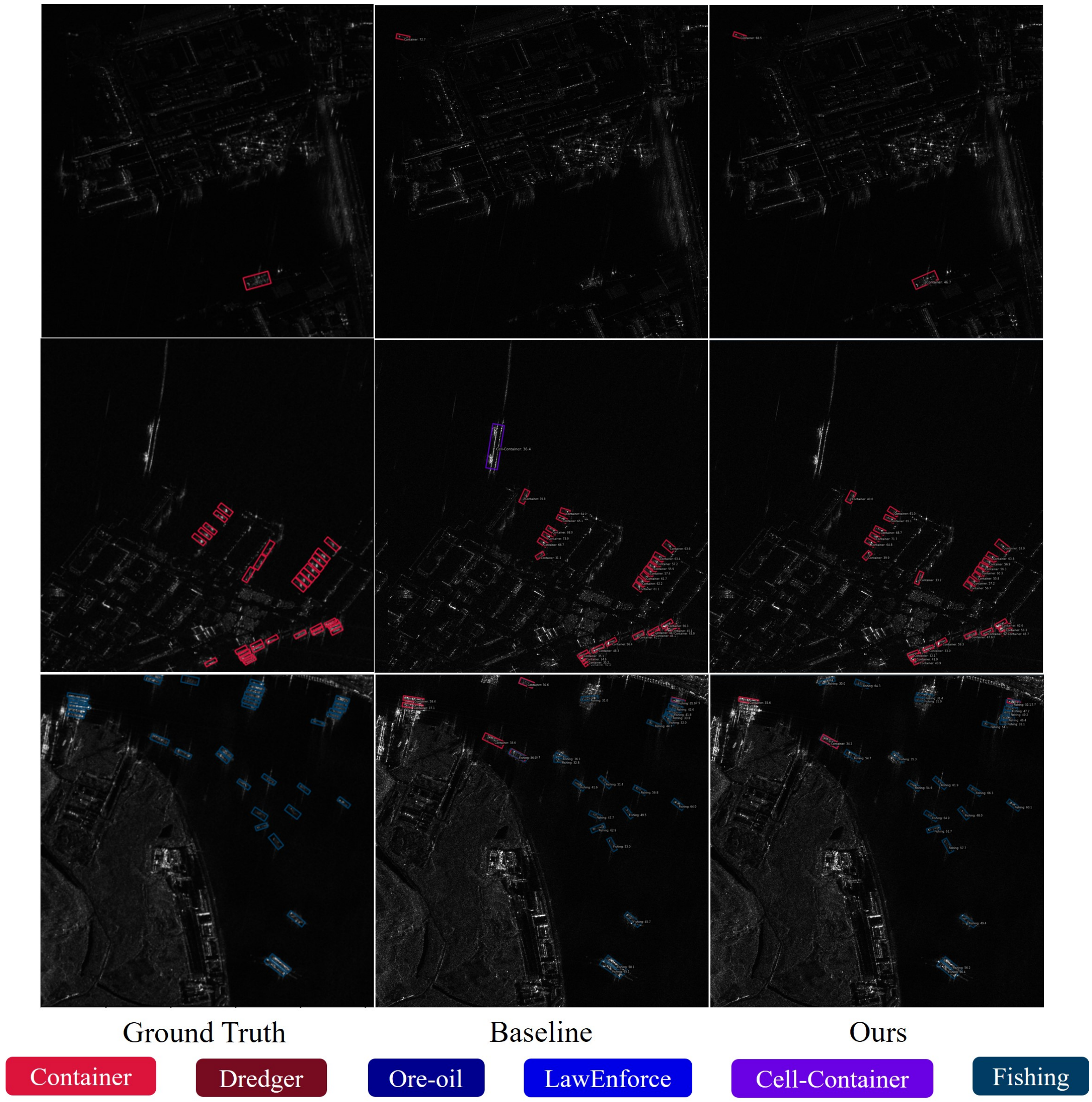

Figure 8. We have presented a sequence of detection outcomes obtained by our proposed MFCANet on the SRSDD dataset. These outcomes emphasize MFCANet's capability to accurately extract target features despite complex backgrounds near coastal and marine areas, ultimately yielding precise results. The initial column portrays the dataset's authentic annotations, while the second column displays the baseline results, and the third column exhibits our method's outcomes. Each row corresponds to three sets of results for a single image. The rectangular boxes at the bottom, each in a different color, represent the bounding box colors corresponding to different categories.

4.3.3. Results on HRSC

The HRSC dataset encompasses vessels with high aspect ratios navigating in different directions, posing significant challenges for precise target localization. Our proposed model showcases robust capabilities in feature extraction, emphasizing global information within the feature maps and effectively identifying class-specific features, resulting in exceptional performance. As illustrated in Table 3, our method has achieved remarkable performance, securing evaluation scores of 90.48% and 97.84% for the VOC2007 and VOC2012 benchmarks, respectively. Figure 9 displays the visual outcomes of implementing our method on the HRSC dataset. From the images, it's apparent that compared to the baseline, our model can more accurately identify results. For instance, in the first row, the second column, and the third column, the baseline incorrectly identifies the object as a vessel, whereas our model adeptly avoids this misidentification. Similarly, when correctly identifying an object, our model expresses higher confidence in the identification. In the case of the last row where the vessel is not recognized, it might be due to the image cropping that retains only a small portion of the vessel, hindering the model from effectively extracting the vessel's features.

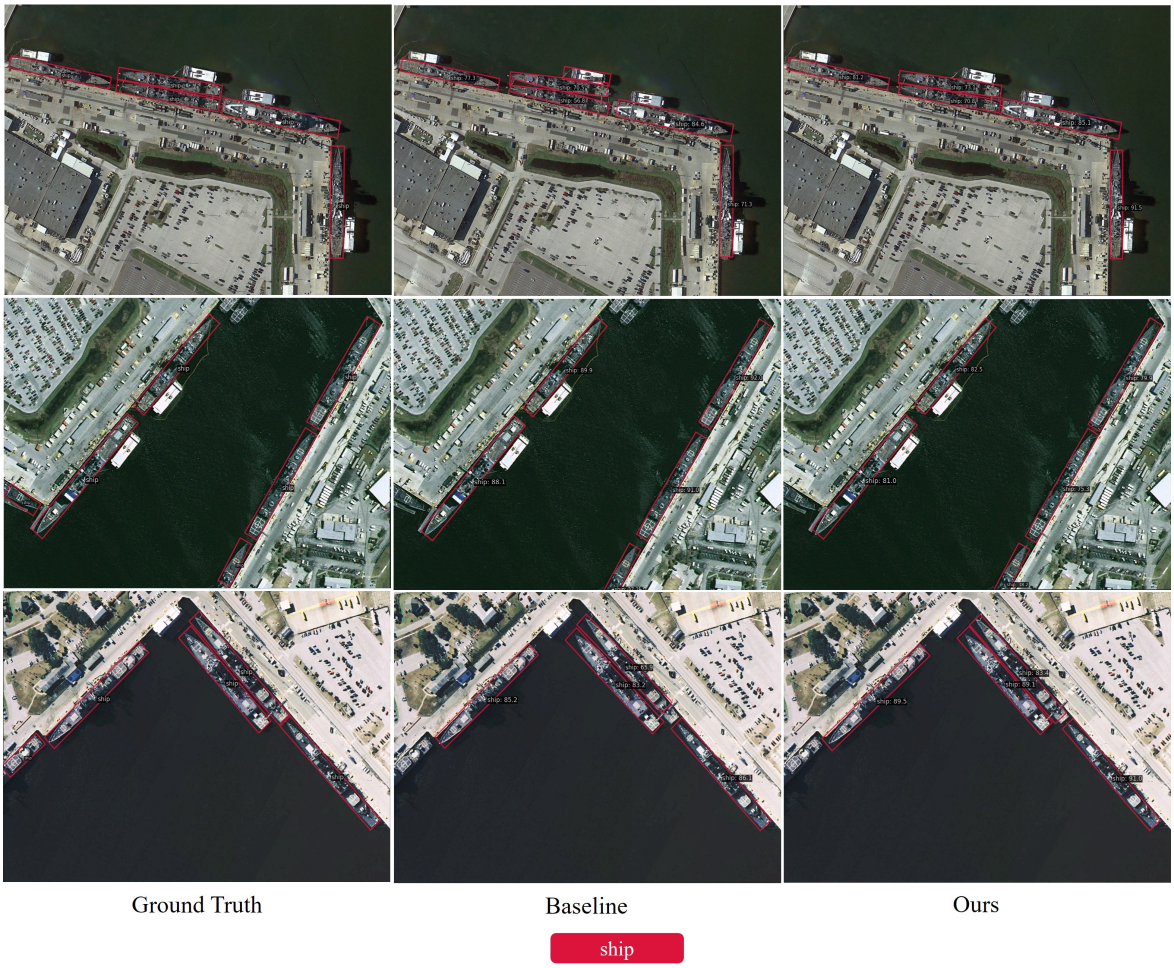

Figure 9. We display a subset of detection outcomes achieved using our MFCANet on the HRSC dataset. The initial column depicts actual images, the second column exhibits predictions from the baseline model, and the third column illustrates predictions from our model. Our approach demonstrates outstanding performance by producing precise and high-quality detection outcomes, especially in identifying densely clustered ships with challenging high aspect ratios.

Table 3. Detection Accuracy of Different Detection Methods on the HRSC Dataset. The numerical value in black bold represents the maximum.

Method	Backbone	mAP (07)(%)	mAP (12)(%)
S^2ANet [61]	R-101	90.17	95.01
AOGC [68]	R-50	89.80	95.20
MSSDet [62]	R-101	76.60	95.30
$R^3Det-KLD$ [25]	R-101	89.97	95.57
MSSDet [62]	R-152	77.30	95.80
R^3Det [60]	R-101	89.26	96.01
DCFPN [25]	R-101	89.98	96.12
RTMDet [18]	CSPNext-52	89.69	96.38
Ours	CSPNext-52	**90.48**	**97.84**

4.4. Ablation Study

4.4.1. Ablation study with different feature fusion methods in MFFM

To deeply analyze how the original features are enhanced during the fusion process with PAFPN features, we conduct an ablation experiment focusing on the skip connections within the Multi-Feature Fusion Module (MFFM). Figure 3 displays skip connections of different colors utilized as modules for the ablation experiment, specifically identified as red and orange. We compare how original features fuse with PAFPN in contrast to the baseline RTMDet on the MAR20 dataset. The experimental results, as depicted in Figure 10, indicate that solely incorporating the yellow skip connection leads to a slight improvement. This could be attributed to the yellow skip connection primarily operating in the middle layer, responsible for fusing the original features, while the other two layers simply replicate the original features. Better results are observed when employing both multi-scale feature fusion methods simultaneously, notably enhancing detection accuracy. This improvement can be attributed to the effective re-fusion of original features with the already fused ones via the red skip connection, compensating for previously overlooked features and thereby enhancing the overall outcome.

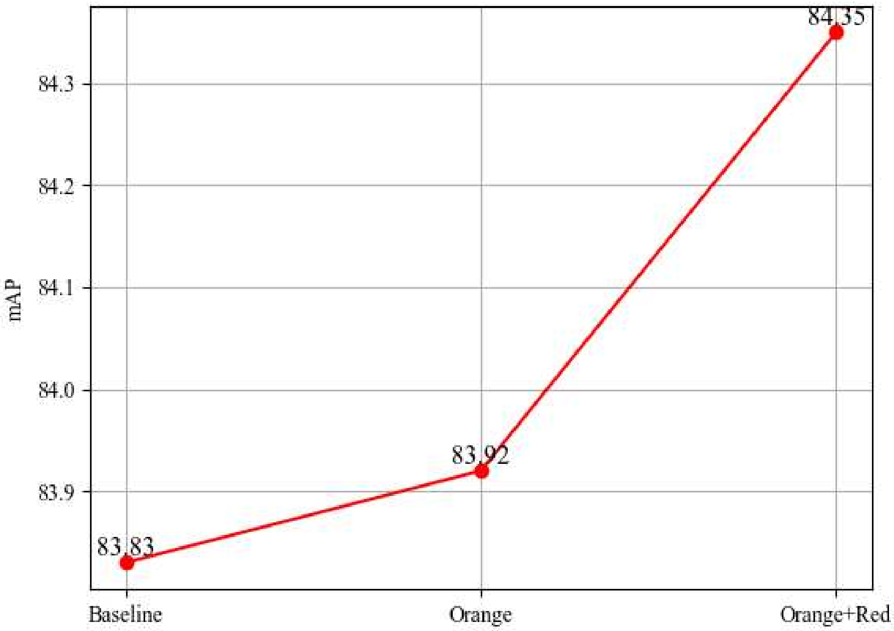

Figure 10. The line chart below illustrates the Baseline, Orange, and Orange+Red, representing the baseline result, the inclusion of yellow skip connections, and the simultaneous inclusion of yellow and red skip connections, respectively. The vertical axis indicates the mAP for each method on the MAR20 dataset.

4.4.2. Ablation study on ODCLayer modules.

For a comprehensive understanding of the enhanced functionality of our proposed ODCLayer module (Figure 4), we conduct an ablation experiment involving the components within the ODCLayer module. Specifically, we employ 3x3 and 5x5 ODConv kernels as individual sets and perform ablation experiments using sets of three, four, and five such combinations. Furthermore, we conduct ablation experiments with and without channel attention. The results from the ablation experiments, as shown in Figure 11, demonstrate that employing four sets of ODConv with the addition of attention achieves optimal performance. Analyzing the outcomes in Figure 11 leads to the following observations: When the set count "Number" equals 3, the features are incompletely integrated, resulting in suboptimal aggregation of contextual feature information and consequently poor results. However, when the set count "Number" is 5, the outcomes degrade compared to "Number" 4, as it aggregates background and noise information during feature context fusion, leading to worsened results. Due to the diverse impacts of distinct channel weights on the outcomes, channel attention integration mitigates the adverse effects of specific channel information on the results. Consequently, incorporating channel attention further enhances the results when the set count "Number" is 4, yielding the most favorable outcomes.

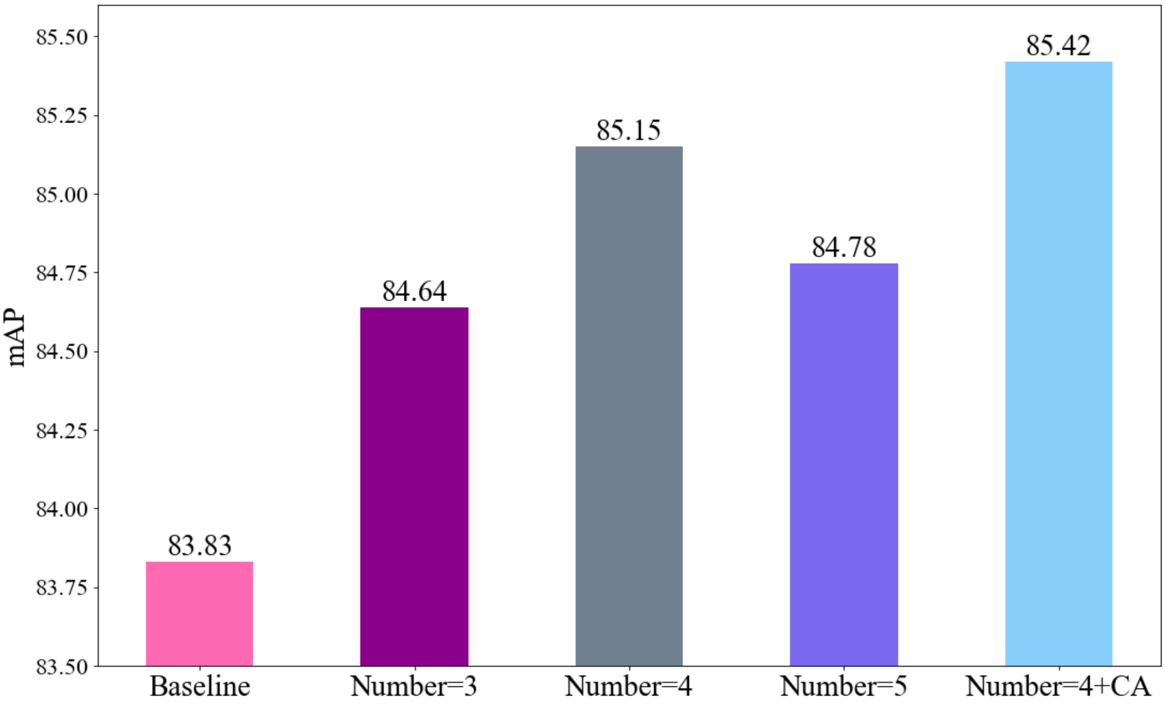

Figure 11. From left to right, each bar in the bar chart represents combinations of three, four, and five sets of 3x3 and 5x5 ODCLayer configurations. The last column in the bar chart corresponds to the addition of channel attention to the configurations of four sets of 3x3 and 5x5 ODCLayers. The vertical axis represents the mAP of each method on the MAR20 dataset.

4.4.3. Ablation study on MFCANet

To assess the efficacy of each proposed module, we compared the baseline with the individual enhancement modules using the MAR20 dataset, using RTMDet as the baseline for detection. The assessment primarily centers on the Average Precision (AP) and mean Average Precision (mAP) of standard object categories, such as A4, A5, A11, A13, A14, A15, A16, A18, and A20. Due to the similarity among fine-grained objects in remote sensing images and the complexity of backgrounds under various seasons and lighting conditions, their detection presents challenges.

Meticulous ablation experiments have been conducted on each enhancement module, and the results, presented in Table 4, highlight the recognition outcomes for some particularly challenging targets. These experiments unequivocally show the effectiveness of the FFCA Module in significantly boosting the backbone network's ability to extract features across various scales. Simultaneously, the ODCLayer module, employing a multidimensional attention mechanism and broader receptive fields through extensive kernel convolutions, adeptly captures comprehensive contextual information. This strategic approach effectively reduces background interference while enhancing the nuances of target features, thus increasing the model's sensitivity to target identification. Furthermore, the skip connections network skillfully utilizes original feature information, preventing information loss during the fusion process. The synergistic interaction among these three modules vividly showcases the exceptional capability of our multi-scale feature context aggregation network.

Table 4. The table clearly shows that adding each module independently enhances the detection performance of the baseline model. This suggests that our methods facilitate aggregating features and their contextual information within the baseline model at their respective positions. Moreover, the combination of any two modules exceeds the detection results achieved by a single module, illustrating the mutual enhancement among our method modules. Remarkably, integrating all three modules simultaneously significantly improves the detection results. Although certain individual module methods exhibit minor decreases in specific categories compared to the baseline, these variations stem from the diverse focal points of the respective module methods. Overall, the collective integration of our module methods produces a significant enhancement.

Baseline	M1	M2	M3	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	mAP
✓				87.7	84.0	82.5	77.4	77.7	90.7	90.5	90.0	90.5	90.6	84.5	87.7	69.2	86.9	71.7	85.7	90.5	82.9	81.5	74.4	83.83
✓	✓			85.4	80.5	85.4	81.0	82.7	90.8	90.8	90.1	90.5	90.8	82.8	85.3	72.9	85.9	72.7	88.1	90.4	84.4	81.8	74.4	84.32
✓		✓		87.1	81.2	83.2	84.5	80.0	90.5	89.8	87.1	90.6	90.9	85.0	88.8	68.3	88.2	63.8	87.1	90.4	86.9	83.8	79.8	84.35
✓			✓	87.5	87.7	85.9	83.0	81.1	90.8	90.8	90.1	90.6	90.9	83.1	84.7	78.7	88.5	69.9	87.5	90.4	84.8	83.4	79.2	85.42
✓	✓	✓		84.6	85.3	88.9	85.9	79.2	90.7	90.5	87.6	89.2	90.9	83.6	89.6	69.8	88.6	61.3	87.3	90.5	86.4	83.4	76.8	84.51
✓		✓	✓	88.7	84.7	84.3	85.1	81.5	90.6	90.1	90.4	90.6	90.8	85.3	88.3	72.5	88.6	71.0	88.9	90.4	88.0	82.9	79.3	85.61
✓	✓		✓	88.3	85.0	89.9	87.3	83.1	90.8	90.5	89.4	90.7	90.9	85.1	88.6	71.6	86.2	73.9	88.7	90.5	82.9	83.8	78.6	85.79
✓	✓	✓	✓	86.7	83.5	83.0	84.5	81.2	90.5	90.9	89.4	90.8	90.7	85.7	88.3	78.1	88.9	76.1	88.2	90.4	88.5	83.8	79.8	85.96

5. Conclusion

To address the complex task of detecting targets in intricate backgrounds within remote sensing images, we propose a novel target detection network tailored for remote sensing imagery. By combining three modules synergistically, we efficiently extract more precise features of interest. Following this, we devise a module dedicated to comprehensively fusing multi-level and multi-dimensional features, thereby enriching valuable features across each layer of PAFPN. Ultimately, we fuse the original feature map information with the results obtained from PAFPN. We conduct thorough validation and ablation studies on three publicly accessible datasets. Our experimental results establish the superiority of our method compared to existing detection networks on these challenging datasets, affirming the effectiveness and versatility of the introduced modules. Nevertheless, it's important to note that our approach still faces limitations in detecting densely occluded small targets. We suppose that MFCANet does not fully mine the unobvious features of small samples. ***For future research***, we plan to investigate scenarios involving dense occlusion of small targets and refine our network model for improved handling of such scenarios.

Author Contributions: Conceptualization, Honghui Jiang, and Tingting Luo.; methodology, Honghui Jiang.; software, Honghui Jiang.; validation, Honghui Jiang., Tingting Luo. and Guozheng Zhang.; formal analysis, Honghui Jiang, Hu Peng, and Guozheng Zhang.; investigation, Honghui Jiang, and Hu Peng; resources, Hu Peng.; data curation, Honghui Jiang, and Tingting Luo.; writing—original draft preparation, Honghui Jiang, and Tingting Luo.; writing—review and editing, Honghui Jiang, and Tingting Luo.; visualization, Tingting Luo.; supervision, Honghui Jiang.; project administration, Honghui Jiang.; funding acquisition, Honghui Jiang. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the 2023 Anhui Province Higher Education Research Project and Universities under Grant 2023AH052692, as well as by the 2022 Anhui Province Higher Education Research Project, Grant Number 2022AH052361.

References

1. Barmpoutis, P.; Papaioannou, P.; Dimitropoulos, K.; Grammalidis, N. A review on early forest fire detection systems using optical remote sensing. *Sensors* **2020**, *20*, 6442.
2. Mohan, A.; Singh, A.K.; Kumar, B.; Dwivedi, R. Review on remote sensing methods for landslide detection using machine and deep learning. *Transactions on Emerging Telecommunications Technologies* **2021**, *32*(7), e3998.
3. Hu, J.; Huang, Z.; Shen, F.; He, D.; Xian, Q. A Robust Method for Roof Extraction and Height Estimation. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*; IEEE: 2023.
4. Weng, W.; Ling, W.; Lin, F.; Ren, J.; Shen, F. A Novel Cross Frequency-domain Interaction Learning for Aerial Oriented Object Detection. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*; Springer: 2023.
5. LI, B.; XIE, X.; WEI, X.; TANG, W. Ship detection and classification from optical remote sensing images: A survey. *Chinese Journal of Aeronautics* **2021**, *34*(3), 145–163.
6. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2014; pp. 580–587.
7. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*; 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*.
9. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems* **2016**, *29*.
10. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
11. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
12. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF international conference on computer vision*; 2021; pp. 2778–2788.
13. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; others. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976* **2022**.
14. 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2023; pp. 7464–7475.
15. Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. *arXiv preprint arXiv:2310.06313* **2023**.
16. Shen, F.; He, X.; Wei, M.; Xie, Y. A competitive method to vipriors object detection challenge. *arXiv preprint arXiv:2104.09059* **2021**.
17. Han, J.; Ding, J.; Li, J.; Xia, G.-S. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–11.
18. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784* **2022**.
19. 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*; 2017; pp. 2117–2125.

20. 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*; 2018; pp. 8759–8768.
21. Zheng, S.; Wu, Z.; Xu, Y.; Wei, Z. Instance-Aware Spatial-Frequency Feature Fusion Detector for Oriented Object Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
22. Zheng, S.; Wu, Z.; Xu, Y.; Wei, Z.; Plaza, A. Learning orientation information from frequency-domain for oriented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–12.
23. Chen, S.; Zhao, J.; Zhou, Y.; Wang, H.; Yao, R.; Zhang, L.; Xue, Y. Info-FPN: An Informative Feature Pyramid Network for object detection in remote sensing images. *Expert Systems with Applications* **2023**, *214*, 119132.
24. Zhang, G.; Lu, S.; Zhang, W. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*(12), 10015–10024.
25. LI, Y.; WANG, H.; FANG, Y.; WANG, S.; LI, Z.; JIANG, B. Learning power Gaussian modeling loss for dense rotated object detection in remote sensing images. *Chinese Journal of Aeronautics* **2023**, Elsevier.
26. Zhang, W.; Jiao, L.; Li, Y.; Huang, Z.; Wang, H. Laplacian feature pyramid network for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–14.
27. 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 Sensing* **2023**, *15*(14), 3497.
28. Chen, J.; Hong, H.; Song, B.; Guo, J.; Chen, C.; Xu, J. MDCT: Multi-Kernel Dilated Convolution and Transformer for One-Stage Object Detection of Remote Sensing Images. *Remote Sensing* **2023**, *15*(2), 371.
29. Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; Zeng, H. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing* **2023**.
30. Shen, F.; Peng, X.; Wang, L.; Zhang, X.; Shu, M.; Wang, Y. HSGM: A Hierarchical Similarity Graph Module for Object Re-identification. In *2022 IEEE International Conference on Multimedia and Expo*; 2022.
31. Shen, F.; Shu, X.; Du, X.; Tang, J. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval. In *Proceedings of the 31th ACM International Conference on Multimedia*; 2023.
32. 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*; 2020; pp. 11534–11542.
33. Zhang, Q.-L.; Yang, Y.-B. Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; 2021; pp. 2235–2239.
34. Lee, Y.; Park, J. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*; 2020; pp. 13906–13915.
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
36. Hu, J.; Huang, Z.; Shen, F.; He, D.; Xian, Q. A Bag of Tricks for Fine-Grained roof Extraction. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*; 2023.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*; 2021; pp. 10012–10022.
38. Wu, H.; Shen, F.; Zhu, J.; Zeng, H.; Zhu, X.; Lei, Z. A sample-proxy dual triplet loss function for object re-identification. *IET Image Processing* **2022**, *16*(14), 3781–3789.
39. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–15.
40. Qiao, C.; Shen, F.; Wang, X.; Wang, R.; Cao, F.; Zhao, S.; Li, C. A Novel Multi-Frequency Coordinated Module for SAR Ship Detection. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*; 2022; pp. 804–811.
41. Shen, F.; Wei, M.; Ren, J. HSGNet: Object Re-identification with Hierarchical Similarity Graph Network. *arXiv preprint arXiv:2211.05486* **2022**.
42. Liu, F.; Chen, R.; Zhang, J.; Ding, S.; Liu, H.; Ma, S.; Xing, K. ESRTMDet: An End-to-End Super-Resolution Enhanced Real-Time Rotated Object Detector for Degraded Aerial Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2023**, *IEEE*.

43. Shen, F.; Zhu, J.; Zhu, X.; Xie, Y.; Huang, J. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems* **2021**, *23*(7), 8793–8804.
44. Min, L.; Fan, Z.; Lv, Q.; Reda, M.; Shen, L.; Wang, B. YOLO-DCTI: Small Object Detection in Remote Sensing Base on Contextual Transformer Enhancement. *Remote Sensing* **2023**, *15*(16), 3970.
45. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–11.
46. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *44*(4), 1922–1933.
47. Wang, X.; Girdhar, R.; Yu, S. X.; Misra, I. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2023; pp. 3124–3134.
48. Yang, X.; Zhang, G.; Li, W.; Wang, X.; Zhou, Y.; Yan, J. H2RBox: Horizonal Box Annotation is All You Need for Oriented Object Detection. *arXiv preprint arXiv:2210.06742* **2022**.
49. Sun, X.; Cheng, G.; Pei, L.; Li, H.; Han, J. Threatening patch attacks on object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *IEEE*.
50. Wan, D.; Lu, R.; Wang, S.; Shen, S.; Xu, T.; Lang, X. YOLO-HR: Improved YOLOv5 for Object Detection in High-Resolution Optical Remote Sensing Images. *Remote Sensing* **2023**, *15*(3), 614.
51. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking rotated object detection with gaussian wasserstein distance loss. In *International conference on machine learning*; 2021; pp. 11830–11841.
52. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* **2021**.
53. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*; 2022; pp. 11976–11986.
54. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*; 2022; pp. 11963–11975.
55. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2018; pp. 7794–7803.
56. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*; 2022; pp. 7331–7334.
57. Qian, X.; Zhang, N.; Wang, W. Smooth giou loss for oriented object detection in remote sensing images. *Remote Sensing* **2023**, *15*(5), 1259.
58. Zhang, Y.; Wang, Y.; Zhang, N.; Li, Z.; Zhao, Z.; Gao, Y.; Chen, C.; Feng, H. RoI Fusion Strategy with Self-Attention Mechanism for Object Detection in Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2023**, *IEEE*.
59. Ding, J.; Xue, N.; Long, Y.; Xia, G.; 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*; 2019; pp. 2849–2858.
60. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. *Proceedings of the AAAI conference on artificial intelligence* **2021**, *35*(4), 3163–3171.
61. Han, J.; Ding, J.; Li, J.; Xia, G. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing* **2021**, *60*, 1–11.
62. Chen, W.; Han, B.; Yang, Z.; Gao, X. MSSDet: Multi-Scale Ship-Detection Framework in Optical Remote-Sensing Images and New Benchmark. *Remote Sensing* **2022**, *14*(21), 5460.
63. 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*; 2021; pp. 3520–3529.
64. Yu, W.; Cheng, G.; Wang, M.; Yao, Y.; Xie, X.; Yao, XW.; Han, JW. MAR20: A Benchmark for Military Aircraft Recognition in Remote Sensing Images. *National Remote Sensing Bulletin* **2022**.
65. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*; 2017; Volume 2; pp. 324–331.
66. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing* **2020**, *159*, 296–307.

67. Shen, F.; Du, X.; Zhang, L.; Tang, J. Triplet Contrastive Learning for Unsupervised Vehicle Re-identification. *arXiv preprint arXiv:2301.09498* **2023**.
68. Wang, Z.; Bao, C.; Cao, J.; Hao, Q. AOGC: Anchor-Free Oriented Object Detection Based on Gaussian Centerness. *Remote Sensing* **2023**, *15*(19), 4690.
69. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **2015**, *37*(9), 1904–1916.
70. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *40*(4), 834–848.
71. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In *Proceedings of the European conference on computer vision (ECCV)*; 2018; pp. 385–400.
72. Li, C.; Zhou, A.; Yao, A. Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947* **2022**.
73. Lei, S.; Lu, D.; Qiu, X.; Ding, C. SRSDD-v1.0: A high-resolution SAR rotation ship detection dataset. *Remote Sensing* **2021**, *13*(24), 5104.
74. Faster RCNN. Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *9199*, 2969239–2969250.
75. 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*; 2017; pp. 2980–2988.
76. Yi, J.; Wu, P.; Liu, B.; Huang, Q.; Qu, H.; Metaxas, D. Oriented object detection in aerial images with box boundary-aware vectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; 2021; pp. 2150–2159.
77. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*; 2018; pp. 734–750.
78. Xu, Y.; Fu, M.; Wang, Q.; Wang, Y.; Chen, K.; Xia, G.; Bai, X. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *43*(4), 1452–1459.
79. Shao, Z.; Zhang, X.; Zhang, T.; Xu, X.; Zeng, T. RBFA-net: a rotated balanced feature-aligned network for rotated SAR ship detection and classification. *Remote Sensing* **2022**, *14*(14), 3345.

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.