IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 January 2024, accepted 21 March 2024, date of publication 27 March 2024, date of current version 2 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381539

RESEARCH ARTICLE

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

HONGHUI JIANG[1], (Member, IEEE), TINGTING LUO[2], (Member, IEEE), HU PENG[3], (Member, IEEE), AND GUOZHENG ZHANG[4], (Member, IEEE)

[1]School of Internet and Communication, Anhui Technical College of Mechanical and Electrical Engineering, Wuhu 241003, China
[2]State Gride Wuhu Power Supply Company, Wuhu 230061, China
[3]School of Instrument Science and Opto-Electronics Engineering, Hefei University of Technology, Hefei 230009, China
[4]School of Mechanical Engineering, Anhui Technical College of Mechanical and Electrical Engineering, Wuhu 241003, China

Corresponding author: Guozheng Zhang (jenuel@163.com)

This work was supported in part by the 2023 Anhui Province Higher Education Research Project under Grant 2023AH052692, and in part by the 2022 Anhui Province Higher Education Research Project under Grant 2022AH052361.

ABSTRACT Rotated object detection in remote sensing images presents a highly challenging task due to the extensive fields of view and complex backgrounds. While Convolutional Neural Networks (CNNs) and Transformer networks have made progress in this area, there is still a lack of research on extracting and fusing features for small targets in complex backgrounds. To address this gap, we have extended the RTMDet framework by introducing three modules: the Focused Feature Context Aggregation Module, the Feature Context Information Enhancement Module, and the Multi-scale Feature Fusion Module. In the Focused Feature Context Aggregation Module, we replaced the Spatial Pyramid Pooling Bottleneck (SPPFBottleneck) to better extract small target features by focusing on contextual information. The Feature Context Information Enhancement Module enhances the model's perception of multi-dimensional temporal and spatial information. Finally, we combined the original features with the fused ones to prevent the loss of specific features during the fusion process. Our proposed model, named the Multi-scale Feature Context Aggregation Network (MFCANet), was evaluated on four challenging remote sensing datasets (MAR20, SRSDD, HRSC, and DIOR-R). The experimental results demonstrate that our method outperforms baseline models, achieving improvements of 2.13%, 10.28%, 1.46%, and 1.13% in mAP for the MAR20, SRSDD, HRSC, and DIOR-R datasets, respectively.

INDEX TERMS Object detection, complex backgrounds, remote sensing images, context information, multiscale feature fusion.

I. INTRODUCTION

Object detection in remote sensing images [1], [2], [3], [4], [5], [6] plays a crucial role in various applications such as environmental monitoring, military operations, national security, transportation, forestry, and oil and gas activities. It aims to identify the location and category of objects of interest.

To address challenges related to background overlap and offer a more accurate delineation of target boundaries, rotated bounding boxes have been employed. However, detecting dense objects in complex backgrounds remains a challenging issue, highlighting the importance of effective extraction of multidimensional target feature context information. To achieve improved detection results, advanced remote sensing object detection networks typically enhance the two-stage R-CNN detector [7], [8], [9], [10], [11]. This network consists of a backbone network, neck, and detection head. The detection head includes a Region Proposal Network

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

(RPN) and R-CNN detection head. In the typical pipeline, the backbone network extracts valuable target information, the neck performs top-down multiscale feature fusion, and the features are then fed into the detection head for regression and classification. However, most detection networks currently only utilize single-dimensional features from the fused features after multiscale fusion [8], [12], [13], [14], [15], [16]. This approach inadequately exploits the information from the fused features. Additionally, re-extracting information from the fused multiscale features may lead to the loss of smaller targets, which contain crucial high-level semantic and low-level geometric information. Therefore, it is crucial to maximize the utilization of fused feature information for enhancing performance.

Single-stage object detectors are commonly used for rotated object detection in remote sensing images. These detectors, similar to two-stage detectors, consist of three parts but do not include a Region Proposal Network. Instead, they directly perform classification and bounding box regression on the fused multiscale feature maps. Although this architecture is computationally efficient, it often lacks accuracy compared to two-stage detectors. Therefore, the challenges faced by two-stage detectors also apply to single-stage detectors [14], [17], [18], [19]. In single-stage detectors, the C4 and C5 parts of the backbone network usually incorporate a module to enhance the semantic information of features(As shown in Figure 5). This ensures that the deepest feature map contains rich semantic information while maintaining scale invariance in the output [20], [21], [22]. However, this approach overlooks the importance of contextual information between different features, which is crucial for object classification and regression. Considering the contextual information is a key factor in improving dense object detection performance, especially in complex backgrounds.

To address the aforementioned issues, this paper proposes a method for detecting rotated objects in remote-sensing images. The method, called Multiscale Feature Context Aggregation Network for Oriented Object Detection in Remote-Sensing Images (MFCANet), is based on RTMDet. The proposed method consists of three main modules: a Focused Feature Context Aggregation Module, a Feature Context Information Enhancement Module (ODCLayer Module), and a Multiscale Feature Fusion Module. The Focused Feature Context Aggregation Module is designed to extract semantic information and context information related to the focus feature. The ODCLayer Module extracts fused feature information from four different dimensions using various-sized convolutional kernels. The Multiscale Feature Fusion Module integrates the original features with the features after the ODCLayer Module to prevent the loss of small target features. The effectiveness of the proposed method is evaluated on the MAR20, HRSC, SRSDD, and DIOR-R datasets. In summary, the contributions of this work include:

- 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.

II. RELATED WORKS

A. 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 [23], RetinaNet [24], R^2ANet [13], the YOLO series [14], [17], [25], [26], [27], RTMDet [28], among others) and two-stage models (R-CNN [29], Fast R-CNN [7], Faster R-CNN [8], R-FCN [12], 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 [30], [31] 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 [17], [20], [21], [22]. However, they have not fully considered the impact of feature context information on detection results.

Moreover, some researchers have introduced attention mechanisms into CNNs [27], [32], [33], [34], 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 [10], [35], 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 [36], [37], [38], all of which have demonstrated favorable results.

In recent years, Transformer-based models [39], [40], [41], [42] have shown promising results in the field of object detection. The Vision Transformer (ViT) [39]demonstrated that Transformers can be applied to computer vision with minimal modifications and achieve excellent performance. The DETR [40] 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 [41] 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.

B. 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 the multiscale detection challenges arising from different object sizes in remote sensing imagery, mSODANet [43] constructs a hierarchical dilation network using parallel dilated convolutions in PAFPN to enhance the extraction of feature targets. This network facilitates contextual information learning for diverse object types across various scales and fields of view. The Super-Yolo model [18] integrates multimodal data and incorporates auxiliary super-resolution learning to tackle multiscale detection challenges for high-resolution objects, balancing detection accuracy and computational cost. MFAF [44] proposes an adaptive multiscale feature fusion approach, utilizing multiscale feature integration modules and spatial attention weight modules to create a feature fusion module, facilitating flexible fusion of multiscale features. MDCT [31] introduces a single-stage object detection model in FPN, relying on multi-kernel dilated convolution blocks and Transformer blocks to improve the intrinsic and neighboring spatial features of small objects. ANSDA [45] employs 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 [30] integrates 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 [46] 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. HFAN [47] 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 [19] 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 [27] incorporates the Swin-Transformer into PAFPN to more effectively detect objects of various scales.

Additionally, researchers have designed Simplified Spatial Pyramid Pooling Fast [17] to replace Spatial Pyramid Pooling Fast [27], speeding up feature extraction. Inspired by SPP, the Atrous Spatial Pyramid Pooling module [21] is proposed in DeepLabv2, utilizing multiple parallel atrous convolution layers with different sampling rates. Features extracted at each sampling rate are further processed in separate branches and fused to generate the final result. This module constructs convolutional kernels with different dilation rates to build convolutions with different receptive fields, enabling the acquisition of multiscale object information in Atrous Spatial Pyramid Pooling. The Receptive Field Block [22] strengthens the network's feature extraction capability by simulating the receptive field of human vision. Structurally borrowing from the Inception concept, RFB incorporates atrous convolution to effectively increase the receptive field. Despite the significant progress made by spatial pooling methods in improving feature extraction speed, these approaches have not fully addressed the importance of contextual information for target feature extraction. In the construction of feature pyramids, there are still unresolved issues in multiscale feature fusion and feature extraction, necessitating further

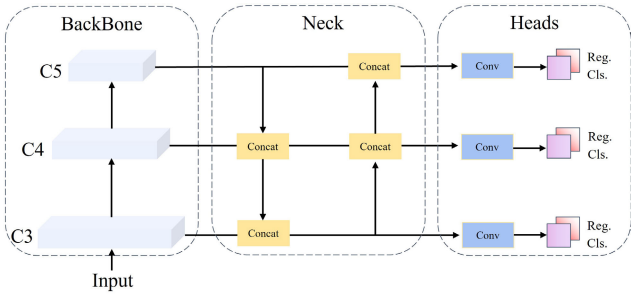

FIGURE 1. The fundamental macro-architecture comprises three segments: the backbone, neck, and heads. Input images are processed through the backbone network to extract features, resulting in three sets of feature maps at varying scales. The neck section utilizes PAFPN for the bidirectional merging of 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.

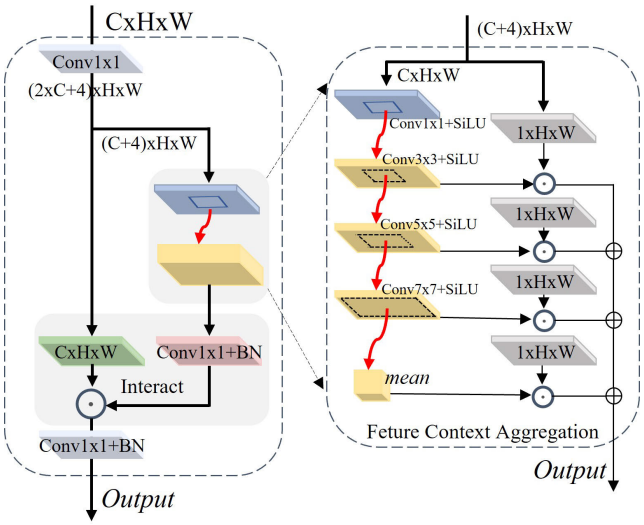

FIGURE 2. The FFCA module is specifically designed to acquire multi-scale focal feature context information. C, H, and W respectively denote the channel, height, and width of the feature map. 'mean' represents the tensor's mean operation, ⊙ denotes tensor multiplication, and ⊕ signifies tensor addition.

in-depth research and exploration. Therefore, more emphasis should be placed on achieving more effective multiscale feature fusion in feature pyramids, as well as a more detailed and comprehensive consideration of information during the fused feature extraction process.

III. METHODS

A. BASIC ROTATED DETECTION METHOD AS BASELINE

Previous approaches have commonly relied on horizontal bounding boxes for object delineation, overlooking the detection of rotated bounding boxes [15], [48]. However, remote-sensing images often contain objects with complex backgrounds, and traditional horizontal bounding boxes include background information that can hinder precise object localization. In contrast, rotated bounding boxes enable precise object localization while minimizing background interference. Additionally, rotated bounding boxes have minimal overlap, ensuring clear object delineation. Therefore, it is crucial to explore and implement more accurate 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 [49], requiring the width w to be greater than the height h. We utilize the one-stage rotation object detector RTMDet [28] to detect sparse and dense objects with complex backgrounds in remote sensing images. RTMDet is an improved version derived from YOLOX [16], having a comparable overall macro-architecture to the YOLO series. The complete model structure is depicted in Figure 1.

Specifically, RTMDet comprises CSPNeXt, CSPNeXtPAFPN, and SepBNHead, which share convolutional weights while performing batch normalization independently. Furthermore, it takes cues from ConvNeXt [50] and RepLKNet [51], 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 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.

B. FOCUSED FEATURE CONTEXT AGGREGATION MODULE

Spatial feature pooling and its variations [20], [21], [22] are commonly employed in the backbone network to extract multi-scale features for 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 propose a novel method called the Focal Feature Context Aggregation Module. This module first adjusts the channel dimensions of the input tensor, which has a channel dimension of C, through convolution to yield a tensor with dimensions 2C+4. The tensor is then partitioned into three groups along the channel dimension, with channel dimensions of C, C, and 4 respectively. The part with channel dimensions of 4 is further decomposed into four tensors, each with a channel dimension of 1, as shown in the gray-boxed convolution in Figure 2. One of the tensors with

a channel dimension of C undergoes sequential convolution operations with kernels of sizes 1, 3, 5, and 7. After each convolution, mean operations are applied. The outcomes of these convolutions are then element-wise multiplied with the corresponding tensors obtained earlier (channel dimension of 1). The results are summed and passed through activation functions. Finally, the resulting tensor, which still has a channel dimension of C, is multiplied by another tensor with a channel dimension of C. This product is refined through a 1×1 convolution operation to extract specific target features along with their corresponding contextual information. The entire process is visually represented in Figure 2. This module is embedded within the FFCA Module section of the backbone network, as illustrated in Figure 5.

The FFCA Module is mathematically described as follows:

$$\boldsymbol{Y} = F(\boldsymbol{X}) \odot (\sum_{i=0}^{3} H_i(\boldsymbol{X}^{'}) \odot G_i(\boldsymbol{X}^{''})). \tag{2}$$

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.

C. 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 [52] 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

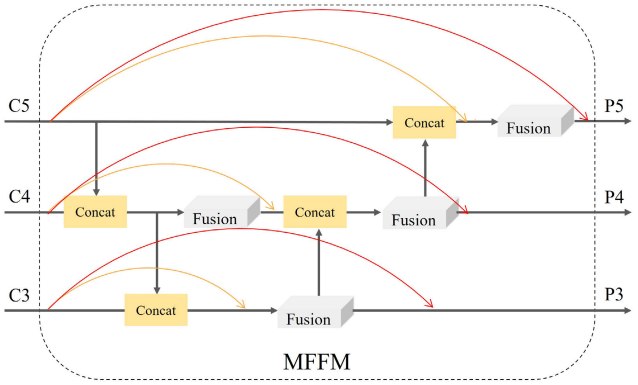

FIGURE 3. The multiscale feature fusion network integrates intermediate and final outputs from PAFPN with the original output features using a red solid line as a residual connection. The fusion of intermediate-level information with deep-layer information is denoted by a deep yellow dashed line, employing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

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. This module corresponds to the MFFM section in the Neck of Figure 5. The entire process is outlined as follows.

$$\boldsymbol{P_3} = F_3(G_3(G_4(\boldsymbol{C_4}, \boldsymbol{C_5}), \boldsymbol{C_3}) + \boldsymbol{C_3}) + \boldsymbol{C_3}. \tag{3}$$

$$\boldsymbol{P_4} = F_4(G_4(F_4(G_4(\boldsymbol{C_4}, \boldsymbol{C_5})) + \boldsymbol{C_4}, \boldsymbol{P_3} - \boldsymbol{C_3})) + \boldsymbol{C_4}. \tag{4}$$

$$\boldsymbol{P_5} = F_5(G_5(\boldsymbol{C_5}, \boldsymbol{P_4} - \boldsymbol{C_4}) + \boldsymbol{C_5}) + \boldsymbol{C_5}. \tag{5}$$

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.

D. FEATURE CONTEXT INFORMATION ENHANCEMENT MODULE

Traditional convolution methods typically utilize fixed kernels that are independent of input samples. However, dynamic convolution technology integrates attention mechanisms and deformable convolution to enhance the model's perception of both time and space information. ODConv [53] employs a parallel strategy and combines multi-dimensional attention mechanisms to achieve flexible attention learning of cross-convolution nuclear space. As depicted in Figure 4, ODConv incorporates different attention values for space position, input channels, convolution filters, and overall convolution kernels. These diverse types of attention complement each other, facilitating various convolutional operations based on positions, channels, filters, and kernels. This, in turn, enhances the capture of context information across

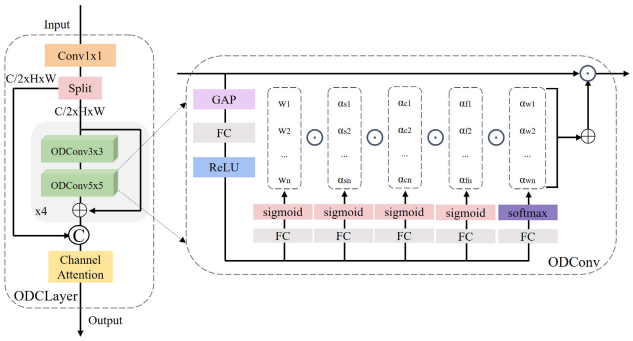

FIGURE 4. The ODCLayer begins by integrating input features using a 1-sized convolutional kernel. The integrated features are then split into two segments. ODConv modules with kernel sizes of 3 and 5 are concatenated in series while maintaining residual connections. This sequence is repeated four times. Afterward, the other segment is concatenated along the channel dimension. Finally, channel attention mechanisms assign different weights to different channels.

different dimensions. Consequently, ODConv significantly enhances the feature extraction capacity of convolution operations. Building upon the powerful performance of ODConv, we have developed the ODCLayer. Please refer to Figure 5 for detailed information and its application in the 'neck' part of our model for feature fusion in PAFPN. The ODCLayer module can be observed in the neck part of Figure 4.

ODCLayer is mathematically described as follows:

$$\boldsymbol{Y} = H(\boldsymbol{concat}(F(C_1(X)), C_2(X))). \tag{6}$$

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 ***concat*** denotes tensor concatenation across the channel dimension.

E. 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.

IV. EXPERIMENTS

This section evaluates the performance of our proposed model by training and testing it on four widely used datasets: MAR20, SRSDD, HRSC, and DIOR-R. We provide a comprehensive overview of our experiments, including details on the experimental design, parameter configurations, comparisons with state-of-the-art (SOTA) models, and the outcomes of our experiments. In addition, we conducted an ablation study on the MAR20 dataset to demonstrate the effectiveness of each module. Our software environment consists of 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 evaluated using DotaMetric. We utilize the AdamW optimizer 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.

A. DATASETS AND EVALUATION METRICS

1) DATASETS

The ***MAR20*** [54] 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 MAR20 dataset encompasses diverse scenarios and conditions, facilitating the evaluation of the model's generalization capabilities. Validating the network on such a dataset provides a more nuanced understanding of the model's performance in specific scenarios.

The ***SRSDD*** [55] 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

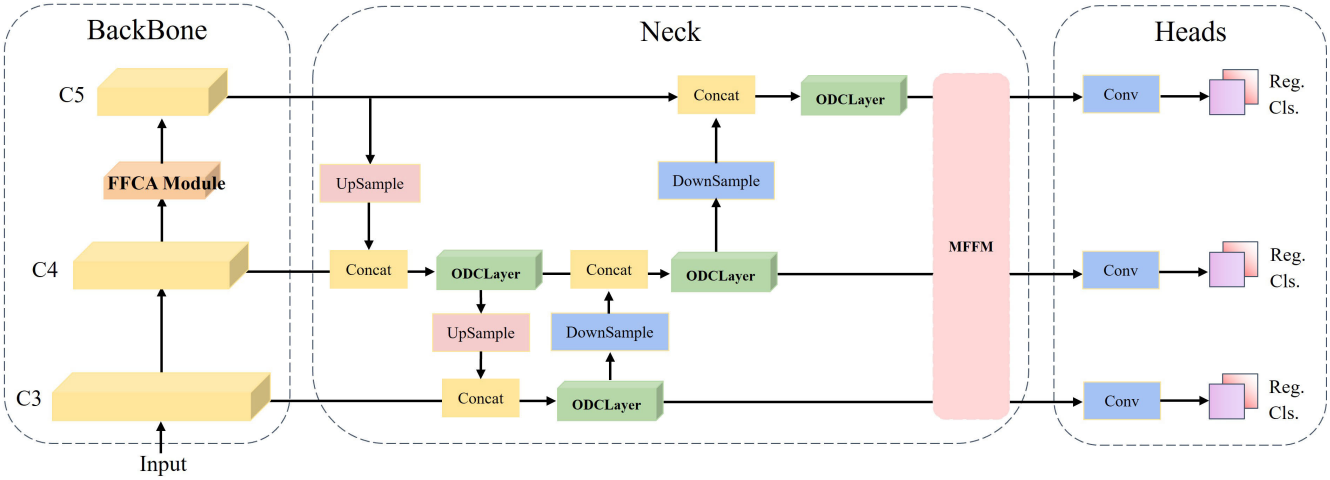

FIGURE 5. The architectural components of MFCANet consist of crucial modules. Initially, we employ the FFCA module to replace the SPPFBottleneck in the backbone network, capturing multi-scale feature context information related to focal targets. Subsequently, utilizing the MFFM module enhances the utilization 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.

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 [56] 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.

The ***DIOR-R*** [57] dataset serves as an extended iteration of the DIOR dataset, featuring reannotation with directional attributes. This dataset holds a prominent position as a standard benchmark for the evaluation of rotated object detection capabilities within remote sensing applications. The DIOR-R dataset is systematically organized into training, validation, and testing subsets. It comprises 20 distinct categories, each denoted by specific labels such as Expressway-Toll-Station (ETS), Chimney (CHI), Baseball-Field (BF), Vehicle (VE), Harbor (HA), Basketball-Court (BC), Golf-Field (GF), Tennis-Court (TC), Storage-Tank (ST), Windmill (WM), Train-Station (TS), Bridge (BR), Ground-Track-Field (GTF), Ship (SH), Airport (APO), Airplane (APL), Expressway-Service-Area (ESA), Dam (DA), Stadium (STA), and Overpass (OP). In total, the DIOR-R dataset encompasses 23,463 images, collectively representing the 20 designated categories, amounting to 192,472 distinct instances. The training and validation datasets jointly consist of 11,725 images, incorporating 68,073 individual instances. Meanwhile, the test dataset comprises 11,738 images and encompasses 124,445 distinct instances. All images adhere to a consistent size of 800 × 800 pixels, with pixel resolutions ranging from 0.5 meters to 30 meters.

2) EVALUATION METRICS

In experiments, it is common practice to employ various evaluation metrics to assess the effectiveness of remote-sensing object detection models. In this paper, we utilize Average Precision (AP) as a performance measure for the object detection model. The calculation formula for AP is as follows:

$$\mathrm{P} = \frac{TP}{TP + FP}. \tag{7}$$

$$\mathrm{r} = \frac{TP}{TP + FN}. \tag{8}$$

$$\mathrm{AP} = \int_0^1 p(r)\, dr. \tag{9}$$

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:

$$\mathrm{mAP} = \frac{1}{N} \sum_{i=1}^{N} \int_0^1 P_i(R_i)\, dR_i. \tag{10}$$

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.

Furthermore, we applied the frames-per-second (FPS) to evaluate the detection efficiency of different methods as follows:

$$\mathrm{FPS} = \frac{1}{\boldsymbol{T_{per-img}}}. \tag{11}$$

where $T_{per-img}$ represents the inference time per image. As for the algorithm complexity, we use the number of parameters, model size, and floating-point operations per second (FLOPS) to evaluate the different methods.

B. IMPLEMENTATION DETAILS

We perform experiments utilizing RTMDet [28] within the MMRotate toolbox [58]. 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 800 × 800 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 1024 × 1024 and 800 × 800 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 and DIOR-R dataset, 144 epochs for the SRSDD dataset, and 108 epochs for the HRSC dataset to derive the inference model.

C. COMPARISONS WITH SOTA

We compare our proposed method with several other state-of-the-art (SOTA) approaches on the MAR20, SRSDD, HRSC, and DIOR-R datasets using the mAP metric. Particularly, we provide the frames-per-second (FPS) for the evaluation of the SRSDD dataset. First, we introduce the characteristics of the compared methods, and then we present the comparative results.

S^2A Net [13] enhances classification scores and localization accuracy in aerial image target detection by utilizing the Feature Alignment Module (FAM) and Directional Detection Module (ODM), addressing the misalignment issue between anchor boxes and convolution features. Faster R-CNN [8], a two-stage object detector, improves accuracy by generating proposals before detection. Oriented R-CNN [54] introduces a simple and universal oriented Region Proposal Network (oriented RPN) for direct high-quality proposal generation in rotation object detection in remote sensing images. The RoI Transformer module [9] is a two-stage object detector comprising RRoI Learner and RRoI Deform. It achieves more accurate RRoI by learning the transformation from HRoI to RRoI without increasing anchor points. RRoI Deform extracts rotation-invariant features from RRoI for subsequent classification and regression tasks. RetinaNet [24] achieves effective rotation object prediction by designing a balanced sample loss function and adding angle prediction. R^3det [59] improves detection speed and recall by initially detecting horizontal boxes in the first stage and refining rotation boxes in the refinement stage to adapt to dense object detection. It addresses feature misalignment with a feature refinement module using feature interpolation to obtain refined position information and reconstruct feature maps for alignment. BBAVectors [60], a single-stage anchor-free detection method, predicts object positions through center and corner points without pre-setting anchors. It incorporates angle information prediction based on CenterNet. Gliding Vertex [61], utilizing the structure of Faster RCNN, predicts rotation rectangles along with classification results and horizontal box coordinates. FR-O [62] augments Faster RCNN with angle prediction and adds a Feature Pyramid Network (FPN) for multi-scale feature fusion. The Region Proposal Network (RPN) still uses horizontal boxes for initial filtering in the first stage, offering advantages in training and testing speed. In the second stage, it adds angle information prediction based on the first stage. RBFA Net [63] designs three target networks: Balanced Attention Feature Pyramid Network (BAFPN), Anchored Feature Alignment Network (AFAN), and Rotation Detection Network (RDN). BAFPN, an improved FPN, reduces the negative impacts of multi-scale ship feature differences. AFAN adopts aligned convolution layers to adaptively align convolution features based on rotated anchor boxes, addressing misalignment. RDN includes a Task Decoupling Module (TDM) to separately adjust feature maps, resolving conflicts between regression and classification tasks. AOGC [64]proposes an anchor-free oriented object detection method based on Gaussian centrality, addressing challenges in anchor-based methods such as high computational load and low accuracy. MSSDet [65] designs a Joint Recursive Feature Pyramid (JRFP) for generating semantic-rich and spatially fine multi-scale features, enhancing detection accuracy. DCFPN [46], based on dilated convolutions, designs a Dense Context Feature Pyramid Network and α-Gaussian loss for improved rotation object detection accuracy. It introduces a simple and effective bounding box representation, integrated into two detection stages, avoiding discontinuity issues and inconsistencies between two-stage regression schemes. In the first stage, it specifically initializes four quadrant points as regression starting points to generate high-quality oriented candidates. In the second stage, it refines the final localization results using the proposed novel bounding box representation, achieving a good balance between accuracy

and speed. AOPG [66], using a coarse localization module (CLM), produces coarse-oriented boxes in an anchor-free manner, refined into high-quality orientation proposals. Following AOPG, a convolutional neural network (R-CNN) head based on a fast region proposal is applied to generate the final detection results. DODet [67] proposes a two-stage oriented object detection method to address spatial and feature misalignment issues. The first stage is the Oriented Proposal Network (OPN), generating high-quality proposals using a new representation scheme for oriented targets. The second stage is the Location Guided Detection Head (LDH), aiming to alleviate feature misalignment between classification and localization. RoIF-Net [68] proposes a dense object determination method for Oriented Bounding Boxes (OBB), determining dense objects in the dataset based on class inter-distance, class intra-distance, object minimum distance, and object minimum side length. To fully utilize target features, it introduces a two-stage detection head, extracting regions of interest from input images and merging them with RoI extracted from feature maps to add detailed features. It constructs a feature induction module based on a self-attention mechanism for position regression and category classification. This structure can be applied to any two-stage network to enhance detection capability. AOPG SGIoU [69] introduces a new bounding box regression loss named Smooth Generalized Intersection over Union (SGIoU) loss. Firstly, the Smooth GIoU loss can adopt more appropriate learning intensities within different GIoU value ranges to address the mentioned issues. The design scheme of the Smooth GIoU loss can be extended to other IoU-based bounding box regression (BBR) losses. Secondly, existing GIoU loss computation schemes can be modified to suit rotation object detection.

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.

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 figures 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.

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	
S^2A Net[56]	82.6	81.6	86.2	80.8	76.9	90.0	84.7	85.7	88.7	90.8	
Faster R-CNN[56]	85.0	81.6	87.5	70.7	79.6	90.6	89.7	89.8	90.4	91.0	
Oriented R-CNN[56]	86.1	81.7	**88.1**	69.6	75.6	89.9	90.5	89.5	89.8	**90.9**	
RoI Trans[56]	85.4	81.5	87.6	78.3	80.5	90.5	90.2	87.6	87.9	**90.9**	
RTMDet(Baseline)	**87.7**	**84.0**	82.5	77.4	77.7	**90.7**	90.5	**90.0**	90.5	90.6	
Ours	86.7	83.5	83.0	**84.5**	**81.2**	90.5	**90.9**	89.4	**90.8**	90.7	
Method	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	mAP
S^2A Net[56]	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[56]	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[56]	**87.6**	88.4	67.5	88.5	46.3	88.3	**90.6**	70.5	78.7	**80.3**	81.9
RoI Trans[56]	82.7	85.9	**89.3**	67.2	88.2	47.9	89.1	90.5	74.6	81.3	80.0
RTMDet(Baseline)	84.5	87.7	69.2	86.9	71.7	85.7	90.5	82.9	81.5	74.4	83.83
Ours	85.7	88.3	**78.1**	**88.9**	**76.1**	88.2	90.4	**88.5**	**83.8**	79.8	**85.96**

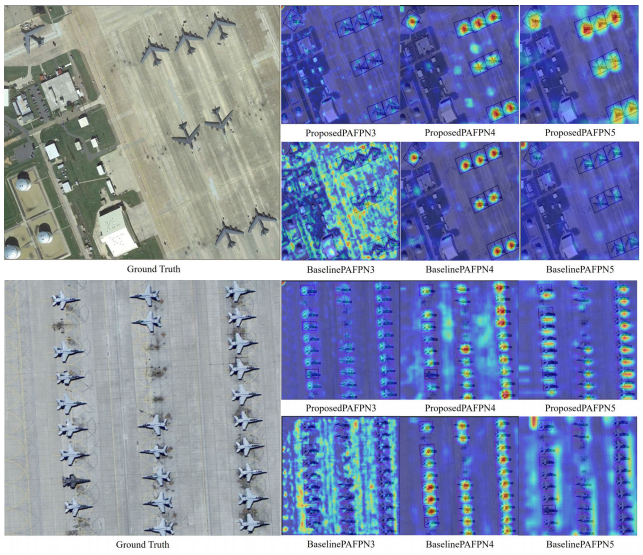

FIGURE 7. Each image's top row represents the output results of our method, while the second row showcases the baseline's output results. The first column corresponds to the real image, and the subsequent columns, from the second to the fourth, display the output features from the P3, P4, and P5 levels 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.

2) RESULTS ON SRSDD

The SRSDD dataset serves as a dataset for detecting rotated objects within complex backgrounds. There is a significant imbalance in the quantity among different categories in this dataset, posing a noticeable issue of data imbalance. Simultaneously, the complex background of the dataset contains considerable noise, presenting significant challenges for the detection task. Most algorithms exhibit relatively low detection results, as shown in Table 2. Our model has been compared against various state-of-the-art methods on the SRSDD dataset, demonstrating a 10.28% improvement over the baseline model. Additionally, we have compared the trade-off between accuracy and speed with other state-of-the-art methods in Figure 8, where our method, with an input size of 1024 × 1024, provides the highest detection accuracy (66.28% mAP) while achieving a high speed of up to 18.3 frames per second. Considering the accuracy/speed trade-off, our method achieves the state-of-the-art.

Specifically, our model achieves the best results in two categories: Ore-oil vessels with distinct features that make them easier to detect across various algorithms, and Law-enforce vessels, which are scarce and usually poorly detected by most algorithms. The improvement in this category stems from our model's ability to capture the specific features and contextual information related 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. However, issues persist in our network, such as missed detections when numerous vessels are in proximity and classification errors for vessels with less distinct features. Figure 9 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[70]	30.4	11.5	2.1	67.7	35.8	48.9	32.73
R^3Det[59]	44.6	18.3	1.1	54.3	43.0	73.5	39.12
BBAVeectors[60]	54.3	21.0	1.1	82.2	34.8	78.5	45.33
R-FCOS[71]	54.9	25.1	5.5	**83.0**	47.4	81.1	49.49
Glid Vertex[61]	43.4	34.6	27.3	71.3	52.8	79.6	51.50
FR-O[62]	55.6	30.9	27.3	77.8	46.7	**85.3**	53.93
ROI[9]	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[63]	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**

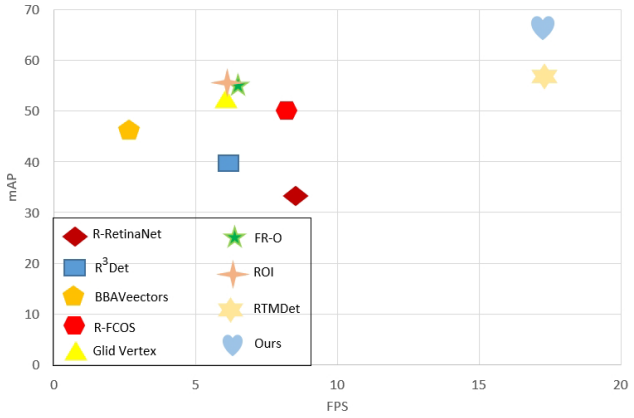

FIGURE 8. Speed versus accuracy on the SRSDD testing set. Our method is extremely fast and accurate.

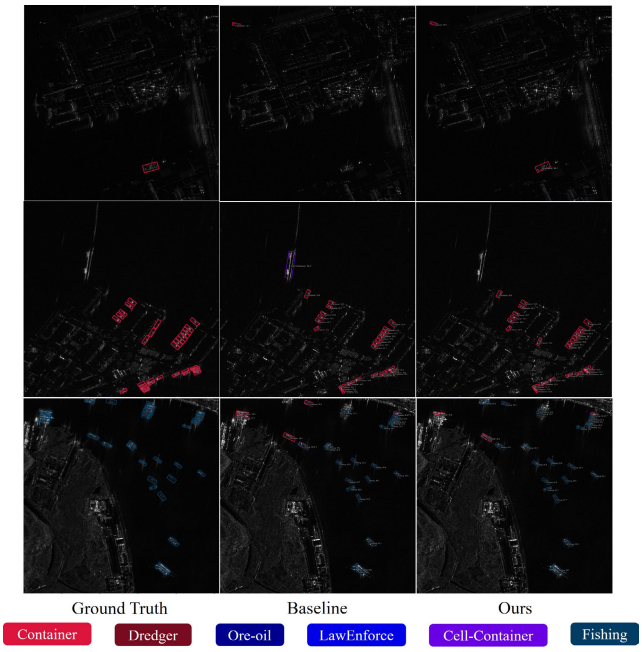

FIGURE 9. We have presented a sequence of detection outcomes obtained by our proposed MFCANet on the SRSDD dataset. These outcomes emphasize MFCANet's capability to accurately extract target features despite complex backgrounds near coastal and marine areas, ultimately yielding precise results. The initial 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.

TABLE 3. The comparison results between our method and the baseline regarding the number of parameters and floating-point operations per second (FLOPS) on the SRSDD dataset are provided.

Method	Params	GFLOPs
RTMDet(Baseline)	52.26M	204.16
Ours	74.47M	184.4

The baseline achieves a mAP of 56.00% with 204.16 GFLOPs and 52.26 M parameters, demonstrating the reliability of our baseline. Our method shows a slight decrease in the number of parameters and floating-point operations per second (FLOPS) compared to the baseline but with only a marginal decrease in FPS. Notably, the mAP improves by 10.28% compared to the baseline, indicating that our method achieves a better trade-off between speed and accuracy.

From Figure 9, 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.

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 4, our method has achieved remarkable performance, securing evaluation scores of 90.48% and 97.84% for the VOC2007 and VOC2012 benchmarks, respectively. Figure 10 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

TABLE 4. 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^2A Net[13]	R-101	90.17	95.01
AOGC[64]	R-50	89.80	95.20
MSSDet[65]	R-101	76.60	95.30
$R^3Det-KLD$[46]	R-101	89.97	95.57
MSSDet[65]	R-152	77.30	95.80
R^3Det[59]	R-101	89.26	96.01
DCFPN[46]	R-101	89.98	96.12
RTMDet(Baseline)	CSPNext-52	89.69	96.38
Ours	CSPNext-52	**90.48**	**97.84**

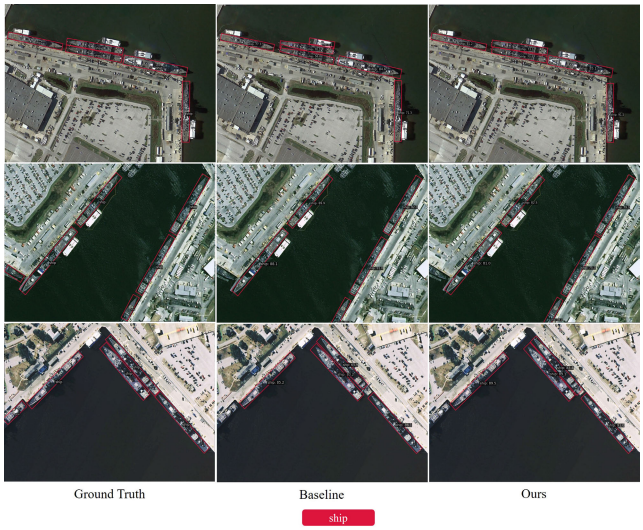

FIGURE 10. We display a subset of detection outcomes achieved using our MFCANet on the HRSC dataset. The initial column depicts actual images, the second column exhibits predictions from the baseline model, and the third column illustrates predictions from our model. Our approach demonstrates outstanding performance by producing precise and high-quality detection outcomes, especially in identifying densely clustered ships with challenging high aspect ratios.

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.

4) RESULTS ON DIOR-R

DIOR-R is a large-scale dataset characterized by an extensive array of categories and complex scenes. We compared our approach with several state-of-the-art detectors on the DIOR-R dataset, revealing that our model can extract high-quality feature maps, enabling effective category recognition and precise learning of object bounding boxes.

We selected a subset of images with complex backgrounds for visualization, and the detection results are illustrated in Figure 11. It can be observed from the figures that our proposed method accurately detects target objects in complex backgrounds. Our method effectively avoids false positives

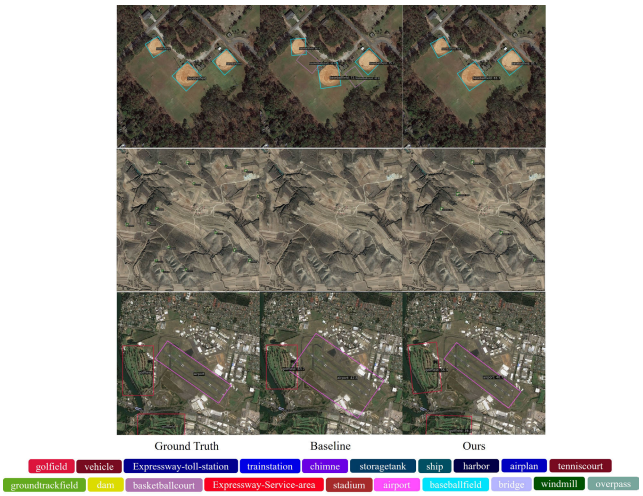

FIGURE 11. We present a series of detection results obtained by our proposed MFCANet on the DIOR-R dataset. These results highlight the capability of MFCANet to accurately extract target features in complex backgrounds, ultimately yielding precise outcomes. The initial column displays the authentic annotations of the dataset, the second column shows the baseline results, and the third column exhibits the outcomes of our method. 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.

and false negatives compared to the baseline. Although there are still some challenging scenarios where targets are difficult to observe, especially those with extremely complex backgrounds or targets not easily visible to the human eye, our method outperforms the baseline, indicating the effectiveness of our improvements.

Table 5 presents specific performance metrics for each object category. Our approach demonstrates significant improvements over the baseline, particularly in categories such as DA, APL, and APO, where the limited number of training instances per class (less than 1500) still leaves room for substantial improvement. Similarly, some small object categories (e.g., BR and VE) did not achieve optimal performance due to their small size (less than 80 pixels), presenting a challenge for accurate detection.

Overall, our method exhibits superior performance across most categories, achieving outstanding results with an accuracy of 69.82%.

D. ABLATION STUDY

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 12,

TABLE 5. Detection accuracy of different detection methods on the DIOR-R dataset. The numerical value in black bold represents the maximum.

Method	GF	VE	ETS	TS	CHI	ST	SH	HA	APL	TC	
RoI Trans[9]	69.0	43.3	78.7	54.9	72.6	70.3	81.2	47.7	63.3	81.6	
QPDet[72]	70.1	43.4	**78.9**	58.1	72.6	72.7	81.2	47.8	63.2	89.1	
AOPG[66]	73.2	52.4	65.4	60.0	72.5	71.3	81.2	42.3	62.4	81.5	
DODET[67]	74.2	51.9	78.8	55.5	72.6	71.6	85.4	48.0	63.4	81.5	
RoIF-Net[68]	78.6	50.6	74.9	**63.2**	72.7	71.2	81.3	51.1	**72.2**	**89.8**	
RTMDet(Baseline)	77.6	51.9	74.4	57.3	**77.8**	75.2	88.2	46.1	68.7	88.6	
AOPG SGIoU[69]	**79.5**	**55.9**	72.9	62.6	77.4	**78.3**	**89.7**	**52.6**	69.6	81.5	
Ours	82.5	78.0	52.8	76.8	62.3	77.4	76.1	87.8	49.3	68.7	
Method	GTF	DA	BC	ESA	STA	APO	BF	BR	WM	OP	mAP
RoI Trans[9]	82.7	26.9	87.5	68.1	78.2	37.9	71.8	40.7	65.5	55.6	63.87
QPDet[72]	83.0	28.8	88.6	69.0	72.1	41.4	72.0	41.2	65.4	55.5	64.20
AOPG[66]	81.9	31.1	87.6	78.0	72.7	37.8	71.6	40.9	70.0	54.5	64.41
DODET[67]	75.5	33.3	81.3	70.8	74.0	43.4	72.1	43.1	66.4	59.3	65.10
RoIF-Net[68]	**84.7**	34.1	89.7	88.7	**83.0**	44.0	72.2	43.9	66.5	57.5	68.49
RTMDet(Baseline)	82.2	33.8	90.1	88.0	74.1	47.0	**79.8**	**47.6**	65.1	60.2	68.69
AOPG SGIoU[69]	82.5	36.1	88.7	82.8	75.6	**53.0**	71.7	46.6	**71.0**	59.6	69.37
Ours	89.3	**37.5**	**90.1**	**89.2**	74.9	51.6	78.0	47.4	65.0	**61.8**	**69.82**

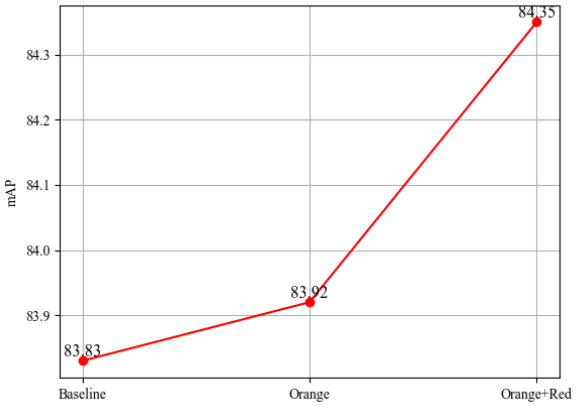

FIGURE 12. The line chart below illustrates the baseline, orange, and orange+red, representing the baseline result, the inclusion of yellow skip connections, and the simultaneous inclusion of yellow and red skip connections, respectively. The vertical axis indicates the mAP for each method on the MAR20 dataset.

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.

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 3 × 3 and 5 × 5 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 13, demonstrate that employing four sets of ODConv with the addition of attention achieves optimal performance. Analyzing the outcomes in Figure 13 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.

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

TABLE 6. 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	
✓				87.7	84.0	82.5	77.4	77.7	90.7	90.5	90.0	90.5	90.6	
✓	✓			85.4	80.5	85.4	81.0	82.7	90.8	90.8	90.1	90.5	90.8	
✓		✓		87.1	81.2	83.2	84.5	80.0	90.5	89.8	87.1	90.6	90.9	
✓			✓	87.5	87.7	85.9	83.0	81.1	90.8	90.8	90.1	90.6	90.9	
✓	✓	✓		84.6	85.3	88.9	85.9	79.2	90.7	90.5	87.6	89.2	90.9	
✓		✓	✓	88.7	84.7	84.3	85.1	81.5	90.6	90.1	90.4	90.6	90.8	
✓	✓		✓	88.3	85.0	89.9	87.3	83.1	90.8	90.5	89.4	90.7	90.9	
✓	✓	✓	✓	86.7	83.5	83.0	84.5	81.2	90.5	90.9	89.4	90.8	90.7	
Baseline	M1	M2	M3	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	mAP
✓				84.5	87.7	69.2	86.9	71.7	85.7	90.5	82.9	81.5	74.4	83.83
✓	✓			82.8	85.3	72.9	85.9	72.7	88.1	90.4	84.4	81.8	74.4	84.32
✓		✓		85.0	88.8	68.3	88.2	63.8	87.1	90.4	86.9	83.8	79.8	84.35
✓			✓	83.1	84.7	78.7	88.5	69.9	87.5	90.4	84.8	83.4	79.2	85.42
✓	✓	✓		83.6	89.6	69.8	88.6	61.3	87.3	90.5	86.4	83.4	76.8	84.51
✓		✓	✓	85.3	88.3	72.5	88.6	71.0	88.9	90.4	88.0	82.9	79.3	85.61
✓	✓		✓	85.1	88.6	71.6	86.2	73.9	88.7	90.5	82.9	83.8	78.6	85.79
✓	✓	✓	✓	85.7	88.3	78.1	88.9	76.1	88.2	90.4	88.5	83.8	79.8	85.96

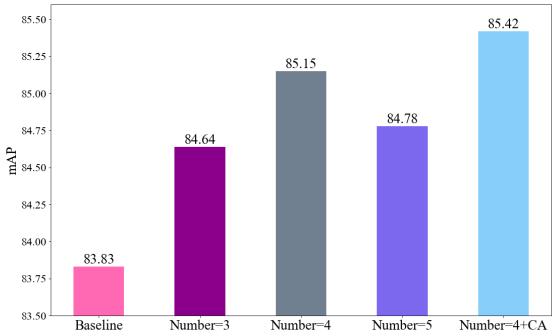

FIGURE 13. From left to right, each bar in the bar chart represents combinations of three, four, and five sets of 3 × 3 and 5 × 5 ODCLayer configurations. The last column in the bar chart corresponds to the addition of channel attention to the configurations of four sets of 3 × 3 and 5 × 5 ODCLayers. The vertical axis represents the mAP of each method on the MAR20 dataset.

in Table 6, 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.

V. 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 four 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***, Despite the commendable performance and competitiveness of our proposed method in rotated object detection in remote sensing images, it is important to note that our validation efforts have been limited to CNN architectures and a subset of datasets. Future work will focus on expanding our methodology to Transformer architectures and conducting validation on a more extensive set of datasets. This expansion aims to contribute significantly to the enhancement of accuracy in rotated object detection in remote sensing images. Incorporating Transformer architectures and testing on diverse datasets is expected to provide a more comprehensive

understanding of the method's capabilities and effectiveness across various scenarios, further solidifying its potential for real-world applications in remote sensing.

REFERENCES

[1] Z. Yang, T. Dan, and Y. Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.

[2] L. Khelifi and M. Mignotte, "Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis," *IEEE Access*, vol. 8, pp. 126385–126400, 2020.

[3] F. Lv, M. Han, and T. Qiu, "Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder," *IEEE Access*, vol. 5, pp. 9021–9031, 2017.

[4] Y. Yang, W. Wan, S. Huang, F. Yuan, S. Yang, and Y. Que, "Remote sensing image fusion based on adaptive IHS and multiscale guided filter," *IEEE Access*, vol. 4, pp. 4573–4582, 2016.

[5] B. Li, X. Xie, X. Wei, and W. Tang, "Ship detection and classification from optical remote sensing images: A survey," *Chin. J. Aeronaut.*, vol. 34, no. 3, pp. 145–163, Mar. 2021.

[6] H. Jiang, H. Peng, and G. Zhang, "CGSNet: Channel group shuffling network for remote sensing image fusion," *IEEE Access*, vol. 11, pp. 121387–121398, 2023.

[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[9] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2844–2853.

[10] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[11] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3500–3509.

[12] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[13] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3062048.

[14] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOV7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[15] D. Wan, R. Lu, S. Wang, S. Shen, T. Xu, and X. Lang, "YOLO-HR: Improved YOLOV5 for object detection in high-resolution optical remote sensing images," *Remote Sens.*, vol. 15, no. 3, p. 614, Jan. 2023.

[16] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[17] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOV6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[18] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3258666.

[19] L. Min, Z. Fan, Q. Lv, M. Reda, L. Shen, and B. Wang, "YOLO-DCTI: Small object detection in remote sensing base on contextual transformer enhancement," *Remote Sens.*, vol. 15, no. 16, p. 3970, Aug. 2023.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[22] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 385–400.

[23] F. Shen, H. Ye, J. Zhang, C. Wang, X. Han, and W. Yang, "Advancing pose-guided image synthesis with progressive conditional diffusion models," 2023, *arXiv:2310.06313*.

[24] F. Shen, X. He, M. Wei, and Y. Xie, "A competitive method to VIPriors object detection challenge," 2021, *arXiv:2104.09059*.

[25] J. Redmon and A. Farhadi, "YOLOV3: An incremental improvement," 2018, *arXiv:1804.02767*.

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOV4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[27] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOV5: Improved YOLOV5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2778–2788.

[28] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "RTMDet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[30] Y. Li, H. Wang, L. M. Dang, H.-K. Song, and H. Moon, "ORCNN-X: Attention-driven multiscale network for detecting small objects in complex aerial scenes," *Remote Sens.*, vol. 15, no. 14, p. 3497, Jul. 2023.

[31] J. Chen, H. Hong, B. Song, J. Guo, C. Chen, and J. Xu, "MDCT: Multi-kernel dilated convolution and transformer for one-stage object detection of remote sensing images," *Remote Sens.*, vol. 15, no. 2, p. 371, Jan. 2023.

[32] F. Shen, Y. Xie, J. Zhu, X. Zhu, and H. Zeng, "GiT: Graph interactive transformer for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 1039–1051, 2023.

[33] F. Shen, X. Peng, L. Wang, X. Hao, M. Shu, and Y. Wang, "HSGM: A hierarchical similarity graph module for object re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.

[34] F. Shen, X. Shu, X. Du, and J. Tang, "Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 8922–8931.

[35] S. Zheng, Z. Wu, Y. Xu, and Z. Wei, "Instance-aware spatial-frequency feature fusion detector for oriented object detection in remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3265025.

[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[37] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2235–2239.

[38] Y. Lee and J. Park, "CenterMask: Real-time anchor-free instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13903–13912.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[40] J. Hu, Z. Huang, F. Shen, D. He, and Q. Xian, "A bag of tricks for fine-grained roof extraction," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2023, pp. 678–680.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[42] F. Shen, X. Du, L. Zhang, X. Shu, and J. Tang, "Triplet contrastive representation learning for unsupervised vehicle re-identification," 2023, *arXiv:2301.09498*.

[43] H. Wu, F. Shen, J. Zhu, H. Zeng, X. Zhu, and Z. Lei, "A sample-proxy dual triplet loss function for object re-identification," *IET Image Process.*, vol. 16, no. 14, pp. 3781–3789, Dec. 2022.

[44] C. Qiao, F. Shen, X. Wang, R. Wang, F. Cao, S. Zhao, and C. Li, "A novel multi-frequency coordinated module for SAR ship detection," in *Proc. IEEE 34th Int. Conf. Tools Artif. Intell. (ICTAI)*, Oct. 2022, pp. 804–811.

[45] F. Shen, M. Wei, and J. Ren, "HSGNet: Object re-identification with hierarchical similarity graph network," 2022, *arXiv:2211.05486*.

[46] Y. Li, H. Wang, Y. Fang, S. Wang, Z. Li, and B. Jiang, "Learning power Gaussian modeling loss for dense rotated object detection in remote sensing images," *Chin. J. Aeronaut.*, vol. 36, no. 10, pp. 353–365, Oct. 2023.
[47] F. Shen, J. Zhu, X. Zhu, Y. Xie, and J. Huang, "Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8793–8804, Jul. 2022.
[48] X. Sun, G. Cheng, L. Pei, H. Li, and J. Han, "Threatening patch attacks on object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3273287.
[49] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.
[50] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.
[51] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31×31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11963–11975.
[52] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[53] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," 2022, *arXiv:2209.07947*.
[54] W. Yu, G. Cheng, M. Wang, Y. Yao, X. Xie, X. Yao, and J. Han, "MAR20: A benchmark for military aircraft recognition in remote sensing images," *Nat. Remote Sens. Bull.*, vol. 27, no. 12, pp. 2688–2696, 2023.
[55] S. Lei, D. Lu, X. Qiu, and C. Ding, "SRSDD-V1.0: A high-resolution SAR rotation ship detection dataset," *Remote Sens.*, vol. 13, no. 24, p. 5104, Dec. 2021.
[56] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.
[57] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.
[58] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu, W. Zhang, and K. Chen, "MMRotate: A rotated object detection benchmark using PyTorch," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 7331–7334.
[59] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, May 2021, pp. 3163–3171.
[60] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2149–2158.
[61] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
[62] R. Faster, "Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, vol. 9199, no. 10, pp. 2969239–2969250.
[63] Z. Shao, X. Zhang, T. Zhang, X. Xu, and T. Zeng, "RBFA-Net: A rotated balanced feature-aligned network for rotated SAR ship detection and classification," *Remote Sens.*, vol. 14, no. 14, p. 3345, Jul. 2022.
[64] Z. Wang, C. Bao, J. Cao, and Q. Hao, "AOGC: Anchor-free oriented object detection based on Gaussian centerness," *Remote Sens.*, vol. 15, no. 19, p. 4690, Sep. 2023.
[65] W. Chen, B. Han, Z. Yang, and X. Gao, "MSSDet: Multi-scale ship-detection framework in optical remote-sensing images and new benchmark," *Remote Sens.*, vol. 14, no. 21, p. 5460, Oct. 2022.
[66] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3183022.
[67] G. Cheng, Y. Yao, S. Li, K. Li, X. Xie, J. Wang, X. Yao, and J. Han, "Dual-aligned oriented detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618111.
[68] Y. Zhang, Y. Wang, N. Zhang, Z. Li, Z. Zhao, Y. Gao, C. Chen, and H. Feng, "RoI fusion strategy with self-attention mechanism for object detection in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Obser. Remote Sens.*, vol. 16, pp. 5990–6006, 2023.
[69] X. Qian, N. Zhang, and W. Wang, "Smooth GIoU loss for oriented object detection in remote sensing images," *Remote Sens.*, vol. 15, no. 5, p. 1259, Feb. 2023.
[70] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[71] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[72] Y. Yao, G. Cheng, G. Wang, S. Li, P. Zhou, X. Xie, and J. Han, "On improving bounding box representations for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 3231340.

HONGHUI JIANG (Member, IEEE) received the B.S. degree in biomedical engineering and the M.S. degree in control engineering from Hefei University of Technology, in 2015 and 2018, respectively. From 2018 to 2020, he was with ZTE Communications Company Ltd. He is currently with Anhui Technical College of Mechanical and Electrical. He has published three articles. His research interests include remote sensing image processing and deep learning.

TINGTING LUO (Member, IEEE) received the bachelor's degree in electrical engineering and automation and the master's degree in electrical engineering and control from Hefei University of Technology, in 2015 and 2018, respectively. Since 2018, she has been with State Grid Wuhu Power Supply Company. She has authored three research articles. Her research interests include electrical interlocking fault risk assessment and deep learning.

HU PENG (Member, IEEE) received the B.S. degree in radio engineering from Anhui University, in 1984, and the M.S. degree in circuits and systems and the Ph.D. degree in biomedical engineering from the University of Science and Technology of China, in 1990 and 1997, respectively. He has published more than a dozen articles and more than seven patent applications. His research interests include ultrasound imaging, electrocardiogram modeling and simulation, electroencephalography, and deep learning.

GUOZHENG ZHANG (Member, IEEE) received the B.S. degree in mechanical design, manufacturing, and automation from Anhui University of Science and Technology, in 2004, and the M.S. degree in mechatronics engineering and the Ph.D. degree in mechanical design, manufacturing, and automation from Hefei University of Technology in 2009 and 2018, respectively. He has published more than ten articles and more than ten patent applications. His research interests include modern integrated manufacturing systems and precision CNC gear machining and deep learning.