MFCA: Multiscale Feature Context Aggregation Detector for Oriented Object Detection in Remote-Sensing Images

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

Posted Date: 26 October 2023

doi: 10.20944/preprints202310.1631.v1

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

Article

MFCA: Multiscale Feature Context Aggregation Detector for Oriented Object Detection in Remote-Sensing Images

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

[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 rotational objects in remote sensing imagery is a significant challenge. These images typically encompass a broad field of view, featuring diverse and intricate backgrounds, with ground objects of various sizes densely scattered. As a result, identifying objects of interest within these images is a daunting task. While the integration of Convolutional Neural Networks (CNN) and Transformer networks leads to some advancements in rotational object detection, there is still room for improvement, particularly in enhancing the extraction and utilization of information related to smaller objects. To address this, our paper presents a multi-scale feature fusion module and a global feature context aggregation module. Initially, we fuse original, shallow, and deep features to reduce the loss of shallow feature information, thereby improving the detection performance of small objects in complex backgrounds. Subsequently, we compute the correlation of contextual information within feature maps to extract valuable insights. We name the newly proposed model the "Multiscale Feature Context Aggregation Module" (MFCA). We evaluate our proposed methodology on three challenging remote sensing datasets: DIOR-R, HRSC, and MAR20. Comprehensive experimental results show that our approach surpasses baseline models by 2.07% mAP, 1.02% mAP, and 1.98% mAP on the DIOR-R, HRSC2016, and MAR20 datasets, respectively.

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

1. Introduction

In the realm of remote sensing imagery, object detection [1–5] plays a crucial role by aiming to ascertain the presence and precise locations of objects of interest within a given image. Its applications hold substantial promise across diverse domains, including environmental monitoring, military applications, national security, transportation, forestry, and the detection of oil and gas activities. These remote-sensing images are acquired from various sources, such as aerial, satellite, and unmanned aerial vehicle platforms. However, the complexities inherent in remote sensing imagery, including intricate backgrounds, arbitrary object orientations, varying object densities, and differences in object size ratios, pose formidable challenges for small object detection. As opposed to the conventional use of horizontal bounding boxes, employing rotating bounding boxes can significantly reduce background overlap and offer a more precise delineation of object boundaries. Consequently, there is a growing imperative for research in the field of rotating object detection within remote sensing imagery.

In the context of remote sensing images, it's noteworthy that the same object can exhibit substantial differences in appearance depending on the background, leading to considerable intra-class variability. This is particularly pertinent in the case of fine-grained remote sensing images, where distinctions between object classes are less pronounced. In such scenarios, fully leveraging feature information becomes critical for achieving effective detection. To address the challenge of object scale variations

 doi:10.20944/preprints202310.1631.v1

in remote sensing images, multi-level feature pyramid networks are widely employed. Within the Feature Pyramid Network (FPN) [6] framework, higher-level feature maps contain richer semantic information but have smaller scales, making them less adept at detecting small objects. In contrast, lower-level feature maps have larger scales but lack distinctive object representations. To bridge this gap, FPN adopts a top-down lateral connection structure, facilitating the propagation of semantic information from higher-level to lower-level features, thereby enabling the detection of objects at various scales. Consequently, numerous research efforts are dedicated to further improving FPN to better accommodate the requirements of object detection in remote sensing images.

The DCFPN [7] leverages densely connected multi-path dilated layers to cover objects of various sizes in remote sensing scenes. This allows for the dense and accurate extraction of multi-scale information, further enhancing the detection capabilities for objects of varying sizes. The LFPN [8] considers both low-frequency and high-frequency features, utilizing trainable Laplacian operators to extract high-frequency object features from Laplacian pathways. Additionally, an attention mechanism is introduced within the feature pyramid network to accentuate more pronounced multi-scale object features. SPH-YOLOv5 [9] incorporates an attention mechanism into FPN, facilitating the acquisition of semantic information between features to emphasize crucial spatial features while suppressing redundant ones. Info-FPN [10] introduces a PixelShuffle-based lateral connection module (PSM) designed to fully retain channel information within the feature pyramid. Simultaneously, to alleviate confusion resulting from feature misalignment, a feature alignment module (FAM) is proposed. FAM employs template matching and learns feature offsets during the feature fusion stage to achieve feature alignment. However, existing FPN-based methods often overlook the shortcomings of the feature pyramid network structure. In particular, they do not fully leverage the original feature information and the performance issues introduced by attention mechanisms. These limitations result in a decreased feature representation capacity, which becomes more apparent when handling objects with significant scale variations in remote sensing images.

In summary, we identify several issues with current pyramid networks:

- Original features play a reinforcing role in fused features, enhancing residual functions and facilitating stable gradient propagation during backpropagation. However, feature pyramids fail to fully exploit the most original feature information.
- Convolutional neural networks are unable to aggregate information between distant pixels in the spatial domain, resulting in underutilization of long-range correlated information that adversely impacts detection results.

In this paper, we present robust solutions to address the aforementioned issues. Leveraging the RTMDet model as our baseline, we propose a multi-scale feature fusion feature pyramid to maximize information flow across all layers in the network. Additionally, we design a feature context aggregation module for fusing spatial context in feature maps, enabling comprehensive learning of inter-feature relationships. These solutions can be seamlessly integrated into object detectors, enhancing detection performance without increasing training complexity. In summary, our contributions are as follows:

- Within the feature pyramid, we efficiently harness original feature information to process multi-scale features more effectively. We introduce a multi-scale fusion pyramid network that connects original features and fused features while shortening the information transmission paths. This connection extends from large-scale features to fused small-scale features, enabling the module to optimally utilize features at each stage.
- Drawing inspiration from attention mechanisms, we design a global feature context aggregation module to aggregate feature information within feature maps and weight them adaptively for each pixel. Through iterative learning of semantic information between features, we fuse useful global information into local regions, resulting in improved pixel-level attention for objects of interest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

- We introduce a novel object detector and conduct extensive experiments on three challenging datasets: the DIOR-R dataset, the HRSC2016 dataset, and the MAR20 dataset confirming the effectiveness of our approach. Experimental results demonstrate outstanding performance.

2. Releated Work

2.1. Object Detection in General Scenarios

Computer vision technology has witnessed rapid development over the past decade, and the continuous iteration of large-scale annotated datasets has further propelled advancements in object detection tasks. These methods can be broadly categorized into two major groups: those based on convolutional neural networks and those leveraging attention mechanisms. Within CNN models, we have one-stage detection models (such as SSD [11], RetinaNet [12], R^2ANet [13], YOLO series [9,14–17], RTMDet [18], etc.) and two-stage models (R-CNN [19], Fast R-CNN [20], Faster R-CNN [21], R-FCN [22], etc.). These models have achieved commendable results; however, models based on CNNs can render very small objects undetectable due to downsampling during the process. To address the issue of detecting small objects, FPN and their variants [23,24] were introduced, which improved small object detection. Nonetheless, this introduced new challenges, including increased computational complexity, the need for parameter adjustments in FPN, and the introduction of cross-level connections that may lead to incomplete feature map matching, resulting in inaccurate predictions at the boundaries. In addition, some researchers have introduced attention mechanisms into CNNs [9,25–27], to some extent enhancing the accuracy of object detection. Methods combining attention with convolution capture both static and dynamic contextual information in images. They possess self-attention learning capabilities while incorporating contextual information. Furthermore, some researchers have transformed temporal information into the frequency domain through techniques like wavelet and Fourier transforms [8,28], subsequently extracting frequency domain features, which have yielded promising results. Various approaches have been proposed from different perspectives, designing a series of channel weight-solving methods to adaptively learn the importance of each channel and weight each channel feature map [29–31], all of which have demonstrated favorable results.

In recent years, Transformer-based models [32–35] have shown promising results in the field of object detection. The Vision Transformer (ViT) [32]demonstrated that Transformers can be applied to computer vision with minimal modifications and achieve excellent performance. The DETR [33] 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 [34] 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 currently widely applied in the field of object detection in remote sensing imagery. A series of CNN-based remote sensing object detection approaches have emerged and have yielded promising results.

 doi:10.20944/preprints202310.1631.v1

To address the challenge of multiscale detection due to varying object sizes in remote sensing imagery, mSODANet [36] employs parallel dilated convolutions to explore a hierarchical dilation network, enabling the learning of contextual information for different object types across multiple scales and fields of view. The introduced hierarchical dilation network effectively captures visual information in aerial images, enhancing the model's detection capabilities. The Super-Yolo model [37] integrates multimodal data, utilizes auxiliary super-resolution learning, and considers both detection accuracy and computational cost for high-resolution object detection of multiscale objects. MFAF [38] proposes a multiscale feature-adaptive fusion method, utilizing multiscale feature integration modules and spatial attention weight modules to construct a feature fusion module, enabling adaptive fusion of multiscale features. MDCT [24] introduces a single-stage object detection model based on multi-kernel dilated convolution blocks and Transformer blocks. This enhances the intrinsic and neighboring spatial features of small objects, and Transformer blocks are integrated into the model's neck network to prevent the loss of object information in complex backgrounds and dense scenes. ANSDA [39] leverages NASFPN for feature extraction and introduces context enhancement modules and channel attention modules to enhance the feature extraction capabilities for shallow-level features and small object semantics. ORCNN-X [23] adopts a dynamic attention module and an efficient feature fusion mechanism in a multiscale feature extraction network to enhance the model's perception capabilities and handle scale and orientation variations. DCFPN [7] designs a Dense Context Feature Pyramid Network and Gaussian loss for rotation object detection. It uses dense multi-path dilated layers to densely and accurately extract multiscale information, addressing the discontinuity issues in boundary regression through the Gaussian loss function, resulting in favorable performance. ESRTMDet [40] designs a lightweight embedded feature map super-resolution module, embedding it into PAFPN to enhance and magnify the backbone's output features, making it easier for the detection head to detect small objects. HFAN [41] 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 [42] addresses the challenge of global modeling of pixel-level information for small objects by designing a context transformer framework and embedding it into the detection head for small object detection. SPH-Yolo [9] incorporates the Swin-Transformer into PAFPN to more effectively detect objects of different scales.

In addition, Some researchers explore anchor-free mechanisms as alternatives to anchors based on rotation object detection. AOPG [43] generates coarse-oriented boxes in an anchor-free manner using a coarse localization module and then refines them into high-quality-oriented proposals. FCOS [44] proposes a fully convolutional single-stage object detector that solves object detection in a per-pixel prediction manner, completely avoiding the complex computations associated with predefined anchor boxes. CLU [45] introduces a method for training unsupervised object detection, leveraging the characteristics of self-supervised models to "discover" objects without supervision. H2RBOX [46], employ weakly supervised training using horizontal bounding box annotations to achieve rotation box object detection. Specifically, they use weakly supervised learning and self-supervised learning to predict the object's angle by exploiting the consistency between two different views, yielding promising results.

Sparse and dense small objects in remote sensing images occupy a significant proportion, placing high demands on feature extraction networks. Typically, CNNs extract features with translational invariance, excelling at capturing local information. However, they fall short in extracting contextual information from features. On the other hand, attention mechanisms excel at global modeling to acquire contextual information for feature maps. Therefore, combining these two approaches can harness their respective strengths and yield features more conducive to detection. Building upon the insights mentioned above, we propose an improved PAFPN-based single-stage object detection model, leveraging the foundation of RTMDet. We aspire that our work will contribute to the advancement of object detection in remote sensing imagery.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

3. Methodology

3.1. Basic Rotated Detection Method as Baseline

In previous works, the detection of rotated bounding boxes was not considered, and horizontal bounding boxes were commonly used to delineate objects [47,48]. However, in remote sensing images, there is a significant proportion of small objects. Traditional horizontal bounding box annotations introduce background information that is not conducive to accurate object localization. Rotated bounding boxes, on the other hand, enable precise object localization with minimal background inclusion. Furthermore, rotated bounding boxes rarely overlap, allowing for clear delineation of the objects within them. Therefore, it is imperative to investigate and utilize more accurate rotated bounding box representations for object detection in remote sensing images. The representation of rotated bounding boxes (RBB) is typically defined as follows:

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

Where, $\theta \in [-\pi/2, \pi/2]$, represents the clockwise rotation angle from the image coordinate system's direction X to the bounding box's relative coordinate system's direction X. We adopt the long-edge-based format [49], where the width w must be greater than the height h. We employ the one-stage rotation object detector RTMDet [18] for detecting both sparse and dense small objects in remote sensing images. RTMDet is an enhancement based on YOLOX [50], sharing a similar overall macro-architecture with the YOLO series. RTMDet employs CSPDarkNet [15] as its baseline and utilizes CSPPAFPN, composed of the same building units, for multi-scale feature fusion. Subsequently, features are fed into different detection heads to perform tasks such as object detection, instance segmentation, and rotation bounding box detection. The overall model structure is illustrated in Figure 1.

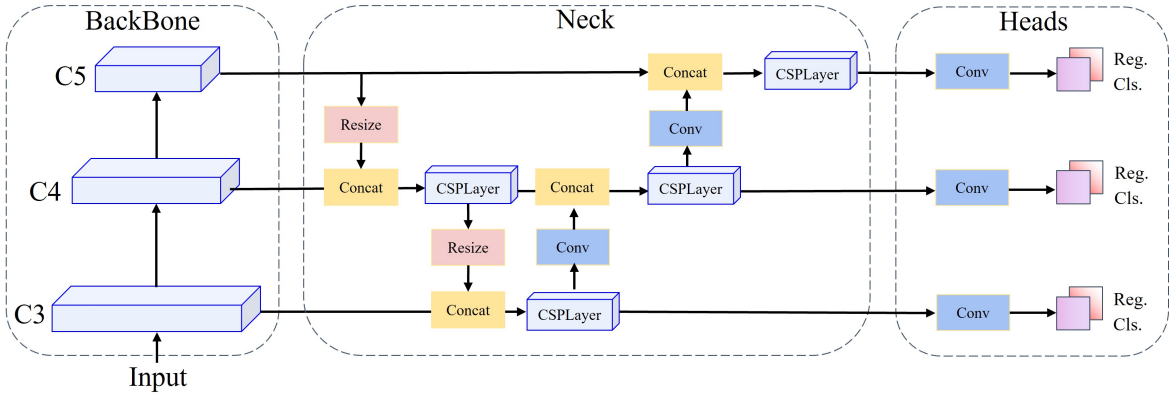

Figure 1. Baseline macro-architecture.

In particular, RTMDet consists of CSPNeXt, CSPNeXtPAFPN, and SepBNHead, which share convolutional weights but compute batch normalization separately. Additionally, it draws inspiration from the practices of ConvNeXt [51] and RepLKNet [52], enhancing feature extraction capabilities with large kernel convolutions in the Basic Block. The authors also employ a dynamic SimOTA approach for rotation object detection, using DistanceAnglePointCoder for Bbox encoding and decoding. RTMDet introduces a Dynamic Soft Label Assigner to implement a dynamic matching strategy for labels. This method primarily includes the use of prior position information loss, sample regression loss, and sample classification loss, with soft processing applied to these three losses for parameter tuning to achieve the optimal dynamic matching effect. After calculating the sum of these three losses to obtain the final cost matrix, SimOTA is then used to determine the number of matched samples for each ground truth (GT) and determine the final samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

3.2. Multiscale Feature Fusion Network

In remote sensing images, objects often vary significantly in size, necessitating that the feature maps output by neural networks cover a range of receptive field scales to extract comprehensive object features. PAFPN [53] first employs a bottom-up structure to extract feature maps at different scales and then upsamples these feature maps using a top-down structure. Finally, it combines the downsampled and upsampled results through lateral connections, ultimately outputting feature maps at higher pyramid levels to incorporate stronger semantic information. However, the PAFPN model has certain limitations in detecting small objects. Due to the small feature regions of such objects, the PAFPN model partitions the image into multiple scales via feature pyramids, potentially leading to the neglect or misclassification of small objects during feature extraction. Moreover, multiple fusions can dilute crucial features since feature fusion reduces the clarity of feature maps. Diluted features cannot provide sufficient information for small object detection. Hence, there is a need to optimize and adjust the feature fusion mechanism of the PAFPN model to enhance its performance.

Figure 2 illustrates the model structure we propose. Firstly, this model introduces lateral skip connections to establish direct connections between the original features and the fused feature maps, enabling more effective utilization of features from the original feature maps to enhance the model's performance. Secondly, we introduce two connections to fuse top and bottom pyramid information, reducing the path length for information transfer and effectively extracting feature information from low-resolution feature maps. Since both of these methods are based on feature fusion, combining them essentially does not increase computational costs. The whole process is described as follows.

$$\begin{cases} P_3 = f_3(g_3(r_4(f_4(g_4(r_5(C_5) + C_4))) + C_3) + C_3) + C_3 \\ P_4 = f_4(g_4(f_4(g_4(r_5(C_5) + C_4)) + h_3(P_3)) + C_4) + C_4 \\ P_5 = f_5(g_5(C_5 + h_4(P_4)) + C_5) + C_3 + C_5 \end{cases} \tag{2}$$

Where C3, C4, and C5 represent the features extracted by the backbone, while P3, P4, and P5 correspond to the results of feature fusion. Function f signifies the CSPLayer operation, function g represents channel-wise concatenation, function r represents 2x nearest-neighbor upsampling and function h denotes downsampling achieved using a 3x3 convolution kernel with a stride of 2. The subscripts accompanying 'f,' 'g,' 'r,' and 'h' denote the respective layers, with values ranging from 3 to 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

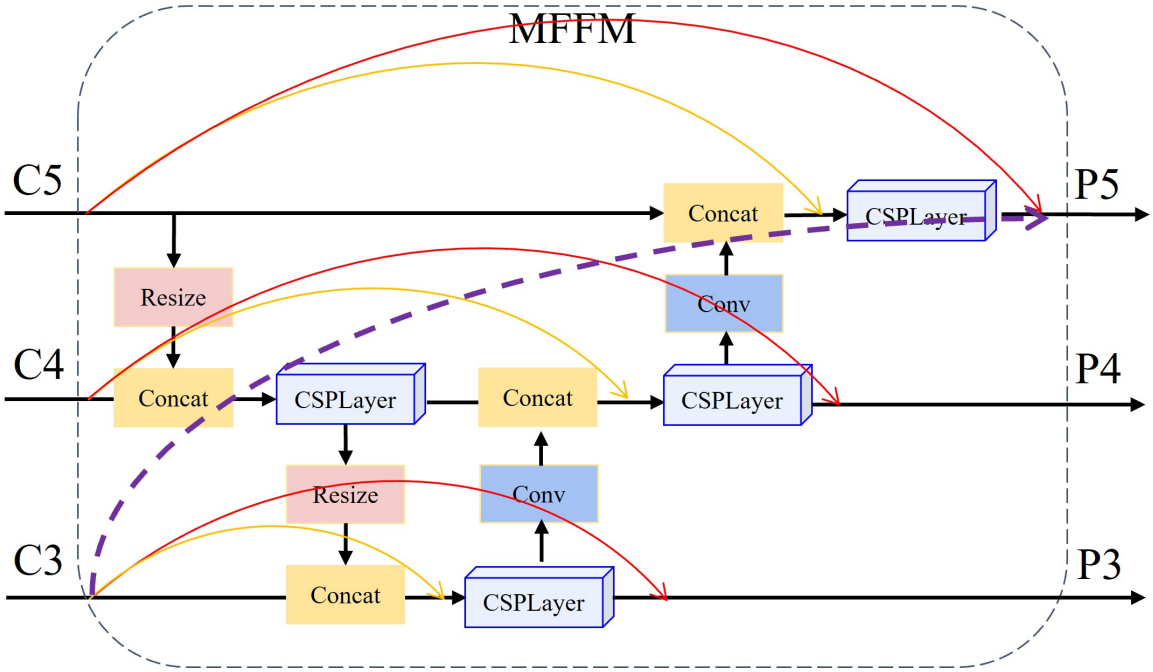

Figure 2. The multiscale feature fusion network. The red solid line represents the residual connection, fusing the original output features with the intermediate and final outputs of PAFPN. The deep red dashed line indicates the fusion of shallow information from bottom to top with deep-layer information. The deep yellow dashed line represents the fusion of intermediate-level information with deep-layer information. 1x1 convolutional kernels are used for channel dimension adjustment. 4x4/4 convolutional kernels perform downsampling with a stride of 4. 3x3/2 convolutional kernels perform downsampling with a stride of 2.

3.3. Global Feature Content Aggregation Module

The feature pyramid retains local information after aggregating feature maps from different levels. To address this, the Non-local neural networks [54] incorporate attention modules into the convolution to achieve a global receptive field. However, in the context of small object detection in remote sensing images, this design may introduce some irrelevant background information, thus increasing the detection difficulty. Therefore, we have devised the Global Feature Context Aggregation Module (GFCAM), as illustrated in Figure 3. This module employs three 1x1 convolutions to obtain three matrices from the input features, followed by feature context relevance calculation based on attention mechanisms. This process enhances the feature by learning the global feature context within each level. Given the effectiveness of residual structures in models such as ResNet and DenseNet, we have incorporated residual connections into the structure to effectively fuse local and global features while reducing information confounding.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

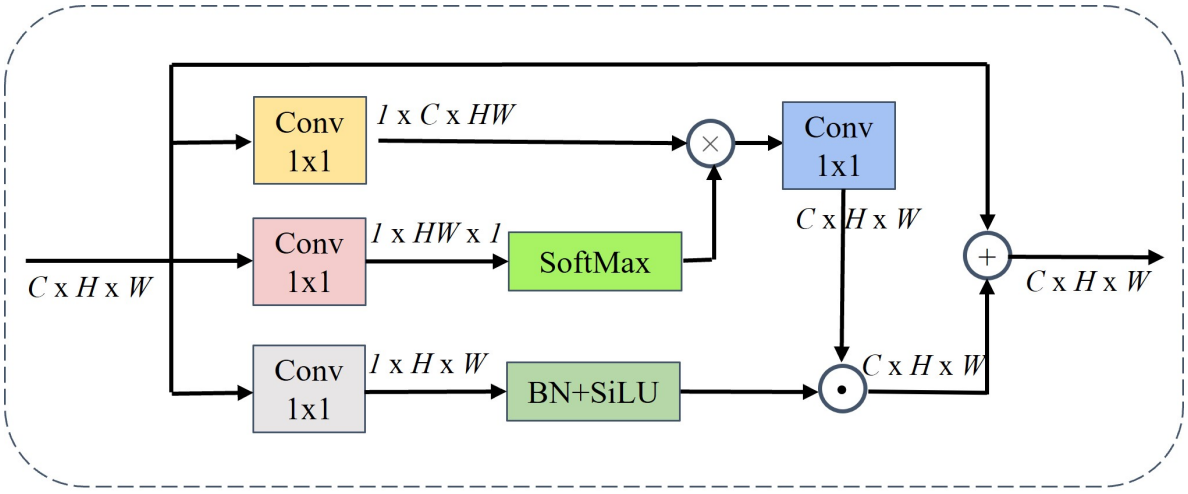

Figure 3. GFCAM, conv1x1 represents a convolutional kernel of size 1, which reshapes the tensor dimensions after the convolution operation. Softmax, Batch Normalization (BN), and SiLU are employed for normalization purposes. Three 1x1 convolutions are used to adjust the tensor dimensions and perform normalization, followed by matrix operations to obtain relevant information on the global feature context.

GFCAM is mathematically described as follows:

$$Y = X + f_1^{1x1}(f_2^{1x1}(X) \otimes g(f_3^{1x1}(X))) \odot \sigma(BN(f_4^{1x1}(X))) \tag{3}$$

Where, X represents input feature values, while C, H, and W respectively denote the number of channels, height, and width of the feature map. The symbol $\sigma(\cdot)$ represents the activation function SiLU, and $BN(\cdot)$ stands for Batch Normalization. The function f signifies the convolution operation, where the superscript indicates the kernel size, and the subscript distinguishes convolution operations aimed at generating different dimensions. The symbol $\otimes$ denotes matrix multiplication, $\odot$ represents the standard multiplication operation within tensors, and the '+' symbol signifies tensor addition operations.

3.4. MFCA

Figure 4 illustrates the overall architecture of our proposed multi-scale feature context aggregation network, which is based on RTMDet. In essence, it consists of a feature extraction module, a feature pyramid module, and prediction heads. The backbone network extracts features at three different scales to handle objects of various sizes in the context of object detection. We integrate the original features with the output features on the basis of PAFPN. Additionally, to minimize feature information loss during propagation, connections from C3 to C5 and from C4 to C5 are introduced. Finally, the fused feature information undergoes context aggregation through our designed GFCAM, aiming to obtain information that better reflects real features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

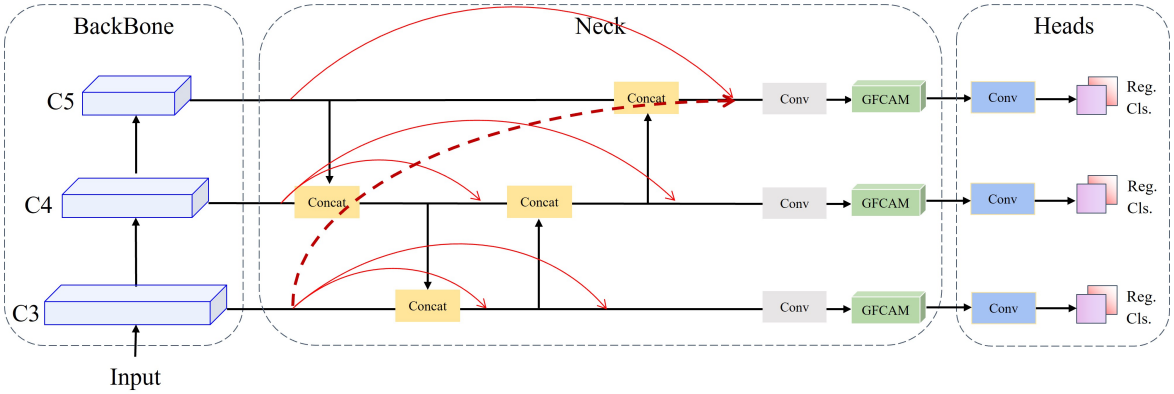

Figure 4. The architecture of MFCA. In comparison with the original version, the red and yellow lines represent the fusion of original features and the reduction of information loss during the propagation process. GFCAM is employed to perform context aggregation on the fused global feature information.

4. Experiments

In this subsection, we assess the effectiveness of our proposed model through training and testing on three widely used datasets: DIOR-R, HRSC2016, and MAR20. We present a comprehensive overview of our experiments, including experimental design and parameter configurations, and compare our model with current state-of-the-art models and experimental results. Furthermore, we conducted an ablation study on the DIOR-R dataset to demonstrate the effectiveness of each module. Our software environment comprises CUDA11.8, Python 3.8.10, PyTorch 2.0, mmdetection3.1.0, and mmrotate1.x, while our hardware setup includes an Intel(R) Xeon(R) Platinum 8350C @ 2.60GHz, NVIDIA GeForce RTX 3090, and 80GB of memory. We employ a two-stage training approach, initially using Mosaic and MixUp [12] without rotation for training. In the final 10 epochs, we fine-tune the model with a smaller learning rate under weaker augmentation. All experiments utilize the AdamW optimizer with a base learning rate of 0.00025, a momentum of 0.9, and a weight decay of 0.05.

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

DIOR-R: [55] The DIOR-R 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 800x800 pixels, with pixel resolutions ranging from 0.5 meters to 30 meters.

HRSC2016 [56] is another widely-used arbitrary-oriented object detection benchmark. It contains 1061 images with sizes ranging from 300×300 to 1500×900. The training set (436 images) and validation set (181 images) are used for training and the remaining for testing. For the evaluation metrics on the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

HRSC2016, we report the COCO style means average precision (mAP) as well as the average precision under the 0.5 and 0.75 thresholds (AP50 and AP75).

MAR20 [57] is currently the largest publicly available dataset for military aircraft target recognition in remote sensing imagery. It comprises 3842 images, featuring 20 different military aircraft models with a total of 22341 instances. The majority of the images have a resolution of 800×800 pixels. These instances were collected from 60 military airfields located in the United States, Russia, and other countries, utilizing Google Earth imagery. The MAR20 dataset includes a specific set of 20 aircraft models. Among them, six are Russian aircraft, including 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 aircraft models consist of U.S. aircraft, such as 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 represented by abbreviations A1 to A20. The training set consists of 1331 images and 7870 instances, while the test set comprises 2511 images and 14471 instances.

4.1.2. Evaluation Metrics

In the experiment, various commonly used RSOD (Remote Sensing Object Detection) metrics are employed to assess the effectiveness of the proposed model. This paper employs Average Precision (AP) as the performance evaluation metric for the object detection model. The calculation formula for AP is as follows:

$$\begin{cases} P = \frac{TP}{TP+FP} \\ r = \frac{TP}{TP+FN} \\ AP = \int_0^1 p(r)\, dr \end{cases} \tag{4}$$

TP represents the number of correctly classified targets, FP is the count of background identifications as targets, and FN signifies the number of object identifications misclassified as background. Precision (p) indicates the ratio of correctly identified targets among all detected results. Recall (r) represents the ratio of correctly identified targets to the true values of all targets. The area enclosed by the curve with p on the vertical axis, r on the horizontal axis, and the coordinate axes is the AP value. The AP metric takes into account both precision and recall, and a higher AP value indicates higher detection accuracy. The mean Average Precision (mAP) for each class is calculated using the following formula:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} \int_0^1 P_i(R_i)\, dR_i \tag{5}$$

Here, N refers to the number of object categories. mAP@0.5 denotes the mean average precision for all classes at an Intersection over the Union (IoU) threshold of 0.5. mAP@0.5:0.95 signifies the average mAP across IoU thresholds ranging from 0.5 to 0.95. IoU, which stands for Intersection over Union, is a metric used to assess the degree of overlap between two regions. The computation formula is as follows:

$$IoU = \frac{area(X) \cap area(Y)}{area(X) \cup area(Y)} \tag{6}$$

In the equation, X represents the object box predicted by the model, and Y represents the real object box in the image.

4.2. Implementation Details

We conduct experiments using RTMDet [18] from the MMRotate toolbox [58]. Our experiments adhere to the configuration employed in RTMDet, where CSPNetXtBlock serves as the backbone and CSPNetXt-PAFPN functions as the neck. During the initial stages of model training, we apply various

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

data augmentation techniques, including random flipping, rotation, CachedMosaic, and CachedMixUp. In the final 10 epochs, we modify the augmentation strategy, eliminating CachedMosaic and CachedMixUp while retaining the remaining original RTMDet model training configuration. No augmentation techniques are used during the testing and inference phases. In comparative experiments, we maintain consistent hyperparameter settings throughout the training process to ensure a fair comparison with other state-of-the-art methods. The learning rate undergoes linear decay in the first half of training and cosine decay in the second half.

Regarding the processing of the HRSC2016 and MAR20 datasets, we crop the original images into patches of 800x800 pixels with a 200-pixel overlap between adjacent patches. We use the training portion for training, the test portion for validation, and inference. As for the DIOR-R dataset, the image sizes remain unchanged, all being 800x800 pixels. We utilize the trainval portion for training, the val portion for validation, and the test portion for inference. We conduct training for 50 epochs on the DIOR-R and MAR20 datasets, while the HRSC2016 dataset is trained for 100 epochs to obtain the inference model.

4.3. Comparisons with State-of-the-Art

We compare our proposed method with other SOTA approaches on the DIOR-R, HRSC2016, and MAR20 datasets. As shown in the table, without unnecessary elaboration, our method demonstrates superior performance compared to the SOTA approaches.

4.3.1. Results on DIOR-R

DIOR-R is a large-scale dataset characterized by an extensive array of categories and complex scenes. We have compared our approach to several SOTA detectors on the DIOR-R dataset. The proposed model extracts high-quality feature maps, enabling effective category recognition and precise learning of object bounding boxes. We have chosen various categories of objects at different scales and scenes with both dense and sparse object arrangements for visualization. The detection results are illustrated in the Figures 5. 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. Thanks to the utilization of large kernel convolutions in RTMDet, CSPNextBlock, and data augmentation strategies during training, our baseline accuracy surpasses the current SOTA by a significant margin. For individual categories like DA, TS, and ST, the detection results are still subject to considerable improvement due to the limited number of training instances for each class, which is less than 1500. Similarly, some small object categories (e.g., BR and VE) have not achieved optimal performance due to their small size, which is less than 80 pixels, making accurate detection challenging. Overall, our approach outperforms most categories and achieves an outstanding performance of 74.51%.

Table 1. Detection Accuracy of Different Detection Methods on the DIOR-R Dataset. The color red is indicative of the highest value, while blue represents the second-highest value.

Method	Backbone	GF	VE	ETS	TS	CHI	ST	SH	HA	APL	TC	mAP
RoI Trans [59]	R-50	69.0	43.3	78.7	54.9	72.6	70.3	81.2	47.7	63.3	81.6	
AOPG [43]	R-50	73.2	52.4	65.4	60.0	72.5	71.3	81.2	42.3	62.4	81.5	
ROIF [60]	R-50	74.7	49.4	69.5	55.0	73.8	63.9	82.4	47.4	72.1	82.7	
ROIF [60]	ConvNext-50	78.6	50.6	74.9	63.2	72.7	71.2	81.3	51.1	72.2	89.8	
AOPG SGIoU [61]	R-50	79.5	55.9	72.9	62.6	77.4	78.3	89.7	52.6	69.6	81.5	
RTMDet [18]	CSPNext-52	75.8	57.3	76.1	63.8	79.8	79.6	89.8	53.2	90.4	90.5	
Ours	CSPNext-52	77.9	61.1	79.1	64.9	80.7	80.2	90.1	54.3	90.7	90.7	
Method	Backbone	GTF	DA	BC	ESA	STA	APO	BF	BR	WM	OP	
RoI Trans [59]	R-50	82.7	26.9	87.5	68.1	78.2	37.9	71.8	40.7	65.5	55.6	63.87
AOPG [43]	R-50	81.9	31.1	87.6	78.0	72.7	37.8	71.6	40.9	70.0	54.5	64.41
ROIF [60]	R-50	84.0	29.2	82.6	78.1	80.7	39.0	72.9	40.8	67.4	55.5	65.12
ROIF [60]	ConvNext-50	84.7	34.1	89.7	88.7	83.0	44.0	72.2	43.9	66.5	57.5	68.49
AOPG SGIoU [61]	R-50	82.5	36.1	88.7	82.8	75.6	53.0	71.7	46.6	71.0	59.6	69.37
RTMDet [18]	CSPNext-52	84.6	35.8	90.3	89.2	85.0	49.0	84.8	46.3	65.9	61.7	72.44
Ours	CSPNext-52	84.8	42.0	90.5	89.1	86.6	53.0	88.5	50.2	73.2	62.8	74.51

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

Figure 5. Here are some detection results of our proposed MFCA model on DIOR-R. Each color represents a distinct category, and the displayed results include six classes: ships, harbors, airplanes, vehicles, windmills, and expressway toll stations. It is evident that the MFCA module excels at identifying dense small targets amidst complex backgrounds.

As shown in Figure 6, the CSPPAFPN model in the baseline fails to extract sufficient features for the objects of interest. In contrast, the inclusion of our proposed module results in an enhancement of its feature extraction capabilities, thereby conferring a distinct advantage in the detection of various targets within remote sensing images. The extracted features exhibit greater prominence, enhanced spatial clarity, and improved localization precision, substantiating the efficacy of our approach in acquiring more robust feature information and achieving performance improvements.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

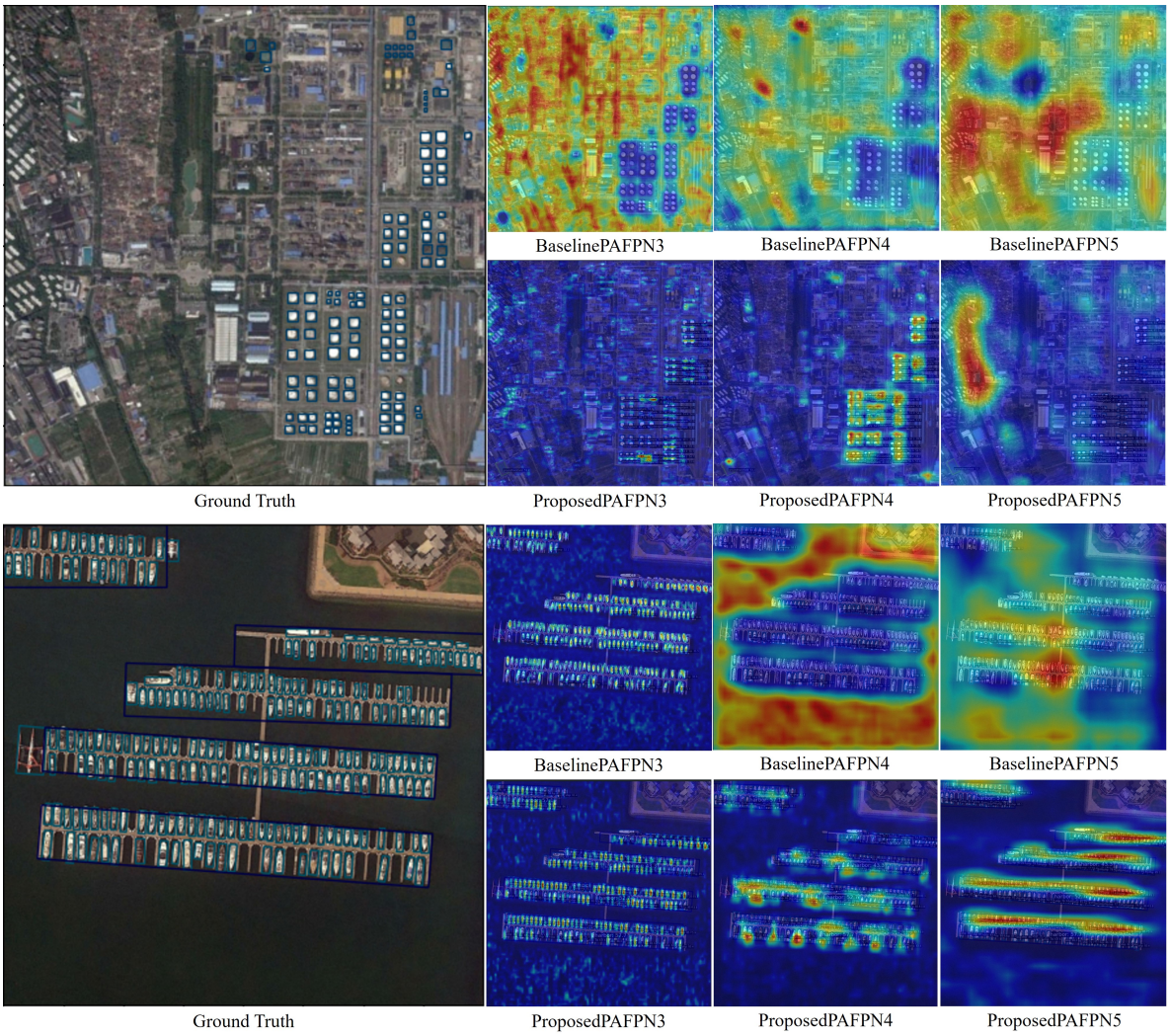

Figure 6. In DIOR-R, we conduct a comparative visualization of feature maps between the baseline and the MFCA module. After the Backbone extracts features and PAFPN fuses them, we visualize feature maps from various layers. The color blue represents the background, while brighter regions in red and yellow indicate heightened attention responses. A comparison reveals that the MFCA model, as proposed, effectively suppresses background information and focuses more on densely packed small target regions.

4.3.2. Results on HRSC2016

HRSC2016 dataset consists of vessels with high aspect ratios, sailing in arbitrary directions, presenting a significant challenge for precise object localization. Our proposed model possesses robust feature extraction capabilities, emphasizing global information within the feature maps, effectively identifying class-specific features, thus yielding superior performance. As demonstrated in the Table 2, our approach achieves commendable results, attaining evaluation scores of 90.05% and 97.53% for the VOC2007 and VOC2012 benchmarks, respectively. Figure 7 showcases the visual outcomes of our method on the HRSC2016 dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

Table 2. Detection Accuracy of Different Detection Methods on the HRSC2016 Dataset. The color red is indicative of the highest value, while blue represents the second-highest value.

Method	Backbone	mAP (07)(%)	mAP (12)(%)
S^2ANet[62]	R-101	90.17	95.01
AOGC [63]	R-50	89.80	95.20
MSSDet [64]	R-101	76.60	95.30
$R^3Det-KLD$[7]	R-101	89.97	95.57
MSSDet [64]	R-152	77.30	95.80
R^3Det[65]	R-101	89.26	96.01
DCFPN [7]	R-101	89.98	96.12
RTMDet [18]	CSPNext-52	89.10	96.51
Ours	CSPNext-52	90.05	97.53

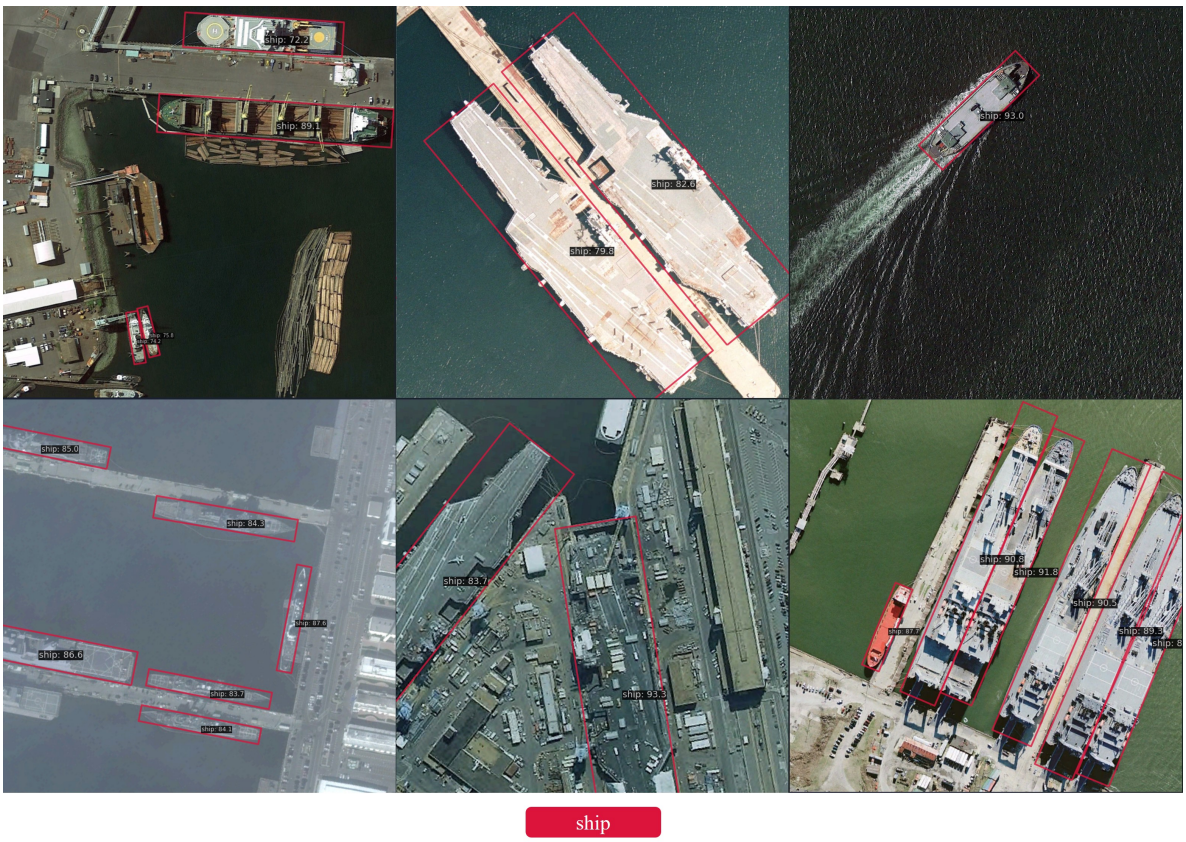

Figure 7. We present a selection of detection results achieved by our proposed MFCA on the HRSC2016 dataset. These outcomes emphasize MFCA's capacity to accurately extract target features, even when dealing with complex backgrounds, ultimately leading to precise results.

4.3.3. Results on MAR20

MAR20 is a fine-grained dataset designed for military aircraft detection, encompassing a wide range of target scales. This dataset contains remote sensing images captured under various climatic conditions, different seasons, and varying lighting scenarios. Thanks to the robust feature extraction and information learning capabilities within our proposed model, our model's inference outperforms all existing detectors, achieving a top mAP of 92.41%. The results of our approach to the MAR20 dataset are presented in Table 3. The detection results are illustrated in the Figure 8.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

Table 3. Detection Accuracy of Different Detection Methods on the MAR20 Dataset. The color red is indicative of the highest value, while blue represents the second-highest value.

Method	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	mAP
S^2A-Net[57]	82.6	81.6	86.2	80.8	76.9	90.0	84.7	85.7	88.7	90.8	
Faster R-CNN [57]	85.0	81.6	87.5	70.7	79.6	90.6	89.7	89.8	90.4	91.0	
Oriented R-CNN [57]	86.1	81.7	88.1	69.6	75.6	89.9	90.5	89.5	89.8	90.9	
RoI Trans [57]	85.4	81.5	87.6	78.3	80.5	90.5	90.2	87.6	87.9	90.9	
RTMDet [18]	85.5	96.0	94.6	90.9	86.0	90.9	95.1	98.7	90.9	90.9	
Ours	88.6	98.7	98.4	90.7	87.5	95.1	94.9	99.2	90.9	99.0	
Method	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	
S^2A-Net[57]	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 [57]	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 [57]	87.6	88.4	67.5	88.5	46.3	88.3	90.6	70.5	78.7	80.3	81.9
RoI Trans [57]	85.9	89.3	67.2	88.2	47.9	89.1	90.5	74.6	81.3	80.0	82.7
RTMDet [18]	82.8	90.7	88.8	90.1	84.6	90.5	90.7	94.8	86.6	89.4	90.43
Ours	89.6	90.7	89.7	90.3	89.1	90.5	90.6	97.6	87.2	89.9	92.41

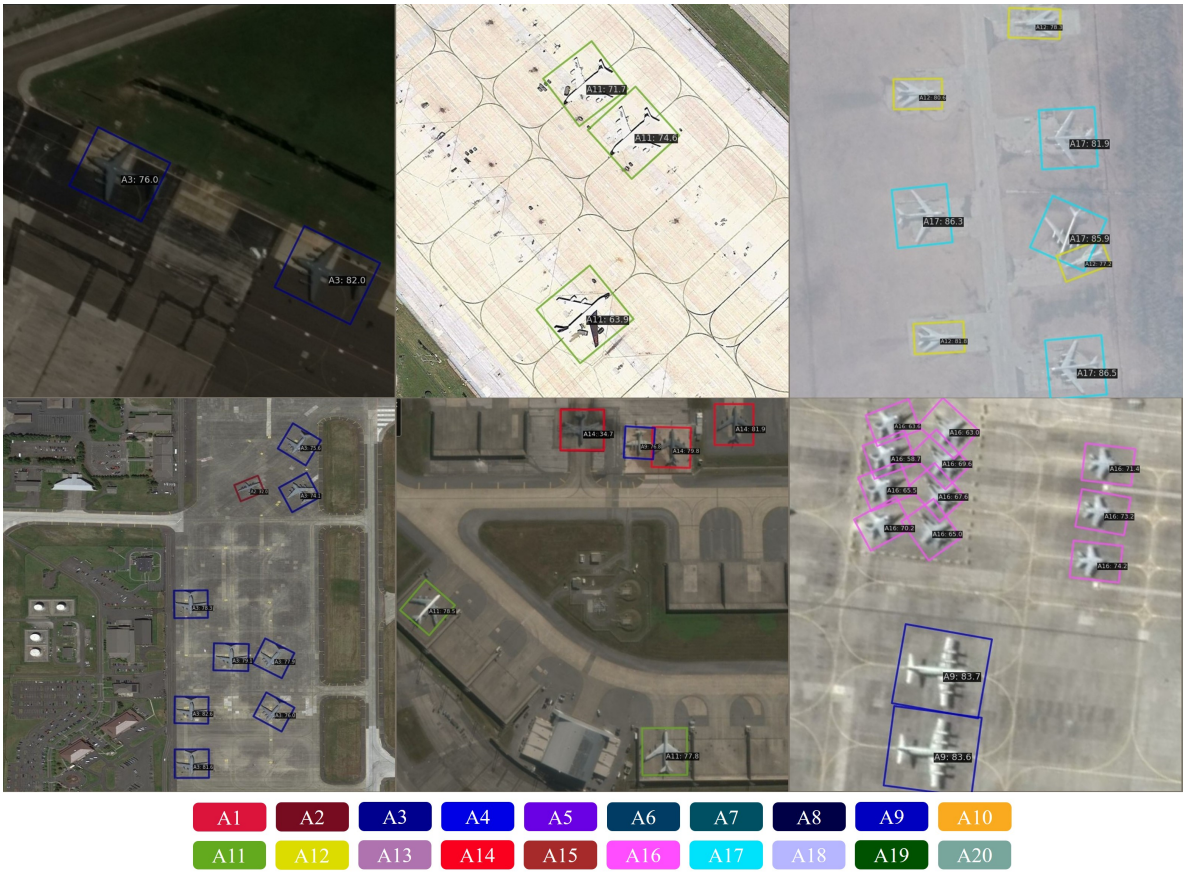

Figure 8. The results predicted by our proposed method on the MAR20 dataset, which comprises 20 different categories, are presented in the image. The displayed detection results pertain to classes A2, A3, A9, A11, A12, A14, A16, and A17, with their corresponding bounding box colors and categories as illustrated in the figure.

4.4. Ablation Study

4.4.1. Ablation Test with Different Feature Fusion Methods in MFFM

To offer an in-depth analysis of the augmented function of original features within the fusion process involving PAFPN features, we undertake an ablation experiment on the skip connections within the Multi-Feature Fusion Module (MFFM). In Figure 2, skip connections of varying colors serve as modules for the ablation experiment, designated in red, orange, and purple. We compare the variances in the fusion of original features with PAFPN concerning the baseline RTMDet on the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1631.v1

MAR20 dataset.

4.4.2. Ablation Test of MFCA

To validate the effectiveness of each module proposed in this study, we compared the original CSPNext-PAFPN with each enhancement module on the MAR20 dataset, considering RTMDet as the baseline for detection. The evaluation primarily focuses on the Average Precision (AP) and mean Average Precision (mAP) of typical object categories, including A3, A6, A10, A11, A15, and A18. Given the similarity among fine-grained objects in remote sensing images and the complexity of backgrounds under different seasons and lighting conditions, their detection becomes challenging.

Through the ablation experiments conducted for each enhancement, the recognition performance of some challenging targets is presented in Table 5. By comparing experiments one and two, our proposed multi-scale feature fusion network demonstrates a superior ability to represent multi-scale target features compared to the baseline PAFPN. This superiority stems from our comprehensive utilization of original features for enhanced feature fusion, resulting in improved model performance. Comparing experiments one and three, it becomes evident that GFCAM significantly enhances the model's detection capabilities. The Global Feature Context Aggregation Module filters out background interference and enriches target feature information, thus augmenting the model's sensitivity to targets.

Table 4. In our research, we conduct an ablation study on various fusion methods for combining original features with PAFPN within the MFFM framework, utilizing the MAR20 dataset. The distinctive fusion methods are represented by the colors red, orange, and purple. These colors align with the fusion methods illustrated in Figure 2. As each fusion method is introduced independently, it results in an improvement in detection accuracy when compared to the baseline. Remarkably, the collective combination of all fusion methods leads to a noteworthy enhancement in detection accuracy.

Baseline	Red	Orange	Purple	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	mAP
✓				85.5	96.0	94.6	90.9	86.0	90.9	95.1	98.7	90.9	90.9	
✓	✓			87.7	98.1	93.5	90.9	86.2	90.8	93.7	99.1	90.8	94.3	
✓	✓	✓		86.3	90.8	98.7	90.6	87.2	92.2	95.3	99.3	90.9	99.7	
✓	✓	✓	✓	88.1	90.7	97.0	90.8	86.5	97.7	95.9	99.3	90.9	98.6	
Baseline	Red	Orange	Purple	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	
✓				82.8	90.7	88.8	90.1	84.6	90.5	90.7	94.8	86.6	89.4	90.43
✓	✓			88.6	90.8	89.5	90.3	87.9	90.5	90.6	94.2	86.8	89.2	91.17
✓	✓	✓		85.4	90.4	89.7	90.5	83.4	90.5	90.8	96.0	90.0	90.3	91.40
✓	✓	✓	✓	88.3	90.7	89.6	90.0	86.7	90.3	90.8	95.9	88.2	89.2	91.76

Table 5. Ablation studies of the individual modules proposed by us on the MAR20 dataset. When each module is added independently, the detection accuracy is improved compared to the baseline. Notably, when all modules are combined, there is a substantial increase in detection accuracy.

Baseline	MFFN	GFCAM	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	mAP
✓			85.5	96.0	94.6	90.9	86.0	90.9	95.1	98.7	90.9	90.9	
✓	✓		88.1	90.7	97.0	90.8	86.5	97.7	95.9	99.3	90.9	98.6	
✓		✓	87.7	97.1	93.3	90.8	86.4	90.9	92.5	98.6	90.9	99.9	
✓	✓	✓	88.6	98.7	98.4	90.7	87.5	95.1	94.9	99.2	90.9	99.0	
Baseline	MFFN	GFCAM	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	
✓			82.8	90.7	88.8	90.1	84.6	90.5	90.7	94.8	86.6	89.4	90.43
✓	✓		88.3	90.7	89.6	90.0	86.7	90.3	90.8	95.9	88.2	89.2	91.76
✓		✓	87.0	90.8	89.7	90.1	82.4	90.6	90.6	97.0	89.9	90.1	91.32
✓	✓	✓	89.6	90.7	89.7	90.3	89.1	90.5	90.6	97.6	87.2	89.9	92.41

5. Conclusions

In addressing the challenging problem of detecting densely distributed small targets in remote sensing images with complex backgrounds, we propose a novel algorithm for remote sensing image target detection. Leveraging our devised multiscale feature fusion method, we effectively integrate

 doi:10.20944/preprints202310.1631.v1

information from the original feature maps with the results of the FPN. This integration mitigates the issue of shallow feature information loss, consequently enhancing the detection capability of small targets in complex backgrounds. Additionally, we introduce a global feature space context aggregation module designed to augment valuable features in each layer of the FPN. Extensive validation and ablation studies are conducted on three publicly available datasets. Experimental results demonstrate that the proposed approach outperforms existing detectors on these three challenging datasets, substantiating the effectiveness and generalizability of the introduced modules. However, it is worth noting that our approach still has limitations in detecting densely occluded targets. In future research, we intend to explore scenarios involving dense target occlusion and refine our network model to better handle such cases.

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 research was funded by the 2023 Anhui Province Higher Education Research Project grant number 2023AH052692.

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. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e3998.
3. Karthikeyan, L.; Chawla, I.; Mishra, A.K. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. *J. Hydrol.* **2020**, *586*, 124905.
4. Chawla, I.; Karthikeyan, L.; Mishra, A.K. A review of remote sensing applications for water security: Quantity, quality, and extremes. *J. Hydrol.* **2020**, *585*, 124826.
5. Bo, L.; Xiaoyang, X.; Xingxing, W.; Wenting, T. Ship detection and classification from optical remote sensing images: A survey. *Chin. J. Aeronaut.* **2021**, *34*, 145–163.
6. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
7. Yang, L.; Haining, W.; Yuqiang, F.; Shengjin, W.; Zhi, L.; JIANG, B. Learning power Gaussian modeling loss for dense rotated object detection in remote sensing images. *Chin. J. Aeronaut.* **2023**.
8. Zhang, W.; Jiao, L.; Li, Y.; Huang, Z.; Wang, H. Laplacian feature pyramid network for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14.
9. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2778–2788.
10. 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 Syst. Appl.* **2023**, *214*, 119132.
11. Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. *arXiv* **2023**, arXiv:2310.06313.
12. Shen, F.; He, X.; Wei, M.; Xie, Y. A competitive method to vipriors object detection challenge. *arXiv* **2021**, arXiv:2104.09059.
13. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11.
14. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.

15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
16. 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* **2022**, arXiv:2209.02976.
17. 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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.
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* **2022**, arXiv:2212.07784.
19. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
20. Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
22. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
23. Li, Y.; Wang, H.; Dang, L.M.; Song, H.K.; Moon, H. ORCNN-X: Attention-Driven Multiscale Network for Detecting Small Objects in Complex Aerial Scenes. *Remote Sens.* **2023**, *15*, 3497.
24. 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 Sens.* **2023**, *15*, 371.
25. Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; Zeng, H. Git: Graph interactive transformer for vehicle re-identification. *IEEE Trans. Image Process.* **2023**.
26. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368.
27. Shen, F.; Shu, X.; Du, X.; Tang, J. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval **2023**.
28. Zheng, S.; Wu, Z.; Xu, Y.; Wei, Z. Instance-Aware Spatial-Frequency Feature Fusion Detector for Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**.
29. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.
30. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 2235–2239.
31. Lee, Y.; Park, J. Centermask: Real-time anchor-free instance segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13906–13915.
32. 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* **2020**, arXiv:2010.11929.
33. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. European conference on computer vision. Springer, 2020, pp. 213–229.
34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
35. Shen, F.; Du, X.; Zhang, L.; Tang, J. Triplet Contrastive Learning for Unsupervised Vehicle Re-identification. *arXiv* **2023**, arXiv:2301.09498.
36. Chalavadi, V.; Jeripothula, P.; Datla, R.; Ch, S.B.; others. mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions. *Pattern Recognit.* **2022**, *126*, 108548.
37. Zhang, J.; Lei, J.; Xie, W.; Fang, Z.; Li, Y.; Du, Q. SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15.

38. Qiao, C.; Shen, F.; Wang, X.; Wang, R.; Cao, F.; Zhao, S.; Li, C. A Novel Multi-Frequency Coordinated Module for SAR Ship Detection **2022**. pp. 804–811.
39. Li, J.; Li, Z.; Chen, M.; Wang, Y.; Luo, Q. A new ship detection algorithm in optical remote sensing images based on improved R3Det. *Remote Sens.* **2022**, *14*, 5048.
40. 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 J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**.
41. Shen, F.; Zhu, J.; Zhu, X.; Xie, Y.; Huang, J. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 8793–8804.
42. 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 Sens.* **2023**, *15*, 3970.
43. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free oriented proposal generator for object detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11.
44. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1922–1933.
45. Wang, X.; Girdhar, R.; Yu, S.X.; Misra, I. Cut and learn for unsupervised object detection and instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3124–3134.
46. 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* **2022**, arXiv:2210.06742.
47. Sun, X.; Cheng, G.; Pei, L.; Li, H.; Han, J. Threatening patch attacks on object detection in optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**.
48. 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 Sens.* **2023**, *15*, 614.
49. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking rotated object detection with gaussian wasserstein distance loss. International conference on machine learning. PMLR, 2021, pp. 11830–11841.
50. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
51. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.
52. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11963–11975.
53. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768.
54. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
55. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307.
56. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. International conference on pattern recognition applications and methods. SciTePress, 2017, Vol. 2, pp. 324–331.
57. Yu, W.; Cheng, G.; Wang, M.; Yao, Y.; Xie, X.; Yao, X.; Han, J. MAR20: A Benchmark for Military Aircraft Recognition in Remote Sensing Images. *Natl. Remote Sens. Bull.* **2022**.
58. Zhou, Y.; Yang, X.; Zhang, G.; Wang, J.; Liu, Y.; Hou, L.; Jiang, X.; Liu, X.; Yan, J.; Lyu, C.; others. Mmrotate: A rotated object detection benchmark using pytorch. Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7331–7334.
59. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI transformer for oriented object detection in aerial images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2849–2858.
60. 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 J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**.

61. Qian, X.; Zhang, N.; Wang, W. Smooth giou loss for oriented object detection in remote sensing images. *Remote Sens.* **2023**, *15*, 1259.
62. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11.
63. Wang, Z.; Bao, C.; Cao, J.; Hao, Q. AOGC: Anchor-Free Oriented Object Detection Based on Gaussian Centerness. *Remote Sens.* **2023**, *15*, 4690.
64. Chen, W.; Han, B.; Yang, Z.; Gao, X. MSSDet: Multi-Scale Ship-Detection Framework in Optical Remote-Sensing Images and New Benchmark. *Remote Sens.* **2022**, *14*, 5460.
65. 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, Vol. 35, pp. 3163–3171.